ISBN 978-82-326-1606-0 (printed ver.)
ISBN 978-82-326-1607-7 (electronic ver.)
ISSN 1503-8181

Doctoral thesis

Hans Moen

NTNU
Norwegian University of
Science and Technology

Hans Moen

Doctoral theses at NTNU, 2016:134

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of Philosophiae Doctor
Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Computer and Information Science
University of Turku
Faculty of Mathematics and Natural Sciences
Department of Information Technology

NTNU

Doctoral theses at NTNU, 2016:134

Hans Moen

Distributional Semantic Models for Clinical Text Applied to Health Record Summarization

NTNU
Norwegian University of
Science and Technology

Turun yliopisto
University of Turku

Hans Moen

Distributional Semantic Models for Clinical Text Applied to Health Record Summarization

Thesis for the Degree of Philosophiae Doctor

Trondheim, May 2016

Norwegian University of Science and Technology
Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Computer and Information Science

University of Turku
Faculty of Mathematics and Natural Sciences
Department of Information Technology

NTNU
Norwegian University of
Science and Technology

Supervisors

Norwegian University of Science and Technology, Norway

Associate Professor Øystein Nytrø, PhD
Department of Computer and Information Science

Professor Björn Gambäck, PhD
Department of Computer and Information Science

Associate Professor Pinar Öztürk, PhD
Department of Computer and Information Science

University of Turku, Finland

Professor Sanna Salanterä, PhD, RN
Department of Nursing Science

Professor Tapio Salakoski, PhD
Department of Information Technology

Others

Doctor Laura Slaughter, Senior Researcher, PhD
The Intervention Centre, Oslo University Hospital, Norway

Joint degree between:

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and Electrical
Engineering Department of Computer and Information Science

University of Turku
Faculty of Mathematics and Natural Sciences
Department of Information Technology

Thesis for the Degree of Philosophiae Doctor

ISBN 978-82-326-1606-0 (printed ver.)
ISBN 978-82-326-1607-7 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2016:134

Printed by NTNU Grafisk senter

Abstract

As information systems in the health sector are becoming increasingly computerized, large amounts of care-related information are being stored electronically. In hospitals clinicians continuously document treatment and care given to patients in electronic health record (EHR) systems. Much of the information being documented is in the form of clinical notes, or narratives, containing primarily unstructured free-text information. For each care episode, clinical notes are written on a regular basis, ending with a discharge summary that basically summarizes the care episode. Although EHR systems are helpful for storing and managing such information, there is an unrealized potential in utilizing this information for smarter care assistance, as well as for secondary purposes such as research and education. Advances in clinical language processing are enabling computers to assist clinicians in their interaction with the free-text information documented in EHR systems. This includes assisting in tasks like query-based search, terminology development, knowledge extraction, translation, and summarization.

This thesis explores various computerized approaches and methods aimed at enabling automated semantic textual similarity assessment and information extraction based on the free-text information in EHR systems. The focus is placed on the task of (semi-)automated summarization of the clinical notes written during individual care episodes. The overall theme of the presented work is to utilize resource-light approaches and methods, circumventing the need to manually develop knowledge resources or training data. Thus, to enable computational semantic textual similarity assessment, word distribution statistics are derived from large training corpora of clinical free text and stored as vector-based representations referred to as distributional semantic models. Also resource-light methods are explored in the task of performing automatic summarization of clinical free-text information, relying on semantic textual similarity assessment. Novel and experimental methods are presented and evaluated that focus on: a) distributional semantic models trained in an unsupervised manner from statistical information derived from large unannotated clinical free-text corpora; b) representing and computing semantic similarities between linguistic items of different granularity, primarily words, sentences and clinical notes; and c) summarizing clinical free-text information from individual care episodes.

Results are evaluated against gold standards that reflect human judgements. The results indicate that the use of distributional semantics is promising as a resource-light approach to automated capturing of semantic textual similarity relations from unannotated clinical text corpora. Here it is important that the semantics correlate with the clinical terminology, and with various semantic similarity assessment tasks. Improvements over classical approaches are achieved when the underlying vector-based representations allow for a broader range of semantic features to be captured and represented. These are either distributed over multiple semantic models trained with different features and training corpora, or use models that store multiple sense-vectors per word. Further, the use of structured meta-level information accompanying care episodes is explored as training features for distributional semantic models, with the aim of capturing semantic relations suitable for care episode-level information retrieval. Results indicate that such models performs well in clinical information retrieval. It is shown that a method called Random Indexing can be modified to construct distributional semantic models that capture multiple sense-vectors for each word in the training corpus. This is done in a way that retains the original training properties of the Random Indexing method, by being incremental, scalable and distributional. Distributional semantic models trained with a framework called Word2vec, which relies on the use of neural networks, outperform those trained using the classic Random Indexing method in several semantic similarity assessment tasks, when training is done using comparable parameters and the same training corpora. Finally, several statistical features in clinical text are explored in terms of their ability to indicate sentence significance in a text summary generated from the clinical notes. This includes the use of distributional semantics to enable case-based similarity assessment, where cases are other care episodes and their “solutions”, i.e., discharge summaries. A type of manual evaluation is performed, where human experts rates the different aspects of the summaries using a evaluation scheme/tool. In addition, the original clinician-written discharge summaries are explored as gold standard for the purpose of automated evaluation. Evaluation shows a high correlation between manual and automated evaluation, suggesting that such a gold standard can function as a proxy for human evaluations.

Preface

The research was part of a four-year PhD program at the Department of Computer and Information Science, Faculty of Information Technology, Mathematics and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Norway. During this period, 25 per cent of the time was devoted to teaching duties at NTNU. The research was also part of a cotutelle collaboration with the Department of Information Technology, University of Turku (UTU), Finland, where I spent a total of one year. The PhD period lasted approximately five years in total.

The research at NTNU was part of, and funded by, a national project called Evidence-based Care Processes: Integrating Knowledge in Clinical Information Systems (EviCare). This was financed by the Research Council of Norway (NFR), project no. 193022.

The stay and research at UTU was partly funded by the Academy of Finland, project no. 140323. At UTU I participated in a consortium named Information and Language Technology for Health Information and Communication (IKITIK[1]).

During the first three years of the PhD period, I participated in a Nordic and Baltic collaboration network named Health Text Analysis Network in the Nordic and Baltic Countries (HEXAnord[2]), funded by Nordforsk (Nordic Council).

[1] http://www.utu.fi/en/units/med/units/hoitotiede/research/projects/ikitik (accessed 1st March 2016)

[2] http://dsv.su.se/en/research/research-areas/health/hexanord-1.113078 (accessed 1st March 2016)

Acknowledgements

First I want to thank Øystein Nytrø for providing me with the opportunity to pursue the PhD degree and for his guidance in the process. Then I want to thank Björn Gambäck for his thorough and inspiring supervision and collaboration. I am grateful to Pinar Öztürk for her advice, for introducing me to the topic of distributional semantics and for providing me with the foundation and inspiration to start the PhD studies through her supervision of my Master's thesis. Thanks also go to Laura Slaughter for her help and advice.

A special and generous thanks to Erwin Marsi at NTNU for his invaluable support, supervision, collaboration and inspiration through the entire process, despite not being my supervisor on paper.

I want to thank Sanna Salanterä and Tapio Salakoski for inviting me to come to UTU, for the warm welcome provided for me there, and for the following supervision and collaboration, which has been a very important part of my PhD studies. I also would like to thank the others affiliated with UTU that I have collaborated with in terms of co-authorship. These are primarily Filip Ginter, Laura-Maria Peltonen (formerly Murtola), Antti Airola, Tapio Pahikkala, Juho Heimonen, Sampo Pyysalo, Heljä Lundgren-Laine, Riitta Danielsson-Ojala and Virpi Terävä.

I would also like to acknowledge the sources of funding that made this work possible: The Research Council of Norway (NFR) which funded the EviCare project; The Department of Computer and Information Science at NTNU; The Academy of Finland, University of Turku; Nordforsk (Nordic Council) which funded the HEXAnord project.

I want to thank those involved in the EviCare project. I am grateful to Thomas Brox Røst for early on in the process providing me with useful and street smart advice about being a PhD student.

The HEXAnord project was a excellent forum for meeting other researchers in the Nordic region with similar interest areas. I was fortunate to establish a close connection to many of the involved partners. This again resulted in international research collaboration with several partners. I want to thank everyone involved in the HEXAnord project, in particular Aron Henriksson, Maria Skeppstedt and Martin Duneld (formerly Hassel) (Stockholm University, Sweden), Vidas Daudaravičius (Vytautas Magnus University, Lithuania) and Ann-Marie Eklund (University of Gothenburg, Sweden) for the productive collaboration and co-authorship. I also like to thank Marie Lindbergh (Linköping University, Sweden) for the collaboration in the HEXAnord PhD courses. Thanks also go to Hercules Dalianis (Stockholm University, Sweden) for his work as the project manager.

I am grateful for the assistance from Jari Björne (UTU) for his advice and input during the writing of the PhD thesis and Stewart Clark (NTNU) who assisted with the final editing.

I want to thank my office mates at NTNU, Simone Mora, Elena Parmiggiani and Francesco Gianni for including me in the Italian society there. Another generous thanks go to the “lunch people” at NTNU, primarily being Gleb Sizov, Erik Smistad, Kai Olav Ellefsen, Lars Bungum, Axel Tidemann, Lester Solbakken and Boye Annfelt Høverstad for all the interesting and entertaining discussions. At the Department of Nursing Science at UTU I shared office with Eriikka Siirala and Hannakaisa Niela-Vilén who I would like to thank for contributing to a nice work atmosphere.

I am grateful to my friends who have played an important role through these years. Special thanks go to Tove Sveen for her helpful suggestions and to Kevin Andre Vatn for contributing to solving certain hardware- and Java-related issues.

Finally, and most importantly, I would like to thank my parents, Jorunn Iversen and Vidar Moen, and my siblings, Martin Moen, Olve Moen and Kjersti Moen, for their constant support, love and encouragement.

Contents

Part I

Research Overview

Chapter 1

Introduction

The work conducted in this thesis approaches the task of automated summarization of clinical free text in care episodes. Focus is placed on methods that mainly exploit distributional statistics in clinical notes and care episodes, thus avoiding manual labor in constructing semantic knowledge resources to support this task. A set of different distributional semantic methods, i.e the models they construct, are first evaluated in the following separate tasks: *synonym extraction* (word similarity assessment), *sentence similarity classification* (sentence similarity assessment), and *care episode retrieval* (care episode similarity assessment). Each of these represents tasks related to supporting clinical work, while also directly or indirectly representing sub-tasks in a intended text summarization system. Finally these models are used in a set of methods for performing *automatic summarization of care episodes*. The work touches upon a number of fields related to *natural language processing*, primarily *computational semantics*, *information retrieval* and *automatic text summarization*.

1.1 Motivation

The development, adoption and implementation of health information technology, such as *electronic health record* (EHR) systems, are strategic focuses of health policies globally European Commission (2012), Blumenthal and Tavenner (2010), Jha (2010), Bartlett et al. (2008). The amount of electronically documented health information is increasing as health records are becoming computerized. In addition, the ongoing advances in diagnostic and health sciences contribute to an increase in the amount of information accumulated for each patient. The large amounts of computerized health information complicate its management and increase the risk of information overload for clinicians (Hall and Walton 2004, Farri et al. 2012). This causes other problems in the clinical work, such as errors, frustration, inefficiency, and communication failures (Lissauer et al. 1991, Suominen and Salakoski 2010). At the same time, this creates opportunities for technological solutions to support clinical care and research.

In hospitals, much of the information that clinicians document are notes in the form of free text that they write in relation to patient care. During a patient's hospital stay, i.e., a *care episode*, clinicians with various specializations write *clinical notes* on a regular basis to document the ongoing care process (status, reasoning, plans, findings, operations, etc.). In the end, normally when the patient is leaving the hospital, a *discharge summary* is written that summarizes the hospital stay. The free text in clinical notes may contain valuable information that is not found or documented elsewhere, such as in the structured, numerical and image data stored in EHRs.

Natural language processing (NLP) (Hirschberg and Manning 2015) tools and resources have the potential to assist clinicians in their interaction with this free-text information. This includes assisting in tasks like automatic event detection in health records (Mendonça et al. 2005), automatic concept indexing (Berman 2004), medication support (Xu et al. 2010), decision support (Demner-Fushman et al. 2009, Velupillai and Kvist 2012), query-based search (Grabar et al. 2009)) and automated summarization (Pivovarov and Elhadad 2015).

These tasks require a certain amount of understanding of the "meaning" of the linguistic items, such as words, sentences and documents. Here, methods in *computational semantics* can be used that focus on how to automate the process of constructing and reasoning with meaning representations of linguistic items. An active area in computational semantics focuses on methods for doing automated *semantic similarity* assessment, which utilizes a similarity metric to calculate a numeric value reflecting the likeness of the meaning, or semantic content, between pairs of linguistic items.

Clinical language has a highly domain-specific terminology, thus specialized NLP tools and resources are commonly used to enable computerized analysis, interpretation and management of written clinical text (Kvist et al. 2011, Meystre et al. 2008, Pradhan et al. 2014). This includes tasks involving computerized semantic similarity assessment. As an example, we have the following two sentences, both referring to the same event and patient, written by two different clinicians:

- *"The patient has broken his right foot during a football match."*
- *"Pt fractured his right ankle when playing soccer."*

This example illustrates how two sentences that barely contain any of the same words can describe one and the same event. A straightforward string matching approach is not adequate to determine that they have a similar meaning, thus a more advanced approach is needed.

There are several lexical resources that enable various degrees of computational semantic textual similarity assessment to be made between words and concepts found in the clinical terminology. Examples of such resources are: the Systematized Nomenclature of Medicine, Clinical Terms (SNOMED-CT) ontology (NLM b); the Medical Subject Headings (MeSH) thesaurus (NLM a); the International Classification of Diseases[1] (ICD) medical classification lists (World Health Organization 1983); the Unified Medical Language System (UMLS) compendium (NLM c) where all these resources are part of or originated from; the generic WordNet ontology (Miller 1995).

Unfortunately, the above lexical resources exist primarily for the English language and/or have limited generalizability in terms of language and coverage, which limits their area of use. Developing new resources, or adapting existing resources to other languages, is costly and time consuming as it requires labor by experts with both linguistic and domain knowledge — medical and clinical knowledge. Such efforts are often done through extensive national or international collaboration projects, one example being the translation of MeSH into Norwegian (Aasen 2012). Further, even though thesauri and ontologies (and such a manual modeling approach in general) are well suited for modeling (semantic) relations on a conceptual level, modeling all possible semantic relations between, e.g., words and concepts used in clinical text would be very difficult and costly to achieve. To enable fine-grained computerized semantic similarity assessment between all possible linguistic items found in clinical text, one would need to develop or adapt semantic resources to the point where they (in sum) capture the totality of the localized terminology used by clinicians in the language(s), region(s), hospital(s) and ward(s) of interest. On top of this come the potential problems related to legal restrictions in terms of distributing clinical information due to its potentially sensitive content. This limits the number of researchers and developers accessing relevant data in the first place. In addition, this is a limiting factor with respect to the amount and coverage of openly available clinical language resources relevant to enable semantic similarity assessment.

An alternative approach focuses on enabling automated, data-driven, learning of semantic similarity in the vocabulary in a text corpus. Such methods are commonly referred to as *distributional semantic methods* (see Turney and Pantel (2010) for

[1] International Statistical Classification of Diseases and Related Health Problems

an overview). Central here is the process of capturing semantic relations based on statistics about word usage in the corpus, and storing this as a computerized vector-based representation, typically referred to as a model — a *distributional semantic (similarity) model*. In applying such methods one can potentially circumvent the need to manually develop lexical resources, in particular those focusing on semantic similarity, or reduce the need for manual labour through hybrid approaches. Pedersen et al. (2007) showed that distributional semantic methods — that exploit statistical distribution patterns in unannotated, unstructured, free text in a *training corpus* of clinical text — are suited for the modeling of semantic similarities between medical concepts on the same level as using SNOMED-CT, WordNet and other available resources. These methods rely on the *distributional hypothesis* (Harris 1954) (see Section 2.1), and their underlying representations are vector-based — vector space models (VSMs) (see Section 2.1.3). They produce distributional semantic models, which are also referred to as *distributional semantic spaces*, in the form of a vector-based representation that enables the computer to calculate similarity between linguistic items (e.g., words, sentences, documents) as a distance measure. Training of such models is commonly done using an unannotated, unstructured, free-text corpora, thus this type of methods can be said to be language independent and "resource light". As the training is data-driven, the resulting models tend to reflect the semantic relations in the language and terminology used in the utilized training corpus. However, there are numerous ways of constructing distributional semantic models with respect to what features to use for training, how to weight the features, how to represent the semantic information, how to calculate similarities between the constituent vectors (i.e., what similarity metric to use), and so on. This necessitates exploration of various ways of capturing and calculating the desired semantics from a training corpus that best match the similarity assessment task at hand (see, e.g., Kolb (2009), Baroni and Lenci (2010), Lenci and Benotto (2012)).

A possible application is related to the discharge summaries that clinicians write when summarizing patients' hospitalization periods (i.e., care episodes). Due to factors such as limited time and information overload, discharge summaries are often produced late, and the information they contain tends to be insufficient (Kripalani et al. 2007). Thus, clinicians would potentially benefit from having a system that supports information summarization through (semi-)automatic text summarization, not only during the discharge process, but also at any point during an ongoing care episode, and for summarizing information from earlier care episodes (Pivovarov and Elhadad 2015). Ultimately such a system could help in saving time and improving the quality of documentation in hospitals. Figure 1.1 illustrates a care episode consisting of several clinical notes, and ends with a discharge summary.

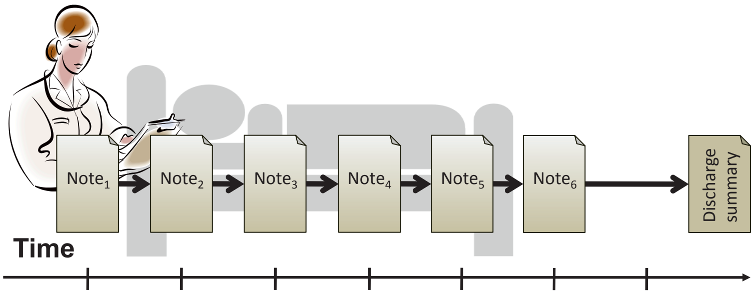


Figure 1.1: A care episode, consisting of a set of clinical notes and ending with a discharge summary.

Automatic text summarization is the computerized process of taking some text from one or more documents and constructing a shortened version that retains the most important information (Luhn 1958). The task of summarizing multiple clinical notes from one care episode is a matter of summarizing multiple documents — i.e., *multi-document summarization* — involving the following goals: include the most important or relevant information; avoid redundant information; produce coherent text. There are many ways to approach this task. (Jones 1999) presents *factors* that one has to be taken into account in order to make a summarization system achieve its task. These mainly concerns *input*, *purpose* and *output*. Others have later discussed and elaborated upon these factors Hahn and Mani (2000), Afantenos et al. (2005). Through a study conducted early on in the PhD process, we identified the following properties and requirements for a text summarization system intended for clinical free-text notes (see Section 1.2): It concerns *multiple documents*; few tailored lexical and knowledge *resources* exist; the content selection is to be done in an *extraction-based* fashion; the produced summaries should contain *indicative* information; the system should be able to produce both *generic* and *user-oriented* summaries. The *output* should be a single piece of text, with similar structure as the notes written by clinicians, which would arguably make *evaluation* more convenient compared to other alternatives, such as graph- or time-line-based visualization. See Section 2.3 for more details.

Selecting what information to include when summarizing the textual content in a care episode is a complex and challenging task. A recent review by Mishra et al. (2014) found that most text summarization techniques used in the biomedical domain can be classified as "knowledge rich" as they depend on (and the quality of) manually developed lexical resources, such as ontologies and annotated training corpora and gold standards[2] available for training and testing. The same seems to apply to techniques and methods in existing summarization systems designed for EHRs (Pivovarov and Elhadad 2015). Typically such knowledge resources are used to first explicitly classify the information in the text that is to be summarized, such as words and concepts, and then used to assess similarities and ultimately significance. This however implies that the systems have restricted generalizability in terms of languages and (sub-)domains.

Textual similarity assessment, particularly on a sentence level, is an important aspect of automatic text summarization (Ferreira et al. 2016). Pivovarov and Elhadad (2015) observed that, in clinical summarization, there has been relatively little work on similarity identification between textual concepts in (sequences of) clinical notes, including the exploration of such information for the purpose of automated summarization. Further, this is identified as an important direction for future EHR summarization methodology. In the approach presented in this thesis (Paper E), a set of techniques and methods are explored and evaluated that uses various types of statistically derived information and features found within the care episode that are to be summarized, and/or in large collections of care episodes. Although this makes the methods/techniques arguably “knowledge poor”, they are easily adaptable to different languages and sub-domains within the health sector. One example is to explore various textual features found internally in a care episode, such as word usage and repeated information; another example is to look at other care episodes with similar content, selected using *information retrieval* (Manning et al. (2008), Chapter 6) (Paper C), and then look at the statistical probability for some information to be found in a discharge summary given that it occurs in one or more of its accompanying clinical notes. These examples depend upon the ability to measure semantic similarity between linguistic items, such as words sentences and documents, which motivates the use of *distributional semantics* (Paper A, B and C).

1.2 Research Objectives

The present research was driven by a set of goals (RG1–RG4), each leading to the next. The initial goal (RG1) was provided by the EviCare project:

RG1: *Explore approaches for conducting summarization of the free text in care episodes, emphasizing approaches and underlying methods that are resource light in terms of adaptation to the domain and different languages.*

[2] A *gold standard*, or *reference standard*, is here defined as being the optimal/ideal solution for the task at hand.

This led to:

RG2: *Explore various (vector-based) distributional semantic methods with respect to their ability to capture semantic similarity between linguistic items in clinical (free) text.*

This again led to:

RG3: *Explore ways to enable distributional semantic methods/models to capture domain- and task-specific semantic information from clinical free text, focusing on the following two tasks:*

- *Sentence similarity classification.*
- *Care episode similarity assessment.*

When pursuing the above goals, conducting a proper evaluation became yet another goal:

RG4: *Find how to automatically and reliably evaluate the various text summarization approaches and the underlying semantic methods in the sub-tasks they are intended for.*

With these goals in mind, RG1 in particular, I had the following research questions that I intended to answer through a set of experiments:

RQ1: *How can the distributional hypothesis be utilized in constructing semantic similarity models suited for clinical text?*

RQ2: *What sentence-level features of clinical text in care episodes are indicative of relevancy for inclusion in a clinical free-text summary?*

RQ3: *How can the evaluation of distributional semantic models and text summaries generated from clinical text be done in a way that is fast, reliable and inexpensive?*

In the work on addressing these research questions, four sets of experiments were conducted. The first set focuses on synonym extraction, the second concerns sentence similarity classification, then care episode retrieval, and finally automatic summarization of care episodes. The clinical text used is mainly from a Swedish and a Finnish hospital, as detailed in Section 1.4.4. The utilized methods are considered language independent (when not taking into consideration the text lemmatization), but are arguably somewhat biased towards the clinical documentation procedures and structure that is common in this region (Allvin et al. 2010). These sets of experiments are presented in five separate papers, as shown in Table 1.1. They build on each other and Figure1.2 visualizes these relations, starting from word-level similarity assessment. The relations between *Research Goals*, *Research Questions* and *Papers* is shown in Table 1.2.

Experiments	Papers
Synonym extraction	A: *Synonym extraction and abbreviation expansion with ensembles of semantic spaces*
Sentence similarity classification	B: *Towards dynamic word sense discrimination with Random Indexing*
Care episode retrieval	C: *Care episode retrieval: distributional semantic models for information retrieval in the clinical domain*
Automatic summarization of care episodes (1)	D: *On evaluation of automatically generated clinical discharge summaries*
Automatic summarization of care episodes (2)	E: *Comparison of automatic summarization methods for clinical free-text notes*

Table 1.1: Experiments and accompanying papers.

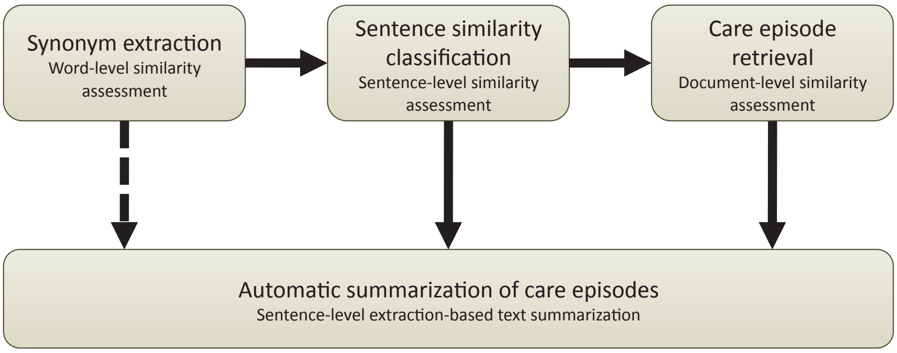

Synonym extraction
Word-level similarity assessment
Sentence similarity classification
Sentence-level similarity assessment
Care episode retrieval
Document-level similarity assessment
Automatic summarization of care episodes
Sentence-level extraction-based text summarization

Figure 1.2: An overview of the conducted research.

1.3 Research Methodology

This thesis work touches upon a number of fields related to NLP. Primarily these are computational semantics, information retrieval and automatic text summarization. As most of the tasks and experiments directly or indirectly focuses on sup-

Research Goals	Research Questions	Papers
1	1, 2	A, B, C, D, E
2	1	A
3	1, 3	B, C
4	3	A, C, D, E

Table 1.2: The relation between *Research Goals*, *Research Questions* and *Papers*.

porting health care, this work is also related to the field of health informatics. The work also involved various degrees of collaboration with clinical professionals, which provided invaluable insight and understanding of their work practice and needs. Innovation was a keyword for the EviCare project and the IKITIK consortium, which is also reflected in this work.

The overall research approach can be viewed as *design science* (Hevner et al. 2004). In the design science paradigm the aim is to design and apply new and/or innovative artifacts aimed at human and organizational use. Knowledge and understanding about the underlying domain and possible solutions are gained through the design, application and evaluation process of the artifact(s), often performed in iterations. As emphasized by Cohen and Howe (1988; 1989), artificial intelligence research should be driven by evaluation. When developing a system or program in this field, evaluation should not only cover performance measures, but also reveal the behavior of the system, limitations, generalizability and prospects for future development.

Starting from RG1, the general direction of the research was set relatively early on in the process. The research questions were then defined in the process of deciding on a general level what techniques and methods that I wanted to explore when approaching RG1. From there the various research goals following RG1 emerged. The various techniques and methods utilized in the different experiments reflects the underlying hypotheses.

Primarily an iterative process was used when conducting the experiments, where each iteration typically included design (software design and implementation), application and evaluation. Mainly a *quantitative* approach (see, e.g., VanderStoep and Johnson (2008), page 7) was used for evaluation. Performance scores were calculated based on gold standards and further compared to scores achieved by various related approaches (baselines and state-of-the-art). In that sense the process was guided by the gold standards used in the various experiments. Through analysing the evaluation scores and identifying problems that arose during the implementation and application, increased understanding was gained regarding the utilized methods in terms of their potential applications, strengths and weaknesses. When developing the manual evaluation scheme related to the automatic text summarization work, the use of open-ended questions was also explored. The latter provided some *qualitative* feedback (see, e.g., VanderStoep and Johnson (2008), page 7) from clinical experts about the direction of that work.

The presented results and methods could potentially contribute to approaches and software methods for others to use and expand upon when pursuing similar goals. Results of this work have been published in conference/workshop proceedings and journals. The experiments and utilized resources are explained in ways that should enable others to replicate the experiments. However, the clinical corpora used are not openly available due to the sensitive nature of clinical text.

1.4 Research Papers and Contributions

1.4.1 List of Papers Included in the Thesis

Paper A: Henriksson, Aron; Moen, Hans; Skeppstedt, Maria; Daudaravičius, Vidas, and Duneld, Martin. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, 5(1):25, 2014.

Paper B: Moen, Hans; Marsi, Erwin, and Gambäck, Björn. Towards dynamic word sense discrimination with Random Indexing. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 83–90, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

Paper C: Moen, Hans; Ginter, Filip; Marsi, Erwin; Peltonen, Laura-Maria; Salakoski, Tapio, and Salanterä, Sanna. Care episode retrieval: distributional semantic models for information retrieval in the clinical domain. *BMC Medical Informatics and Decision Making*, 15(Suppl 2):S2, 2015.

Paper D: Moen, Hans; Heimonen, Juho; Murtola, Laura-Maria; Airola, Antti; Pahikkala, Tapio; Terävä, Virpi; Danielsson-Ojala, Riitta; Salakoski, Tapio, and Salanterä, Sanna. On evaluation of automatically generated clinical discharge summaries. In *Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI 2014)*, pages 101–114, Trondheim, Norway, 2014. CEUR Workshop Proceedings.

Paper E: Moen, Hans; Peltonen, Laura-Maria; Heimonen, Juho; Airola, Antti; Pahikkala, Tapio; Salakoski, Tapio, and Salanterä, Sanna. Comparison of automatic summarisation methods for clinical free text notes. *Artificial In-*

telligence in Medicine, 67:25–37, 2016.

1.4.2 List of Related Papers Not Directly Included in the Thesis

Öztürk, Pinar; Prasath, R. Rajendra, and Moen, Hans. Distributed representations to detect higher order term correlations in textual content. In *Rough Sets and Current Trends in Computing - 7th International Conference, RSCTC 2010*, volume 6086 of *Lecture Notes in Computer Science*, pages 740–750, Warsaw, Poland. Springer, 2010.

Moen, Hans and Marsi, Erwin. Towards retrieving and ranking clinical recommendations with Cross-lingual Random Indexing. In *Proceedings of CLEFeHealth 2012, CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*, volume 1178 of *CEUR Workshop Proceedings*, numpages 4, Rome, Italy, 2012.

Henriksson, Aron; Moen, Hans; Skeppstedt, Maria; Eklund, Ann-Marie; Daudaravičius, Vidas, and Hassel, Martin. Synonym extraction of medical terms from clinical text using combinations of word space models. In *Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine (SMBM 2012)*, pages 10–17, Zurich, Switzerland, 2012. Zurich Open Repository and Archive.

Moen, Hans and Marsi, Erwin. Towards cross-lingual information retrieval using Random Indexing. In *NIK: Norsk Informatikkonferanse, volume 2012*, pages 259–262, Bodø, Norway, 2012. Akademika forlag.

Marsi, Erwin; Moen, Hans; Bungum, Lars; Sizov, Gleb; Gambäck, Björn, and Lynum, André. NTNU-CORE: Combining strong features for semantic similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 66–73, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics.

Moen, Hans and Marsi, Erwin. Cross-lingual Random Indexing for information retrieval. In *Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 164–175. 2013.

Murtola, Laura-Maria; Moen, Hans; Kauhanen, Lotta; Lundgrén-Laine, Heljä; Salakoski, Tapio, and Salanterä, Sanna. Using text mining to explore concepts associated with acute confusion in cardiac patients documentation. In *Proceedings of CLEFeHealth 2013: Student Mentoring Track*, numpages 2, Valencia, Spain, 2013. CLEF online working notes.

Pyysalo, Sampo; Ginter, Filip; Moen, Hans; Salakoski, Tapio, and Ananiadou, Sophia. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LMB 2013)*, pages – 5, Tokyo, Japan, 2013. Database Center for Life Science.

Moen, Hans; Marsi, Erwin; Ginter, Filip; Murtola, Laura-Maria; Salakoski, Tapio, and Salanterä, Sanna. Care episode retrieval. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi 2014) @ EACL*, pages 116–124, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

1.4.3 Contributions

The main contributions of the work presented in the thesis are:

- Paper A: Evaluation of several vector-based distributional semantic models for automated synonym extraction from clinical text, including exploration of combined model pairs trained on clinical text or medical journal articles.

- Paper B: Introduction of a novel method for constructing multi-sense semantic models, evaluated in a task concerning sentence similarity assessment.

- Paper C: Evaluation of a set of information retrieval methods that utilize the distributional hypothesis. The resulting models are evaluated in the task of care episode retrieval. These experiments include novel methods utilizing the ICD-10 codes attached to care episodes to better induce domain-specificity in the resulting models.

- Papers A & C: Proposals for how to evaluate semantic models used for clinical synonym extraction and care episode similarity.

- Paper E: Exploration of a set of resource-light automatic text summarization methods tailored for sequences of clinical (free-text) notes in care episodes.

- Papers D & E: Proposals for how to evaluate clinical text summaries; in an automatic and manual way.

1.4.4 Clinical Corpora

It is typically very difficult for researchers to enquire access to collections of personal health documents of significant size. An asset in the present work is that relatively large corpora of clinical text are used in the experiments.

The corpus used in Paper A is a subset of the *Stockholm EPR Corpus* (Dalianis et al. 2009), extracted from a Swedish hospital and contains clinical notes written primarily in Swedish by physicians, nurses and other health care professionals over a period of six months. It consists of 268 727 notes and approximately 42.5 million words. The use of this corpus for research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/834-31/5. In Papers C, D and E the corpus used is extracted from a Finnish hospital, over a period of four years, consisting of clinical notes written in primarily Finnish by physicians for patients with any type of heart-related problems. It consists of 398 040 notes and approximately 64 million words. Ethical approval for the research was obtained from the Ethics Committee of the Hospital District (17.2.2009 §67), and permission to conduct the research was obtained from the Medical Director of the Hospital District, permission number 2/2009. These corpora are stored in compliance to local regulations concerning sensitive data management.

1.5 Thesis Structure

The remainder of the thesis is structured as follows.

- **Chapter 2** introduces the main concepts and methods that are needed in order to understand the work in the included papers.
- **Chapter 3** contains the main results in the form of paper summaries and retrospective discussions.
- **Chapter 4** discusses the main contributions in relation to the research questions and discusses some directions for future work.
- **Part II** contains the papers of the thesis.

Chapter 2

Background

This chapter provides an overview of the related background for the work presented in the thesis.

2.1 Computational Semantics and Distributional Semantics

Human language is very complex as it reflects high-level cognitive processes of the human brain. To fully understand the true or intended "meaning" and/or content of a word, sentence or document, one needs understanding and knowledge about: The language and underlying grammar and syntax; The meaning of each word and what information they are meant to convey, alone and in context with other words and word phrases, posterior and anterior ones; How each word and word phrase relates to concepts or objects in the real world; The domain, topic and ongoing event(s) or case(s). Even the concept "meaning" itself is rather defuse, as elaborated by Sahlgren (2006).

Tasks or problems requiring *artificial intelligence* (AI) for solving on the same level as human intelligence are commonly referred to as being "AI-complete" (Yampolskiy 2013)[1]. On the one hand, there is a large gap between the cognitive processes underlying language understanding native to human intelligence and that which is achieved by today's computers. On the other hand, many tasks involving processing of natural language do not necessary require a deep understanding of the information it is meant to convey. Today we see that rather shallow approaches can be of great assistance in multiple *natural language processing* (NLP) tasks — approaches that exploit the computational power of computers and the exis-

[1] An AI-complete problem means that its optimal solution implies solving the problem of developing AI in computers that are as intelligent as humans.

tence of large amounts of accumulated digital (textual) information (Hirschberg and Manning 2015). This is reflected in state-of-the-art machine translation systems, search engines and question answering systems, e.g., IBM's Watson system (Ferrucci et al. 2010).

NLP is a field that concerns the interaction between computers and human natural languages (Hirschberg and Manning 2015). An example of a computer system that includes NLP is one that takes human language as input, written or spoken, then tries to interpret and "understand" it in a way by, e.g., converting it into a representation that the computer can further process. This can be to convert free-text queries into a form or representation in an Internet search engine that is used to find web pages that most closely match the information content of the query. This is often referred to as *natural language understanding*. Also going in the other direction is considered NLP, i.e., generating or constructing human-language information from some computerized representation. The latter is often referred to as *natural language generation*. A unifying example is a machine translation system that includes both some type of language understanding and language generation components, together with a translation component, for translating a text phrase from one language into another.

2.1.1 Semantics

Semantics concerns the study of the meaning of natural language expressions and the relationships between them. In *computational semantics* the focus is on automatically constructing and reasoning with the meaning of natural language expressions. A common task in computational semantics is to calculate how similar, or related, linguistic items are based on their *semantic* meaning or content. We will refer to this as *semantic similarity assessment*. For instance, "pain" and "ache" have a rather high degree of semantic similarity since both refer to a type of painful sensation. This differs from, e.g., string similarity, where "pain" is a lot more similar to, lets say, "paint" than to "ache".

A *semantic similarity method/algorithm* usually relies on, and potentially produces, a representation that contains semantic similarity information in a way that enables the computer to reason with it, i.e., compute *semantic similarities*: *"A measure of semantic similarity takes as input two concepts, and returns a numeric score that quantifies how much they are alike."* (Pedersen et al. 2007). The utilized representation may be based on sets of (fixed) features that describes the concepts (see e.g., Smith and Medin (1981) Chapter 3), logical forms, graphs, or some type of combination. Nowadays feature sets are commonly treated as vectors of numeric elements, where each dimension represents a discrete feature, or where features are potentially distributed over multiple dimensions in a more "sub-symbolic" representational manner (c.f. neural networks). Numeric vectors are convenient from a computer perspective in that it allow for the use of geometric algebra operations to, e.g., compute the likeness of vector pairs (see Section 2.1.3 about vector similarity). A vector-based representation is commonly referred to as a *vector space model* (VSM).

2.1.2 Language Processing Resources

Several approaches exist for manually developing lexical resources that model semantic similarity relations. These can be based on constructing rules (e.g., Appelt and Onyshkevych (1998)), thesauri (e.g., McCray et al. (1994)), ontologies (e.g., Bateman et al. (1995), Miller (1995)), and annotated data designed for machine learning (ML) algorithms (e.g., Pyysalo et al. (2007), Velupillai and Kvist (2012)). The more complex the task at hand is, the more manual labor is usually required in the development process. Thus, manually developed lexical resources tend to have very specific and restricted coverage in terms of what they represent, e.g., gene–protein relationships (Ashburner et al. 2000, Lord et al. 2003, Pyysalo et al. 2007).

By far the most comprehensive approach to modeling the terminology used in the medical domain is the development of the UMLS (NLM c) compendium. It consists of various lexical resources comprising primarily the vocabulary in medical research literature and clinical documentation, it also contains a mapping between the different vocabularies therein. SNOMED-CT (NLM b) represents medical terms in an ontological representation. SNOMED-CT originated as a resource for the English language, but has later been, or is currently being, translated into several other languages, primarily Spanish, Danish, Swedish, Dutch and French. MeSH (NLM a) is a thesaurus developed to index health research literature. It was originally made for English, but has later been translated or mapped to several other languages. ICD (the latest version being the 10th — ICD-10) (World Health Organization 1983) is a hierarchical medical classification, containing codes for diseases, signs and symptoms, etc., used primarily to classify diagnoses and treatments given to patients. Today the ICD classification has been translated into multiple languages.

The approach of manually developing lexical resources is well suited for modeling (semantic) relations on a conceptual level. However, with the vast information variety and complexity that natural language (free) text enables, it is very costly and challenging to conduct such modeling manually in a way that covers the language in its entirety. The same goes for enabling mappings between modeled concepts and the vocabulary — including the correct meaning of words and phrases as they are used in the text. An example illustrating some of the underlying challenges is found in Suominen (2009), page 30, where it is reported that a single medicine has over 350 different spellings in clinical notes. This is one reason why normalization of the vocabulary in clinical notes has been the focus of several shared tasks, such as in the ShARe/CLEF eHealth Evaluation Lab 2013 (Suominen et al. 2013).

An alternative approach to manual lexical resource development is to model textual semantics in an automated and corpus-driven way. Methods in *distributional semantics* focus on learning/inducing semantic similarity from statistical information about word usage in a large corpus of unannotated text. In this thesis such a corpus is referred to as *training corpus* — typically being a collection of thousands or millions of unannotated documents. These methods are based on the *distributional hypothesis* (Harris 1954), which states that *linguistic items with similar distributions in language — in the sense that they co-occur with overlapping context — have similar meanings*. Two linguistic items, e.g., two words, having a similar “meaning” according to this hypothesis implies that they, statistically speaking, have been commonly used with the same or similar contextual features in the training corpus. For instance, they have co-occurred with the same neighboring words, or they have been often used within the same documents. The study of utilizing statistical approaches in computational semantics is sometimes referred to as *statistical semantics* (Weaver 1955, Furnas et al. 1983). The goal is to utilize statistical features in text to calculate a semantic similarity score between linguistic items that agrees with human judgement regarding the similarity of the items in a given context (c.f. “pain” and “ache”).

Intuitively, relations between certain textual concept are difficult to obtain through purely statistical approaches, in particular those requiring complex implicit knowledge. For example, the relationships between known genes and proteins (Pyysalo et al. 2007). However, several hybrid approaches have been introduced that combine distributional information with lexical resources (Turney and Pantel 2010). For instance, Chute (1991) used Latent Semantic Analysis (LSA) to construct a distributional semantic model from the UMLS metathesaurus that enables matching of free-text inquiries with UMLS concepts; Henriksson et al. (2013b) constructed semantic models using Random Indexing, constructed from a corpus of clinical text, to extract synonyms for SNOMED-CT concepts/classes. Faruqui et al. (2015) performed retrofitting of word context vectors in various semantic models using lexical information from resources such as WordNet.

Distributional semantic methods have become popular due to their purely statistical approach to computational semantics and semantic similarity computation (Turney and Pantel 2010). Underlying factors are the increasing availability of large corpora and computational power. These methods enable rapid construction of new *semantic models* reflecting the semantic similarities in the languages and domains in the utilized training corpus. Thus no costly manual labor is required for constructing annotated training data or lexical resources. As the training phase is “data-/corpus-driven”, it is usually executed in a fully unsupervised manner. Also since the underlying training mechanisms are not dependent on the language of the training corpora, such methods can be classified as being language independent.

A training corpus used with distributional semantic methods consist commonly of only unannotated text that has been *pre-processed* to a certain degree. Preprocessing aims to improve the desired semantic representations in the resulting semantic model in various ways. *Tokenization* is, in its simplest form, about first splitting documents into sentences and finally into tokens or words/terms. We will be using ’terms’ and ’words’ rather indistinguishably, the main difference is that terms may contain multiple words (e.g., “car wheel” and “Yellowstone National Park”). Such multi-word terms and expressions can be recognized through a dictionary, rules, statistical co-occurrence information (*collocation segmentation*), annotated training corpora, or hybrid solutions (see e.g., Smadja (1993)). *Lemmatization* or *stemming* is a way to normalize a corpus by reducing the number of unique words. This is done by changing each word into their root form by removing and/or replacing words or suffixes (e.g., when using lemmatization “vocabularies” becomes “vocabulary”, while with stemming “vocabularies” becomes “vocabulari”). Further, this tends to result in an increased distributional statistical basis for the remaining words since the vocabulary is reduced. The same also becomes a consequence of *lowercasing* words (e.g., “She” becomes “she”). Such normalization can be seen as a trade-off between precision and recall. E.g., lowercasing means you can no longer distinguish between proper nouns like “Apple” and common nouns like “apple”. However, this will often improve recall since capitalized sentence-initials will not be confused with proper nouns. As distributional semantic models tend to emphasize high-frequent words and word co-occurrences, it is common to exclude “stop words” by filtering the corpus through a stop word list consisting of words that have little discriminative power in general or in a specific domain (e.g., “a” and “the”). *Part-of-speech tagging* and *dependency parsing* is something one can do to enrich the text with additional linguistic knowledge.

2.1.3 The Vector Space Representation

Vector spaces, or vector space models (VSMs), are by far the most common underlying representations in distributional semantics. VSMs were first introduced in text processing for the purpose of information retrieval by Salton et al. (1975). The underlying principle is to let each textual unit in a *training corpus*, such as words, sentences and documents, be represented as a multidimensional vector, or tensor. These vectors are referred to as *context vectors* (e.g., word context vector), representing the “contextual meaning” of the corresponding textual unit in the utilized training corpus. A collection of these vectors constitutes the content of a vector space — a vector space model. Multidimensional vectors have the capacity to encode a lot of language information (e.g., semantics), where each element/dimension encodes a certain feature of the textual unit it represents. In many VSMs, particularly those whose dimensions have been compressed or reduced in some way (e.g., through some explicit dimension reduction operation or indirect feature prediction), these vectors' dimensions do not necessarily correspond to any known features of the language. Thus such vectors can be a composition of "subsymbolic" features.

In addition to being an efficient way of representing textual information, VSMs allow for efficient ways of calculating similarities between vectors. To measure the similarity/dissimilarity between two vectors, it is common to use some type of algebraic distance function or metric. Different metrics tend to emphasize various types of semantic similarity (Lenci and Benotto 2012). In the work presented in this thesis the *cosine similarity metric* has been used. It calculates the cosine of the angle between two vectors. They can be called $\overrightarrow{x}$ and $\overrightarrow{y}$, thus turning the angle into a value between 0 and 1[2].

$$CosSim(\overrightarrow{x}, \overrightarrow{y}) = \frac{x \cdot y}{||x||\,||y||} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

The cosine similarity metric is often used because it is insensitive to the magnitude of vectors. When comparing two word context vectors using the cosine similarity metric, if their cosine similarity is 1 or close to 1, they have a high similarity. The opposite is the case if the cosine similarity is low. However, cosine similarity values in between 0 and 1 should only be interpreted relative to each other within a single VSM because it is often difficult to map these to any absolute truth outside a VSM. What these similarity values reflect according to a linguistic definition, if any, depends on how context vectors are constructed — what features they contain — and what similarity metric is used. A model may for example reflect that two words with a high cosine similarity value are similar because they are (near) synonyms, antonyms, metonyms, or morphological variants of the same word. It could also reflect that one or both words are abbreviations pointing to a concept that has the same or similar meanings, or that one or both are misspellings of the same or similar words. On the phrase/sentence level, it is common to use notions like paraphrasing and entailment, while topic and structural similarities may be used for document level similarities. Then again, such classifications may not be

[2] The cosine similarity metric can potentially return values between -1 and 1 if the vectors contain negative values.

of much importance in various tasks in computational semantics, where instead the main concern is the intra-model (semantic) similarities in relation to a specific task.

A common use-case for a VSM is to find how similar its constituent linguistic items, e.g., words or documents, are in relation to a given query. This is done by first retrieving, or possibly constructing, a vector representing the query, then compute cosine similarity between it and all other context vectors in the model. In this way we can rank the constituent context vectors according to their (semantic) similarity to the query. For example, if we have a VSM of word context vectors, and we query the model with the word "foot", we calculate the cosine similarity value between the context vector belonging to "foot" and all other context vectors in the model. This will give us a list of similarity values of each word relative to the query. By sorting this list (based on cosine similarity values) we can rank the words based on how similar they are to "foot", as illustrated together with the query word "pain" in Table 2.1. VSMs containing word context vectors are sometimes referred to as *word spaces* or *word space models* (WSMs).

foot	(jalka)	$CosSim$	pain	(kipu)	$CosSim$
lower limb	(alaraaja)	0.5905	pain sensation	(kiputuntemus)	0.5097
ankle	(nilkka)	0.3731	ache	(särky)	0.4835
limb	(raaja)	0.3454	pain symptom	(kipuoire)	0.4173
shin	(sääri)	0.3405	chest pain	(rintakipu)	0.4042
peripheral	(periferisia)	0.3112	dull pain	(jomotus)	0.4000
callus	(känsä)	0.3059	backpain	(selkäkipu)	0.3953
top of the foot	(jalkapöytä)	0.2909	pain seizure/attack	(kipukohtaus)	0.3904
upper limb	(yläraaja)	0.2879	pain status	(kiputila)	0.3685
peripheral	(perifer)	0.2875	abdominal pain	(vatsakipu)	0.3653
in lower limb	(alaraajassa)	0.2707	discomfort	(vaiva)	0.3614

Table 2.1: Top 10 most similar words to the query words "foot" and "pain", together with the corresponding cosine similarity scores. The results are derived from a distributional semantic model trained using W2V on a corpus of clinical text. The words have been translated from Finnish to English.

There are endless ways of generating context vectors in terms of what features define the semantic relations they capture and how these are weighted. A method introduced by Salton et al. (1975) works by deriving term-by-document statistics from a document corpus, generating a term-by-document matrix/model. The rows of the term-by-document matrix represent word context vectors, and the columns represent document context vectors. Here each dimension of a word context vector reflects how many times that word has occurred in each document, each dimension corresponding to one document. As an intuitive example, let us assume that we have the following three short documents:

D1: The patient is suffering from an aching neck.

D2: The patient is experiencing pain in the neck.

D3: Take a taxi to the station.

By *preprocessing* these, through stemming, lowercasing and removal of stop words, they become:

D1: patient suffer ache neck

D2: patient experience pain neck

D3: take taxi station

Further, one can now create a term-by-document matrix based on the frequency of each word in each document, as illustrated in Table 2.2.

	D1	D2	D3
patient	1	1	0
suffer	1	0	0
ache	1	0	0
neck	1	1	0
experiencing	0	1	0
pain	0	1	0
take	0	0	1
taxi	0	0	1
station	0	0	1

Table 2.2: VSM example, term-by-document matrix, generated from three documents.

Statistically, words that occur in many of the same documents, i.e, occur in the same contexts, will have context vectors (rows) of high similarity to each other according to the cosine similarity metric. Likewise, documents containing many of the same words will have corresponding document context vectors (columns) of high similarity. Figure 2.1 illustrates on an intuitive level how similarities between the above documents (their context vectors) can be viewed according to vector angles (left); or as relative distances in a 2D semantic space (right). Such term-by-document models are particularly common in *information retrieval* (IR) (Manning et al. (2008), Chapter 6).

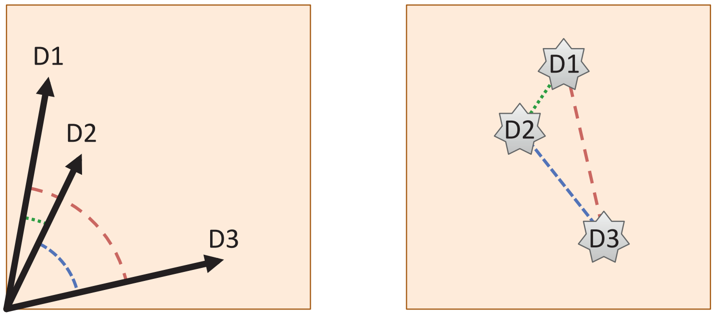


Figure 2.1: Intuitive illustration of how similarities between the three documents in Table 2.2 (their context vectors) can be viewed according to vector angles (left); or as relative distances in a 2D semantic space (right).

One approach to generating word context vectors is through constructing word-by-context models. Lund and Burgess (1996) use the neighboring words as context in their *hyperspace analogue to language* (HAL) method, which defines the semantic meaning of a word based on its neighboring words throughout a corpus. In this way, two words that co-occur with many of the same word neighbors, statistically throughout the training corpus, will have a high semantic similarity. HAL also applies a dimension reduction method to post-compress the matrix based on discarding the columns with lowest variance. Constructing a model from word co-occurrence information is typically done using the *sliding window* technique, where a window of a fixed size is slid across each sentence in the training corpus, iteratively updating each word based on the neighboring words. The size of the sliding window will naturally have an effect on the resulting semantic space, but the exact influence of this parameter seems to be task specific. For example, a window of size ten (5+5, left and right sides of the target word) has been shown to work well for modeling synonymy from clinical corpora (Henriksson et al. 2013a). The sliding-window approach is illustrated in Figure 2.2 where the size of the window is four (2+2). Table 2.3 shows how the resulting VSM becomes from the three example documents/sentences above.

Word context vectors, being the rows of a term-by-context matrix, or model, has what can be referred to as *second-order* co-occurrence relations between them since vector similarity is based on having similar neighbors. By measuring the cosine similarity between each word pairs, we can create a *similarity matrix* as in Table 2.4.

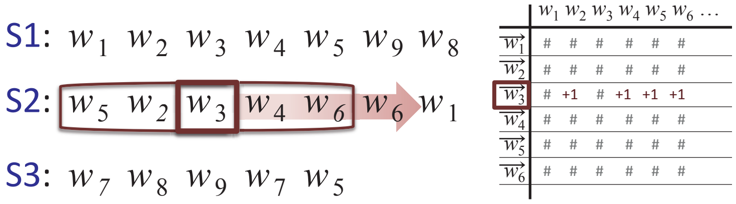

S1: w_1 w_2 w_3 w_4 w_5 w_9 w_8
S2: w_5 w_2 w_3 w_4 w_6 w_6 w_1
S3: w_7 w_8 w_9 w_7 w_5
w_1 w_2 w_3 w_4 w_5 w_6 …
$\vec{w_1}$ # # # # # #
$\vec{w_2}$ # # # # # #
$\vec{w_3}$ # +1 # +1 +1 +1
$\vec{w_4}$ # # # # # #
$\vec{w_5}$ # # # # # #
$\vec{w_6}$ # # # # # #

Figure 2.2: Illustrating training of a word-level co-occurrence distributional semantic similarity model using a "sliding window" with a size of four (2+2).

	patient	suffer	ache	neck	experience	pain	take	taxi	station
patient	0	1	1	0	1	1	0	0	0
suffer	1	0	1	1	0	0	0	0	0
ache	1	1	0	1	0	0	0	0	0
neck	0	1	1	0	1	1	0	0	0
experience	1	0	0	1	0	1	0	0	0
pain	1	0	0	1	1	0	0	0	0
take	0	0	0	0	0	0	0	1	1
taxi	0	0	0	0	0	0	1	0	1
station	0	0	0	0	0	0	1	1	0

Table 2.3: Word-by-context matrix, constructed using a sliding window with a size of four (2+2). Each row represents a word context vector.

	patient	suffer	ache	neck	experience	pain	take	taxi	station
patient	–	0.2887	0.2887	1.0	0.2887	0.2887	0.0	0.0	0.0
suffer	0.2887	–	0.6667	0.2887	0.6667	0.6667	0.0	0.0	0.0
ache	0.2887	0.6667	–	0.2887	0.6667	0.6667	0.0	0.0	0.0
neck	1.0	0.2887	0.2887	–	0.2887	0.2887	0.0	0.0	0.0
experience	0.2887	0.6667	0.6667	0.2887	–	0.6667	0.0	0.0	0.0
pain	0.2887	0.6667	0.6667	0.2887	0.6667	–	0.0	0.0	0.0
take	0.0	0.0	0.0	0.0	0.0	0.0	–	0.5	0.5
taxi	0.0	0.0	0.0	0.0	0.0	0.0	0.5	–	0.5
station	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5	–

Table 2.4: Similarity matrix derived from the word context vectors in Table 2.3.

The similarity matrix in Table 2.4 can potentially be visualized as vectors or points in a semantic space, where their similarities are represented by their relative angles or distances, similarly to the illustration in Figure 2.1, with words instead of documents.

The semantic relations, represented in distributional semantic models, tend to be greatly influenced by common and frequent words occurring in many documents, words that often add little or nothing to the semantic meaning of a document. To counter this, one can re-weight the vector/matrix elements of a term-by-document matrix, using some weighting function. A common weighting method is to multiply the term/word frequencies by their corresponding inverse document frequency (TF*IDF) (Sparck Jones 1972).

$$\mathit{tfidf}(t, d, D) = \mathit{freq}(t, d) \times \mathit{idf}(t, D)$$

$$\mathit{idf}(t, D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

Where:

- t is the term/word in question.
- d is a document.
- D are all documents in the corpus.
- N is the total document count, $|D|$.
- $|\{d \in D : t \in d\}|$ is the number of documents in which t occurs.

The purpose of TF*IDF weighting is to reduce the influence (weight) of words that occur in almost all documents and therefore have little value in discriminating one document from another. At the same time it increases the importance of words that are more rare and limited to a few documents, as these are potentially important to the topic of a document. TF*IDF weighted term-by-document matrices/models are used in various popular search engines, such as Apache Lucene (Cutting 1999).

As mentioned earlier, Latent Semantic Analysis (LSA) is a popular method for constructing (distributional) semantic models. LSA reduces the dimensionality of the VSM, while also having it emphasize latent correlations between words (and documents) through discovering higher order co-occurrence relations within a corpus (second order and above). Landauer and Dumais (1997) achieved human-level performance scores using LSA on the Test of English as a Foreign Language (TOEFL), a test where one has to choose the correct (closest) synonym among four alternatives for each query word. These scores have later been improved upon by others, through using LSA or other methods (see, e.g., Bullinaria and Levy (2012)). Examples of more recent VSM-based distributional semantic methods are: Holographic Reduced Representations (HRR) (Plate 1991); Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999); Non-negative matrix vectorization (NMF) (Lee and Seung 1999); Random Indexing (RI) (Kanerva et al. 2000); Latent Dirichlet allocation (LDA) (Blei et al. 2003); various neural network-based language models, most popular being the Word2vec (W2V) implementation (Mikolov et al. 2013b). RI and W2V are used in various ways in the experiments in this thesis, and will be described in more detail in Sections 2.1.4 and 2.1.5.

2.1.4 Random Indexing

RI (Kanerva et al. 2000) is a method for building a *compressed* VSM with a fixed (reduced) dimensionality, and is done in an incremental fashion. This technique was originally intended as a way of overcoming the performance issues associated with LSA implementations at that time (computational complexity, scalability and memory requirements). Due to its computational efficiency, RI remains popular today for training on large corpora, such as MEDLINE/PubMed abstracts or articles (Cohen 2008, Jonnalagadda et al. 2012, Pyysalo et al. 2013), and social media (Sahlgren and Karlgren 2009). It has been shown to perform well, and is comparable to other methods, such as LSA, in a range of semantic similarity assessment tasks, including the TOEFL synonym test (Sahlgren and Swanberg 2000, Karlgren and Sahlgren 2001).

RI involves the following two steps:
Step 1 - Initialization: first, each word/term in the training corpus is assigned an *index vector* as its unique signature in the VSM. Index vectors have a predetermined dimensionality and consist mostly of zeros together with a small number of randomly distributed 1's and -1's — uniquely distributed for each unique word. This is based on the *Johnson Lindenstrauss Lemma* (Johnson and Lindenstrauss 1984), as discussed by Cohen et al. (2010), stating that distances between points in a high-dimensional space will be approximately preserved when projected into a lower-dimensional subspace. In other words, vectors being orthogonal in the high-dimensional space are assumed to be "near orthogonal" in the lower-dimensional subspace. Thus index vectors will have pairwise similarities, according to the cosine similarity metric, close to 0.
Step 2 - Training: the second step is the training step where context vectors are generated/induced for each unique word in the corpus. This is most commonly done using a sliding window of a fixed size (e.g., 2+2) to traverse the training corpus, inducing context vectors by superimposing the index vectors of the neighboring words in the window, as illustrated in Figure 2.3.

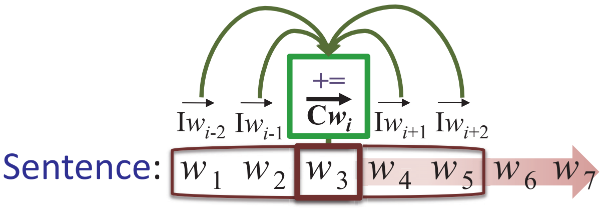

+=
$\overrightarrow{Cw_i}$
$\overrightarrow{Iw_{i-2}}$
$\overrightarrow{Iw_{i-1}}$
$\overrightarrow{Iw_{i+1}}$
$\overrightarrow{Iw_{i+2}}$
Sentence:
w_1 w_2 w_3 w_4 w_5 w_6 w_7

Figure 2.3: Illustrating how the training in RI works. The word context vector $\overrightarrow{Cw_i}$ is updated by adding the index vectors of its neighbors, i.e., through superimposing it with the neighbouring word index vectors $\overrightarrow{Iw_{i-2}}$, $\overrightarrow{Iw_{i-1}}$, $\overrightarrow{Iw_{i+1}}$ and $\overrightarrow{Iw_{i+2}}$.
This is a slightly modified version of Figure 3 in Moen et al. (2015), Paper C in this thesis.

As the dimensionality of the index vectors is fixed, the dimensionality of the vector space will not grow beyond the size $W \times Dim$, where W is the number of unique words in the vocabulary, and Dim being the pre-selected dimensionality to use for the index vectors, and ultimately the context vectors. As a result, RI models are significantly smaller than a full term-by-context model, which again make them a lot less computationally expensive in terms of storage and similarity computation. Additionally, the method is fully incremental in that additional training data can be added at any given time without having to retrain the model. It is also parallelizable and scalable, meaning that it allows for rapid training on very large corpora in a distributed on-line fashion. Using the JavaSDM implementation [3], with default parameters except a dimensionality of 800, the training on a Finnish clinical corpus (see Section 1.4.4) consisting of about 64 million words has an execution time of about 25 minutes. This is on a computer with the following hardware: Intel Core i7-3770 CPU @ 3.40GHz, 4 cores, 16GB RAM.

Some variants of the RI-based approach have been introduced, such as Random Permutations (RP) (Sahlgren et al. 2008) and Reflective Random Indexing (RRI) (Cohen et al. 2010), as well as cross-lingual variants (Sahlgren and Karlgren 2005).

2.1.5 Word2vec — Semantic Neural Network Models

The Word2vec (W2V) (Mikolov et al. 2013a) method/framework relies on using an artificial neural network to construct *neural network language models*. The models it constructs are vector-based and have been found to perform well in a range of semantic similarity assessment tasks (Baroni et al. 2014). Through training the network on a corpus, distributionally similar words are given similar vector

[3] `http://www.nada.kth.se/~xmartin/java/` (accessed 1st March 2016)

representations (i.e., context vectors).

W2V stems from the field *Deep Learning* (LeCun et al. 2015, Collobert et al. 2011). It uses a somewhat simplified neural network model, consisting of an *input layer* with as many input nodes as there are unique words (vocabulary items), a *hidden linear projection layer* with node count equal to the predefined dimensionality of the vector space, and finally a *hierarchical soft-max output layer* predicting the same words as the input layer (Morin and Bengio 2005, Mnih and Hinton 2009).

The context used for training is typically a sliding window. W2V has two training procedures/architectures: “continuous bag-of-words” (CBOW), and “continuous skip-gram model” (Skip-gram). The CBOW training approach aims to predict each word in the training corpus based on its context (co-occurring words). For each target word, the words corresponding to its context are activated in the input layer sequentially, i.e., the values of their corresponding input notes are set to 1 while the rest are 0. The expected/correct output for each training case is the correct target word in the output layer. Each target word and its connected weights are subsequently adjusted to decrease the error between the network outputs (normalized with soft-max) and the training cases using the *back-propagation* procedure (McClelland et al. 1986). This procedure is repeated for all training pairs, often in several passes over the entire training corpus, until the network converges and the error does not decrease any further. Now each word of the input layer has a context vector given by the set of weights connecting its corresponding input node to the hidden layer, as illustrated in Figure 2.4. The Skip-gram training approach predicts each individual context word (output layer) given the corresponding target word (input layer).

To understand on an intuitive level why the network learns efficient representations, i.e., distributional semantic models, we can consider the two-step process of the prediction: first, the input layer is used to activate the hidden, representation layer; and second, the hidden layer is used to activate the output layer and predict the context word. To maximize the performance on this task, the network is thus forced to assign similar hidden layer representations to words that tend to have similar contexts. Since these representations form the resulting model, distributionally similar words are given similar vector representations (c.f. context vectors).

One of the main practical advantages of the W2V method (CBOW/Skip-gram) lies in its relatively low complexity, giving it great scalability and allows for training on billions of words of input text in the matter of several hours. Using the optimized version in the Gensim implementation of Word2vec[4], with default pa-

[4] `https://radimrehurek.com/gensim/models/word2vec.html` (ac-

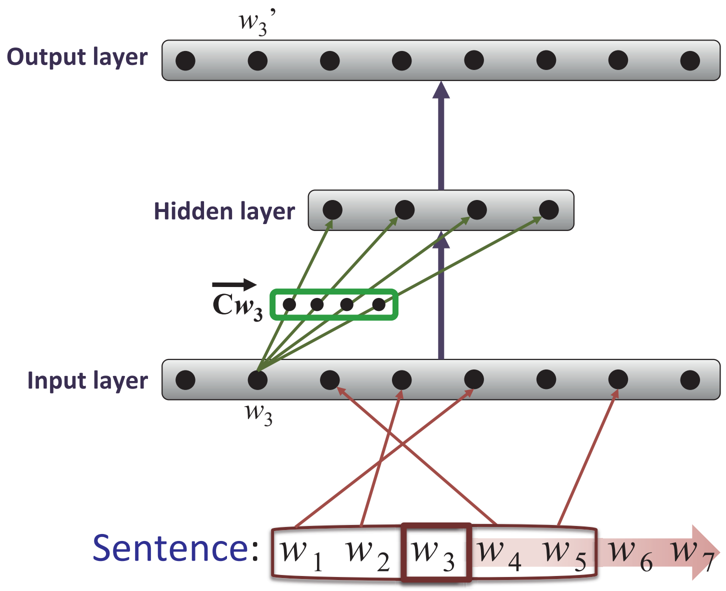


Figure 2.4: Illustration of how training happens in the W2V implementation of CBOW. A sliding window with the size of four (2+2) is moved over the text, word by word. The input layer nodes of the network corresponding to the words in the context window of the word w_3 are activated. The error in the output layer prediction and the expected prediction for the focus word w_3 is back-propagated through the network. When the training is completed, the context vector $\overrightarrow{Cw_3}$ constitutes the set of weights connecting the input layer node for w_3 and the hidden layer.
This is a slightly modified version of Figure 6 in Moen et al. (2015), Paper C in this thesis.

rameters except a dimensionality of 800, the training on a Finnish clinical corpus (see Section 1.4.4) consisting of about 64 million words has an execution time of about 20 minutes. This is on a computer with the following hardware: Intel Core i7-3770 CPU @ 3.40GHz, 4 cores, 16GB RAM. Shorter execution times are achieved when using a fully C-based implementation/package[5].

Neural network-based methods are based on *predicting* the target word or its context features. This differs from *count-based* methods such as RI and LSA that more directly *count* co-occurrences. Baroni et al. (2014) showed that prediction-based models, represented by W2V CBOW, achieved better results than a set of count-based ones in a range of tasks focusing on word-level semantic similarity

cessed 1st March 2016)

[5] https://code.google.com/p/word2vec/ (accessed 1st March 2016)

assessment, including synonym detection/extraction in the TOEFL test and semantic similarity/relatedness classification. However, Levy et al. (2015) later showed that this performance advantage is likely due to smart parameter use and post-processing. An attractive property of W2V-based models is that they seem to preserve syntactic and semantic regularities (Mikolov et al. 2013c), e.g., $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman}$ result in a vector similar to $\overrightarrow{queen}$. Levy and Goldberg (2014) revealed that these same regularities are also preserved, to the same extent, in count-based models when using some alternative similarity metric (i.e., not the cosine similarity metric).

2.1.6 Compositionality in Vector Space Models

So far this chapter has mainly discussed ways to construct distributed semantic models of words, representing word meaning by context vectors. However, in the experiments presented in this thesis, context vectors representing sentences, clinical notes, and care episodes have also been used. This is accomplished through performing some type of *compositionality* (Frege 1892, Montague and Thomason 1976) to ensemble such vectors from the constituent word context vectors. The idea is that a composition of word context vectors will result in a vector that captures the combined meaning of these words. Partee et al. (1990) explains *"The Principle of Compositionality"* as follows: *"... The meaning of a complex expression is a function of the meaning of its parts and of the syntactic rules by which they are combined."*.

Landauer et al. (1997) conclude that much information regarding the semantic similarity of texts, essays in this case, is carried by the semantic similarity between constituent words independently of their order. Thus one straightforward approach to representing multi-word items in VSMs is to treat a collection of words, e.g., a sentence, as simply a "Bag of Words" (BoW), where their order is irrelevant. In this approach, a composed vector, e.g., a sentence context vector, is generated through simply pointwise summing its constituent word context vectors (also referred to as superimposing). In addition to vector *addition*, other alternatives includes vector *multiplication* and the use of *circular convolution* in *holographic reduced representations* (Plate 1991). To reduce the influence of the magnitude of each individual vector, it is common to first normalize vectors to unit length. Further re-weighting can be done by applying TF*IDF weighting (see Section 2.1.3).

There are also ways to compose vectors that incorporate some information about the constituent word order. Such approaches typically focus on constructing context vectors for n-grams, short phrases or larger linguistic items that in some way emphasize the order of the constituent words (see, e.g., Guevara (2011), Mikolov et al. (2013b), Le and Mikolov (2014)). Such a VSM may, as an example, contain vectors representing the phrases “dog eats rabbit”, “rabbit eats dog”, and “dog eats dinner”. Here the phrase “dog eats rabbit” should intuitively be closer to “dog eats dinner” than “rabbit eats dog” in the semantic space. Another way to retain word order information when performing similarity assessment is to simply view each sentence, document, etc., as a list of their constituent word context vectors. In this approach the order information is not actually modeled into the semantic model, but is calculated at retrieval time. When computing the similarity of two sentences one may apply some sequence aligning algorithm, e.g., *Needleman-Wunsch* (Needleman and Wunsch 1970), to compute a total similarity score based on word alignment and word pairwise cosine similarity scores (see, e.g., Feng et al. (2008)).

2.2 Distributional Semantic Models and Clinical Language Processing

Clinical language in this thesis is defined as the language clinicians use when documenting patient care, mainly in the form of written text notes stored in the patients’ health records. As the focus is on clinical text written in hospitals, we refer to physicians/doctors and nurses involved in clinical care in hospitals as *clinicians*. There are often physicians with different medical specializations, located at different wards within the hospital, involved in the treatment during a care episode, such as internal medicine, cardiology and surgery. Thus the clinical notes, or narratives, that they write tend to reflect the different tasks being performed at the respective wards. In this thesis the term ‘clinical note’ refers to any of the different notes that the various specialists write to document patient care. During a care episode, a sequence of clinical notes are written (as illustrated in Figure 1.1). These are stored in the patient’s health record, which again is stored digitally in an electronic health record (EHR) system.

Clinical notes contain highly domain-specific terminology (Rector 1999, Friedman et al. 2002, Allvin et al. 2010), and clinical language can be regarded as a *scientific sub-language* (Meystre et al. 2008). Some features of the written clinical language/notes are:

- The different professions and individual clinicians tend to have their own way of documenting — documenting observations, symptoms, diagnoses, and their reasoning and speculations.
- Each note may have a different author, including those belonging to the same care episode.
- The texts usually contain ungrammatical language, incomplete sentences,

abbreviations, and medical jargon.

- Authors do not necessarily utilize any common note structure.
- The written information tend to be highly domain- and case-specific, including a fair share of implicit information.
- All notes in one care episode are related to the care and treatment given to the same patient, meaning that they all are linked to a series of related events and often contain repeated and overlapping information. This is also the case when looking at the full health record belonging to a patient, since some information from one care episode could be relevant to a later one.

Figure 2.5 shows an example of a clinical note.

English translation:
61-years old female with Crohn's disease. Attended cycling event in Salo, flu prirorlry. Arfter cycling, experienced breathing difficulties and went to the emergency department and elevated herart enzymes and incompensation were found. Was admitted to the ICU for care of incompensation and pneumonia. In UKG 2.6. ef 30%. In coronary angiography, significrant stenoses in RCA, LCX and mrarin. Trordray, elective quadrurple bypass LITA-LAD, Ao-LOM-LPL and Ao-RBD, in which goord flow. Pre.op. left ventricle, the posteriror myocardium and septum contract lamely, ef about 35%, mitral valve 1-2/4 leak. Aortic cross-clamp time 1 h 32 min. Post.op. ef over 40%. On basis of the UKG-finding pre.op. Simdax-infusion was initiated. On arrival to ICU, haemodynamics was stable, norepinephrine administered. Cardiac index 3,2. Warming-up and weaning in ventilator.

Finnish original:
61-vuotias nainen jolla Crohnin tauti. Salossa ollessaan osallistunut pyöräilytapahtumaan, edelträvrästi flunssaa. Pyöräilyn jrälkeen hengitysvaikeuksien takia TYKSin ensiapuun ja todettu sydränentsyymit kohonneiksi ja inkompensaatiota. Otettu teho-osastolle inkompensaation ja pneumonian hoitoon. UKG:ssa 2.6. ef 30%. Koronaariangiossa merkitträvrät stenoosit RCA:ssa, LCX:ssa ja prärärungossa. Tränrärän elektiivinen neljrän suonen ohitus LITA-LAD, Ao-LOM-LPL ja Ao-RBD, joihin hyvrät virtaukset. Pre.op. vasemman kammion takaseinrä ja septum supistuvat vaisusti, ef noin 35%, mitraaliläpässä 1-2/4 vuoto. Aortan sulkuaika 1 t 32 min. Post.op. ef yli 40%. UKG-löydöksen perusteella potilaalle aloitettu jo pre.op. Simdax-infuusio. Teho-osastolle saapuessa hemodynamiikka stabiilia, noradrenaliini menossa. Cardiac index 3,2. Lämmitys ja vieroitus respiraattorissa.

Figure 2.5: Example of a clinical note. This is a fake case originally created in Finnish by domain experts, then translated into English. Common misspellings are included intentionally.
This is the same example as in Figure 2 in Moen et al. (2016), Paper E in this thesis.

Developing methods and systems for clinical NLP is hard for a number of reasons. Some of the main challenges are: lack of data available to software developers and researchers, primarily due to the sensitivity of clinical information/text; lack of existing and robust text processing resources that support the broad range of (sub-) languages and applications, such as text parsers, computerized thesauri and ontologies; the cost of developing new or customized methods and resources for processing the text; the issues related to integrating NLP software into the clinical practice through existing and new information systems (Chapman 2010, Rector 1999, Kate 2012, Friedman et al. 2013, Meystre et al. 2008, Grabar et al. 2009, Kvist et al. 2011).

NLP has been applied to clinical text for a variety of tasks. Some examples are automatic event detection in health records (Mendonça et al. 2005), automatic concept indexing (Berman 2004), medication support (Xu et al. 2010), decision support (Demner-Fushman et al. 2009, Velupillai and Kvist 2012), query-based search (Grabar et al. 2009)) and automated summarization (Pivovarov and Elhadad 2015). Openly available tools designed for clinical NLP typically dependent on specialized and extensive knowledge resources to classify and reason with concepts in the text. Such resources that are commonly built in a manual fashion (see Section 2.1.2 for more information). Further, due to the domain specificity of clinical text, generic resources for computational semantics tend to be of limited use. However, the use of distributional semantic methods in this domain is promising due to their focus on learning semantic relations directly from unannotated corpora. This enables acquisition of semantic resources in an resource-lean manner.

Distributional semantic methods utilize statistics that are derived from the corpus used for training, thus the resulting models and its constituent context vectors will reflect the semantic similarity relations that are found in the utilized training corpus, which again reflects the domain of the corpus. Koopman et al. (2012) show that the domain-specificity of the corpora used for training such distributional semantic models is important for the content and quality of the resulting model with respect to the intended task. Pedersen et al. (2007) explore a set of measures for automatically judging semantic similarity and relatedness among medical concept pairs whose semantic similarity have been pre-assessed by human experts. These range from various measures based on lexical resources (WordNet, SNOMED-CT, UMLS, Mayo Clinic Thesaurus) to one based on a distributional semantic model trained on a corpus of unannotated clinical text. The latter measure uses the LSA method, and the semantic similarity between concept pairs is calculated using the cosine similarity measure applied to *concept context vectors* — assembled from the corresponding words. Pedersen et al. (2007) find that this measure performed at least as well as any of the other measures. Related work has also shown that distributional semantic models, induced automatically from large corpora of clinical text, or other types of medical text, are well suited as a fast and cost-efficient approach to capturing and representing domain-specific terminology (Koopman et al. 2012, Cohen and Widdows 2009, Cohen et al. 2014, De Vine et al. 2014).

As an example, a semantic model trained with W2V CBOW on a fairly large corpus of clinical free-text notes is able to detect that the words “pain” and “discomfort” have a similar meaning or contextual use because they have a relatively high cosine similarity value — relative to the other words in the model. However, if some other type of corpus was used for training, e.g., a collection of newspaper articles, the resulting model would not necessarily contain the same semantic relations. The same goes for other domain-specific terms that clinicians use when documenting care. A number of these would not even be present in a newspaper corpus, let alone abbreviations and spelling mistakes that are common in clinical text. Paper A explores various combinations of distributional semantic models, trained on one of two different corpora — one containing clinical text and the other medical research literature — and evaluates these on the tasks of automatic extraction of synonyms and abbreviation-expansion pairs.

Karlgren and Sahlgren (2001) argue that text models and methods for constructing/training them still have a way to go in terms of capturing the actual language use, rather than the language in abstract. Most distributional semantic methods construct word space models that contain one context vector per unique word. This means that each word will have one semantic meaning, representing its “prototypicality”, relative to the others in the semantic model, accumulated from all its occurrences with the utilized training features, e.g., neighboring words, in the training corpus. However, in reality the meaning of a word may vary greatly based on the context of its use, thus each word could potentially have multiple meanings or senses. Such words are referred to as polysemes or homonyms (Panman 1982). As an example, the word “discomfort” may refer to some type of physical pain, or it may refer to psychological/social inconvenience. Another example is the word “rock”, which may refer to a type of music or a material. One direction in distributional semantics concerns training vector spaces that allow words to potentially have more than one representation, i.e., multiple context vectors (Reisinger and Mooney 2010, Schütze 1998, Neelakantan et al. 2014). This would intuitively be beneficial in a range of semantic similarity assessment tasks, including tasks in the clinical domain. Arguably these type of methods may enhance the information complexity represented in the resulting models, as well as their discriminative capabilities — discrimination between different local meanings of words found in the training corpus. This may further enhance the task and domain specificity of semantic models.

Having potentially multiple context vectors for each word from the training corpus means that a large(r) number of vectors have to be stored in the computer memory, in particular during training. In the approaches by Reisinger and Mooney (2010) and Schütze (1998) every contextual occurrence of each word throughout the training corpus is stored in memory before applying some type of clustering. Paper B explores a novel "multi-sense", or "multi-prototype", distributional semantic method that performs incremental clustering as part of the training phase. It builds on the RI method and retains the properties of RI concerning reduced dimensionality and on-line training. We evaluate this method at a semantic textual similarity task (Agirre et al. 2013), where the goal is to automatically assess and classify similarities between sentence pairs.

In Friedman et al. (2013) it is suggested that future work in clinical NLP should aim to exploit existing knowledge bases about medications, treatments, diseases, symptoms, and care plans, despite these not having been explicitly built for the purpose of clinical NLP. One way to potentially improve the task- and domain-specificity of distributional semantic models is to exploit more domain specific features of the training corpus for constriction. This may assist in forming a semantic space that better reflects the semantic relations of interest. Paper C explores the use of ICD-10-code labels (see Section 2.1) as training features in an attempt to induce the underlying relations into a distributional semantic model. This is used in a set of experiments that explores various ways of constructing distributional semantic models for the task of *care episode retrieval*, using only the free-text information in clinical notes for the retrieval process.

2.3 Automatic Text Summarization of Clinical Text

(Jones 1999) presents *factors* that one has to take into account in order to make a summarization system achieve its task. The three main categories of factors described are *input*, *purpose* and *output*. These factors have also been discussed and elaborated upon by (Hahn and Mani 2000, Afantenos et al. 2005). Following are the factors and their underlying properties and recommendations that we have identified as being most suited relative to the research goal (c.f. RG1)[6]. For further details about these factors, please see the mentioned papers.

Input Factors

- *Single document or multiple documents*: As the task is to summarize clinical free text notes written during care episodes, the system input would primarily be multiple documents — sequences of multiple clinical notes constituting care episodes — one care episode per summary that is to be produced.

[6] These properties and recommendations are the result of a literature study conducted relatively early on in the PhD period, but is currently unpublished material.

- *Structure*: Information about the structure of the document or documents can help in classifying the content. As each clinical note are parts of a continuous patient story, the approach should have a scope that covers the full care episode when assessing what the most relevant information is. If a predictable document/note structure is used by clinicians, this should be exploited.

- *Language*: The language specificity of the system is commonly determined by the underlying resources and tools that it relies on. Further, a knowledge-poor approach would potentially enable easy adaptation to a broad range of languages and sub-languages at a low cost.

Purpose Factors

- *Indicative, informative, and/or critical*: We strive towards a system that is able to provide an indicative overview of the free text documented for care episodes. Together with structured data (such as laboratory test results, images, diagnostic codes and personal information) it could help clinicians to quickly familiarize themselves with the content of individual care episodes and the patients problems, which is particularly useful if such information is needed urgently.

- *Generic or user-oriented*: The system should be both generic and user-oriented in order to meet the specialized information needs of clinicians. However, for (automated) evaluation purposes, we believe that producing generic summaries is the first thing to aim for.

Output Factors

- *Extracts or abstracts*: We aim for a extraction-based summarization approach, in which the summary is generated by selecting a subset of sentences from the relevant text. This approach is viable because a sizeable portion of clinical text summaries, such as discharge summaries, are created by copying or deriving information from clinical notes (Van Vleck et al. 2007, Sørby and Nytrø 2005, Meng et al. 2005, Wrenn et al. 2010).

- *Available domain knowledge*: It is common to distinguish between "knowledge-rich" and "knowledge-poor" systems based on the availability of data and domain knowledge resources for the system to exploit. As already mentioned, in particular for small (clinical) languages, few such specialized resources exists.

- *Output format*: The output should, at least as an initial approach, have a format that is similar to the notes that clinicians produce themselves the care process to enable automated evaluation against existing summaries (c.f. gold standard).
- *Quality (evaluation)*: The quality of a summarization system is commonly measured by its content-selection capability, presented as its output. Using manually created summaries — a so called *gold standard* — for comparison is a common way to evaluate the quality of a summarization system. However, creating manual summaries is a expensive and time-consuming process. We suggest exploring summaries constructed/written by clinicians during the care process for this purpose (see Paper D and E).

The most central issue in text summarization is to determine what information to include in a summary. In *extraction-based summarization* this concerns selecting a subset of sentences from the text that is to be summarized. Common techniques here are: Topic-based extraction (see, e.g., Carbonell and Goldstein (1998), Goldstein et al. (2000), Steinberger and Křišt'an (2007)), where relevance scores for sentences are computed with respect to one or more topics of interest; Centrality-based extraction (Patil and Brazdil 2007, Chatterjee and Mohan 2007, Erkan and Radev 2004, Mihalcea and Tarau 2004), where typically some underlying graph-based representation is used to calculate sentence significance based on the document coverage of the sentences relative to the other sentences. An important sub-task when applying these techniques is to avoid redundant information in the produced summaries. For this purpose it is common to apply some ways of checking for textual similarity overlap based on the Maximal Marginal Relevance (MMR) criterion (Carbonell and Goldstein 1998) or similar techniques. Distributional semantic models, in various forms, have been quite extensively used in the field of text summarization, see e.g., Luhn (1958), Chatterjee and Mohan (2007), Hassel and Sjöbergh (2007), Nenkova and McKeown (2011).

Several pieces of work have been identified focusing on the task of automatically generating summaries from the text in clinical notes. Liu (2009) uses the MEAD summarization toolkit. Van Vleck et al. (2007) perform structured interviews to identify and classify phrases that clinicians considered relevant to explaining the patient's history. Meng et al. (2005) use an annotated training corpus together with tailored semantic patterns to determine what information that should be repeated in a new clinical note or summary. Velupillai and Kvist (2012) focus on recognizing diagnostic statements in clinical text, learned from an annotated training corpus, and further to classify these based on the level of certainty. Extracted diagnostic statements are then used to produce a text summary. Bashyam et al. (2009), Hirsch et al. (2015) reports the work on extensive clinical summarization systems. These apply various information extraction tools and resources to identify, classify and reason with entities and information in free text. Visualization is also an important part of these systems, including timeline-based visualization in Hirsch et al. (2015). Others have worked on more conceptual models for understanding and supporting generation of information summaries in the clinical domain (Sarkar et al. 2011, Abulkhair et al. 2013). However, to the best of my knowledge, summarization of clinical free-text information has been pursued by relatively few researchers. This is not surprising given the challenges related to clinical NLP and the task. This is also a prominent issue considering recent reviews and related work (Mishra et al. 2014, Pivovarov and Elhadad 2015, Kvist et al. 2011).

In extractive multi-document summarization there is a need to have the computer "understand" the *terminology* of, or semantic similarities between, the candidate sentences to determine if some information is repeated, redundant, or similar to some topic or query (Ferreira et al. 2016). For this task, distributional semantic models are commonly used. In Paper E we explore various distributional semantic models at the task of summarizing clinical notes for individual care episodes. We focus on exploring a resource-light approach that circumvents the need for manually developed knowledge and training resources tailored for the task.

Computer generated summaries are typically evaluated by comparing the summary with a *gold standard*, being one or more reference summaries that has been constructed manually, often in relation to a shared task[7]. To perform this comparison in an automated fashion, a computerized similarity metric is used. The *ROUGE evaluation package* (Recall Oriented Understudy for Gisting Evaluation) (Lin 2004) evaluates text similarity based on N-gram overlap. ROUGE metrics are commonly used in text summarization evaluation because its scores have shown to correlate well with human judgements (Lin 2004). Liu (2009) performs automatic evaluation of computer generated summaries of clinical notes by using the original discharge reports as gold summaries. An alternative type of evaluation is to do the content assessment manually. Lissauer et al. (1991) evaluate computer generated discharge summaries from neonate's reports by analysing if they contain the required information according to a guideline. In Papers D and E we apply the ROUGE evaluation package for evaluation of automatically generated clinical free-text summaries where the original discharge summaries are used as gold standard. Evaluation scores are then compared to manual evaluation, conducted in a

[7] A 'shared task' is here defined as a specific task proposed and organized by a dedicated committee that provide the necessary data and evaluation setup to the participants. A shared task is typically held in relation to a conference where there are multiple participating research groups who are competing to achieve the best results.

similar fashion as in the work by Lissauer et al. (1991).

Chapter 3

Paper Summaries

The first experiments focused on automated assessment of word-level similarities, which resulted in Paper A. More precisely this study concerned automatic detection of synonymic relations between words, including between full form words and their abbreviations. One motivation behind these experiments was to get an insight into what methods and parameters that produces models that best captures synonymic relations on a word level. Further, it is likely that a similar setup would also apply to sentence-level semantics — a central part of (sentence-level) extraction-based text summarization (c.f. Paper E). The next set of experiments, presented in Paper B, focused on sentence-level semantic textual similarity assessment. Here a method that performs automatic word-sense discrimination was evaluated. In relation to the work on automatic text summarization of clinical free-text notes, I wanted a way to retrieve care episodes that are similar to a target care episode, belonging to other (patients') hospital stays. This resulted in the work on care episode retrieval, presented in Paper C. The last set of experiments is related to automatic generation of summaries from clinical free text, from one care episode at a time, and evaluation of the generated summaries. This work is presented in Papers D and E. The work in Paper E is to a large extend based on the lessons learned from previous papers/experiments and includes many of the approaches, methods and models therein.

Below is an overview of the five papers in this thesis. For each paper there is: a *summary* of the main content; followed by a *retrospective view* discussing the work from a possibly more enlightened perspective.

3.1 Paper A: Synonym extraction and abbreviation expansion with ensembles of semantic spaces

Authors: *Aron Henriksson, Hans Moen, Maria Skeppstedt, Vidas Daudaravičius and Martin Duneld*

3.1.1 Summary

Terminologies that account for variation in language use by linking synonyms and abbreviations to their corresponding concepts are important for enabling automated semantic similarity assessment and high-quality information extraction from clinical texts. Due to the use of specialized sub-languages in the clinical domain, manual construction of semantic resources that accurately reflect language use is both costly and challenging.

In this paper we explore the use of distributional semantic models for automated extraction of synonymic relations from clinical/medical text, including abbreviations (abbreviations to long forms and long forms to abbreviations). Models are trained on one or both of two corpora, one corpus consisting of Swedish clinical text and another consisting of Swedish medical journal articles. Two approaches to constructing the models are used, classic Random Indexing (RI) and the Random Permutations (RP) variant. Various model training parameters and ways of combining the retrieved candidate words/synonyms are explored.

Evaluation is done using two gold standards. For the synonym extraction task, MeSH terms and associated synonyms are used. And for the other two tasks (abbreviations to long forms and long forms to abbreviations), we used a list of medical abbreviation-expansion pairs. Only single-word terms were used for evaluation. For each query a list of ten candidate words are retrieved by the model(s) being evaluated. The results are measured primarily as recall among these ten (R, top 10), calculated from the proportion of expected candidate words that are among these. When combining the results from two or more models, their retrieved lists of scored candidate words are combined through summing or averaging over matching words.

We also explore the use of some post-processing filtering. For abbreviation-expansion extraction, the filtering is based on word-length threshold filtering and on checking for overlapping letters and their matching order. For synonym extraction, the filtering rule checks if the retrieved candidates has a cosine similarity above a given thresholds, together with checking if their rank (in the retrieved list) are above a given threshold. Also different word frequency thresholds are explored, i.e. words below a given threshold are removed from the gold standards.

We found that a combination of the two models (RI + RP), trained on a single corpus outperforms the use of one model in isolation. Furthermore, combining semantic models induced from the two different types of corpora further improved the results (RI$_{clinical}$ + RI$_{medical}$ + RP$_{clinical}$ + RP$_{medical}$), also outperforming the use of a conjoint corpus (RI$_{clinical+medical}$ + RP$_{clinical+medical}$). A combination strategy that simply sums the cosine similarity scores of candidate words — retrieved from each model — is generally the best performing one. Finally, applying the post-processing filtering rules yielded substantial performance gains on the tasks of extracting abbreviation-expansion pairs, but this is not the case for synonym extraction. A word frequency threshold in the range of 30-50 seems to be optimal. Best results achieved R = 0.39 for abbreviations to long forms (R$_{baseline}$ = 0.23), R = 0.33 for long forms to abbreviations (R$_{baseline}$ = 0.24), and R = 0.47 for synonyms (R$_{baseline}$ = 0.39).

This study demonstrates that ensembles of semantic models can yield improved performance on the tasks of automatically extracting synonyms and abbreviation-expansion pairs — improvements compared to using a single model. Further, this encourages further exploration in utilizing and combining different semantic models, trained with different parameters and context features, and/or trained on different types of corpora. This also includes exploring different ways of combining the model outputs during the retrieval phase. The methods, models and model combinations in this study could potentially be used in (semi-) automated terminology development in the clinical/medical domain, as well as in a range of other NLP tasks.

3.1.2 Retrospective View and Results/Contributions

This study gave valuable directions and insight into how to generate semantic models from clinical/medical free text that capture word-level similarities, reflecting similarity in terms of having the same or closely related synonymic meaning. Experienced gained here were important for further work on distributional semantic similarity models.

From a retrospective view, it would have been interesting to see how these methods, RI and RP, fare at the given tasks in comparison to, or in combination with, other distributed semantic methods/models such as LSA and more recent neural network-based methods, such as W2V CBOW and Skip-gram. Fundamentally different methods that does not rely on distributional semantics, such as more rule-based methods (e.g., Ao and Takagi (2005)), are likely to perform well when it comes to detecting relations between terms and their abbreviations.

3.2 Paper B: Towards Dynamic Word Sense Discrimination with Random Indexing

Authors: *Hans Moen, Erwin Marsi and Björn Gambäck*

3.2.1 Summary

Most distributional models of word similarity represent a word type by a single vector of contextual features, even though words often have more than one lexical sense (Reisinger and Mooney 2010, Huang et al. 2012). In this paper we present a novel method for learning and constructing a distributional semantic model that may contain more than one context vector, or "sense vector", for each unique word in the utilized training corpus. A common way of capturing multiple senses per word with the distributional semantic approach is to first construct and store one vector for each occurrence of a word in the training corpus — storing the features of each single word use. Then these vectors may be clustered in some way to create sense vectors. However, storing and clustering these vectors can be expensive as it generates a set of vectors equal to the word count of the training corpus. As an alternative, we introduce *Multi-Sense Random Indexing*, that performs on-the-fly incremental clustering of word senses, allowing multiple senses per unique word in the training corpus. A range of different measures for sentence similarity are explored that focus primarily on deriving these from the maximum bipartite similarities between the underlying words and their different senses. Various measures for word-sense alignment are illustrated in Figure 3.1.

For training the semantic models we use the CLEF 2004–2008 English corpus (CLE 2004). We use the STS 2012 and STS 2013 shared tasks' evaluation data (Agirre et al. 2012; 2013) for sentence similarity assessment, where the task is to score the similarity between sentence pairs with a number between 0 and 5. A *support vector regressor*[1] is trained/optimized using the training data accompanying the STS task for the purpose of mapping the cosine similarity scores to a final score between 0 and 5. The various multi-sense based performance scores are compared to those from using a classic (single sense) RI model. Performance scores are calculated using the mean Pearson product-moment correlation coefficient (PPMCC) (Lehman 2005), same as in the mentioned STS tasks.

Our experimental results did not show a clear systematic difference between single-prototype and multi-prototype models. The highest scores were achieved on the STS 2013 evaluation data, with a mean PPMCC = 0.46 with the multi-sense Hungarian Algorithm-based similarity measure, compared to a mean PPMCC = 0.45

[1] `http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html` (accessed 1st March 2016)

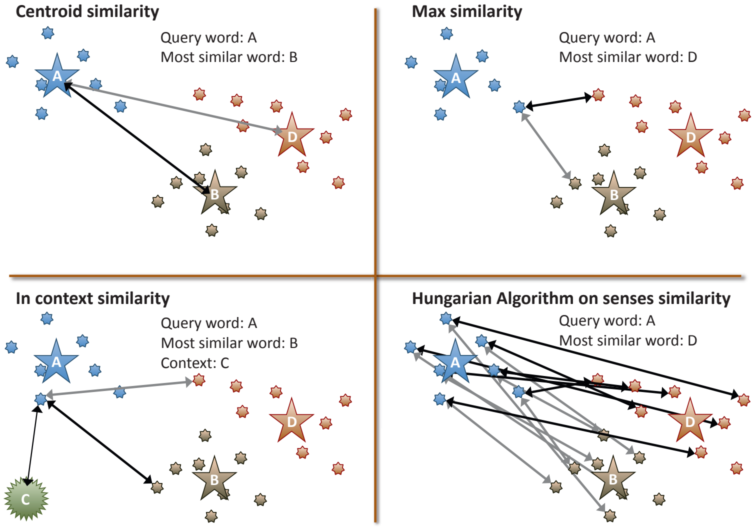


Figure 3.1: Various similarity measures tested with the multi-sense vector space model. In this 2D illustration the relative distances between words and senses reflect how similar they are. Large stars represent the centroid location of words, and the small stars represent their underlying senses.

achieved with single-sense Hungarian Algorithm-based similarity measure (see Kuhn (1955) for more information on the Hungarian Algorithm).

3.2.2 Retrospective View and Results/Contributions

The motivation for the multi-sense method introduced in this paper was to see if we could better capture the meaning of words by creating semantic models that learn potentially more than one context vector per word, i.e., “sense”. At the same time we wanted to retain the incremental and compressed dimensionality features of the RI method. When calculating the similarity between two sentences, the aim was to select the most appropriate sense vector for each word based on a) the context defined by the other words in the same sentence, or b) the words (and their senses) in the other sentence. Then we calculated sentence similarity from word-pairs using the Hungarian Algorithm for word alignment, or composed two sentence context vectors before calculating their similarity. The results from the latter approach was not included in this paper since the Hungarian Algorithm-based approach generally performed better at the task. However, results using such sentence context vectors with TF*IDF weighting have been reported in Marsi et al. (2013).

Table 3 in Marsi et al. (2013) shows that the top three strongest single similarity features, individually trained using a support vector regressor (Vapnik et al. 1997), are those based on character n-gram overlap. Although the approach rely on manually classified training data for the regressor, it is worth noting that this rather simple approach performs well when compared to the more complex ones.

This study, together with that presented in Marsi et al. (2013), provided us with valuable insight into sentence similarity assessment. This was also an opportunity to compare a variety of different approaches, including some not relying on distributional semantic models. In the experiments conducted in Marsi et al. (2013), a support vector regressor was used to learn a function that, for each sentence pair, take a range of different sentence similarity features as input, and the output is a single similarity score that is optimized to reflect some given training instances (manually scored with values in the range 0 to 5). Here it became clear that individual similarity features, also those achieving relatively weak scores individually, would contribute to increasing the final score when used in combination with others. Further, the lessons learned here were important to the sentence similarity calculations and clustering used in the summarization task presented in Paper E.

Despite that the results presented in this paper did not show systematic improvements over the evaluation scores gained by *not* using the multi-sense method, this approach calls for further research. The more recent publication by Neelakantan et al. (2014), who applied a similar training algorithm as ours in their multiple-sense (W2V) Skip-gram-based method, indicates that this direction in distributional semantics has a certain actuality. Also, after publishing our paper, we did more exploration with various parameters, and were able to improve the mean PPMCC of the multi-sense Hungarian Algorithm-based similarity measure to 0.49, up from 0.46, on the STS 2013 evaluation data. In the future, it would also be interesting to evaluate this method on other tasks than sentence-level similarity assessment.

There are still many unresolved and open questions regarding parameters and training features to use during training and how to do the word and sense extraction/retrieval and similarity calculations. Finally, if such an approach is able to improve upon the existing state-of-the-art in automated sentence similarity assessment, there is little doubt that it should also have a positive influence on tasks such as automatic terminology development and (clinical) text summarization.

3.3 Paper C: Care Episode Retrieval: Distributional Semantic Models for Information Retrieval in the Clinical Domain

Authors: *Hans Moen, Filip Ginter, Erwin Marsi, Laura-Maria Peltonen, Tapio Salakoski and Sanna Salanterä*

3.3.1 Summary

Electronic health records (EHRs) are used throughout the health care sector by professionals, administrators and patients, primarily for clinical purposes, but they are also used for secondary purposes such as decision support and research. The vast amounts of information in EHR systems complicate information management and increase the risk of information overload. Therefore, clinicians and researchers need new tools to manage the information stored in the EHRs. A common use case is, given a — possibly unfinished — care episode, to retrieve the most similar care episodes among the records.

This paper presents several methods for information retrieval, focusing on the task of *care episode retrieval*. The experimental setup is illustrated in Figure 3.2. Care episode similarity is calculated based on their textual content. This is achieved through constructing different distributional semantics models from a corpus of clinical text, and then applying the cosine similarity measure. Methods used to construct these models include variants of RI and W2V. A novel model construction approach is introduced that utilize the ICD-10 codes attached to care episodes as training features to better induce domain-specificity in the resulting distributional semantic model. When calculating the similarity between care episode pairs, we explore a set of different approaches for aligning and comparing them in terms of their underlying clinical notes.

We report on experimental evaluation of care episode retrieval that circumvents the lack of human judgements regarding episode relevance. Results are reported as: precision among the top-10 retrieved care episodes (P@10); precision at the R-th position in the results (Rprec), where R is the number of correct entries in the gold standard; mean of the average precision over all queries (MAP). The results suggest that several of the proposed methods outperform a state-of-the-art search engine (Lucene) on the retrieval task. The best results were achieved when using the ICD-10-based semantic model, constructed using a modification of the W2V Skip-gram algorithm (W2V-ICD), and when treating care episodes as single conjoint text documents (i.e., not as a series individual clinical notes). On this setup, the best performing method, W2V-ICD, achieved: MAP = 0.2666, P@10 = 0.3975, Rprec = 0.2874. In comparison, on the same setup, Lucene achieved: MAP = 0.1210, P@10 = 0.2800, Rprec = 0.1527. And for the random baseline the

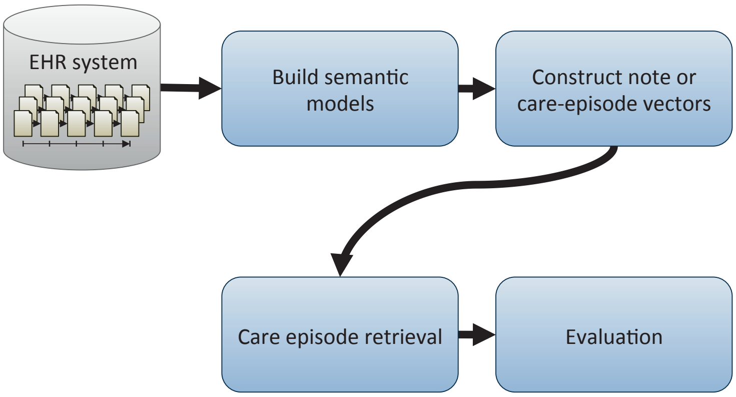


Figure 3.2: Illustration of the care episode retrieval experiments in Paper C. This is Figure 2 in Moen et al. (2015), Paper C in this thesis.

scores were: MAP = 0.0178, P@10 = 0.0175, Rprec = 0.0172.

3.3.2 Retrospective View and Results/Contributions

This paper was a continuation of the work presented in Moen et al. (2014b). This study focuses on exploring a set of distributional semantic models for use in retrieval of care episodes, only relying on the free-text information therein. One motivation for conducting this research was that we wanted to use this type of care episode retrieval in the summarization task presented in Paper E.

Such multi-document (multi-note) information retrieval — where also the query is a care episode, i.e., a collection of documents/notes — is a rather unique task as far as I know. Naturally, finding a reliable way of conducting automated evaluation was a central issue here. The number of result tables became fairly large due to the fact that we were evaluating eight different models/systems on two different evaluation setups. The results indicate that using ICD-10 codes as context for training the semantic models seems to be a promising direction for information retrieval on the level of care episodes. It is likely that using other training features from clinical practice, commonly documented in EHRs, could potentially produce even better semantic spaces for this task, and in general — models more suited for information access and NLP in the clinical domain. The use of such domain-specific context features in constructing semantic models would also be interesting to evaluate on the level of word synonyms, e.g., similar to the experiment in Paper A.

3.4 Paper D: On Evaluation of Automatically Generated Clinical Discharge Summaries

Authors: *Hans Moen, Juho Heimonen, Laura-Maria Murtola, Antti Airola, Tapio Pahikkala, Virpi Terävä, Riitta Danielsson-Ojala, Tapio Salakoski and Sanna Salanterä*

3.4.1 Summary

Proper evaluation is crucial for developing high-quality computerized text summarization systems. In the clinical domain, the specialized information needs of the clinicians complicate the task of evaluating automatically constructed text summaries — constructed from the free-text information that clinicians document in relation to patients' care episodes. The focus of this paper is on evaluation. We propose an automated and manual evaluation approach. We are not interested in the actual performance of some summarization method, instead the focus is on determining if, and to what degree, there is a correlation between how the automated and manual evaluation approaches rank various summarization methods. We assume that the manual evaluation scores are good indicators for relative performance, however this is not clear for the automated evaluation measures in question. Further, if such a correlation is observed, we may rely on the much faster automated evaluation when further developing the summarization methods. The experiment setup is illustrated in Figure 3.3.

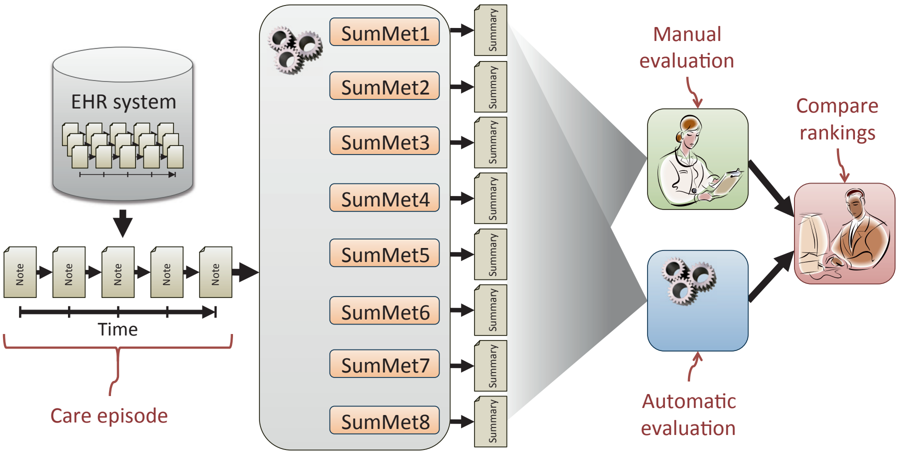


Figure 3.3: Illustration of the evaluation experiment conducted in Paper D. This is a slightly modified version of Figure 1 in Moen et al. (2014a), Paper D in this thesis.

Four different evaluation measures in the *ROUGE evaluation toolkit* are explored in the automated evaluation approach, where the utilized gold standard for each summarized care episode is the accompanying discharge summary. The manual evaluation is performed by domain experts who use an evaluation scheme/tool that we developed as part of this study. The scores from the manual evaluation is calculated from the average summarization method scores from five care episodes. The scores from the automatic evaluation is based on the average scores from 156 care episodes. To identify which of the automatic evaluation metrics that best follows the manual evaluation, Pearson product-moment correlation coefficient (PPMCC) and Spearman's rank correlation coefficient (Spearman's rho) were calculated between the normalized manual evaluation scores and each of the automatic evaluation scores.

We find that all ROUGE measures correlate well with that of the manual evaluation, where the ROUGE-SU4 measure correlates the most. It achieves: PPMCC = 0.9510 (p-value = 0.00028), and Spearman's rho score = 0.8571 (p-value = 0.00653). The agreement among the manual evaluators is "good" according to guidelines on interrater agreement. These preliminary results indicate that the utilized automatic evaluation setup can be used as an automated and reliable way to rank clinical summarization methods internally in terms of their performance. This allows us to rely on the presented automatic evaluation approach when further developing automatic text summarization for clinical text.

3.4.2 Retrospective View and Results/Contributions

A central issue in the works related to this thesis has been to enquire evaluation data suited for evaluating the different methods that have been introduced along the way. Automatic text summarization is in itself a very complex and challenging task, and to assess what is a good or poor summary is heavily influenced by the perspective of the judging subject and the underlying task. This is also the case when it comes to the task of generating and evaluating clinical free-text summaries. To *automate* such evaluation complicates things further. These are the reasons why in this paper we chose to try to generate a summary that is comparable to discharge summaries from care episodes — which clinicians construct/write manually as a part of the patient discharge process. This enables the use of hospital guidelines for manual judgement of the content and quality of a summary (c.f. the utilized manual evaluation scheme/tool). Further, this enables us to use the original discharge summaries as gold standard when performing automated evaluation. However, the evaluation approach does not provide any absolute truth when it comes to how the summarization methods performs, but primarily how they perform in relation to each other.

The main results of this experiment was that a correlation was found between: a) how human evaluators rank a set of different summarization methods, and b) how some automated evaluation metrics rank these same summarization methods. With this knowledge we could rely on the automated evaluation approach for rapid evaluation during further experimentation and development of summarization methods. This was a important step in the process of developing some of the summarization methods presented in Paper E.

3.5 Paper E: Comparison of automatic summarisation methods for clinical free text notes

Authors: *Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola,*
Tapio Pahikkala, Tapio Salakoski and Sanna Salanterä

3.5.1 Summary

Managing the information in EHR systems tends to be time consuming for clinicians (Farri et al. 2012, Hirsch et al. 2015). Automatic text summarization could assist in providing an overview of the free-text information in ongoing or finished care episodes, as well as in writing the final discharge summaries. This work focuses on summarization of the clinical free text written by clinicians (physician) in care episodes. We evaluated eight different automatic text summarization methods. Among these are four novel extraction-based text summarization methods, tailored for summarizing the free-text content of care episodes. A key feature of these methods is that they try to take into account the sequential and repetitive nature of the documented text/information. Most of them rely on the use of distributional semantic models, exploiting various textual features found in care episodes.

Care episodes used in this study are from EHRs belonging to heart patients admitted to a university hospital in Finland. The performance of the summarization methods are evaluated both automatically and manually. We utilized the *ROUGE evaluation toolkit* for automatic evaluation with discharge summaries used as gold standard, while a rating-based evaluation scheme/tool is used for the manual evaluation. By comparing how the automatic and manual evaluations correlates in terms of how they rank the different summarization method, we are able to perform a meta-evaluation of these ROUGE evaluation measures. Figure 3.4 shows an overview of the experimental setup.

The results show that there is a good agreement between the manual evaluators.

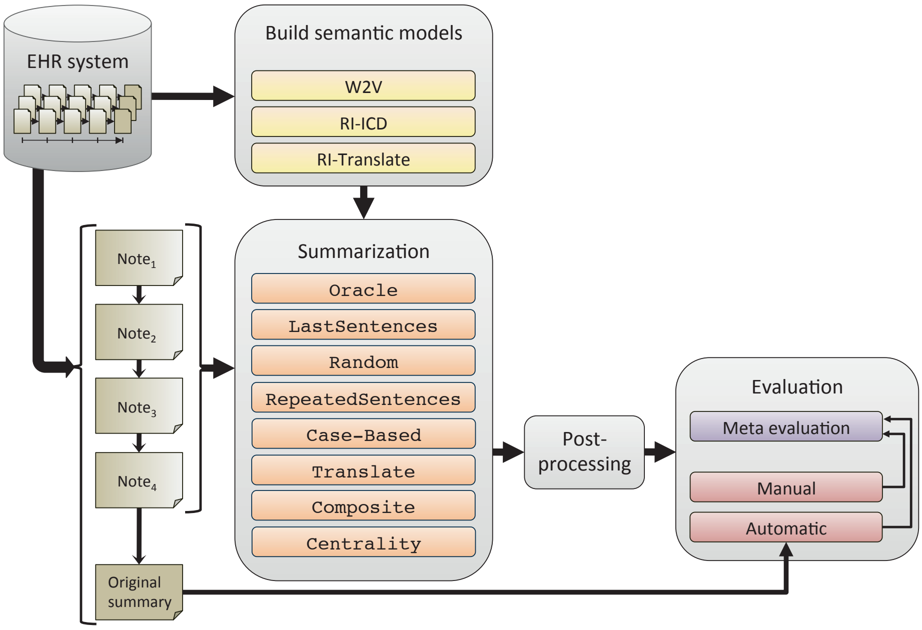


Figure 3.4: Illustration of the text summarization experiments conducted in Paper E. This is a slightly modified version of Figure 1 in Moen et al. (2016), Paper E in this thesis.

There is also a high correlation between how the manual evaluators and the automated evaluation rank the various summarization methods. Here the ROUGE-N2 and ROUGE-l metrics have the highest correlation with the manual evaluators. The high correlation between manual and automated evaluations suggests that the less labor-intensive automated evaluations can be used as a proxy for human evaluations when developing summarization methods. This is of significant practical value for summarization method development aimed at this task. Both the automated and manual evaluations agree in that a proposed composition based summarization method outperforms all the other considered methods.

3.5.2 Retrospective View and Results/Contributions

Here the focus was on exploring a set of different methods, exploring various features of clinical free-text information, for performing the automated summarization. They all rely solely on exploiting statistical features that are found in EHRs/care episodes, without the use of any manual annotation or similar manual work tailored for this task. This again reflects the underlying motivation for our approach; to explore ways of conducting such summarization while surpassing the need for developing NLP resources tailored for the task.

The utilized automatic evaluation approach was the same as that used in Paper D. However, for the manual evaluation, a somewhat simplified evaluation scheme (compared to that used in Paper D) was used by the human evaluators. The evaluators found the original evaluation scheme from Paper D to be very time-consuming to use due to its complex , thus a simplification was done in order to allow for evaluation of more summaries within certain time and resource limits. Yet, the results showed that a correlation between the manual and automatic evaluation was present also in this experiment.

With today's hospital practice, the optimal text summary generated from care episodes through sentence-level extraction-based text summarization will hardly ever become identical to a corresponding discharge summary written by a clinician. One reason for this is that much of the information that clinicians write in a discharge summary is never written in the clinical (daily) notes that constitute a care episode. However, we demonstrate here that there are certain sentence-level textual features that can be indicative of sentences inclusion potential in a (discharge) summary.

Future work on this task includes further developing these methods so that they can be used for assisting clinicians through semi-automatic, user-guided, discharge summary writing. There are of course also other summarization approaches and methods that should be explored, including exploiting other (statistical) features of care episodes and clinical text.

The summarization methods explored in this paper are potentially suited for presenting an *indicative overview* of the free-text content written in a — possible ongoing — care episode, that clinicians could read in situations where they do not have time to read through all previously written clinical notes. This would be supplementary to a comprehensive overview/visualization of the more structured and coded data in EHRs, such as images, laboratory test results, medications, diagnosis codes (see Roque et al. (2010), Pivovarov and Elhadad (2015)). Further, situations where they could find use would be where the need for such an overview outweighs the possible patient safety issues that may be caused by lack of relevant information in the generated summary.

Future work should focus more towards conducting *extrinsic evaluation* — evaluating how the use of automatic free-text summarization systems in a (simulated) clinical setting will impact documentation speed and quality, as well as health care quality and patient outcomes.

Chapter 4

Conclusions and Recommendations for Future Work

This thesis has focused on distributional semantic methods used to construct domain and task specific semantic models from primarily clinical text. Five sets of experiments have been presented, published as separate papers, that focused on different applications of distributional semantic models. Three of these focus on textual similarity assessment on different granularity levels: words (Paper A), sentences (Paper B), and clinical notes (Paper C), where some novel ways of training the utilized distributional semantic models were presented and evaluated. The other two papers, Papers D and E, focus on applications of semantic models in free-text summarization methods tailored for clinical text, as well as evaluations of these. Both existing and novel approaches and methods have been applied and evaluated in these experiments.

4.1 Conclusions

Three research questions were presented in Section 1.2. Here these are linked to the experiments in the various papers, discussed and concluded.

Research Question 1

How can the distributional hypothesis be utilized in constructing semantic similarity models suited for clinical text?

In Paper A we found that combining distributional semantic models trained differently can yield improved performance in terms of modeling word synonym relations. First, combining the retrieved candidate words from two models (Random Indexing (RI) and Random Permutation (RP)), both trained on the same corpus seemingly enhances the synonymic relations between the query and the resulting top extracted candidate words, outperforming the use of one model alone. Second, combining four models, trained using RI and RP on one of two different training corpora — one consisting of clinical text and the other medical research literature — improves the results further (RI$_{clinical}$ + RI$_{medical}$ + RP$_{clinical}$ + RP$_{medical}$). This also outperforms the approach of using a conjoint corpus for training. This suggests that such a multi-model approach allows for a broader range of semantic similarity features to be captured from free-text corpora, and that combining the models in various ways (their cosine similarity scores) may elevate certain desirable semantic similarity features.

In Paper B the use of multiple sense-vectors per word is explored as a possible approach to enhancing the spectrum of semantic relations captured in the resulting semantic model. Although the presented method demonstrated only minor improvements over the classical RI training approach, the work on multi-sense semantic similarity methods and models calls for further research. One possible use is in sentence similarity assessment and sentence topic clustering for use in extraction-based text summarization, similar to how it is done in Paper E. This could for example assist in providing more fine-grained discrimination between which sentence-topic cluster each sentence should belong to. Also, relatively little work is published on the use of such word-level multi-sense semantic similarity models in compositionality for applications including composing sentence context vectors and document context vectors.

Training of the distributional semantic models used in Papers A and B is done using a sliding window approach, where the context is defined by the neighboring words in the training corpora. However, in Paper C we explore also the use of other contextual features for training. Most notable are the methods relying on labeled ICD-10 codes and their internal hierarchy as context for inducing word-level semantics. This achieves the best results compared to the other evaluated methods and systems. As the experiment focused on care episode similarity (care episode retrieval), it is arguably natural that the use of such domain-specific meta-information as training features results in semantic spaces that better reflect semantic similarity relations suited for this task. However, the use of domain-specific (meta) features for constructing semantic models is something that deserves to be explored further, possibly evaluated on other granularity levels, such as word and sentence similarity assessment. For instance, this may include medications, allergies, age, lab tests, relevant clinical practice guidelines, SNOMED-CT concepts and so on, alone or in combination with neighboring words (c.f. sliding window). Paper C also includes the use of word2vec (W2V) as an alternative to RI for constructing distributional semantic models. The results show that W2V-based models outperform RI-based ones when constructed/trained using comparable context features and identical corpora.

Research Question 2

What sentence-level features of clinical text in care episodes are indicative of relevancy for inclusion in a clinical free-text summary?

Based on the experiments in Paper E we can conclude the following:

- Clustering sentences into topics that span across clinical notes is a seemingly desirable way of reducing redundancy with respect to what is of interest to the reader (clinician).
- The importance of a sentence in a clinical note is related to how many times the same/similar information has been mentioned throughout a care episode.
- By looking at discharge summaries from other similar care episodes, one can assess the importance of a sentence based on whether or not the same or similar information has been written there.
- If, using a VSM-based translation system, a sentence (its vector representation) can be "translated" into a vector representation that is similar to how this same sentence would look like in the translated vector space, it should be considered for inclusion in the final summary.

The experiments in the thesis represents some initial steps towards the goal of enabling summarization of care episodes in an fully unsupervised and resource light manner. It is not evident just how far one can go with the selected approach. A more user-centered evaluation is needed in order to shed additional light on the strengths, weaknesses and limitations of the explored summarization methods.

Research Question 3

How can the evaluation of distributional semantic models and text summaries generated from clinical text be done in a way that is fast, reliable and inexpensive?

Evaluation setups that allow for rapid automated evaluation are crucial when developing new computerized methods and algorithms. The most common way to do this is to manually construct gold standards, such as those made in relation to various shared tasks. In the synonym extraction task in Paper A, we used a set of MeSH terms and their synonyms as a gold standard. However, when conducting a manual analysis of extracted samples, we found that the semantic models not only extract synonyms that are present in the gold standard, but also other equally valid synonyms not present there. This indicates that constructing a complete list of synonyms for a word/term is challenging, especially when its usage is not clearly defined in terms of context and (sub-)domain. Further, this indicates that distributional semantic models can be used to improve coverage of lexical resources.

In Paper C we conclude that experiments conducted in most of the related work, including ours, are based on evaluation through pure retrieval performance according to a predefined gold standard. This is normally referred to as *intrinsic evaluation* (Hirschberg and Manning 2015). Future research on information retrieval in the clinical domain should arguably focus more on *extrinsic evaluation* — evaluating information retrieval systems in terms of support for health care and patient outcomes, as also argued in Mishra et al. (2014), Hirschberg and Manning (2015).

The evaluations conducted in Papers D and E are also defined as *intrinsic* in that pre-defined gold standards are used as evaluation criteria. Here we suggest that future work on such a task should (also) incorporate extrinsic evaluation — evaluating how the use of automatic text summarization systems in a clinical setting will impact documentation speed and quality, as well as health care quality and patient outcomes.

4.2 Future Work

Through the various experiments presented in this thesis, a number of possibilities for future work have been unveiled. The following are suggestions for future work in the context of the three research questions and related experiments

A major focus has been the training and use of distributed semantic models, where several novel methods for constructing such models with various properties have been proposed and evaluated. The main use of these models has been to compute semantic textual similarity between linguistic items of various granularity (words, sentences, clinical notes/care episodes).

For training there are multiple interdependent parameters and training features involved, these are mainly: the dimensionality of the VSM (fixed or post-training reduced); what training approach to use for inducing the vector space (RI vs RP vs W2V CBOW vs W2V Skip-gram); what contextual features to use and how to weight these (e.g., using sliding window with a certain window size); non-zero elements for the index vectors used in the RI approach; thresholds concerning clustering, such as the thresholds used in the sentence-topic clustering for text summarization and in the sense clustering in the multi-sense RI method; lower and upper frequency filters, e.g., filtering out words that occur less or more than some given thresholds. An additional factor is the utilized training corpus: the type of text and domain, or domains; the(ir) size; how pre-processing should be done (tokenization, lemmatization/stemming, stop-word removal, etc.). There is little doubt that improvements can be gained through optimizing these parameters and training features.

However, it is not evident how such optimization should be done since there is virtually an infinite number of different parameter values, possibly training features and combinations that can be tested and explored. Nor is it completely evident how different parameter settings and training features are related or how they may affect the resulting models and task performance. This is linked to the unsupervised nature of the underlying data-driven training approach together with the vast complexity of the training data (natural language text) and its size (millions of documents). Further, it is likely that there are no universally optimal settings, but optimal settings are instead task-specific. Thus there is a relatively long “distance” between setting the initial parameter values to having a fully trained distributional semantic model that has been properly evaluated on a given task. This is particularly the case when the task has a certain complexity level to it (e.g. text summarization). Henriksson and Hassel (2013) explore various vector dimensionalities using a fixed set of value alternatives. A somewhat similar type of exploration was conducted in Moen and Marsi (2013). Future work in this direction could focus more on evaluating the different parameter values as a multi-variable optimization task, possibly using some type of *gradient*-based or *hill-climbing* search algorithm (Russell and Norvig (2005), page 139 and 149). A possible outcome may include suggestions for how such parameter optimization should be approached.

For languages where comprehensive lexical knowledge resources are available, such as the SNOMED-CT and WordNet ontologies for English, hybrid approaches that combine statistical semantics with lexical resources, e.g., using the approach by Faruqui et al. (2015), could potentially contribute to producing semantic spaces that more correctly reflects the clinical terminology. Also for tasks where proper evaluation data is available, some type of task-specific retrospective fitting of pre-constructed distributional semantic models could be performed. For example, one may explore a similar approach as that used by Chen and Manning (2014), where they explore the use of a pre-trained neural network language model as the starting point for training a neural network classifier for use in a dependency parser. This general direction implies “moving” some of the context vectors in a semantic model so that they better reflect the (semantic) similarity relations expected by the ontologies or training examples. Ideally this would also result in movement of the context vectors belonging to neighboring words and concepts that are not found in the ontologies or training examples.

Semantic vectors representing sentences, notes or care episodes have here primarily been composed from word context vectors. This was done essentially through pointwise summation of the constituent (normalized) word context vectors. One obvious drawback with this approach is that word order is not taken into consideration. For instance, given a sentence, it is obvious that word order is of significance to the meaning it is meant to convey (e.g., “dog eats rabbit”). However, Landauer et al. (1997) concludes that when grading similarity between texts (essays), word order is seemingly not of great importance when relying on a distributional semantic model (LSA) to compute similarities. However, given the task of computerized semantic textual similarity assessment, where our strategy is to compose sentence context vectors for judging semantic similarity between sentence pairs. Here we would expect to see that improvements over classic distributional semantic models will be achieved by models and composition techniques that to some degree are able to produce sentence vectors whose point in the semantic space is adjusted based on the order of its constituent words (e.g., $\overrightarrow{dog} + \overrightarrow{eats} + \overrightarrow{rabbit}$ VS $\overrightarrow{rabbit} + \overrightarrow{eats} + \overrightarrow{dog}$). This would differ from approaches where training is done based on, e.g., co-occurrence information of pre-defined phrases, or where some type of convolution or shifting is used to construct completely new vectors. Perhaps a plausible approach would share certain similarities with multi-sense modeling techniques.

In looking at the work by Tversky (1977), one may argue that the use of distributional semantics for semantic textual similarity assessment is somewhat comparable to how humans judge similarity between concepts or objects. That is, similarity between two concepts is calculated based on measuring the likeness of, and the lack of, common features. However, Tversky also shows that the context in which the objects are evaluated has an impact on human similarity assessment. Further, the order in which two objects are compared may have an impact on how similarity is judged, i.e., $sim(x, y) \neq sim(y, x)$. The latter two properties of similarity are today not well explored in work on distributional semantics. I am not aware of any published work on bi-directional VSMs or distance metrics. Enabling this, as well as general improvements to automatically induced semantics, it may require that new ways of training and representing semantic models have to be introduced. This includes representations that can model (domain-/corpus-specific) bi-directionality and training algorithms that captures these properties from distributional statistics in text. Such a representation may, for example, be in the form of a multi-dimensional (hyper-)cube, or some type of graph. Regarding capturing these properties, it may be so that we have to identify and explore alternatives to the *distributional hypothesis*, alternatives that are based on some other fundamental properties of language and text, yet applicable to fully unsupervised methods for learning semantics.

In the present work on automatic text summarization, a set of features were explored in terms of their (statistical) indication of sentences relevance in clinical free-text summaries. The explored features are primarily based on statistics about (sequential) information redundancy and what other clinicians have deemed important in comparable cases. An exploratory approach was used when investigating potential features, motivated primarily by: more and less obvious patterns observable in clinical notes; observations reported in related work; feedback and suggestions from domain experts. Better understanding of such features could potentially be gained through conducting a thorough observation of domain experts during their work, in particular the actual process of writing or constructing summarized information and discharge summaries.

It is still unclear how far one can go with such an unsupervised approach, compared to some type of knowledge modeling. The latter could, e.g., include manual identification and classification of the significance of pre-defined concepts and features in clinical text, and the extraction of these to construct a summary, similarly to how it is done by Velupillai and Kvist (2012). However, this introduces the need for extensive manual labor. Hybrid approaches — that combines significance scoring of concepts derived from both supervised and unsupervised methods — is something to explore in future work.

Another important part of information summarization is how the summarized information is presented to the user. This has not been a focus in the work in this thesis, but will be a natural part of future work.

The evaluation criteria used in the automatic text summarization work was mainly based on discharge summaries. For automated evaluation, the gold standard consisted of the original discharge summaries, while for manual evaluation we used hospital guidelines concerning discharge summary content. This evaluation gave some indications of how well we are able to "reproduce" the content of a discharge summary. However, both Mishra et al. (2014) and Hirschberg and Manning (2015) conclude that future research should focus more on evaluating the impact of introducing text summarization systems in (simulated) hospital settings. I believe that this is the natural next step with respect to evaluation of this text summarization work. Such an evaluation, with clinicians as users of the system, is likely to provide more qualitative feedback regarding performance and user needs. This could also provide indications for features to use in terms of assessing information significance relative to the summarization process.

Also, I believe that a more user-guided summarization system is required, in particular when it comes to supporting the process of writing discharge summaries. One approach could include having the summarization system iteratively suggest sentences for inclusion based on first analyzing what content the user has, at any point, written in, or extracted into, the summary under creation.

When looking at the evaluation conducted in the other presented experiments (Papers A, B and C), the used gold standards consist of classifications done by humans. However, in Paper A it was revealed that the gold standard used for synonyms was lacking in terms of coverage since it lacked true positives. It is likely that this is also the case for the evaluation setup used in Paper C. This shows that a complete and clear cut gold standard is challenging to produce when dealing with natural language text. However, there are few alternative evaluation approaches available that also support rapid evaluation during method development. Chapman (2010) argue that it is important to involve end-users early (earlier) in the development process of NLP applications designed for assisting in patient care. Likewise, I believe that future work on development and evaluation of (distributional) semantic models for use in clinical NLP applications, would benefit from incorporating the end-users at various stages in the process. This could assist the developers greatly when it comes to understanding the domain and what (con)textual features that could potentially be utilized to achieve the desired semantic relations and properties in the resulting model.

4.3 Final Remarks

The work presented in this thesis could be of inspiration to others when it comes to constructing distributional semantic similarity models for use in the clinical domain. Several methods have been presented and evaluated at various textual semantic similarity assessment tasks, primarily using clinical text. We have also seen an approach to automated summarization of clinical free-text information that primarily relies on the (re-)use of statistical information derived from clinical corpora. This direction, focusing on the reuse of the large amounts of digitally stored clinical data being accumulated in hospitals nowadays, could facilitate the development of resource-lean software tools able to support and ease the information access and management work for clinicians and others in the health sector.

References

Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, volume 3491 of *Lecture Notes in Computer Science*. Springer-Verlag, Bath, England, 2004.

Aasen, Sigrun Espelien. News from the MeSH special interest group; MeSH speaks Norwegian in 2013! *Journal of the European Association for Health Information and Libraries*, 9(1):38–40, 2012.

Abulkhair, Maysoon; ALHarbi, Nora; Fahad, Amani; Omair, Seham; ALHosaini, Hadeel, and AlAffari, Fatimah. Intelligent integration of discharge summary: A formative model. In *Intelligent Systems Modelling & Simulation (ISMS 2013), 4th International Conference on Intelligent Systems, Modelling and Simulation*, pages 99–104, Bangkok, Thailand, 2013. Institute of Electrical and Electronics Engineers.

Afantenos, Stergos; Karkaletsis, Vangelis, and Stamatopoulos, Panagiotis. Summarization from medical documents: A survey. *Artificial Intelligence in Medicine*, 33(2):157–177, 2005.

Agirre, Eneko; Cer, Daniel; Diab, Mona, and Gonzalez-Agirre, Aitor. SemEval-2012 Task 6: A pilot on semantic textual similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pages 385–393, Montreal, Canada, June 2012. Association for Computational Linguistics.

Agirre, Eneko; Cer, Daniel; Diab, Mona; Gonzalez-Agirre, Aitor, and Guo, Weiwei. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint*

*Conference on Lexical and Computational Semantics (*SEM)*, volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pages 32–43, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

Allvin, Helen; Carlsson, Elin; Dalianis, Hercules; Danielsson-Ojala, Riitta; Daudaravičius, Vidas; Hassel, Martin; Kokkinakis, Dimitrios; Lundgren-Laine, Heljä; Nilsson, Gunnar; Nytrø, Øystein; Salanterä, Sanna; Skeppstedt, Maria; Suominen, Hanna, and Velupillai, Sumithra. Characteristics and analysis of Finnish and Swedish clinical intensive care nursing narratives. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 53–60, Los Angeles, California, USA, 2010. Association for Computational Linguistics.

Ao, Hiroko and Takagi, Toshihisa. ALICE: An Algorithm to Extract Abbreviations from MEDLINE. *Journal of the American Medical Informatics Association*, 12 (5):576–586, 2005.

Appelt, Douglas E and Onyshkevych, Boyan. The common pattern specification language. In *Proceedings of the TIPSTER workshop*, pages 23–30, Baltimore, Maryland, USA, 1998. Association for Computational Linguistics.

Ashburner, Michael; Ball, Catherine A; Blake, Judith A; Botstein, David; Butler, Heather; Cherry, J Michael; Davis, Allan P; Dolinski, Kara; Dwight, Selina S; Eppig, Janan T; Harris, Midori A; Hill, David P; Issel-Tarver, Laurie; Kasarskis, Andrew; Lewis, Suzanna; Matese, John C; Richardson, Joel E; Ringwald, Martin; Rubin, Gerald M, and Sherlock, Gavin. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Baroni, Marco and Lenci, Alessandro. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

Baroni, Marco; Dinu, Georgiana, and Kruszewski, Germán. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247. Association for Computational Linguistics, 2014.

Bartlett, Christopher; Boehncke, Klaus, and Haikerwal, M. E-health: Enabler for Australia's health reform. *Melbourne: Booz & Co*, 2008.

Bashyam, Vijayaraghavan; Hsu, William; Watt, Emily; Bui, Alex A. T.; Kangarloo, Hooshang, and Taira, Ricky K. Problem-centric organization and visualization of patient imaging and clinical data. *RadioGraphics*, 29(2):331–343, 2009.

Bateman, John A; Magnini, Bernardo, and Fabris, Giovanni. The generalized upper model knowledge base: Organization and use. In *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing*, pages 60–72. IOS Press, Amsterdam, 1995.

Berman, Jules J. Doublet method for very fast autocoding. *BMC Medical Informatics and Decision Making*, 4(1):16, 2004.

Blei, David M.; Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Blumenthal, David and Tavenner, Marilyn. The “meaningful use” regulation for electronic health records. *New England Journal of Medicine*, 363(6):501–504, 2010.

Bullinaria, John A and Levy, Joseph P. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior research methods*, 44(3):890–907, 2012.

Carbonell, Jaime and Goldstein, Jade. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, Melbourne, Australia, 1998. Association for Computing Machinery.

Chapman, WW. Closing the gap between NLP research and clinical practice. *Methods of Information in Medicine*, 49(4):317, 2010.

Chatterjee, Niladri and Mohan, Shiwali. Extraction-based single-document summarization using Random Indexing. In *19th IEEE International Conference on Tools with Artificial Intelligence. ICTAI 2007*, volume 2, pages 448–455, Patras, Greece, 2007. Institute of Electrical and Electronics Engineers.

Chen, Danqi and Manning, Christopher. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October 2014. Association for Computational Linguistics.

Chute, Christopher G. Classification and retrieval of patient records using natural language: An experimental application of Latent Semantic Analysis. In *Engineering in Medicine and Biology Society, Proceedings of the Annual International Conference of the IEEE*, volume 13, pages 1162–1163, Orlando, Florida, USA, 1991. Institute of Electrical and Electronics Engineers.

Cohen, Paul R and Howe, Adele E. How evaluation guides AI research: The message still counts more than the medium. *AI magazine*, 9(4):35, 1988.

Cohen, Paul R and Howe, Adele E. Toward AI research methodology: Three case studies in evaluation. *Systems, Man and Cybernetics, IEEE Transactions on*, 19 (3):634–646, 1989.

Cohen, Raphael; Aviram, Iddo; Elhadad, Michael, and Elhadad, Noémie. Redundancy-aware topic modeling for patient record notes. *PloS one*, 9(2): e87555, 2014.

Cohen, Trevor. Exploring MEDLINE space with random indexing and pathfinder networks. In *AMIA Annual Symposium Proceedings*, volume 2008, pages 126–130, Washington, DC, USA, 2008. American Medical Informatics Association.

Cohen, Trevor and Widdows, Dominic. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2): 390–405, April 2009.

Cohen, Trevor; Schvaneveldt, Roger, and Widdows, Dominic. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, 43(2):240–256, 2010.

Collobert, Ronan; Weston, Jason; Bottou, Léon; Karlen, Michael; Kavukcuoglu, Koray, and Kuksa, Pavel. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

Cutting, Doug. Apache Lucene open source package. `http://lucene.apache.org/`, 1999. (accessed 1st March 2016).

Dalianis, Hercules; Hassel, Martin, and Velupillai, Sumithra. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of ISHIMR 2009, Evaluation and implementation of e-health and health information initiatives: international perspectives. 14th International Symposium for Health Information Management Research, Kalmar, Sweden*, pages 243–249, 2009.

De Vine, Lance; Zuccon, Guido; Koopman, Bevan; Sitbon, Laurianne, and Bruza, Peter. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM International Conference on Conference on Information* *and Knowledge Management*, pages 1819–1822, Shanghai, China, 2014. Association for Computing Machinery.

Demner-Fushman, Dina; Chapman, Wendy W, and McDonald, Clement J. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009.

Erkan, Günes and Radev, Dragomir R. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, pages 457–479, 2004.

European Commission,. eHealth Action Plan 2012-2020: Innovative healthcare for the 21st century. https://ec.europa.eu/digital-single-market/en/news/ehealth-action-plan-2012-2020-innovative-healthcare-21st-century, 2012. (accessed 1st March 2016).

Farri, Oladimeji; Pieckiewicz, David S; Rahman, Ahmed S; Adam, Terrence J; Pakhomov, Serguei V, and Melton, Genevieve B. A qualitative analysis of EHR clinical document synthesis by clinicians. In *AMIA Annual Symposium Proceedings*, pages 1211–1220, Chicago, IL, USA, 2012. American Medical Informatics Association.

Faruqui, Manaal; Dodge, Jesse; Jauhar, Sujay Kumar; Dyer, Chris; Hovy, Eduard, and Smith, Noah A. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

Feng, Jin; Zhou, Yi-Ming, and Martin, Trevor. Sentence similarity based on relevance. In *Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, volume 8, pages 832–839, Málaga, Spain, 2008.

Ferreira, Rafael; Lins, Rafael Dueire; Simske, Steven J.; Freitas, Fred, and Riss, Marcelo. Assessing sentence similarity through lexical, syntactic and semantic analysis. *Computer Speech & Language*, 39:1–28, 2016. ISSN 0885-2308.

Ferrucci, David; Brown, Eric; Chu-Carroll, Jennifer; Fan, James; Gondek, David; Kalyanpur, Aditya A; Lally, Adam; Murdock, J William; Nyberg, Eric; Prager, John; Schlaefer, Nico, and Welty, Chris. Building Watson: An overview of the DeepQA project. *AI magazine*, 31(3):59–79, 2010.

Frege, Gottlob. Über Sinn Und Bedeutung. *Zeitschrift für Philosophie Und Philosophische Kritik*, 100(1):25–50, 1892.

Friedman, Carol; Kra, Pauline, and Rzhetsky, Andrey. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235, 2002.

Friedman, Carol; Rindflesch, Thomas C, and Corn, Milton. Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *Journal of Biomedical Informatics*, 46(5):765–773, 2013.

Furnas, George W; Landauer, Thomas K; Gomez, Louis M, and Dumais, Susan T. Human factors and behavioral science: Statistical semantics: Analysis of the potential performance of key-word information systems. *The Bell System Technical Journal*, 62(6):1753–1806, 1983.

Goldstein, Jade; Mittal, Vibhu; Carbonell, Jaime, and Kantrowitz, Mark. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, volume 4, pages 40–48. Association for Computational Linguistics, 2000.

Grabar, Natalia; Varoutas, Paul-Christophe; Rizand, Philippe; Livartowski, Alain, and Hamon, Thierry. Automatic acquisition of synonym resources and assessment of their impact on the enhanced search in EHRs. *Methods of Information in Medicine*, 48(2):149, 2009.

Guevara, Emiliano. Computing semantic compositionality in distributional semantics. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 135–144. Association for Computational Linguistics, 2011.

Hahn, Udo and Mani, Inderjeet. The challenges of automatic summarization. *Institute of Electrical and Electronics Engineers - Computer*, 33(11):29–36, 2000.

Hall, Amanda and Walton, Graham. Information overload within the health care system: a literature review. *Health Information & Libraries Journal*, 21(2): 102–108, 2004.

Harris, Zellig S. Distributional structure. *Word*, 10:146–162, 1954.

Hassel, Martin and Sjöbergh, Jonas. Navigating through summary space: Selecting summaries, not sentences. In *Resource Lean and Portable Automatic Text Summarization*, pages 109–132. KTH Royal Institute of Technology, 2007.

Henriksson, Aron and Hassel, Martin. Optimizing the dimensionality of clinical term spaces for improved diagnosis coding support. In *Proceedings of the Louhi Workshop on Health Document Text Mining and Information Analysis*, pages 1–6, 2013.

Henriksson, Aron; Conway, Mike; Duneld, Martin, and Chapman, Wendy Webber. Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records. In *AMIA 2013, American Medical Informatics Association Annual Symposium*, Washington, DC, USA, 2013a.

Henriksson, Aron; Skeppstedt, Maria; Kvist, Maria; Duneld, Martin, and Conway, Mike. Corpus-driven terminology development: populating Swedish SNOMED-CT with synonyms extracted from electronic health records. In *Proceedings of BioNLP*, pages 36–44, Sofia, Bulgaria, 2013b. Association for Computational Linguistics.

Hevner, Alan R; March, Salvatore T; Park, Jinsoo, and Ram, Sudha. Design science in information systems research. *MIS quarterly*, 28(1):75–105, 2004.

Hirsch, Jamie S; Tanenbaum, Jessica S; Lipsky Gorman, Sharon; Liu, Connie; Schmitz, Eric; Hashorva, Dritan; Ervits, Artem; Vawdrey, David; Sturm, Marc, and Elhadad, Noémie. Harvest, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274, 2015.

Hirschberg, Julia and Manning, Christopher D. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

Hofmann, Thomas. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. Association for Computing Machinery.

Huang, Eric H.; Socher, Richard; Manning, Christopher D., and Ng, Andrew Y. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Jeju Island, Korea, 2012. Association for Computational Linguistics.

Jha, Ashish K. Meaningful use of electronic health records: The road ahead. *The Journal of the American Medical Association*, 304(15):1709–1710, 2010.

Johnson, William B and Lindenstrauss, Joram. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Jones, Karen Sparck. Automatic summarising: Factors and directions. *Advances in Automatic Text Summarization*, pages 1–12, 1999.

Jonnalagadda, Siddhartha; Cohen, Trevor; Wu, Stephen, and Gonzalez, Graciela. Enhancing clinical concept extraction with distributional semantics. *Journal of Biomedical Informatics*, 45(1):129–140, 2012.

Kanerva, Pentti; Kristofersson, Jan, and Holst, Anders. Random Indexing of text samples for Latent Semantic Analysis. In *Proceedings of 22nd Annual Conference of the Cognitive Science Society*, page 1036, Philadelphia, PA, USA, 2000.

Karlgren, Jussi and Sahlgren, Magnus. From words to understanding. In *Foundations of Real World Intelligence*, CSLI Lecture Notes 125, pages 294–308. CSLI, 2001.

Kate, Rohit J. Unsupervised grammar induction of clinical report sublanguage. *Journal of Biomedical Semantics*, 3(S-3):S4, 2012.

Kolb, Peter. Experiments on the difference between semantic similarity and relatedness. In *Proceedings of the 17th Nordic Conference on Computational Linguistics, NODALIDA'09*, volume 4, pages 81–88, Odense, Denmark, 2009. NEALT Proceedings series.

Koopman, Bevan; Zuccon, Guido; Bruza, Peter; Sitbon, Laurianne, and Lawley, Michael. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 2439–2442, Maui, HI, USA, 2012. Association for Computing Machinery.

Kripalani, Sunil; LeFevre, Frank; Phillips, Christopher O; Williams, Mark V; Basaviah, Preetha, and Baker, David W. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. *Journal of the American Medical Association*, 297(8):831–841, 2007.

Kuhn, Harold W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

Kvist, Maria; Skeppstedt, Maria; Velupillai, Sumithra, and Dalianis, Hercules. Modeling human comprehension of Swedish medical records for intelligent access and summarization systems-future vision, a physician's perspective. In *9th Scandinavian Conference on Health Informatics, SHI 2011*, Oslo, Norway, 2011. Tapir Academic Press.

Landauer, Thomas K and Dumais, Susan T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211, 1997.

Landauer, Thomas K; Laham, Darrell; Rehder, Bob, and Schreiner, Missy E. How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pages 412–417. Cognitive Science Society, Citeseer, 1997.

Le, Quoc and Mikolov, Tomas. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, Beijing, China, 2014.

LeCun, Yann; Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *Nature*, 521 (7553):436–444, 2015.

Lee, Daniel D and Seung, H Sebastian. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Lehman, Ann. *JMP for basic univariate and multivariate statistics: A step-by-step guide*. SAS Institute, 2005.

Lenci, Alessandro and Benotto, Giulia. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 75–79, Montreal, Canada, 2012. Association for Computational Linguistics.

Levy, Omer and Goldberg, Yoav. Linguistic regularities in sparse and explicit word representations. In *Proceedings of Eighteenth Conference on Computational Natural Language Learning (CoNNL-2014)*, Baltimore, Maryland, USA, 2014. Association for Computational Linguistics.

Levy, Omer; Goldberg, Yoav, and Dagan, Ido. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

Lin, Chin-Yew. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Lissauer, T; Paterson, CM; Simons, A, and Beard, RW. Evaluation of computer generated neonatal discharge summaries. *Archives of Disease in Childhood*, 66 (4 Spec No):433–436, 1991.

Liu, Shuhua. Experiences and reflections on text summarization tools. *International Journal of Computational Intelligence Systems*, 2(3):202–218, 2009.

Lord, Phillip W.; Stevens, Robert D.; Brass, Andy, and Goble, Carole A. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

Luhn, Hans Peter. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

Lund, Kevin and Burgess, Curt. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, June 1996.

Manning, Christopher D; Raghavan, Prabhakar, and Schütze, Hinrich. *Introduction to information retrieval*, volume 1. Cambridge University Press, Cambridge, UK, 2008.

Marsi, Erwin; Moen, Hans; Bungum, Lars; Sizov, Gleb; Gambäck, Björn, and Lynum, André. NTNU-CORE: Combining strong features for semantic similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 66–73, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics.

McClelland, James L; Rumelhart, David E, and Group, PDP Research. Parallel distributed processing. *Explorations in the microstructure of cognition*, 2, 1986.

McCray, Alexa T; Srinivasan, Suresh, and Browne, Allen C. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 235–239, Washington, DC, USA, 1994. American Medical Informatics Association.

Mendonça, Eneida A; Haas, Janet; Shagina, Lyudmila; Larson, Elaine, and Friedman, Carol. Extracting information on pneumonia in infants using natural language processing of radiology reports. *Journal of Biomedical Informatics*, 38 (4):314–321, 2005.

Meng, Frank; Taira, Ricky K; Bui, Alex AT; Kangarloo, Hooshang, and Churchill, Bernard M. Automatic generation of repeated patient information for tailoring clinical notes. *International Journal of Medical Informatics*, 74(7):663–673, 2005.

Meystre, Stéphane M; Savova, Guergana K; Kipper-Schuler, Karin C, and Hurdle, John F. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics*, 35:128–44, 2008.

Mihalcea, Rada and Tarau, Paul. TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 404–411, Barcelona, Spain, 2004. Association for Computational Linguistics.

Mikolov, Tomas; Chen, Kai; Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg S., and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119. 2013b.

Mikolov, Tomas; Yih, Wen-tau, and Zweig, Geoffrey. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, USA, June 2013c. Association for Computational Linguistics.

Miller, George A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Mishra, Rashmi; Bian, Jiantao; Fiszman, Marcelo; Weir, Charlene R; Jonnalagadda, Siddhartha; Mostafa, Javed, and Del Fiol, Guilherme. Text summarization in the biomedical domain: A systematic review of recent research. *Journal of Biomedical Informatics*, 2014.

Mnih, Andriy and Hinton, Geoffrey E. A scalable hierarchical distributed language model. In *Advances in Neural Information Processing Systems (NIPS 2009)*, pages 1081–1088, Vancouver, B.C., Canada, 2009. Citeseer.

Moen, Hans and Marsi, Erwin. Cross-lingual random indexing for information retrieval. In *Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 164–175. Springer Berlin Heidelberg, 2013.

Moen, Hans; Heimonen, Juho; Murtola, Laura-Maria; Airola, Antti; Pahikkala, Tapio; Terävä, Virpi; Danielsson-Ojala, Riitta; Salakoski, Tapio, and Salanterä, Sanna. On evaluation of automatically generated clinical discharge summaries. In *Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI 2014)*, pages 101–114, Trondheim, Norway, 2014a. CEUR Workshop Proceedings.

Moen, Hans; Marsi, Erwin; Ginter, Filip; Murtola, Laura-Maria; Salakoski, Tapio, and Salanterä, Sanna. Care episode retrieval. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*, pages 116–124, Gothenburg, Sweden, 2014b. Association for Computational Linguistics.

Moen, Hans; Ginter, Filip; Marsi, Erwin; Peltonen, Laura-Maria; Salakoski, Tapio, and Salanterä, Sanna. Care episode retrieval: distributional semantic models for information retrieval in the clinical domain. *BMC Medical Informatics and Decision Making*, 15(Suppl 2):S2, 2015.

Moen, Hans; Peltonen, Laura-Maria; Heimonen, Juho; Airola, Antti; Pahikkala, Tapio; Salakoski, Tapio, and Salanterä, Sanna. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37, 2016.

Montague, Richard and Thomason, Richmond H. *Formal Philosophy: Selected Papers of Richard Montague; Ed. and with an Introduction by Richmond H. Thomason*. Yale University Press, 1976.

Morin, Frederic and Bengio, Yoshua. Hierarchical probabilistic neural network language model. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, pages 246–252. Citeseer, 2005.

Needleman, Saul B and Wunsch, Christian D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

Neelakantan, Arvind; Shankar, Jeevan; Passos, Alexandre, and McCallum, Andrew. Efficient nonparametric estimation of multiple embeddings per word in vector space. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014.

Nenkova, Ani and McKeown, Kathleen. *Automatic Summarization*. Foundations and Trends in Information Retrieval. Now Publishers Inc., 2011.

NLM, National Library of Medicine. MeSH (Medical Subject Headings), a. URL `http://www.ncbi.nlm.nih.gov/mesh`. (accessed 10th October 2015).

NLM, National Library of Medicine. International Health Terminology Standards Development Organisation: Supporting Different Languages, b. URL `http://www.ihtsdo.org/snomed-ct`. (accessed 10th October 2015).

NLM, National Library of Medicine. Unified Medical Language System, c. URL https://www.nlm.nih.gov/research/umls/. (accessed 10th October 2015).

Panman, Otto. Homonymy and polysemy. *Lingua*, 58(1):105–136, 1982.

Partee, Barbara; ter Meulen, Alice, and Wall, Robert. *Mathematical Methods in Linguistics*, volume 30. Springer Science & Business Media, 1990.

Patil, Kaustubh and Brazdil, Pavel. Sumgraph: text summarization using centrality in the pathfinder network. *International Journal on Computer Science and Information Systems*, 2(1):18–32, 2007.

Pedersen, Ted; Pakhomov, Serguei VS; Patwardhan, Siddharth, and Chute, Christopher G. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, 2007.

Pivovarov, Rimma and Elhadad, Noémie. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, 2015.

Plate, Tony. Holographic Reduced Representations: Convolution Algebra for Compositional Distributed Representations. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI 1991)*, pages 30–35, San Mateo, CA, 1991. Citeseer.

Pradhan, Sameer; Elhadad, Noémie; Chapman, Wendy; Manandhar, Suresh, and Savova, Guergana. Semeval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, Dublin, Ireland, 2014. Association for Computational Linguistics and Dublin City University.

Pyysalo, Sampo; Ginter, Filip; Heimonen, Juho; Björne, Jari; Boberg, Jorma; Järvinen, Jouni, and Salakoski, Tapio. BioInfer: A corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):50, 2007.

Pyysalo, Sampo; Ginter, Filip; Moen, Hans; Salakoski, Tapio, and Ananiadou, Sophia. Distributional semantics resources for biomedical text processing. In *Proceedings of Languages in Biology and Medicine*, 2013.

Rector, Alan L. Clinical terminology: why is it so hard? *Methods of Information in Medicine*, 38(4/5):239–252, 1999.

Reisinger, Joseph and Mooney, Raymond J. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117, Los Angeles, California, USA, June 2010.

Roque, Francisco S; Slaughter, Laura, and Tkatšenko, Alexandr. A comparison of several key information visualization systems for secondary use of electronic health record content. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 76–83, Los Angeles, California, USA, 2010. Association for Computational Linguistics.

Russell, Stuart J. and Norvig, Peter. *Artificial Intelligence - A Modern Approach (Second Edition)*. Pearson Education, 2005.

Sahlgren, Magnus. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. 2006.

Sahlgren, Magnus and Karlgren, Jussi. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11 (03):327–341, 2005.

Sahlgren, Magnus and Karlgren, Jussi. Terminology mining in social media. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 405–414, Hong Kong, China, 2009. Association for Computing Machinery.

Sahlgren, Magnus and Swanberg, David. *Vector based semantic analysis: Modeling linguistic knowledge in computer systems*. PhD thesis, Stockholm University, 2000.

Sahlgren, Magnus; Holst, Anders, and Kanerva, Pentti. Permutations as a means to encode order in word space. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Washington, DC, USA, 2008.

Salton, Gerard; Wong, Anita, and Yang, Chung-Shu. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Sarkar, Kamal; Nasipuri, Mita, and Ghose, Suranjan. Using machine learning for medical document summarization. *International Journal of Database Theory and Application*, 4:31–49, 2011.

Schütze, Hinrich. Automatic word sense discrimination. *Computational Linguistics – Special issue on word sense disambiguation*, 24(1):97–123, 1998.

Smadja, Frank. Retrieving collocations from text: Xtract. *Computational linguistics*, 19(1):143–177, 1993.

Smith, Edward E and Medin, Douglas L. *Categories and concepts*. Harvard University Press Cambridge, MA, 1981.

Sørby, Inger Dybdahl and Nytrø, Øystein. Does the EPR support the discharge process? A study on physicians' use of clinical information systems during discharge of patients with coronary heart disease. *Journal of Healthcare Information Management*, 34(4):112–119, 2005.

Sparck Jones, Karen. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

Steinberger, Josef and Křišt'an, M. LSA-based multi-document summarization. In *Proceedings of 8th International PhD Workshop on Systems and Control*, volume 7, Balatonfured, Hungary, 2007. Citeseer.

Suominen, Hanna. *Machine learning and clinical text. supporting health information flow*. PhD thesis, University of Turku, 2009.

Suominen, Hanna; Salanterä, Sanna; Velupillai, Sumithra; Chapman, Wendy W.; Savova, Guergana; Elhadad, Noemie; Pradhan, Sameer; South, Brett R.; Mowery, Danielle L.; Jones, Gareth J. F.; Leveling, Johannes; Kelly, Liadh; Goeuriot, Lorraine; Martinez, David, and Zuccon, Guido. *Information Access Evaluation. Multilinguality, Multimodality, and Visualization: 4th International Conference of the CLEF Initiative (CLEF 2013)*, chapter Overview of the ShARe/CLEF eHealth Evaluation Lab 2013, pages 212–231. Springer Berlin Heidelberg, Germany, Valencia, Spain, 2013.

Suominen, Hanna J. and Salakoski, Tapio I. Supporting communication and decision making in finnish intensive care with language technology. *Journal of Healthcare Engineering*, 1:595–614, 2010.

Turney, Peter D and Pantel, Patrick. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

Tversky, Amos. Features of similarity. *Psychological Review*, 84:327–352, 1977.

Van Vleck, Tielman T; Stein, Daniel M; Stetson, Peter D, and Johnson, Stephen B. Assessing data relevance for automated generation of a clinical summary. In *AMIA Annual Symposium Proceedings*, volume 2007, pages 761–765, Chicago, IL, USA, 2007. American Medical Informatics Association.

VanderStoep, Scott W and Johnson, Deidre D. *Research methods for everyday life: Blending qualitative and quantitative approaches*, volume 32. John Wiley & Sons, 2008.

Vapnik, Vladimir; Golowich, Steven E., and Smola, Alex. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems*, volume 9, pages 281–287. MIT Press, Cambridge, Massachusetts, 1997.

Velupillai, Sumithra and Kvist, Maria. Fine-grained certainty level annotations used for coarser-grained e-health scenarios. In *Computational Linguistics and Intelligent Text Processing*, pages 450–461. Springer, 2012.

Weaver, Warren. Translation. *Machine Translation of Languages*, 14:15–23, 1955.

World Health Organization, others. International classification of diseases (ICD). 1983.

Wrenn, Jesse O; Stein, Daniel M; Bakken, Suzanne, and Stetson, Peter D. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*, 17(1):49–53, 2010.

Xu, Hua; Stenner, Shane P; Doan, Son; Johnson, Kevin B; Waitman, Lemuel R, and Denny, Joshua C. MedEx: A medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, 2010.

Yampolskiy, Roman V. *Artificial Intelligence, Evolutionary Computing and Metaheuristics: In the Footsteps of Alan Turing*, chapter Turing Test as a Defining Feature of AI-Completeness, pages 3–17. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

Part II

Papers

List of Research Papers

Paper A: Henriksson, Aron; Moen, Hans; Skeppstedt, Maria; Daudaravičius, Vidas, and Duneld, Martin. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, 5(1):25, 2014.

Paper B: Moen, Hans; Marsi, Erwin, and Gambäck, Björn. Towards dynamic word sense discrimination with Random Indexing. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 83–90, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

Paper C: Moen, Hans; Ginter, Filip; Marsi, Erwin; Peltonen, Laura-Maria; Salakoski, Tapio, and Salanterä, Sanna. Care episode retrieval: distributional semantic models for information retrieval in the clinical domain. *BMC Medical Informatics and Decision Making*, 15(Suppl 2):S2, 2015.

Paper D: Moen, Hans; Heimonen, Juho; Murtola, Laura-Maria; Airola, Antti; Pahikkala, Tapio; Terävä, Virpi; Danielsson-Ojala, Riitta; Salakoski, Tapio, and Salanterä, Sanna. On evaluation of automatically generated clinical discharge summaries. In *Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI 2014)*, pages 101–114, Trondheim, Norway, 2014. CEUR Workshop Proceedings.

Paper E: Moen, Hans; Peltonen, Laura-Maria; Heimonen, Juho; Airola, Antti; Pahikkala, Tapio; Salakoski, Tapio, and Salanterä, Sanna. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37, 2016.

Paper A

Synonym extraction and abbreviation expansion with ensembles of semantic spaces

Henriksson, Aron; Moen, Hans; Skeppstedt, Maria; Daudaravičius, Vidas, and Duneld, Martin

Journal of Biomedical Semantics, 5(1):25, 2014.

Henriksson *et al. Journal of Biomedical Semantics* 2014, **5**:6
http://www.jbiomedsem.com/content/5/1/6

JOURNAL OF BIOMEDICAL SEMANTICS

RESEARCH Open Access

Synonym extraction and abbreviation expansion with ensembles of semantic spaces

Aron Henriksson[1*], Hans Moen[2†], Maria Skeppstedt[1†], Vidas Daudaravičius[3] and Martin Duneld[1]

Abstract

Background: Terminologies that account for variation in language use by linking synonyms and abbreviations to their corresponding concept are important enablers of high-quality information extraction from medical texts. Due to the use of specialized sub-languages in the medical domain, manual construction of semantic resources that accurately reflect language use is both costly and challenging, often resulting in low coverage. Although models of distributional semantics applied to large corpora provide a potential means of supporting development of such resources, their ability to isolate synonymy from other semantic relations is limited. Their application in the clinical domain has also only recently begun to be explored. Combining distributional models and applying them to different types of corpora may lead to enhanced performance on the tasks of automatically extracting synonyms and abbreviation-expansion pairs.

Results: A combination of two distributional models – Random Indexing and Random Permutation – employed in conjunction with a single corpus outperforms using either of the models in isolation. Furthermore, combining semantic spaces induced from different types of corpora – a corpus of clinical text and a corpus of medical journal articles – further improves results, outperforming a combination of semantic spaces induced from a single source, as well as a single semantic space induced from the conjoint corpus. A combination strategy that simply sums the cosine similarity scores of candidate terms is generally the most profitable out of the ones explored. Finally, applying simple post-processing filtering rules yields substantial performance gains on the tasks of extracting abbreviation-expansion pairs, but not synonyms. The best results, measured as recall in a list of ten candidate terms, for the three tasks are: 0.39 for abbreviations to long forms, 0.33 for long forms to abbreviations, and 0.47 for synonyms.

Conclusions: This study demonstrates that ensembles of semantic spaces can yield improved performance on the tasks of automatically extracting synonyms and abbreviation-expansion pairs. This notion, which merits further exploration, allows different distributional models – with different model parameters – and different types of corpora to be combined, potentially allowing enhanced performance to be obtained on a wide range of natural language processing tasks.

Background

In order to create high-quality information extraction systems, it is important to incorporate some knowledge of semantics, such as the fact that a concept can be signified by multiple signifiers[a]. Morphological variants, abbreviations, acronyms, misspellings and synonyms – although different in form – may share semantic content to different degrees. The various lexical instantiations of a concept thus need to be mapped to some standard representation of the concept, either by converting the different expressions to a canonical form or by generating lexical variants of a concept's 'preferred term'. These mappings are typically encoded in semantic resources, such as thesauri or ontologies[b], which enable the recall (sensitivity) of information extraction systems to be improved. Although their value is undisputed, manual construction of such resources is often prohibitively expensive and may also result in limited coverage, particularly in the biomedical and clinical domains where language use variability is exceptionally high [1].

*Correspondence: aronhen@dsv.su.se
†Equal contributors
[1] Department of Computer and Systems Sciences (DSV), Stockholm University, Forum 100, SE-164 40 Kista, Sweden
Full list of author information is available at the end of the article

There is thus a need for (semi-)automatic methods that can aid and accelerate the process of lexical resource development, especially ones that are able to reflect real language use in a particular domain and adapt to different genres of text, as well as to changes over time. In the clinical domain, for instance, language use in general, and (ad-hoc) abbreviations in particular, can vary significantly across specialities. Statistical, corpus-driven and language-agnostic methods are attractive due to their inherent portability: given a corpus of sufficient size in the target domain, the methods can be applied with no or little adaptation needed. Models of distributional semantics, building on the assumption that linguistic items with similar distributions in large bodies of linguistic data have similar meanings, fulfill these requirements and have been used to extract semantically similar terms from large corpora; with increasing access to data from electronic health records, their application in the clinical domain has lately begun to be explored. In this paper, we present a method that employs distributional semantics for the extraction of synonyms and abbreviation-expansion pairs from two corpora: a clinical corpus (comprising health record narratives) and a medical corpus (comprising journal articles). We also demonstrate that performance can be enhanced by creating ensembles of (distributional) semantic spaces – both with different model parameter configurations and induced from different genres of text.

The structure of this paper is as follows. First, we present some relevant background literature on synonyms, abbreviations and their extraction/expansion. We also introduce the ideas underlying distributional semantics in general and, in particular, the models employed in this study: Random Indexing and Random Permutation. Then, we describe our method of combining semantic spaces induced from single and multiple corpora, including the details of the experimental setup and the materials used. A presentation of the experimental results is followed by an analysis and discussion of their implications. Finally, we conclude the paper with a summary and conclusions.

Language use variability: synonyms and abbreviations

Synonymy is a semantic relation between two phonologically distinct words with very similar meaning. It is, however, extremely rare that two words have the exact same meaning – perfect synonyms – as there is often at least one parameter that distinguishes the use of one word from another [2]. For this reason, we typically speak of near-synonyms instead; that is, two words that are interchangeable in some, but not all, contexts[c] [2]. Two near-synonyms may also have different connotations, such as conveying a positive or a negative attitude. To complicate matters further, the same concept can sometimes be referred to with different words in different dialects; for a speaker who is familiar with both dialects, these can be viewed as synonyms. A similar phenomenon concerns different formality levels, where one word in a synonym pair is used only as slang and the other only in a more formal context [2]. In the medical domain, there is one vocabulary that is more frequently used by medical professionals, whereas patients often use alternative, layman terms [3]. When developing many natural language processing (NLP) applications, it is important to have ready access to terminological resources that cover this variation in the use of vocabulary by storing synonyms. Examples of such applications are query expansion [3], text simplification [4] and, as already mentioned previously, information extraction [5].

The use of abbreviations and acronyms is prevalent in both medical journal text [6] and clinical text [1]. This leads to decreased readability [7] and poses challenges for information extraction [8]. Semantic resources that also link abbreviations to their corresponding concept, or, alternatively, simple term lists that store abbreviations and their corresponding long form, are therefore as important as synonym resources for many biomedical NLP applications. Like synonyms, abbreviations are often interchangeable with their corresponding long form in some, if not all, contexts. An important difference between abbreviations and synonyms is, however, that abbreviations are semantically overloaded to a much larger extent; that is, one abbreviation often has several possible long forms, with distinct meanings. In fact, 81% of UMLS[d] abbreviations in biomedical text were found to be ambiguous [6].

Identifying synonymous relations between terms

The importance of synonym learning is well recognized in the NLP research community, especially in the biomedical [9] and clinical [1] domains. A wide range of techniques has been proposed for the identification of synonyms and other semantic relations, including the use of lexico-syntactic patterns, graph-based models and, indeed, distributional semantics [10] – the approach investigated in this study.

For instance, Hearst [11] proposes the use of lexico-syntactic patterns for the automatic acquisition of hyponyms[e] from unstructured text. These patterns are hand-crafted according to observations in a corpus. Patterns can similarly be constructed for other types of lexical relations. However, a requirement is that these syntactic patterns are common enough to match a wide array of hyponym pairs. Blondel et al. [12] present a graph-based method that takes its inspiration from the calculation of hub, authority and centrality scores when ranking hyperlinked web pages. They illustrate that the central similarity score can be applied to the task of automatically extracting synonyms from a monolingual dictionary, in this case the Webster dictionary, where the assumption is that synonyms have a large overlap in the words used in their definitions; they also co-occur in the definition of many words. Another possible source for extracting synonyms is the use of linked data, such as Wikipedia. Nakayama et al. [13] also utilize a graph-based method, but instead of relying on word co-occurrence information, they exploit the links between Wikipedia articles (treated as concepts). This way they can measure both the strength (the number of paths from one article to another) and the distance (the length of each path) between concepts: concepts close to each other in the graph and with common hyperlinks are deemed to be more closely related than those farther away.

There have also been some previous efforts to obtain better performance on the synonym extraction task by not only using a single source and a single method. Inspiration for some of these approaches has been drawn from ensemble learning, a machine learning technique that combines the output of several different classifiers with the aim of improving classification performance (see [14] for an overview). Curran [15] exploits this notion for synonym extraction and demonstrates that ensemble methods outperform individual classifiers even for very large corpora. Wu and Zhou [16] use multiple resources – a monolingual dictionary, a bilingual corpus and a large monolingual corpus – in a weighted ensemble method that combines the individual extractors, thereby improving both precision and recall on the synonym acquisition task. Along somewhat similar lines, van der Plas and Tiedemann [17] use parallel corpora to calculate distributional similarity based on (automatic) word alignment, where a translational context definition is employed; synonyms are extracted with both greater precision and recall compared to a monolingual approach. This approach is, however, hardly applicable in the medical domain due to the unavailability of parallel corpora. Peirsman and Geeraerts [18] combine predictors based on collocation measures and distributional semantics with a so-called compounding approach, wherein cues are combined with strongly associated words into compounds and verified against a corpus. This ensemble approach is shown substantially to outperform the individual predictors of strong term associations in a Dutch newspaper corpus. In information retrieval, Diaz and Metzler [19] report increased performance gains when utilizing language models that derive evidence from both a target corpus and an external corpus, compared to using the target corpus alone.

In the biomedical domain, most efforts have focused on extracting synonyms of gene and protein names from the biomedical literature [20-22]. In the clinical domain, Conway and Chapman [23] propose a rule-based approach to generate potential synonyms from the BioPortal ontology – using permutations, abbreviation generation, etc. – after which candidate synonyms are verified against a large clinical corpus. Henriksson et al. [24,25] use models of distributional semantics to induce unigram word spaces and multiword term spaces from a large corpus of clinical text in an attempt to extract synonyms of varying length for SNOMED CT preferred terms. Zeng et al. [26] evaluate three query expansion methods for retrieval of clinical documents and conclude that an LDA-based topic model generates the best synonyms. Pedersen et al. [27] explore a set of measures for automatically judging semantic similarity and relatedness among medical term pairs that have been pre-assessed by human experts. The measures range from ones based on thesauri or ontologies (WordNet, SNOMED-CT, UMLS, Mayo Clinic Thesaurus) to those based on distributional semantics. They find that the measure based on distributional semantics performs at least as good as any of the ontology-dependent measures. In a similar task, Koopman et al. [28] evaluate eight different data-driven measures of semantic similarity. Using two separate training corpora, one containing clinical notes and the other medical literature articles, they conclude that the choice of training corpus has a significant impact on the performance of these measures.

Creating abbreviation dictionaries automatically

There are a number of studies on the automatic creation of biomedical abbreviation dictionaries that exploit the fact that abbreviations are sometimes defined in the text on their first mention. These studies extract candidates for abbreviation-expansion pairs by assuming that either the long form or the abbreviation is written in parentheses [29]; other methods that use rule-based pattern matching have also been proposed [30]. The process of determining which of the extracted candidates that are likely to be correct abbreviation-expansion pairs is then performed either by rule-based [30] or machine learning [31,32] methods. Most of these studies have been conducted for English; however, there is also a study on Swedish medical text [33], for instance.

Yu et al. [34] have, however, found that around 75% of all abbreviations found in biomedical articles are never defined in the text. The application of these methods to clinical text is most likely inappropriate, as clinical text is often written in a telegraphic style, mainly for documentation purposes [1]; that effort would be spent on defining used abbreviations in this type of text seems unlikely. There has, however, been some work on identifying such undefined abbreviations [35], as well as on finding the intended abbreviation expansion among several possible expansions available in an abbreviation dictionary [36].

In summary, automatic creation of biomedical abbreviation dictionaries from texts where abbreviations are defined is well studied. This is also the case for abbreviation disambiguation given several possible long forms in an abbreviation dictionary. The abbreviation part of this study, however, focuses on a task that has not as yet been adequately explored: to find abbreviation-expansion pairs without requiring the abbreviations to be defined in the text.

Distributional semantics: inducing semantic spaces from corpora

Distributional semantics (see [37] for an overview of methods and their application in the biomedical domain) were initially motivated by the inability of the vector space model [38] – as it was originally conceived – to account for the variability of language use and word choice stemming from natural language phenomena such as synonymy. To overcome the negative impact this had on recall in information retrieval systems, models of distributional semantics were proposed [39-41]. The theoretical foundation underpinning such semantic models is the *distributional hypothesis* [42], which states that words with similar distributions in language – in the sense that they co-occur with overlapping sets of words – tend to have similar meanings. Distributional methods have become popular with the increasing availability of large corpora and are attractive due to their computational approach to semantics, allowing an estimate of the semantic relatedness between two terms to be quantified.

An obvious application of distributional semantics is the extraction of semantically related terms. As near-synonyms are interchangeable in at least some contexts, their distributional profiles are likely to be similar, which in turn means that synonymy is a semantic relation that should, to a certain degree, be captured by these methods. This seems intuitive, as, next to identity, the highest degree of semantic relatedness between terms is realized by synonymy. It is, however, well recognized that other semantic relations between terms that share similar contexts will likewise be captured by these models [43]; synonymy cannot readily be isolated from such relations.

Spatial models[f] of distributional semantics generally differ in how vectors representing term meaning are constructed. These vectors, often referred to as *context vectors*, are typically derived from a term-context matrix that contains the (weighted, normalized) frequency with which terms occur in different contexts. Working directly with such high-dimensional (and inherently sparse) data — where the dimensionality is equal to the number of contexts (e.g. the number of documents or the size of the vocabulary, depending on which context definition is employed) — would entail unnecessary computational complexity, in particular since most terms only occur in a limited number of contexts, which means that most cells in the matrix will be zero. The solution is to project the high-dimensional data into a lower-dimensional space, while approximately preserving the relative distances between data points. The benefit of dimensionality reduction is two-fold: on the one hand, it reduces complexity and data sparseness; on the other hand, it has also been shown to improve the coverage and accuracy of term-term associations, as, in this reduced (semantic) space, terms that do not necessarily co-occur directly in the *same* contexts – this is indeed the typical case for synonyms and abbreviation-expansion pairs – will nevertheless be clustered about the same subspace, as long as they appear in *similar* contexts, i.e. have neighbors in common (co-occur with the same terms). In this way, the reduced space can be said to capture higher order co-occurrence relations.

In latent semantic analysis (LSA) [39], dimensionality reduction is performed with a computationally expensive matrix factorization technique known as singular value decomposition. Despite its popularity, LSA has consequently received some criticism for its poor scalability properties. More recently, alternative methods for constructing semantic spaces based on term co-occurrence information have been proposed.

Random indexing

Random indexing (RI) [44] is an incremental, scalable and computationally efficient alternative to LSA in which explicit dimensionality reduction is avoided[g]: a lower dimensionality d is instead chosen *a priori* as a model parameter and the d-dimensional context vectors are then constructed incrementally. This approach allows new data to be added at any given time without having to rebuild the semantic space. RI can be viewed as a two-step operation:

1. Each *context* (e.g. each document or unique term) is first given a static, unique representation in the vector space that is approximately uncorrelated to all other contexts. This is achieved by assigning a sparse, ternary[h] and randomly generated d-dimensional *index vector*: a small number (usually around 1–2%) of $+1$'s and -1's are randomly distributed, with the rest of the elements set to zero. By generating sparse vectors of a sufficiently high dimensionality in this way, the index vectors will be *nearly* orthogonal[i].
2. Each *unique term* is assigned an initially empty *context vector* of the same dimensionality d. The context vectors are then incrementally populated with context information by adding the (weighted) index vectors of the contexts in which the target term appears. With a sliding window context definition, this means that the index vectors of the surrounding terms are added to the target term's context vector. The meaning of a term, represented by its context vector, is effectively the (weighted) sum of all the contexts in which it occurs.

Random permutation

Models of distributional semantics, including RI, generally treat each context as a *bag of words*[j]. Such models are often criticized for failing to account for term order. Recently, methods have been developed for building distributional semantic models that store and emphasize word order information [45-47]. Random permutation (RP) [46] is a modification of RI that encodes term order information by simply *permuting* (i.e., shifting) the elements in the index vectors according to their direction and distance[k] from the target term before they are added to the context vector. For instance, before adding the index vector of a term two positions to the left of the target term, the elements are shifted two positions to the left; similarly, before adding the index vector of a term one position to the right of the target term, the elements are shifted one position to the right. In effect, each term has multiple unique representations: one index vector for each possible position relative to the target term in the context window. Incorporating term order information not only enables order-based retrieval; it also constrains the types of semantic relations that are captured.

Model parameters

There are a number of model parameters that need to be configured according to the task that the induced semantic spaces will be used for. For instance, the types of semantic relations captured depends on the context definition [43,48]. By employing a document-level context definition, relying on direct co-occurrences, one models *syntagmatic* relations. That is, two terms that frequently co-occur in the same documents are likely to be about the same general topic. By employing a sliding window context definition, one models *paradigmatic* relations. That is, two terms that frequently co-occur with similar sets of words – i.e., share neighbors – but do not necessarily co-occur themselves, are semantically similar. Synonymy is a prime example of a paradigmatic relation. The size of the context window also affects the types of relations that are modeled and needs to be tuned for the task at hand. This is also true for semantic spaces produced by RP; however, the precise impact of window size on RP spaces and the internal relations of their context vectors is yet to be studied in depth.

Method

The main idea behind this study is to enhance the performance on the task of extracting synonyms and abbreviation-expansion pairs by combining multiple and different semantic spaces – different in terms of (1) type of model and model parameters used, and (2) type of corpus from which the semantic space is induced. In addition to combining semantic spaces induced from a single corpus, we also combine semantic spaces induced from two different types of corpora: in this case, a clinical corpus (comprising health record notes) and a medical corpus (comprising journal articles). The notion of combining multiple semantic spaces to improve performance on some task is generalizable and can loosely be described as creating *ensembles of semantic spaces*. By combining semantic spaces, it becomes possible to benefit from model types that capture slightly different aspects of semantics, to exploit various model parameter configurations (which influence the types of semantic relations that are modeled), as well as to observe language use in potentially very different contexts (by employing more than one corpus type). We set out exploring this approach by querying each semantic space separately and then combining their output using a number of combination strategies (Figure 1).

The experimental setup can be divided into the following steps: (1) corpora preprocessing, (2) construction of semantic spaces from the two corpora (and from the conjoint corpus), (3) identification of the most profitable single-corpus (and conjoint corpus) combinations, (4) identification of the most profitable (disjoint) multiple-corpora combinations, (5) evaluations of the single-corpus (including the conjoint corpus) and multiple-corpora combinations, (6) post-processing of candidate terms, and (7) frequency threshold experiments. Once the corpora have been preprocessed, ten semantic spaces from each corpus, as well as the conjoint corpus, are induced with different context window sizes (RP spaces are induced with and without stop words). Ten pairs of semantic spaces are then combined using three different combination strategies. These are evaluated on the three tasks – (1) abbreviations → expansions, (2) expansions → abbreviations and (3) synonyms – using the development subsets of the reference standards (a list of medical abbreviation-expansion pairs for 1 and 2 and MeSH synonyms for 3). Performance is mainly measured as recall top 10, i.e. the proportion of expected candidate terms that are among a list of ten suggestions. The pair of semantic spaces involved in the most profitable combination for each corpus is then used to identify the most profitable multiple-corpora combinations, where eight different combination strategies are evaluated. The best single-corpus combinations are evaluated on the evaluation subsets of the reference standards, where using RI and RP in isolation constitute the two baselines. The best multiple-corpora combination is likewise evaluated on the evaluation subsets of the reference standards; here, the results are compared both to (1) semantic spaces induced from a single corpus and the conjoint corpus, and (2) ensembles of semantic spaces induced from a single corpus (and the conjoint corpus). Post-processing rules are then constructed using the development subsets of the reference standards and the outputs of the various semantic space

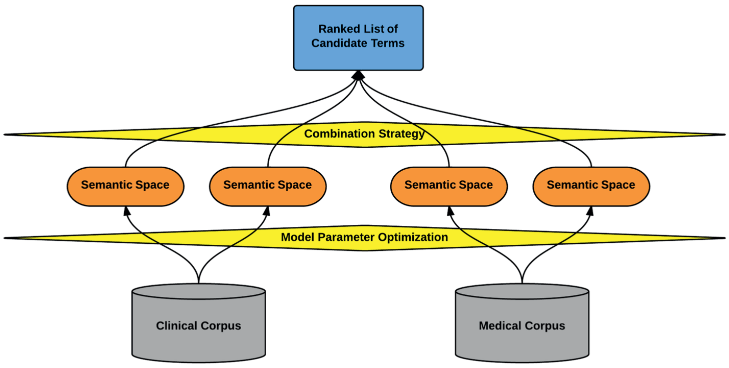


Figure 1 Ensembles of semantic spaces for synonym extraction and abbreviation expansion. Semantic spaces built with different model parameters are induced from different corpora. The output of the semantic spaces are combined in order to obtain better results compared to using a single semantic space in isolation.

combinations. These are evaluated on the evaluation subsets of the reference standards using the most profitable single-corpus and multiple-corpora ensembles. All evaluations on the evaluation subsets of the reference standards also include an evaluation of weighted precision, see Eq. 1:

$$Weighted\ Precision : P_w = \frac{\sum_{i=0}^{j-1} (j-i) \cdot f(i)}{\sum_{i=0}^{j-1} j-i}$$

$$\text{where} \qquad (1)$$

$$f(i) = \begin{cases} 1 \text{ if } i \in \{tp\} \\ 0 \text{ otherwise} \end{cases}$$

and j is the pre-specified number of labels – here, ten, except in the case of a dynamic cut-off – and $\{tp\}$ is the set of true positives. In words, this assigns a score to true positives according to their (reverse) ranking in the list, sums their scores and divides the total score by the maximum possible score (where all j labels are true positives).

Finally, we explore the impact of frequency thresholds (i.e., how many times each pair of terms in the reference standards needs to occur to be included) on performance.

Inducing semantic spaces from clinical and medical corpora

Each individual semantic space is constructed with one model type, using a predefined context window size and induced from a single corpus type. The semantic spaces are constructed with random indexing (RI) and random permutation (RP) using JavaSDM [49]. For all semantic spaces, a dimensionality of 1,000 is used (with 8 non-zero, randomly distributed elements in the index vectors: four 1s and four -1s). When the RI model is employed, the index vectors are weighted according to their distance from the target term, see Eq. 2, where $dist_{it}$ is the distance to the target term. When the RP model is employed, the elements of the index vectors are instead shifted according to their direction and distance from the target term; no weighting is performed.

$$weight_i = 2^{1-dist_{it}} \qquad (2)$$

For all models, window sizes of two (1 + 1), four (2 + 2) and eight (4 + 4) surrounding terms are used. In addition, RI spaces with a window size of twenty (10 + 10) are induced in order to investigate whether a significantly wider context definition may be profitable. Incorporating order information (RP) with such a large context window makes little sense; such an approach would also suffer from data sparseness. Different context definitions are experimented with in order to find one that is best suited to each task. The RI spaces are induced only from corpora that have been stop-word filtered, as co-occurrence information involving high-frequent and widely distributed words contribute very little to the meaning of terms. The RP spaces are, however, also induced from corpora in which stop words have been retained. The motivation behind this is that all words, including function words – these make up the majority of the items in the stop-word lists – are important to the syntactic structure of language and may thus be of value when modeling order information [45]. A stop-word list is created for each corpus by manually inspecting the most frequent word types and removing from the list those words that may be of interest, e.g. domain-specific terms. Each list consists of approximately 150 terms.

The semantic spaces are induced from two types of corpora – essentially belonging to different genres, but both within the wider domain of medicine: (1) a clinical corpus, comprising notes from health records, and (2) a medical corpus, comprising medical journal articles.

The *clinical corpus* contains a subset of the Stockholm EPR Corpus [50], which encompasses health records from the Karolinska University Hospital in Stockholm, Sweden over a five-year period[l]. The clinical corpus used in this study is created by extracting the free-text, narrative parts of the health records from a wide range of clinical practices. The clinical notes are written in Swedish by physicians, nurses and other health care professionals over a six-month period in 2008. In summary, the corpus comprises documents that each contain clinical notes documenting a single patient visit at a particular clinical unit.

The *medical corpus* contains the freely available subset of Läkartidningen (1996–2005), which is the Journal of the Swedish Medical Association [51]. It is a weekly journal written in Swedish and contains articles discussing new scientific findings in medicine, pharmaceutical studies, health economic evaluations, etc. Although these issues have been made available for research, the original order of the sentences has not been retained due to copyright reasons. The sentences thus appear in a randomized order, which means that the original texts cannot be recreated.

Both corpora are lemmatized using the *Granska Tagger* [52] and thereafter further preprocessed by removing punctuation marks and digits. Two versions of each corpus are created: one version in which the stop words are retained and one version in which they are removed[m]. As the sentences in Läkartidningen are given in a random order, a document break is indicated between each sentence for this corpus. It is thereby ensured that context information from surrounding sentences will not be incorporated in the induced semantic space. Statistics for the two corpora are shown in Table 1.

In summary, a total of thirty semantic spaces are induced – ten from each corpus type, and ten from the conjoint corpus. Four RI spaces are induced from each corpus type (12 in total), the difference being the context definition employed (1 + 1, 2 + 2, 4 + 4, 10 + 10). Six RP spaces are induced from each corpus type (18 in total), the difference being the context definition employed (1 + 1, 2 + 2, 4 + 4) and whether stop words have been removed or retained (sw).

Table 1 Corpora statistics

Corpus	With stop words	Without stop words	Segments
Clinical	∼42.5M tokens (∼0.4M types)	∼22.5M tokens (∼0.4M types)	268,727 documents
Medical	∼20.3M tokens (∼0.3M types)	∼12.1M tokens (∼0.3M types)	1,153,824 sentences

The number of tokens and unique terms (word types) in the medical and clinical corpus, with and without stop words.

Combinations of semantic spaces from a single corpus

Since RI and RP model semantic relations between terms in slightly different ways, it may prove profitable to combine them in order to increase the likelihood of capturing synonymy and identifying abbreviation-expansion pairs. In one study it was estimated that the overlap in the output produced by RI and RP spaces is, on average, only around 33% [46]: by combining them, we hope to capture different semantic properties of terms and, ultimately, boost results. The combinations from a single corpus type involve only two semantic spaces: one constructed with RI and one constructed with RP. In this study, the combinations involve semantic spaces with identical window sizes, with the following exception: RI spaces with a wide context definition (10 + 10) are combined with RP spaces with a narrow context definition (1 + 1, 2 + 2). The RI spaces are combined with RP spaces both with and without stop words.

Three different strategies of combing an RI-based semantic space with an RP space are designed and evaluated. Thirty combinations are evaluated for each corpus, i.e. sixty in total (Table 2). The three combination strategies are:

- $RI \subset RP_{30}$
  Finds the top ten terms in the RI space that are among the top thirty terms in the RP space.
- $RP \subset RI_{30}$
  Finds the top ten terms in the RP space that are among the top thirty terms in the RI space.
- $RI + RP$
  Sums the cosine similarity scores from the two spaces for each candidate term.

For the first two strategies ($RI \subset RP_{30}$ and $RP \subset RI_{30}$) a two-stage approach is applied. First one type of model is used (RI or RP) to produce an initial ranking of words according to a given query. The other model type, trained on the same corpus, is then used to re-rank the top 30 words produced by the first model according to its internal ranking. The intuition behind this approach is to see if synonyms and abbreviation-expansion pairs can be detected by trying to ensure that the set of contextually related words also have similar grammatical properties, and vice versa. In the third strategy ($RI + RP$), we apply a straightforward summing of the generated similarity scores.

Table 2 Overview of experiments conducted with a single semantic space

For each of the **2** corpora, **10** semantic spaces were induced.							
RI spaces	RI_20	RI_2		RI_4		RI_8	
RP spaces		RP_2	RP_2_sw	RP_4	RP_4_sw	RP_8	RP_8_sw
The induced semantic spaces were combined in **10** different combinations.							
Combinations							
Identical window size		RI_2, RP_2		RI_4, RP_4		RI_8, RP_8	
Identical window size, stop words		RI_2, RP_2_sw		RI_4, RP_4_sw		RI_8, RP_8_sw	
Large window size		RI_20, RP_2		RI_20, RP_4			
Large window size, stop words		RI_20, RP_2_sw		RI_20, RP_4_sw			
For each combination, **3** combination strategies were evaluated.							
Combination strategies	$RI \subset RP_{30}$		$RP \subset RI_{30}$		$RI + RP$		

For each of the two corpora and the conjoint corpus, 30 different combinations were evaluated. The configurations are described according to the following pattern: *model_windowSize*. For RP, *sw* means that stop words are retained in the semantic space. For instance, *model_20* means a window size of 10+10 was used.

Combinations of semantic spaces from multiple corpora

In addition to combining semantic spaces induced from one and the same corpus, a combination of semantic spaces induced from multiple corpora could potentially yield even better performance on the task of extracting synonyms and abbreviation-expansion pairs, especially if the terms of interest occur with some minimum frequency in both corpora. Such ensembles of semantic spaces – in this study consisting of four semantic spaces – allow not only different model types and model parameter configurations to be employed, but also allow us to capture language use in different genres or domains, in which terms may be used in slightly different contexts. The pair of semantic spaces from each corpus that is best able to perform each of the aforementioned tasks – consisting of two semantic spaces – is subsequently combined using various combination strategies.

The combination strategies can usefully be divided into two sets of approaches: in the first, the four semantic spaces are treated equally – irrespective of source – and combined in a single step; in the other, a two-step approach is assumed, wherein each pair of semantic spaces – induced from the same source – is combined separately before the combination of combinations is performed. In both sets of approaches, the outputs of the semantic spaces are combined in one of two ways: *SUM*, where the cosine similarity scores are merely summed, and *AVG*, where the average cosine similarity score is calculated based on the number of semantic spaces in which the term under consideration exists. The latter is an attempt to mitigate the effect of differences in vocabulary between the two corpora. In the two-step approaches, the *SUM*/*AVG* option is configurable for each step. In the single-step approaches, the combinations can be performed either with or without *normalization*, which in this case means replacing the exact cosine similarity scores of the candidate terms in the output of each queried semantic space with their ranking in the list of candidate terms. This means that the candidate terms are now sorted in ascending order, with zero being the highest score. When combining two or more lists of candidate terms, the combined list is also sorted in ascending order. The rationale behind this option is that the cosine similarity scores are relative and thus only valid within a given semantic space: combining similarity scores from semantic spaces constructed with different model types and parameter configurations, and induced from different corpora, might have adverse effects. In the two-step approach, normalization is always performed after combining each pair of semantic spaces. In total, eight combination strategies are evaluated:

Single-step approaches

- SUM: $RI_{clinical} + RP_{clinical} + RI_{medical} + RP_{medical}$
  Each candidate term's cosine similarity score in each semantic space is summed. The top ten terms from this list are returned.
- SUM, normalized: $norm(RI_{clinical}) + norm(RP_{clinical}) + norm(RI_{medical}) + norm(RP_{medical})$
  The output of each semantic space is first normalized by using the ranking instead of cosine similarity; each candidate term's (reverse) ranking in each semantic space is then summed. The top ten terms from this list are returned.
- AVG: $\dfrac{RI_{clinical} + RP_{clinical} + RI_{medical} + RP_{medical}}{count_{term}}$
  Each candidate term's cosine similarity score in each semantic space is summed; this value is then averaged over the number of semantic spaces in which the term exists. The top ten terms from this list are returned.
- AVG, normalized: $\frac{norm(RI_{clinical}) + norm(RP_{clinical}) + norm(RI_{medical}) + norm(RP_{medical})}{count_{term}}$
  The output of each semantic space is first normalized by using the ranking instead of cosine similarity; each

candidate term's normalized score in each semantic space is then summed; this value is finally averaged over the number of semantic spaces in which the term exists. The top ten terms from this list are returned.

Two-step approaches

- SUM→SUM: $norm(RI_{clinical} + RP_{clinical}) + norm(RI_{medical} + RP_{medical})$
  Each candidate term's cosine similarity score in each pair of semantic spaces is first summed; these are then normalized by using the ranking instead of the cosine similarity; finally, each candidate term's normalized score is summed. The top ten terms from this list are returned.
- AVG→AVG:
  $\frac{norm\left(\frac{RI_{clinical} + RP_{clinical}}{count_{term-source-a}}\right) + norm\left(\frac{RI_{medical} + RP_{medical}}{count_{term-source-b}}\right)}{count_{term-source-a} + count_{term-source-b}}$
  Each candidate term's cosine similarity score for each pair of semantic spaces is first summed; for each pair of semantic spaces, this value is then averaged over the number of semantic spaces in that pair in which the term exists; these are subsequently normalized by using the ranking instead of the cosine similarity; each candidate term's normalized score in each combined list is then summed and averaged over the number of semantic spaces in which the term exists (in both pairs of semantic spaces). The top ten terms from this list are returned.
- SUM→AVG:
  $\frac{norm(RI_{clinical} + RP_{clinical}) + norm(RI_{medical} + RP_{medical})}{count_{term}}$
  Each candidate term's cosine similarity score for each pair of semantic spaces is first summed; these are then normalized by using the ranking instead of the cosine similarity; each candidate term's normalized score in each combined list is then summed and averaged over the number of semantic spaces in which the term exists. The top ten terms from this list are returned.
- AVG→SUM: $norm\left(\frac{RI_{clinical} + RP_{clinical}}{count_{term}}\right) + norm\left(\frac{RI_{medical} + RP_{medical}}{count_{term}}\right)$
  Each candidate term's cosine similarity score for each pair of semantic spaces is first summed and averaged over the number of semantic spaces in that pair in which the term exists; these are then normalized by using the ranking instead of the cosine similarity; each candidate term's normalized score in each combined list is finally summed. The top ten terms from this list are returned.

Post-processing of candidate terms

In addition to creating ensembles of semantic spaces, simple filtering rules are designed and evaluated for their ability to enhance performance further on the task of extracting synonyms and abbreviation-expansion pairs. For obvious reasons, this is easier for abbreviation-expansion pairs than for synonyms.

With regards to abbreviation-expansion pairs, the focus is on increasing precision by discarding poor suggestions in favor of potentially better ones. This is attempted by exploiting properties of the abbreviations and their corresponding expansions. The development subset of the reference standard (see Evaluation framework) is used to construct rules that determine the validity of candidate terms. For an abbreviation-expansion pair to be considered valid, each letter in the abbreviation has to be present in the expansion and the letters also have to appear in the same order. Additionally, the length of abbreviations and expansions is restricted, requiring an expansion to contain more than four letters, whereas an abbreviation is allowed to contain a maximum of four letters. These rules are shown in Eq. 3 and Eq. 4.

For synonym extraction, cut-off values for rank and cosine similarity are instead employed. These cut-off values are tuned to maximize precision for the best semantic space combinations in the development subset of the reference standard, without negatively affecting recall (see Figures 2, 3 and 4). Used cut-off values are shown in Eq. 5 for the clinical corpus, in Eq. 6 for the medical corpus, and in Eq. 7 for the combination of the two corpora. In Eq. 7, *Cos* denotes the combination of the cosine values, which means that it has a maximum value of four rather than one.

$$Exp \rightarrow Abbr = \begin{cases} True,\ if\ (Len < 5) \wedge (Sub_{out} = True) \\ False,\ Otherwise \end{cases} \tag{3}$$

$$Abbr \rightarrow Exp = \begin{cases} True,\ if\ (Len > 4) \wedge (Sub_{in} = True) \\ False,\ Otherwise \end{cases} \tag{4}$$

$$Syn_{clinical} = \begin{cases} True,\ if\ (Cos \geq 0.60) \vee (Cos \geq 0.40 \wedge Rank < 9) \\ False,\ Otherwise \end{cases} \tag{5}$$

$$Syn_{medical} = \begin{cases} True,\ if\ (Cos \geq 0.50) \\ False,\ Otherwise \end{cases} \tag{6}$$

$$Syn_{clinical+medical} = \begin{cases} True,\ if\ (Cos \geq 1.9) \vee (Cos \geq 1.8 \wedge Rank < 6) \vee (Cos \geq 1.75 \wedge Rank < 3) \\ False,\ Otherwise \end{cases} \tag{7}$$

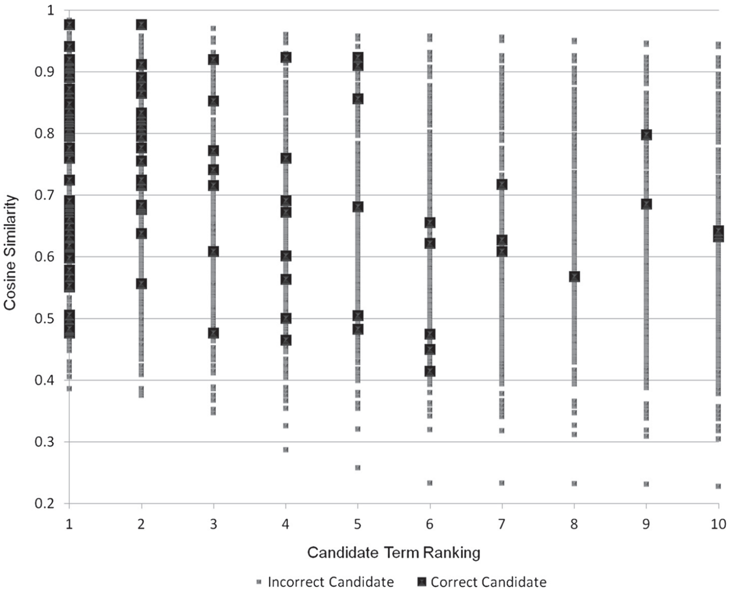


Figure 2 Distribution of candidate terms for the clinical corpus. The distribution (cosine similarity and rank) of candidates for synonyms for the best combination of semantic spaces induced from the clinical corpus. The results show the distribution for query terms in the development reference standard.

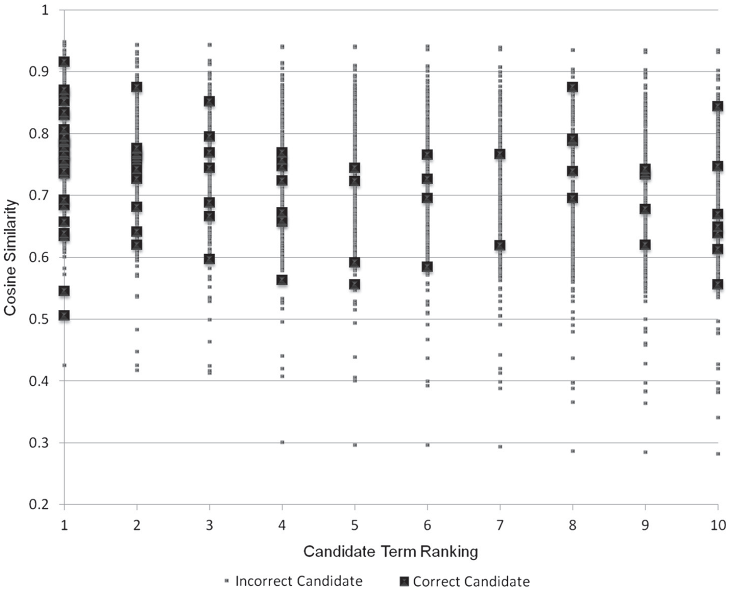


Figure 3 Distribution of candidate terms for the medical corpus. The distribution (cosine similarity and rank) of candidates for synonyms for the best combination of semantic spaces induced from the medical corpus. The results show the distribution for query terms in the development reference standard.

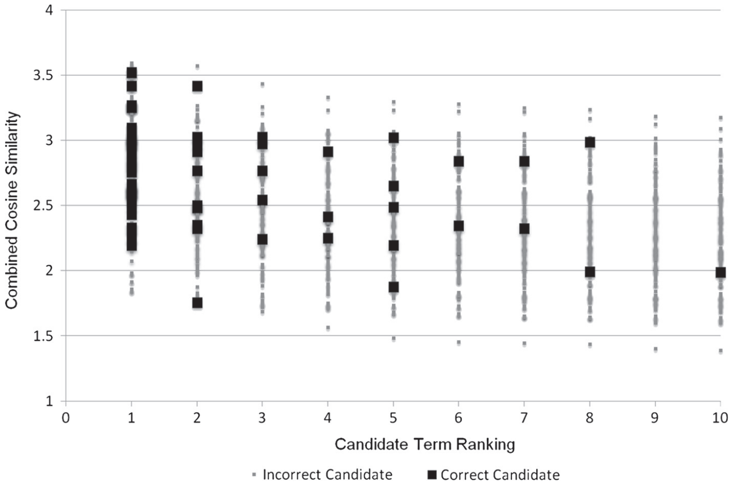


Figure 4 Distribution of candidate terms for clinical + medical corpora. The distribution (combined cosine similarity and rank) of candidates for synonyms for the ensemble of semantic spaces induced from medical and clinical corpora. The results show the distribution for query terms in the development reference standard.

Cos: *Cosine similarity between candidate term and query term.*

Rank: *The ranking of the candidate term, ordered by cosine similarity.*

Sub$_{out}$: *Whether each letter in the candidate term is present in the query term, in the same order and with identical initial letters.*

Sub$_{in}$: *Whether each letter in the query term is present in the candidate term, in the same order and with identical initial letters.*

Len: *The length of the candidate term.*

The post-processing filtering rules are employed in two different ways. In the first approach, the semantic spaces are forced to suggest a predefined number of candidate terms (ten), irrespective of how good they are deemed to be by the semantic space. Candidate terms are retrieved by the semantic space until ten have been classified as correct according to the post-processing rules, or until one hundred candidate terms have been classified. If less than ten are classified as incorrect, the highest ranked discarded terms are used to populate the remaining slots in the final list of candidate terms. In the second approach, the semantic spaces are allowed to suggest a dynamic number of candidate terms, with a minimum of one and a maximum of ten. If none of the highest ranked terms are classified as correct, the highest ranked term is suggested.

Evaluation framework

Evaluation of the numerous experiments is carried out with the use of reference standards: one contains known abbreviation-expansion pairs and the other contains known synonyms. The semantic spaces and their various combinations are evaluated for their ability to extract known abbreviations/expansions (*abbr→exp* and *exp→abbr*) and synonyms (*syn*) – according to the employed reference standard – for a given query term in a list of ten candidate terms (recall top 10). Recall is prioritized in this study and any decisions, such as deciding which model parameters or which combination strategies are the most profitable, are solely based on this measure. When precision is reported, it is calculated as weighted precision, where the weights are assigned according to the ranking of a correctly identified term.

The reference standard for abbreviations is taken from Cederblom [53], which is a book that contains lists of medical abbreviations and their corresponding expansions. These abbreviations have been manually collected from Swedish health records, newspapers, scientific articles, etc. For the synonym extraction task, the reference standard is derived from the freely available part of the Swedish version of MeSH [54] – a part of UMLS – as well as a Swedish extension that is not included in UMLS [55]. As the semantic spaces are constructed only to model unigrams, all multiword expressions are removed from the reference standards. Moreover, hypernym/hyponym and other non-synonym pairs found in the UMLS version of MeSH are manually removed from the reference standard for the synonym extraction task. Models of distributional semantics sometimes struggle to model the meaning of rare terms accurately, as the statistical basis for their representation is insufficiently solid. As a result, we only include term pairs that occur at least fifty times in each respective corpus. This, together with the fact that term frequencies differ from corpus to corpus, means that one separate reference standard is used for the evaluation of the clinical corpus and another is used for the evaluation of the medical corpus. For evaluating combinations of semantic spaces induced from different corpora, a third – common – reference standard is therefore created, in which only term pairs that occur at least fifty times in both corpora are included. Included terms are not restricted to form pairs; in the reference standard for the synonym extraction task, some form larger groups of terms with synonymous relations. There are also abbreviations with several possible expansions, as well as expansions with several possible abbreviations. The term pairs (or n-tuples) in each reference standard are randomly split into a *development set* and an *evaluation set* of roughly equal size. The development sets are used for identifying the most profitable ensembles of semantic spaces (with optimized parameter settings, such as window size and whether to include stop words in the RP spaces) for each of the three tasks, as well as for creating the post-processing filtering rules. The evaluation sets are used for the final evaluation to assess the expected performance of the ensembles in a deployment setting. Baselines for the single-corpus ensembles are created by employing RI and RP in isolation; baselines for the multiple-corpora ensembles are created by using the most profitable clinical and medical ensembles from the single-corpus experiments, as well a single space induced from the conjoint corpus and an ensemble of semantic spaces induced from the conjoint corpus. Statistics for the reference standards are shown in Table 3. The differences in recall between the different semantic spaces/ensembles, when evaluated on the evaluation subset of the reference standards, are tested for statistical significance. The exact binomial sign test is used ([56], pp. 532–535), assuming independence between all query terms.

In addition to the automatic evaluation using the reference standards, a small manual evaluation is also carried out on the synonym task. A random sample of 30 query terms (out of 135 terms in the *Clinical + Medical* reference standard) and their respective ten candidate terms as suggested by the best combination of semantic spaces is investigated and a manual classification of the semantic relation between each of the candidate terms and the target term is carried out. The candidate terms are manually classified as either a synonym, an antonym[n], a hypernym[o], a hyponym or an alternative spelling (for instance *rinitis/rhinitis*) of the target term.

Results

The experimental setup was designed in such a manner that the semantic spaces that performed best in combination for a single corpus would also be used in the subsequent combinations from multiple corpora. Identifying the most profitable combination strategy for each of the three tasks was achieved using the development subsets of the reference standards. These combinations were then evaluated on separate evaluation sets containing unseen data. All further experiments, including the post-processing of candidate terms, were carried out with these combinations on the evaluation sets. This is therefore also the order in which the results will be presented.

Combination strategies: a single corpus

The first step involved identifying the most appropriate window sizes for each task, in conjunction with evaluating the combination strategies. The reason for this is that the optimal window sizes for RI and RP in isolation are not necessarily identical to the optimal window sizes when RI and RP are combined. In fact, when RI is used in isolation, a window size of 2 + 2 performs best on the two abbreviation-expansion tasks, and a window size of 10 + 10 performs best on the synonym task. For RP, a semantic space with a window size of 2 + 2 yields the

Table 3 Reference standards statistics

Reference standard	Clinical corpus			Medical corpus			Clinical + Medical		
	Size	2 Cor	3 Cor	Size	2 Cor	3 Cor	Size	2 Cor	3 Cor
Abbr→Exp (Devel)	117	9.4%	0.0%	55	13%	1.8%	42	14%	0%
Abbr→Exp (Eval)	98	3.1%	0.0%	55	11%	0%	35	2.9%	0%
Exp→Abbr (Devel)	110	8.2%	1.8%	63	4.7%	0%	45	6.7%	0%
Exp→Abbr (Eval)	98	7.1%	0.0%	61	0%	0%	36	0%	0%
Syn (Devel)	334	9.0%	1.2%	266	11%	3.0%	122	4.9%	0%
Syn (Eval)	340	14%	2.4%	263	13%	3.8%	135	11%	0%

Size shows the number of queries, *2 cor* shows the proportion of queries with two correct answers and *3 cor* the proportion of queries with three (or more) correct answers. The remaining queries have one correct answer.

best results on two of the tasks – *abbr*→*exp* and *syn* – while a window size of 4 + 4 is more successful on the *exp*→*abbr* task. These are the model configurations used in the RI and RP baselines, to which the single-corpus combination strategies are compared in the final evaluation.

Using the semantic spaces induced from the clinical corpus, the $RI + RP$ combination strategy, wherein the cosine similarity scores are merely summed, is the most successful on all three tasks: 0.42 recall on the *abbr*→*exp* task, 0.32 recall on the *exp*→*abbr* task, and 0.40 recall on the *syn* task (Table 4). For the abbreviation expansion task, a window size of 2 + 2 appears to work well for both models, with the RP space retaining stop words. On the task of identifying the abbreviated form of an expansion, semantic spaces with window sizes of 2 + 2 and 4 + 4 perform equally well; the RP spaces should include stop words. Finally, on the synonym extraction task, an RI space with a large context window (10 + 10) in conjunction with an RP space with stop words and a window size of 2 + 2 is the most profitable.

Using the semantic spaces induced from the medical corpus, again, the $RI + RP$ combination strategy outperforms the $RI \subset RP_{30}$ and $RP \subset RI_{30}$ strategies: 0.10 recall on the *abbr*→*exp* task, 0.08 recall on the *exp*→*abbr* task, and 0.30 recall on the *syn* task (Table 5) are obtained. This combination outperforms the other two by a large margin on the *exp*→*abbr* task: 0.08 recall compared to 0.03 recall. The most appropriate window sizes for capturing these phenomena in the medical corpus are fairly similar to those that worked best with the clinical corpus. On the *abbr*→*exp* task, the optimal window sizes are indeed identical across the two corpora: a 2 + 2 context window with an RP space that incorporates stop words yields the highest performance. For the *exp*→*abbr* task, a slightly larger context window of 4 + 4 seems to work well – again, with stop words retained in the RP space. Alternatively, combining a large RI space (10 + 10) with a smaller RP space (2 + 2, with stop words) performs comparably on this task and with this test data. Finally, for synonyms, a large RI space (10 + 10) with a very small RP space (1 + 1) that retains all words best captures this phenomenon with this type of corpus.

Using the semantic spaces induced from the conjoint corpus, the $RI \subset RP_{30}$ combination strategy outperforms the other two strategies on the abbr→exp task: 0.30 recall compared to 0.25 and 0.23 (Table 6). On the exp→abbr task, this and the $RI + RP$ combination strategy perform equally well, with 0.18 recall. Finally, on the synonym task, the $RI + RP$ performs best with a recall of 0.46. In general, somewhat larger window sizes seem to work better when combining semantic spaces induced from the conjoint corpus.

The best-performing combinations from each corpus and for each task were then treated as (ensemble) baselines in the final evaluation, where combinations of semantic spaces from multiple corpora are evaluated.

Combination strategies: multiple corpora

The pair of semantic spaces from each corpus that performed best on the three tasks were subsequently employed in combinations that involved four semantic spaces – two from each corpus: one RI space and one RP space. The single-step approaches generally performed better than the two-step approaches, with some exceptions (Table 7). The most successful ensemble was a simple single-step approach, where the cosine similarity scores produced by each semantic space were simply summed (*SUM*), yielding 0.32 recall for *abbr*→*exp*, 0.17 recall for *exp*→*abbr*, and 0.52 recall for *syn*. The *AVG* option, although the second-highest performer on the abbreviation-expansion tasks, yielded significantly poorer results. Normalization, whereby ranking was used instead of cosine similarity, invariably affected performance negatively, especially when employed in conjunction with *SUM*. The two-step approaches performed significantly worse than all non-normalized single-step approaches, with the sole exception taking place on the synonym extraction task. It should be noted that normalization was

Table 4 Results on clinical development set

Strategy	Abbr→Exp			Exp→Abbr			Syn		
	RI	RP	Result	RI	RP	Result	RI	RP	Result
$RI \subset RP_{30}$	RI_8	RP_8_sw	0.38	RI_8	RP_8	0.30	RI_8	RP_8	0.39
$RP \subset RI_{30}$	RI_20	RP_4_sw	0.35	RI_4 RI_20	RP_4_sw RP_4_sw	0.30	RI_8 RI_8 RI_20	RP_8 RP_8_sw RP_2_sw	0.38
$RI + RP$	RI_4	RP_4_sw	**0.42**	RI_4 RI_8	RP_4_sw RP_8_sw	**0.32**	RI_20	RP_4_sw	**0.40**

Results (recall, top ten) of the best configurations for each model and model combination on the three tasks. The configurations are described according to the following pattern: *model_windowSize*. For RP, *sw* means that stop words are retained in the model.

Table 5 Results on medical development set

Strategy	Abbr→Exp			Exp→Abbr			Syn		
	RI	RP	Result	RI	RP	Result	RI	RP	Result
$RI \subset RP_{30}$	RI_4	RP_4_sw	0.08	RI_2	RP_2	0.03	RI_20	RP_4_sw	0.26
	RI_20	RP_2		RI_4	RP_4				
	RI_20	RP_4_sw		RI_4	RP_4_sw				
				RI_8	RP_8				
				RI_20	RP_2				
				RI_20	RP_2_sw				
				RI_20	RP_4				
				RI_20	RP_4_sw				
$RP \subset RI_{30}$	RI_2	RP_2_sw	0.08	RI_2	RP_2	0.03	RI_8	RP_8_sw	0.24
	RI_4	RP_4		RI_2	RP_2_sw				
	RI_4	RP_4_sw		RI_4	RP_4				
	RI_8	RP_8		RI_4	RP_4_sw				
	RI_8	RP_8_sw		RI_8	RP_8				
	RI_20	RP_2_sw		RI_8	RP_8_sw				
	RI_20	RP_4		RI_20	RP_2				
	RI_20	RP_4_sw		RI_20	RP_2_sw				
				RI_20	RP_4				
				RI_20	RP_4_sw				
$RI + RP$	RI_4	RP_4_sw	**0.10**	RI_8	RP_8_sw	**0.08**	RI_20	RP_2_sw	**0.30**
				RI_20	RP_4_sw				

Results (recall, top ten) of the best configurations for each model and model combination on the three tasks. The configurations are described according to the following pattern: *model_windowSize*. For RP, *sw* means that stop words are retained in the model.

always performed in the two-step approaches – this was done after each pair of semantic spaces from a single corpus had been combined. Of the four two-step combination strategies, *AVG*→*AVG* and *AVG*→*SUM* performed best, with identical recall scores on the three tasks.

Final evaluations

The combination strategies that performed best on the development sets were finally evaluated on completely unseen data in order to assess their generalizability to new data and to assess their expected performance in a

Table 6 Conjoined corpus space results on clinical + medical development set

Strategy	Abbr→Exp			Exp→Abbr			Syn		
	RI	RP	Result	RI	RP	Result	RI	RP	Result
$RI \subset RP_{30}$	RI_4	RP_4_sw	**0.30**	RI_4	RP_4_sw	**0.18**	RI_8	RP_8_sw	0.41
	RI_20	RP_4_sw							
$RP \subset RI_{30}$	RI_4	RP_4	0.23	RI_4	RP_4_sw	0.13	RI_8	RP_8	0.36
	RI_4	RP_4_sw		RI_8	RP_8_sw		RI_8	RP_8_sw	
	RI_8	RP_8		RI_20	RP_2_sw		RI_20	RP_2_sw	
	RI_20	RP_2		RI_20	RP_4_sw		RI_20	RP_4_sw	
	RI_20	RP_4							
$RI + RP$	RI_2	RP_2_sw	0.25	RI_4	RP_4_sw	**0.18**	RI_8	RP_8_sw	**0.46**
				RI_8	RP_8_sw				
				RI_20	RP_4_sw				

Results (recall, top ten) of the best configurations for each model and model combination on the three tasks. The configurations are described according to the following pattern: *model_windowSize*. For RP, *sw* means that stop words are retained in the model.

Table 7 Disjoint corpus ensemble results on clinical + medical development set

Strategy	Normalize	Abbr→Exp		Exp→Abbr		Syn	
		Clinical	Medical	Clinical	Medical	Clinical	Medical
		RI_4	*RI_4*	*RI_4*	*RI_8*	*RI_20*	*RI_20*
		RP_4_sw	*RP_4_sw*	*RP_4_sw*	*RP_8_sw*	*RP_4_sw*	*RP_2_sw*
AVG	*True*	0.13		0.09		0.39	
AVG	*False*	0.24		0.11		0.39	
SUM	*True*	0.13		0.09		0.34	
SUM	*False*	**0.32**		**0.17**		**0.52**	
AVG→AVG		0.15		0.09		0.41	
SUM→SUM		0.13		0.07		0.40	
AVG→SUM		0.15		0.09		0.41	
SUM→AVG		0.13		0.07		0.40	

Results (P = weighted precision, R = recall, top ten) of the best models with and without post-processing on the three tasks. Dynamic # of suggestions allows the model to suggest less than ten terms in order to improve precision. The results are based on the application of the model combinations to the development data.

deployment setting. Each evaluation phase involves comparing the results to one or more baselines: in the case of single-corpus combinations, the comparisons are made to RI and RP in isolation; in the case of multiple-corpora combinations, the comparisons are made to semantic spaces induced from a single corpus (as well as the conjoint corpus) and ensembles of semantic spaces induced from a single corpus (and, again, the conjoint corpus).

When applying the single-corpus combinations from the clinical corpus, the following results were obtained: 0.31 recall on *abbr→exp*, 0.20 recall on *exp→abbr*, and 0.44 recall on *syn* (Table 8). Compared to the results on the development sets, the results on the two abbreviation-expansion tasks decreased by approximately ten percentage points; on the synonym extraction task, the performance increased by a couple of percentage points. The RI baseline was outperformed on all three tasks; the RP baseline was outperformed on two out of three tasks, with the exception of the *exp→abbr* task. Finally, it might be interesting to point out that the RP baseline performed better than the RI baseline on the two abbreviation-expansion tasks, but that the RI baseline did somewhat better on the synonym extraction task.

With the medical corpus, the following results were obtained: 0.17 recall on *abbr→exp*, 0.11 recall on *exp→abbr*, and 0.34 recall on *syn* (Table 9). Compared to the results on the development sets, the results were higher for all three tasks. Both the RI and RP baselines were outperformed, with a considerable margin, by their combination. However, the improvement in recall for the combination method compared to the best baseline was only statistically significant for the synonym task. In complete contrast to the clinical corpus, the RI baseline here outperformed the RP baseline on the two abbreviation-expansion tasks, but was outperformed by the RP baseline on the synonym extraction task.

When applying the disjoint corpora ensembles, the following results were obtained on the evaluation sets: 0.30

Table 8 Results on clinical evaluation set

Evaluation configuration	Abbr→Exp		Exp→Abbr		Syn	
	RI_4+RP_4_sw		*RI_4+RP_4_sw*		*RI_20+RP_4_sw*	
	P	R	P	R	P	R
RI Baseline	0.04	0.22	0.03	0.19	0.07	0.39
RP Baseline	0.04	0.23	0.04	0.24	0.06	0.36
Clinical Ensemble	0.05	0.31	0.03	0.20	0.07	**0.44**
+Post-Processing (Top 10)	0.08	**0.42**	0.05	**0.33**	**0.08**	0.43
+Dynamic Cut-Off (Top ≤ 10)	**0.11**	0.41	**0.12**	0.33	0.08	0.42

Results (P = weighted precision, R = recall, top ten) of the best models with and without post-processing on the three tasks. Dynamic # of suggestions allows the model to suggest less than ten terms in order to improve precision. The results are based on the application of the model combinations to the evaluation data. The improvements in recall between the best baseline and the ensemble method for the synonym task and for the abbr→exp task are both statistically significant for a p-value < 0.05. (abbr→exp task: p-value = 0.022 and synonym task: p-value = 0.002.) The improvement in recall that was achieved by post-processing is statistically significant for both abbreviation tasks (p-value = 0.001 for abbr→exp and p-value = 0.000 for exp→abbr).

Table 9 Results on medical evaluation set

Evaluation configuration	**Abbr→Exp**		**Exp→Abbr**		**Syn**	
	RI_4+RP_4_sw		*RI_8+RP_8_sw*		*RI_20+RP_2_sw*	
	P	**R**	**P**	**R**	**P**	**R**
RI baseline	0.02	0.09	0.01	0.08	0.03	0.18
RP baseline	0.01	0.06	0.01	0.05	0.05	0.26
Medical ensemble	0.03	**0.17**	0.01	**0.11**	**0.06**	**0.34**
+Post-processing (top 10)	0.03	0.17	0.02	0.11	0.06	0.34
+Dynamic cut-off (top ≤ 10)	**0.17**	0.17	**0.10**	0.11	0.06	0.34

Results (P = weighted precision, R = recall, top ten) of the best semantic spaces with and without post-processing on the three tasks. Dynamic # of suggestions allows the model to suggest less than ten terms in order to improve precision. The results are based on the application of the model combinations to the evaluation data. The difference in recall when using the ensemble method compared to the best baseline is only statistically significant (p-value < 0.05) for the synonym task (p-value = 0.000).

recall on *abbr→exp*, 0.19 recall on *exp→abbr*, and 0.47 recall on *syn* (Table 10). Compared to the results on the development sets, the results decreased somewhat on two of the tasks, with *exp→abbr* the exception. The p-values for the significance tests of the recall differences in Table 10 are shown in Table 11. The two ensemble baselines were clearly outperformed by the larger ensemble of semantic spaces from two types of corpora on two of the tasks; the clinical ensemble baseline performed equally well on the *exp→abbr* task.

Post-processing

In an attempt to further improve results, simple post-processing of the candidate terms was performed. In one setting, the system was forced to suggest ten candidate terms regardless of their cosine similarity score or other properties of the terms, such as their length. In another setting, the system had the option of suggesting a dynamic number – ten or less – of candidate terms.

This was unsurprisingly more effective on the two abbreviation-expansion tasks. With the clinical corpus, recall improved substantially with the post-processing filtering: from 0.31 to 0.42 on *abbr→exp* and from 0.20 to 0.33 on *exp→abbr* (Table 8). With the medical corpus, however, almost no improvements were observed for these tasks (Table 9). For the combination of semantic spaces from the two corpora, the improvements in recall after applying post-processing on the two abbreviation tasks are not statistically significant (Table 10).

With a dynamic cut-off, only precision could be improved, although at the risk of negatively affecting recall. With the clinical corpus, recall was largely

Table 10 Results on clinical + medical evaluation set

Evaluation configuration	**Abbr→Exp**		**Exp→Abbr**		**Syn**	
	Clinical	**Medical**	**Clinical**	**Medical**	**Clinical**	**Medical**
	RI_4	***RI_4***	***RI_4***	***RI_8***	***RI_20***	***RI_20***
	RP_4_sw	***RP_4_sw***	***RP_4_sw***	***RP_8_sw***	***RP_4_sw***	***RP_2_sw***
	SUM, *False*		**SUM, *False***		**SUM, *False***	
	P	**R**	**P**	**R**	**P**	**R**
Clinical space	0.03	0.17	0.03	0.19	0.05	0.29
Medical space	0.01	0.06	0.01	0.08	0.03	0.18
Conjoint corpus space	0.03	0.19	0.01	0.08	0.05	0.30
Clinical ensemble	0.04	0.24	0.03	0.19	0.06	0.34
Medical ensemble	0.02	0.11	0.01	0.11	0.05	0.33
Conjoint corpus ensemble	0.03	0.19	0.02	0.14	0.07	0.40
Disjoint corpora ensemble	0.05	0.30	0.03	0.19	**0.08**	**0.47**
+Post-processing (top 10)	0.07	**0.39**	0.06	**0.33**	0.08	0.47
+Dynamic cut-off (top ≤ 10)	**0.28**	0.39	**0.31**	0.33	0.08	0.45

Results (P = weighted precision, R = recall, top ten) of the best semantic spaces and ensembles on the three tasks. The results are based on the clinical + medical evaluation set and are grouped according to the number of semantic spaces employed: one, two or four. The disjoint corpus ensemble is performed with and without post-processing. A dynamic cut-off allows less than ten terms to be suggested in an attempt to improve precision. Results for tests of statistical significance are shown in Table 11.

Table 11 P-values for recall results presented in Table 10

P-values, recall (synonym task)	Medical space	Conjoint corpus	Clinical ensemble	Medical ensemble	Conjoint corp. ens.	Disjoint corp. ens.
Clinical space	**0.011**	1.000	0.057	0.885	**0.003**	**0.000**
Medical space	-	**0.004**	**0.000**	**0.000**	**0.000**	**0.000**
Conjoint corpus	-	-	0.210	1.000	**0.001**	**0.000**
Clinical ensemble	-	-	-	0.480	0.189	**0.001**
Medical ensemble	-	-	-	-	**0.047**	**0.000**
Conjoint corp. ens.	-	-	-	-	-	**0.041**

P-values for the differences between the recall results on the synonym task for the semantic spaces/ensembles presented in Table 10. P-values showing a statistically significant difference (p-value < 0.05) are presented in bold-face.

P-values for the post-processing and for the abbr→exp and exp→abbr are not shown in the table. However, for the significance level p-value < 0.05, there were no statistically significant recall difference between the standard Disjoint Corpora Ensemble and the post-processing version for any of the three tasks (p-value = 0.25 for abbr→exp and p-value = 0.062 for exp→abbr). When testing the recall difference between the pairs of semantic spaces/ensembles shown in Table 10 for the abbr→exp task, there was only a significant difference for the pairs Medical Space vs. Clinical Ensemble (p-value = 0.039), Medical Space vs. Disjoint Corpora Ensemble (p-value = 0.004) and Medical Ensemble vs. Disjoint Corpora Ensemble (p-value = 0.039). For the exp→abbr task, there were no statistically significant differences.

unaffected for the two abbreviation-expansion task, while precision improved by 3–7 percentage points (Table 8). With the medical corpus, the gains were even more substantial: from 0.03 to 0.17 precision on *abbr→exp* and from 0.02 to 0.10 precision on *exp→abbr* – without having any impact on recall (Table 9). The greatest improvements on these tasks were, however, observed with the combination of semantic spaces from multiple corpora: precision increased from 0.07 to 0.28 on *abbr→exp* and from 0.06 to 0.31 on *exp→abbr* – again, without affecting recall (Table 10).

In the case of synonyms, this form of post-processing is more challenging, as there are no simple properties of the terms, such as their length, that can serve as indications of their quality as candidate synonyms. Instead, one has to rely on their use in different contexts and grammatical properties; as a result, cosine similarity and ranking of the candidate terms were exploited in an attempt to improve the candidate synonyms. This approach was, however, clearly unsuccessful for both corpora and their combination, with almost no impact on either precision or recall. In a single instance – with the clinical corpus – precision increased by one percentage point, albeit at the expense of recall, which suffered a comparable decrease (Table 8). With the combination of semantic spaces from two corpora, the dynamic cut-off option resulted in a lower recall score, without improving precision (Table 10).

Frequency thresholds

In order to study the impact of different frequency thresholds – i.e., how often each pair of terms had to occur in the corpora to be included in the reference standard – on the task of extracting synonyms, the best ensemble system was applied to a range of evaluation sets with different thresholds from 1 to 100 (Figure 5). With a low frequency threshold, it is clear that a lower performance is obtained. For instance, if each synonym pair only needs to occur at least once in both corpora, a recall of 0.17 is obtained. As the threshold is increased, recall increases too - up to a frequency threshold of around 50, after which no performance boosts are observed. Already with a frequency threshold of around 30, the results seem to level off. With

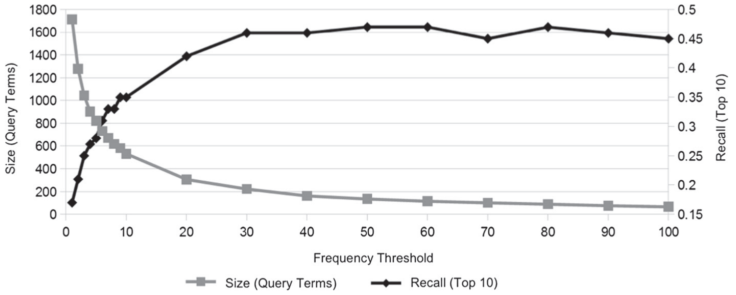


Figure 5 Frequency thresholds. The relation between recall and the required minimum frequency of occurrence for the reference standard terms in both corpora. The number of query terms for each threshold value is also shown.

frequency thresholds over 100, there is not enough data in this case to produce any reliable results.

Discussion

The results clearly demonstrate that combinations of semantic spaces lead to improved results on the synonym extraction task. For the two abbreviation tasks, most of the observed performance gains were not statistically significant. Combining random indexing and random permutation allows slightly different aspects of lexical semantics to be captured; by combining them, stronger semantic relations between terms are extracted, thereby increasing the performance on these tasks. Combining semantic spaces induced from different corpora further improves performance. This demonstrates the potential of distributional ensemble methods, of which this – to the extent of our knowledge – is the primary implementation of its kind, and it only scratches the surface. In this initial study, only four semantic spaces were used; however, with increasing computational capabilities, there is nothing stopping a much larger number of semantic spaces from being combined. These can capture various aspects of semantics – aspects which may be difficult, if not impossible, to incorporate into a single model – from a large variety of observational data on language use, where the contexts may be very different.

Clinical vs. medical corpora

When employing corpus-driven methods to support lexical resource development, one naturally needs to have access to a corpus in the target domain that reflects the language use one wishes to model. Hence, one cannot, without due qualification, state that one corpus type is better than another for the extraction of synonyms or abbreviation-expansion pairs. This is something that needs to be duly considered when comparing the results for the semantic spaces on the clinical and medical corpora, respectively. Another issue concerns the size of each corpus: in fact, the size of the medical corpus is only half as large as the clinical corpus (Table 1). The reference standards used in the respective experiments are, however, not identical: each term pair had to occur at least fifty times to be included – this will differ across corpora. To some extent this mitigates the effect of the total corpus size and makes the comparison between the two corpora fairer; however, differences in reference standards also entail that the results presented in Tables 8 and 9 are not directly comparable. Another difference between the two corpora is that the clinical corpus contains more unique terms (word types) than the medical corpus, which might indicate that it consists of a larger number of concepts. It has previously been shown that it can be beneficial, indeed important, to employ a larger dimensionality when using corpora with a large vocabulary, as is typically the case in the clinical domain [57]; in this study a dimensionality of 1,000 was used to induce all semantic spaces. The results, on the contrary, seem to indicate that better performance is generally obtained with the semantic spaces induced from the clinical corpus.

An advantage of using non-sensitive corpora like the medical corpus employed in this study is that they are generally more readily obtainable than sensitive clinical data. Perhaps such and similar sources can complement smaller clinical corpora and yet obtain similar or potentially even better results.

Combining semantic spaces

Creating ensembles of semantic spaces has been shown to be profitable, at least on the task of extracting synonyms and abbreviation-expansion pairs. In this study, the focus has been on combining the *output* of the semantic spaces. This is probably the most straightforward approach and it has several advantages. For one, the manner in which the semantic representations are created can largely be ignored, which would potentially allow one to combine models that are very different in nature, as long as one can retrieve a ranked list of semantically related terms with a measure of the strength of the relation. It also means that one can readily combine semantic spaces that have been induced with different parameter settings, for instance with different context definitions and of different dimensionality. An alternative approach would perhaps be to combine semantic spaces on a *vector level*. Such an approach would be interesting to explore; however, it would pose numerous challenges, not least in combining context vectors that have been constructed differently and potentially represent meaning in disparate ways.

Several combination strategies were designed and evaluated. In both the single-corpus and multiple-corpora ensembles, the most simple strategy performed best: the one whereby the cosine similarity scores are summed. There are potential problems with such a strategy, since the similarity scores are not absolute measures of semantic relatedness, but merely relative and only valid within a single semantic space. The cosine similarity scores will, for instance, differ depending on the distributional model used and the size of the context window. An attempt was made to deal with this by replacing the cosine similarity scores with ranking information, as a means to *normalize* the output of each semantic space before combing them. This approach, however, yielded much poorer results. A possible explanation for this is that a measure of the semantic relatedness between terms is of much more importance than their ranking. After all, a list of the highest ranked terms does not necessarily imply that they are semantically similar to the query term; only that they are the most semantically similar in this space. For the multiple-corpora ensembles, the AVG strategy was applied with the aim of not penalizing candidate synonyms that only appear in one of the two corpora. It is not surprising that this strategy was not successful given the form of the evaluation, which consisted of suggesting candidate synonyms that were known to occur at least 50 times in both corpora. The two-step approaches for the multiple-corpora ensembles all included a normalizing and/or averaging component, resulting in a lower recall compared to the SUM strategy, probably for the same reasons as when these strategies were applied in the one-step approach.

To gain deeper insights into the process of combining the output of multiple semantic spaces, an error analysis was conducted on the synonym extraction task. This was achieved by comparing the outputs of the most profitable combination of semantic spaces from each corpus, as well as with the combination of semantic spaces from the two corpora. The error analysis was conducted on the development sets. Of the 68 synonyms that were correctly identified as such by the corpora combination, five were not extracted by either of the single-corpus combinations; nine were extracted by the medical ensemble but not by the clinical ensemble; as many as 51 were extracted by the clinical ensemble but not by its medical counterpart; in the end, this means that only three terms were extracted by both the clinical and medical ensembles. These results augment the case for multiple-corpora ensembles. There appears to be little overlap in the top-10 outputs of the corpora-specific ensembles; by combining them, 17 additional true synonyms are extracted compared to the clinical ensemble alone. Moreover, the fact that so many synonyms are extracted by the clinical ensemble demonstrates the importance of exploiting clinical corpora and the applicability of distributional semantics to this genre of text. In Table 12, the first two examples, *sjukhem (nursing-home)* and *depression* show cases for which the multiple-corpora ensemble was successful but the single-corpus ensembles were not. In the third example, both the multiple-corpora ensemble and the clinical ensemble extract the expected synonym candidate.

There was one query term – the drug name *omeprazol* – for which both single-corpus ensembles were able to identify the synonym, but where the multiple-corpora ensemble failed. There were also three query terms for which synonyms were identified by the clinical ensemble, but not by the multiple-corpora ensemble; there were five query terms that were identified by the medical ensemble, but not by the multiple-corpora ensemble. This shows that combining semantic spaces can also, in some cases, introduce noise.

Since synonym pairs were queried both ways, i.e. each term in the pair would be queried to see if the other could be identified, we wanted to see if there were cases where the choice of query term would be important. Indeed, among the sixty query terms for which the expected synonym was not extracted, this was the case in fourteen instances. For example, given the query term *blindtarmsinflammation ("appendix-inflammation")*, the expected synonym *appendicit (appendicitis)* was given as a candidate, whereas with the query term *appendicit*, the expected synonym was not successfully identified.

Models of distributional semantics face the problem of modeling terms with several ambiguous meanings. This is, for instance, the case with the polysemous term *arv* (referring to *inheritance* as well as to *heredity*). Distant synonyms also seem to be problematic, e.g. the pair *rehabilitation/habilitation*. For approximately a third of the synonym pairs that are not correctly identified, however, it is not evident that they belong to either of these two categories.

Post-processing

In an attempt to improve results further, an additional step in the proposed method was introduced: filtering of the candidate terms, with the possibility of extracting new, potentially better ones. For the extraction of abbreviation-expansion pairs, this was fairly straightforward, as there are certain patterns that generally apply to this phenomenon, such as the fact that the letters in an abbreviation are contained – in the same order – in its expansion. Moreover, expansions are longer than abbreviations. This allowed us to construct simple yet effective rules for filtering out unlikely candidate terms for these two tasks. As a result, both precision and recall increased; with a dynamic cut-off, precision improved significantly. Although our focus in this study was primarily on maximizing recall, there is a clear incentive to improve precision as well. If this method were to be used for terminological development support, with humans inspecting the candidate terms, minimizing the number of poor candidate terms has a clear value. However, given the seemingly easy task of filter out unlikely candidates, it is perhaps more surprising that the results were not even better. A part of the reason for this may stem from the problem of semantically overloaded word types, which affects abbreviations to a large degree, particularly in the clinical domain with its telegraphic style and where ad-hoc abbreviations abound. This was also reflected in the reference standard, as in some cases the most common expansion of an abbreviation was not included.

The post-processing filtering of synonyms clearly failed. Although ranking information and, especially, cosine similarity provide some indication of the quality of synonym candidates, employing cut-off values with these features can impossibly improve recall: new candidates will always have a lower ranking and a lower cosine similarity score

Table 12 Examples of extracted candidate synonyms

Query term: sjukhem *(nursing-home)*		
Clinical	**Medical**	**Clinical + Medical**
Heartcenter (*heart-center*)	Vårdcentral (*health-center*)	Vårdcentral (*health-center*)
Bröstklinik (*breast-clinic*)	Akutmottagning (*emergency room*)	Mottagning (*reception*)
Hälsomottagningen (*health-clinic*)	Akuten (*ER*)	**Vårdhem** (*nursing-home*)
Hjärtcenter (*heart-center*)	Mottagning (*reception*)	Gotland (*a Swedish county*)
Län (*county*)	Intensivvårdsavdelning (*ICU*)	Sjukhus (*hospital*)
Eyecenter (*eye-center*)	Arbetsplats (*work-place*)	Gård (*yard*)
Bröstklin (*breast-clin.*)	Vårdavdelning (*ward*)	Vårdavdelning (*ward*)
Sjukhems (*nursing-home's*)	Gotland (*a Swedish county*)	Arbetsplats (*work-place*)
Hartcenter (*"hart-center"*)	Kväll (*evening*)	Akutmottagning (*emergency room*)
Biobankscentrum (*biobank-center*)	Ks (*Karolinska hospital*)	Akuten (*ER*)
Query term: depression *(depression)*		
Clinical	**Medical**	**Clinical + Medical**
Sömnstörning (*insomnia*)	Depressioner (*depressions*)	Sömnstörning (*insomnia*)
Sömnsvårigheter (*insomnia*)	Osteoporos (*osteoporosis*)	Osteoporos (*osteoporosis*)
Panikångest (*panic disorder*)	Astma (*asthma*)	Tvångssyndrom (*OCD*)
Tvångssyndrom (*OCD*)	Fetma (*obesity*)	Epilepsi (*epilepsy*)
Fibromyalgi (*fibromyalgia*)	Smärta (*pain*)	Hjärtsvikt (*heart failure*)
Ryggvärk (*back-pain*)	Depressionssjukdom (*depressive-illness*)	**Nedstämdhet** (*sadness*)
Självskadebeteende (*self-harm*)	Bensodiazepiner (*benzodiazepines*)	Fibromyalgi (*fibromyalgia*)
Osteoporos (*osteoporosis*)	Hjärtsvikt (*heart-failure*)	Astma (*asthma*)
Depressivitet (*"depressitivity"*)	Hypertoni (*hypertension*)	Alkoholberoende (*alcoholism*)
Pneumoni (*pneumonia*)	Utbrändhet (*burnout*)	Migrän (*migraine*)
Query term: allergi *(allergy)*		
Clinical	**Medical**	**Clinical + Medical**
Pollenallergi (*pollen-allergy*)	Allergier (*allergies*)	Allergier (*allergies*)
Födoämnesallergi (*food-allergy*)	Sensibilisering (*sensitization*)	Hösnuva (*hay-fever*)
Hösnuva (*hay-fever*)	Hösnuva (*hay-fever*)	Födoämnesallergi (*food-allergy*)
Överkänslighet (*hypersensitivity*)	Rehabilitering (*rehabilitation*)	Pollenallergi (*pollen-allergy*)
Kattallergi (*cat-allergy*)	Fetma (*obesity*)	**Överkänslighet** (*hypersensitivity*)
Jordnötsallergi (*peanut-allergy*)	Kol (*COPD*)	Astma (*asthma*)
Pälsdjursallergi (*animal-allergy*)	Osteoporos (*osteoporosis*)	Kol (*COPD*)
Negeras (*negated*)	Födoämnesallergi (*food-allergy*)	Osteoporos (*osteoporosis*)
Pollen (*pollen*)	Astma (*asthma*)	Jordnötsallergi (*peanut-allergy*)
Pollenallergiker (*"pollen-allergic"*)	Utbrändhet (*burnout*)	Pälsdjursallergi (*animal-allergy*)

The top ten candidate synonyms for three different query terms with the clinical ensemble, the medical ensemble and the disjoint corpus ensemble. The synonym in the reference standard is in boldface.

than discarded candidate terms. It can, however – at least in theory – potentially improve precision when using these rules in conjunction with a dynamic cut-off, i.e. allowing less than ten candidates terms to be suggested. In this case, however, the rules did not have this effect.

Thresholds

Increasing the frequency threshold further did not improve results. In fact, a threshold of 30 occurrences in both corpora seems to be sufficient. A high frequency threshold is a limitation of distributional methods; thus, the ability to use a lower threshold is important, especially in the clinical domain where access to data is difficult to obtain.

The choice of evaluating recall among ten candidates was based on an estimation of the number of candidate terms that would be reasonable to present to a lexicographer for manual inspection. Recall might improve if more candidates were presented, but it would likely come at the expense of decreased usability. It might instead be more relevant to limit further the number of candidates to present. As is shown in Figure 4, there are only a few correct synonyms among the candidates ranked 6–10. By using more advanced post-processing techniques and/or being prepared to sacrifice recall slightly, it is possible to present fewer candidates for manual inspection, thereby potentially increasing usability. On the other hand, a higher cut-off value could be used for evaluating a system aimed at a user who is willing to review a longer list of suggestions. An option for incorporating this difference in user behavior would be to use an evaluation metrics, such as rank-biased precision [58], that models the persistence of the user in examining additional lower-ranked candidates.

Reflections on evaluation

To make it feasible to compare a large number of semantic spaces and their various combinations, fixed reference standards derived from terminological resources were used for evaluation, instead of manual classification of candidate terms. One of the motivations for the current study, however, is that terminological resources are seldom complete; they may also reflect a *desired* use of language rather than *actual* use. A manual classification on a sample of one of the reference standards, *Medical + Clinical*, was carried out on the synonym task in order to verify this claim. The results in this study thus mainly reflect to what extent different semantic spaces – and their combinations – are able to extract synonymous relations that have been considered relevant according to specific terminologies, rather than to what extent the semantic spaces – and their combinations – capture the phenomenon of synonymy. This is, for instance, illustrated by the query term *depression* in Table 12, in which one potential synonym is extracted by the clinical ensemble – *depressivitet ("depressitivity")* – and another potential synonym by the medical ensemble: *depressionsjukdom (depressive illness).* Although these terms might not be formal or frequent enough to include in all types of terminologies, they are highly relevant candidates for inclusion in terminologies intended for text mining. Neither of these two terms are, however, counted as correct synonyms, and only the multiple-corpora ensemble is able to find the synonym included in the terminology.

Furthermore, a random sample of 30 words (out of 135) was manually classified for the semantic relation between each of the candidate terms in the sample, as suggested by the best combination of semantic spaces (the Disjoint Corpus Ensemble, see Table 10), and the target term. In the reference standard for this sample, 33 synonyms are to be found (only three target words have two synonyms; none have three or more). The best combination finds only 10 of these reference synonyms (exact match), which accounts for the low recall figures in Table 10. However, a manual classification shows that the same combination finds another 29 synonyms that do not occur in the reference standard. Furthermore, the Disjoint Corpus Ensemble also suggests a total of 15 hyponyms, 14 hypernyms and 3 spelling variants as candidate terms, which, depending on the context, can be viewed as synonyms. Among the candidate terms, we also find 3 antonyms, which shows the inability of the models readily to distinguish between different types of semantic relations.

In one instance, we also capture a non-medical sense of a term while completely missing the medical sense. For the target term *sänka* (erythrocyte sedimentation rate), 9 out of 10 candidate terms relate to the more general sense of lowering something (also *sänka* in Swedish), with candidate terms such as *rising, reducing, increasing, halving* and *decreasing*. None of these are included in the reference standard, which for this word only contains the abbreviation *SR* (ESR) as a synonym.

In the case of the target term *variecella*, the reference standard contains only the synonym *vattkoppor* (chickenpox), while the Disjoint Corpus Ensemble correctly suggests the abbreviation *VZV*, as well as *herpes* and the plural form *varicellae* (which is apparently missed by the lemmatizer).

It is important to recognize that this type of manual post-evaluation always bears the risk that you are too generous, believing in your method, and thus (manually) assign too many correct classifications – or, alternatively that you are too strict in your classification in fear of being too generous. Future studies would thus benefit from an extensive manual classification of candidates derived from data generated in clinical practice, beforehand, with the aim of also finding synonyms that are not already included in current terminologies but are in frequent use. These could then be used as reference standards in future evaluations.

The choice of terminological resources to use as reference standards was originally based on their appropriateness for evaluating semantic spaces induced from the clinical corpus. However, for evaluating the extraction of abbreviation-expansion pairs with semantic spaces induced from the medical corpus, the chosen resources – in conjunction with the requirement that terms should occur at least fifty times in the corpus – were less appropriate, as it resulted in a very small reference standard.

This, in turn, resulted in no significant differences for either of the two the abbreviation tasks between the best single space and the combination of medical spaces, or between the conjoint corpus ensemble and the disjoint corpus ensemble. When assessing the potential of using semantic spaces for abbreviation-expansion tasks, more focus should therefore be put on the results from the evaluation on the spaces created from the clinical corpus, as the improvement in recall gained by post-processing was statistically significant for both the abbr→exp task and the exp→abbr task, as was also the improvement gained from using an ensemble of spaces compared to a single corpus space for the abbr→exp task.

For synonyms, the number of instances in the reference standard is, of course, smaller for the experiments with multiple-corpora ensembles than for the single-corpus experiments. However, the differences between the single space and the ensemble of spaces are statistically significant. Moreover, when evaluating the final results with different frequency thresholds, similar results are obtained when lowering the threshold and, as a result, including more evaluation instances. With a threshold of twenty occurrences, 306 input terms are evaluated, which results in a recall of 0.42; with a threshold of thirty occurrences and 222 query terms, a recall of 0.46 is obtained.

Future work

Now that this first step has been taken towards creating ensembles of semantic spaces, this notion should be explored in greater depth and taken further. It would, for instance, be interesting to combine a larger number of semantic spaces, possibly including those that have been more explicitly modeled with syntactic information. To verify the superiority of this approach, it should be compared to the performance of a single semantic space that has been induced from multiple corpora.

Further experiments should likewise be conducted with combinations involving a larger number of corpora (types). One could, for instance, combine a professional corpus with a layman corpus – e.g. a corpus of extracts from health-related fora – in order to identify layman expressions for medical terms. This could provide a useful resource for automatic text simplification.

Another technique that could potentially be used to identify term pairs with a higher degree of semantic similarity is to ensure that both terms have each other as their closest neighbors in the semantic subspace. This is not always the case, as we pointed out in our error analysis. This could perhaps improve performance on the task of extracting synonyms and abbreviation-expansion pairs.

A limitation of the current study – in the endeavor to create a method that accounts for the problem of language use variability – is that the semantic spaces were constructed to model only unigrams. Textual instantiations of the same concept can, however, vary in term length. This needs to be accounted for in a distributional framework and concerns paraphrasing more generally than synonymy in particular. Combining unigram spaces with multiword spaces is a possibility that could be explored. This would also make the method applicable for acronym expansion.

Conclusions

This study demonstrates that combinations of semantic spaces can yield improved performance on the task of automatically extracting synonyms. First, combining two distributional models – random indexing and random permutation – on a single corpus enables the capturing of different aspects of lexical semantics and effectively increases the quality of the extracted candidate terms, outperforming the use of one model in isolation. Second, combining distributional models and types of corpora – a clinical corpus, comprising health record narratives, and a medical corpus, comprising medical journal articles – improves results further, outperforming ensembles of semantic spaces induced from a single source, as well as single semantic space induced from the conjoint corpus. We hope that this study opens up avenues of exploration for applying the ensemble methodology to distributional semantics.

Semantic spaces can be combined in numerous ways. In this study, the approach was to combine the outputs, i.e. ranked lists of semantically related terms to a given query term, of the semantic spaces. How this should be done is not wholly intuitive. By exploring a variety of combination strategies, we found that the best results were achieved by simply summing the cosine similarity scores provided by the distributional models.

On the task of extracting abbreviation-expansion pairs, substantial performance gains were obtained by applying a number of simple post-processing rules to the list of candidate terms. By filtering out unlikely candidates based on simple patterns and retrieving new ones, both recall and precision were improved by a large margin.

Lastly, analysis of a manually classified sample from the synonym task shows that the semantic spaces not only extract synonyms that are present in the reference standard. Equally valid synonyms not present in the reference standard are also found. This serves to show that the reference standards, as most often is the case, lack in coverage, as well as supports the fact that the semantic spaces can be used to enrich and expand such resources.

Endnotes

[a] Signifiers are here simply different linguistic items referring to the same concept.

[b] Ontologies are formal descriptions of concepts and their relationships.

[c] The words *big* and *large* are, for instance, synonymous when describing a house, but certainly not when describing a sibling.

[d] Unified Medical Language System: http://www.nlm.nih.gov/research/umls/

[e] Hyponyms are words that are subordinate to another word, its hypernym. For instance, *dog* is a hyponym of *mammal*, which in turn is a hyponym of *animal*.

[f] There are also probabilistic models, which view documents as a mixture of topics and represent terms according to the probability of their occurrence during the discussion of each topic: two terms that share similar topic distributions are assumed to be semantically related.

[g] Explicit dimensionality reduction is avoided in the sense that an initial term-context matrix is not constructed, the dimensionality of which is then reduced. The high-dimensional data is *prereduced*, if you will, by selecting a much lower dimensionality from the outset (effectively making this a parameter of the model).

[h] Ternary vectors allow three possible values: +1's, 0's and −1's. Allowing negative vector elements ensures that the entire vector space is utilized.

[i] Orthogonal index vectors would yield completely uncorrelated context representations; in the RI approximation, *near*-orthogonal index vectors result in almost uncorrelated context representations.

[j] The bag-of-words model is a simplified representation of a text as an unordered collection of words, where grammar and word order are ignored.

[k] An alternative is to shift the index vectors according to direction only, effectively producing *direction vectors* [46].

[l] This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/834-31/5.

[m] The used stop word lists are available at http://people.dsv.su.se/~mariask/resources/stoppord.txt (clinical corpus) and http://people.dsv.su.se/~mariask/resources/lt_stoppord.txt. (medical corpus)

[n] Antonyms are words that differ in one dimension of meaning, and thus are mutually exclusive in this sense. For instance, something cannot be both *large* and *small* in size at the same time.

[o] Hypernyms are words that are superordinate to another word, its hyponym. For instance, *animal* is a hypernym of *mammal*, which in turn is a hypernym of *dog*.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AH was responsible for coordinating the study and was thus involved in all parts of it. AH was responsible for the overall design of the study and for carrying out the experiments. AH initiated the idea of combining semantic spaces induced from different corpora and implemented the evaluation and post-processing modules. AH also had the main responsibility for the manuscript and drafted parts of the background and results description. HM and MS contributed equally to the study. HM initiated the idea of combining semantic models trained differently (Random Indexing and Random Permutation) and was responsible for designing and implementing strategies for combining the output of multiple semantic models. HM also drafted parts of the method description in the manuscript and surveyed relevant literature. MS initiated the idea of applying the method to abbreviation-expansion extraction and to different types of corpora. MS was responsible for designing the evaluation part of the study, as well as for preparing the reference standards. MS also drafted parts of the background and method description in the manuscript. VD, together with MS, was responsible for designing the post-processing filtering of candidate terms. MD provided feedback on the design of the study and drafted parts of the background and method description in the manuscript. MD also carried out the manual evaluation, and the analysis thereof. AH, HM, MS and MD analyzed the results and drafted the discussion and conclusions in the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was partly (AH) supported by the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection (ref. no. IIS11-0053) at Stockholm University, Sweden. It was also partly (HM) supported by the Research Council of Norway through the project EviCare - Evidence-based care processes: Integrating knowledge in clinical information systems (NFR project no. 193022). We would like to thank the members of our former research network HEXAnord, within which this study was initiated. We would especially like to thank Ann-Marie Eklund for her contributions to the initial stages of this work. We are also grateful to Staffan Cederblom and Studentlitteratur for giving us access to their database of medical abbreviations. Finally, we would like to thank the three reviewers for their insightful comments.

Author details

[1] Department of Computer and Systems Sciences (DSV), Stockholm University, Forum 100, SE-164 40 Kista, Sweden. [2] Department of Computer and Information Science, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway. [3] Faculty of Informatics, Vytautas Magnus University, Vileikos g. 8 - 409, Kaunas, LT-44404, Lithuania.

Received: 4 June 2013 Accepted: 17 January 2014
Published: 5 February 2014

References

1. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF: **Extracting information from textual documents in the electronic health record: a review of recent research.** *Yearb Med Inform* 2008, **47**(1):128–144.
2. Saeed JI: *Semantics*. Oxford: Blackwell Publishers; 1997.
3. Leroy G, Chen H: **Meeting medical terminology needs-the ontology-enhanced Medical Concept Mapper.** *IEEE Trans Inf Technol Biomed* 2001, **5**(4):261–270.
4. Leroy G, Endicott JE, Mouradi O, Kauchak D, Just ML: **Improving perceived and actual text difficulty for health information consumers using semi-automated methods.** In *Proceedings of AMIA Annual Symposium*. Maryland, USA: American Medical Informatics Association; 2012:522–31.
5. Eriksson R, Jensen PB, Frankild S, Jensen LJ, Brunak S: **Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text.** *J Am Med Inform Assoc* 2013, **20**(5):947–953.
6. Liu H, Aronson AR, Friedman C: **A study of abbreviations in MEDLINE abstracts.** In *Proceedings of AMIA Annual Symposium*. Maryland: American Medical Informatics Association; 2002:464–468.
7. Keselman A, Slaughter L, Arnott-Smith C, Kim H, Divita G, Browne A, Tsai C, Zeng-Treitler Q: **Towards consumer-friendly PHRs: Patients' experience with reviewing their health records.** In *Proceedings of AMIA Annual Symposium*. Maryland: American Medical Informatics Association; 2007:399–403.
8. Uzuner Ö, South B, Shen S, DuVall S: **2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text.** *J Am Med Inform Assoc* 2011, **18**(5):552–556.

9. Cohen AM, Hersh WR: **A survey of current work in biomedical text mining.** *Brief Bioinformatics* 2005, **6:**57–71.
10. Dumais S, Landauer T: **A solution to Platos problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge.** *Psychol Rev* 1997, **104**(2):211–240.
11. Hearst M: **Automatic acquisition of hyponyms from large text corpora.** In *Proceedings of COLING 1992*. Stroudsburg: Association for Computational Linguistics; 1992:539–545.
12. Blondel VD, Gajardo A, Heymans M, Senellart P, Dooren PV: **A measure of similarity between graph vertices: applications to synonym extraction and web searching.** *SIAM Rev* 2004, **46**(4):647–666.
13. Nakayama K, Hara T, Nishio S: **Wikipedia mining for an association web thesaurus construction.** In *Web Information Systems Engineering–WISE 2007*. Berlin Heidelberg: Springer-Verlag; 2007:322–334.
14. Dietterich TG: **Ensemble methods in machine learning.** In *Proceedings of the First International Workshop on Multiple Classifier Systems*. Berlin Heidelberg: Springer-Verlag; 2000:1–15.
15. Curran JR: **Ensemble methods for automatic thesaurus extraction.** In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg: Association for Computational Linguistics; 2002:222–229.
16. Wu H, Zhou M: **Optimizing synonym extraction using monolingual and bilingual resources.** In *Proceedings of the Second International Workshop on Paraphrasing - Volume 16, PARAPHRASE '03*. Stroudsburg: Association for Computational Linguistics; 2003:72–79.
17. van der Plas L, Tiedemann J: **Finding synonyms using automatic word alignment and measures of distributional similarity.** In *Proceedings of the COLING/ACL on Main Conference Poster Sessions, COLING-ACL '06*. Stroudsburg: Association for Computational Linguistics; 2006:866–873.
18. Peirsman Y, Geeraerts D: **Predicting strong associations on the basis of corpus data. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09.** *Stroudsburg: Association for Computational Linguistics* 2009:648–656.
19. Diaz F, Metzler D: **Improving the estimation of relevance models using large external corpora.** In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM; 2006:154–161.
20. Yu H, Agichtein E: **Extracting synonymous gene and protein terms from biological literature.** *Bioinformatics* 2003, **1**(19):340–349.
21. Cohen A, Hersh W, Dubay C, Spackman K: **Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts..** *BMC Bioinformatics* 2005, **6:**103.
22. McCrae J, Collier N: **Synonym set extraction from the biomedical literature by lexical pattern discovery.** *BMC Bioinformatics* 2008, **9:**159.
23. Conway M, Chapman W: **Discovering lexical instantiations of clinical concepts using web services, WordNet and corpus resources.** In *Proceedings of AMIA Annual Symposium*. Maryland: American Medical Informatics Association; 2012:1604.
24. Henriksson A, Conway M, Duneld M, Chapman WW: **Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records.** In *Proceedings of AMIA Annual Symposium*. Maryland: American Medical Informatics Association; 2013:600–609.
25. Henriksson A, Skeppstedt M, Kvist M, Conway M, Duneld M: **Corpus-driven terminology development: populating Swedish SNOMED CT with synonyms extracted from electronic health records.** In *Proceedings of BioNLP*. Stroudsburg: Association for Computational Linguistics; 2013.
26. Zeng QT, Redd D, Rindflesch T, Nebeker J: **Synonym, topic model and predicate-based query expansion for retrieving clinical documents.** In *Proceedings of AMIA Annual Symposium*. Maryland: American Medical Informatics Association; 2012:1050–1059.
27. Pedersen T, Pakhomov SV, Patwardhan S, Chute CG: **Measures of semantic similarity and relatedness in the biomedical domain.** *J Biomed Inf* 2007, **40**(3):288–299.
28. Koopman B, Zuccon G, Bruza P, Sitbon L, Lawley M: **An evaluation of corpus-driven measures of medical concept similarity for information retrieval.** In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. New York: ACM; 2012:2439–2442.
29. Schwartz AS, Hearst MA: **A simple algorithm for identifying abbreviation definitions in biomedical text.** In *Proceedings of the 8th Pacific Symposium on Biocomputing*. Singapore: World Scientific; 2003:451–462.
30. Ao H, Takagi T: **ALICE: an algorithm to extract abbreviations from MEDLINE.** *J Am Med Inf Assoc* 2005, **12**(5):576–586.
31. Chang JT, Schütze H, Altman RB: **Creating an online dictionary of abbreviations from MEDLINE.** *J Am Med Inf Assoc* 2002, **9:**612–620.
32. Movshovitz-Attias D, Cohen WW: **Alignment-HMM-based extraction of abbreviations from biomedical text.** In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012)*. Stroudsburg: Association for Computational Linguistics; 2012:47–55.
33. Dannélls D: **Automatic acronym recognition.** In *Proceedings of the 11th conference on European Chapter of the Association for Computational Linguistics (EACL)*. Stroudsburg: Association for Computational Linguistics; 2006:167–170.
34. Yu H, Hripcsak G, Friedman C: **Mapping abbreviations to full forms in biomedical articles.** *J Am Med Inf Assoc:JAMIA* 2002, **9**(3):262–272.
35. Isenius N, Velupillai S, Kvist M: **Initial results in the development of SCAN: a Swedish Clinical Abbreviation Normalizer.** In *Proceedings of the CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis - CLEFeHealth2012*. Sydney: NICTA; 2012.
36. Gaudan S, Kirsch H, Rebholz-Schuhmann D: **Resolving abbreviations to their senses in Medline.** *Bioinformatics* 2005, **21**(18):3658–3664.
37. Cohen T, Widdows D: **Empirical distributional semantics: methods and biomedical applications.** *J Biomed Inform* 2009, **42**(2):390–405.
38. Salton G, Wong A, Yang CS: **A vector space model for automatic indexing.** *Commun ACM* 1975, **11**(18):613–620.
39. Deerwester SC, Dumais ST, Landauer TK, Furnas GW, Harshman RA: **Indexing by latent semantic analysis.** *J Am Soc Inf Sci* 1990, **41**(6):391–407.
40. Schütze H: **Word space.** In *Advances in Neural Information Processing Systems 5*. Burlington, Massachusetts: Morgan Kaufmann; 1993: 895–902.
41. Lund K, Burgess C: **Producing high-dimensional semantic spaces from lexical co-occurrence.** *Behav Res Methods* 1996, **28**(2):203–208.
42. Harris ZS: **Distributional structure.** *Word* 1954, **10:**146–162.
43. Sahlgren M: **The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.** *PhD thesis*, PhD thesis, Stockholm University, 2006.
44. Kanerva P, Kristofersson J, Holst A: **Random indexing of text samples for latent semantic analysis.** In *Proceedings of 22nd Annual Conference of the Cognitive Science Society*. Austin: Cognitive Science Society; 2000:1036.
45. Jones MN, Mewhort DJK: **Representing word meaning and order information in a composite holographic lexicon.** *Psychol Rev* 2007, **1**(114):1–37.
46. Sahlgren M, Holst A, Kanerva P: **Permutations as a means to encode order in word space.** In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*. Austin: Cognitive Science Society; 2008:1300–1305.
47. Symonds M, Bruza PD, Sitbon L, Turner I: **Modelling word meaning using efficient tensor representations.** In *Proceedings of 25th Pacific Asia Conference on Language, Information and Computation*. Tokyo: Digital Enhancement of Cognitive Development; 2011.
48. Symonds M, Zuccon G, Koopman B, Bruza P, Nguyen A: **Semantic judgement of medical concepts: Combining syntagmatic and paradigmatic information with the tensor encoding model.** In *Australasian Language Technology Association Workshop 2012*. Stroudsburg: Association for Computational Linguistics; 2012:15.
49. Hassel M: **JavaSDM package.** 2004. [http://www.nada.kth.se/~xmartin/java/]. [KTH School of Computer Science and Communication; Stockholm, Sweden].
50. Dalianis H, Hassel M, Velupillai S: **The Stockholm EPR corpus - characteristics and some initial findings.** In *Proceedings of ISHIMR 2009, Evaluation and implementation of e-health and health information initiatives: international perspectives. 14th International Symposium for Health Information Management Research, Kalmar, Sweden*. Kalmar: eHealth Institute; 2009:243–249.

51. Kokkinakis D: **The journal of the Swedish Medical Association - a corpus resource for biomedical text mining in Swedish.** In *The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), an LREC Workshop. Turkey*. Paris: European Language Resources Association; 2012.
52. Knutsson O, Bigert J, Kann V: **A robust shallow parser for Swedish.** In *Proceedings of Nodalida*; 2003:2003.
53. Cederblom S: *Medicinska förkortningar och akronymer (In Swedish)*. Lund: Studentlitteratur; 2005.
54. US National Library of Medicine: **MeSH (Medical Subject Headings).** [http://www.ncbi.nlm.nih.gov/mesh]
55. Karolinska Institutet: **Hur man använder den svenska MeSHen (In Swedish, translated as: how to use the Swedish MeSH).** [http://mesh.kib.ki.se/swemesh/manual_se.html] 2012. [Accessed 2012-03-10].
56. Newbold P, Carlson WL, Thorne B: *Statistics for Business and Economics*. 5. ed. Prentice-Hall: Upper Saddle River; 2003.
57. Henriksson A, Hassel M: **Optimizing the dimensionality of clinical term spaces for improved diagnosis coding support.** In *Proceedings of Louhi Workshop on Health Document Text Mining and Information Analysis*. Sydney: NICTA; 2013.
58. Moffat A, Zobel J: **Rank-biased precision for measurement of retrieval effectiveness.** *ACM Trans Inf Syst* 2008, **27:**2:1–2:27.

doi:10.1186/2041-1480-5-6
Cite this article as: Henriksson *et al.*: Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics* 2014 **5**:6.

Paper B

Towards dynamic word sense discrimination with Random Indexing

Moen, Hans; Marsi, Erwin, and Gambäck, Björn

In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 83–90, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

Towards Dynamic Word Sense Discrimination with Random Indexing

Hans Moen, Erwin Marsi, Björn Gambäck
Norwegian University of Science and Technology
Department of Computer and Information and Science
Sem Sælands vei 7-9
NO-7491 Trondheim, Norway
{hansmoe,emarsi,gamback}@idi.ntnu.no

Abstract

Most distributional models of word similarity represent a word type by a single vector of contextual features, even though, words commonly have more than one sense. The multiple senses can be captured by employing several vectors per word in a multi-prototype distributional model, prototypes that can be obtained by first constructing all the context vectors for the word and then clustering similar vectors to create sense vectors. Storing and clustering context vectors can be expensive though. As an alternative, we introduce Multi-Sense Random Indexing, which performs on-the-fly (incremental) clustering. To evaluate the method, a number of measures for word similarity are proposed, both contextual and non-contextual, including new measures based on optimal alignment of word senses. Experimental results on the task of predicting semantic textual similarity do, however, not show a systematic difference between single-prototype and multi-prototype models.

1 Introduction

Many terms have more than one meaning, or sense. Some of these senses are static and can be listed in dictionaries and thesauri, while other senses are dynamic and determined by the contexts the terms occur in. Work in *Word Sense Disambiguation* often concentrate on the static word senses, making the task of distinguishing between them one of classification into a predefined set of classes (i.e., the given word senses); see, e.g., Erk et al. (2013; Navigli (2009) for overviews of current work in the area. The idea of fixed generic word senses has received a fair amount of criticism in the literature (Kilgarriff, 2000).

This paper instead primarily investigates dynamically appearing word senses, word senses that depend on the actual usage of a term in a corpus or a domain. This task is often referred to as Word Sense Induction or *Word Sense Discrimination* (Schütze, 1998). This is, in contrast, essentially a categorisation problem, distinguished by different senses being more or less similar to each other at a given time, given some input data. The dividing line between Word Sense Disambiguation and Discrimination is not necessarily razor sharp though: also different senses of a term listed in a dictionary tend to have some level of overlap.

In recent years, distributional models have been widely used to infer word similarity. Most such models represent a word type by a single vector of contextual features obtained from co-occurrence counts in large textual corpora. By assigning a single vector to each term in the corpus, the resulting model assumes that each term has a fixed semantic meaning (relative to all the other terms). However, due to homonomy and polysemy, word semantics cannot be adequately represented by a single-prototype vector.

Multi-prototype distributional models in contrast employ different vectors to represent different senses of a word (Reisinger and Mooney, 2010). Multiple prototypes can be obtained by first constructing context vectors for all words and then clustering similar context vectors to create a sense vector. This may be expensive, as vectors need to stored and clustered. As an alternative, we propose a new method called Multi-Sense Random Indexing (MSRI), which is based on Random Indexing (Kanerva et al., 2000) and performs an on-the-fly (incremental) clustering.

MSRI is a method for building a multi-prototype / multi-sense vector space model, which attempts to capture one or more senses per unique term in an unsupervised manner, where each sense is represented as a separate vector in the model.

Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality, pages 83–90,
Sofia, Bulgaria, August 9 2013.

This differs from the classical Random Indexing (RI) method which assumes a static sense inventory by restricting each term to have only one vector (sense) per term, as described in Section 2. The MSRI method is introduced in Section 3.

Since the induced dynamic senses do not necessarily correspond to the traditional senses distinguished by humans, we perform an extrinsic evaluation by applying the resulting models to data from the Semantic Textual Similarity shared task (Agirre et al., 2013), in order to compare MSRI to the classical RI method. The experimental set-up is the topic of Section 4, while the results of the experiments are given in Section 5. Section 6 then sums up the discussion and points to ways in which the present work could be continued.

2 Vector Space Models

With the introduction of LSA, Latent Semantic Analysis (Deerwester et al., 1990), distributed models of lexical semantics, built from unlabelled free text data, became a popular sub-field within the language processing research community. Methods for building such semantic models rely primarily on term co-occurrence information, and attempt to capture latent relations from analysing large amounts of text. Most of these methods represent semantic models as multi-dimensional vectors in a vector space model.

After LSA, other methods for building semantic models have been proposed, one of them being Random Indexing (Kanerva et al., 2000). Common to these methods is that they generate a *context vector* for each unique term in the training data which represents the term's "contextual" meaning in the vector space. By assigning a single context vector to each term in the corpus, the resulting model assumes that each term has a fixed semantic meaning (relative to all other terms).

Random Indexing incrementally builds a co-occurrence matrix of reduced dimensionality, by first assigning *index vectors* to each unique term. The vectors are of a predefined size (typically around 1000), and consist of a few randomly placed 1s and -1s. Context vectors of the same size are also assigned to each term, initially consisting of only zeros. When traversing a document corpus using a sliding window of a fixed size, the context vectors are continuously updated: the term in the centre of the window (the target term), has the index vectors of its neighbouring terms (the ones in the window) added to its context vector using vector summation. Then the *cosine similarity measure* can be used on term pairs to calculate their similarity (or "contextual similarity").

Random Indexing has achieved promising results in various experiments, for example, on the TOEFL test ("Test of English as a Foreign Language") (Kanerva et al., 2000). However, it is evident that many terms have more than one meaning or sense, some being static and some dynamic, that is, determined by the contexts the terms occur in. Schütze (1998) proposed a method for clustering the contextual occurrences of terms into individual "prototype" vectors, where one term can have multiple prototype vectors representing separate senses of the term. Others have adopted the same underlying idea, using alternative methods and techniques (Reisinger and Mooney, 2010; Huang et al., 2012; Van de Cruys et al., 2011; Dinu and Lapata, 2010).

3 Multi-Sense Random Indexing, MSRI

Inspired by the work of Schütze (1998) and Reisinger and Mooney (2010), this paper introduces a novel variant of Random Indexing, which we have called "Multi-Sense Random Indexing". MSRI attempts to capture one or more senses per unique term in an unsupervised and incremental manner, each sense represented as an separate vector in the model. The method is similar to classical sliding window RI, but each term can have multiple context vectors (referred to as *sense vectors* here) which are updated separately.

When updating a term vector, instead of directly adding the index vectors of the neighbouring terms in the window to its context vector, the system first computes a separate *window vector* consisting of the sum of the index vectors. The similarity between the window vector and each of the term's sense vectors is calculated. Each similarity score is then compared to a pre-set *similarity threshold*:

- if no score exceeds the threshold, the window vector becomes a new separate sense vector for the term,
- if exactly one score is above the threshold, the window vector is added to that sense vector, and
- if multiple scores are above the threshold, all the involved senses are merged into one sense vector, together with the window vector.

Algorithm 1 MSRI training

```
for all terms $t$ in a document $D$ do
    generate window vector $\vec{win}$ from the neighbouring words' index vectors
    for all sense vectors $\vec{s}_i$ of $t$ do
        $sim(s_i) = CosSim(\vec{win}, \vec{s}_i)$
    end for
    if $sim(s_{i..k}) \geq \tau$ then
        Merge $\vec{s}_{i..k}$ and $\vec{win}$ through summing
    else
        if $sim(s_i) \geq \tau$ then
            $\vec{s}_i + = \vec{win}$
        end if
    else
        if $sim(s_{i..n}) < \tau$ then
            Assign $\vec{win}$ as new sense vector of $t$
        end if
    end if
end for
```

See Algorithm 1 for a pseudo code version. Here τ represents the similarity threshold.

This accomplishes an incremental (on-line) clustering of senses in an unsupervised manner, while retaining the other properties of classical RI. Even though the algorithm has a slightly higher complexity than classical RI, this is mainly a matter of optimisation, which is not the focus of this paper. The incremental clustering that we apply is somewhat similar to what is used by Lughofer (2008), although we are storing in memory only one element (i.e., vector) for each "cluster" (i.e., sense) at any given time.

When looking up a term in the vector space, a pre-set *sense-frequency threshold* is applied to filter out "noisy" senses. Hence, senses that have occurred less than the threshold are not included when looking up a term and its senses for, for example, similarity calculations.

As an example of what the resulting models contain in terms of senses, Table 1 shows four different senses of the term 'round' produced by the MSRI model. Note that these senses do not necessarily correspond to human-determined senses. The idea is only that using multiple prototype vectors facilitates better modelling of a term's meaning than a single prototype (Reisinger and Mooney, 2010).

round$_1$	**round**$_2$	**round**$_3$	**round**$_4$
finish	camping	inch	launcher
final	restricted	bundt	grenade
match	budget	dough	propel
half	fare	thick	antitank
third	adventure	cake	antiaircraft

Table 1: Top-5 most similar terms for four different senses of 'round' using the *Max* similarity measure to the other terms in the model.

3.1 Term Similarity Measures

Unlike classical RI, which only has a single context vector per term and thus calculates similarity between two terms directly using cosine similarity, there are multiple ways of calculating the similarity between two terms in MSRI. Some alternatives are described in Reisinger and Mooney (2010). In the experiment in this paper, we test four ways of calculating similarity between two terms t and t' in isolation, with the Average and Max methods stemming from Reisinger and Mooney (2010).

Let $\vec{s}_{i..n}$ and $\vec{s'}_{j..m}$ be the sets of sense vectors corresponding to the terms t and t' respectively. Term similarity measures are then defined as:

Centroid

For term t, compute its centroid vector by summing its sense vectors $\vec{s}_{i..n}$. The same is done for t' with its sense vectors $\vec{s'}_{j..m}$. These centroids are in turn used to calculate the cosine similarity between t and t'.

Average

For all $\vec{s_{i..n}}$ in t, find the pair $\vec{s}_i, \vec{s'}_j$ with highest cosine similarity:

$$\frac{1}{n}\sum_{i=1}^{n} CosSim_{max}(\vec{s}_i, \vec{s'}_j)$$

Then do the same for all $\vec{s'}_{j..m}$ in t':

$$\frac{1}{m}\sum_{j=1}^{m} CosSim_{max}(\vec{s'}_j, \vec{s}_i)$$

The similarity between t and t' is computed as the average of these two similarity scores.

Max

The similarity between t_i and t'_i equals the similarity of their most similar sense:

$$Sim(t, t') = CosSim_{max_{ij}}(\vec{s}_i, \vec{s'}_i)$$

Hungarian Algorithm

First cosine similarity is computed for each possible pair of sense vectors $\vec{s}_{i..n}$ and $\vec{s'}_{j..m}$, resulting in a matrix of similarity scores. Finding the optimal matching from senses $\vec{s}_i$ to $\vec{s'}_j$ that maximises the sum of similarities is known as the *assignment problem*. This combinatorial optimisation problem can be solved in polynomial time through the Hungarian Algorithm (Kuhn, 1955). The overall similarity between terms t and t' is then defined as the average of the similarities between their aligned senses.

All measures defined so far calculate similarity between terms in isolation. In many applications, however, terms occur in a particular context that can be exploited to determine their most likely sense. Narrowing down their possible meaning to a subset of senses, or a single sense, can be expected to yield a more adequate estimation of their similarity. Hence a context-sensitive measure of term similarity is defined as:

Contextual similarity

Let $\vec{C}$ and $\vec{C'}$ be vectors representing the contexts of terms t and t' respectively. These context vectors are constructed by summing the index vectors of the neighbouring terms within a window, following the same procedure as used when training the MSRI model. We then find $\hat{s}$ and $\hat{s}\,'$ as the sense vectors best matching the context vectors:

$$\hat{s} = \arg\max_i \ CosSim(\vec{s}_i, \vec{C})$$

$$\hat{s}\,' = \arg\max_j \ CosSim(\vec{s}_j, \vec{C'})$$

Finally, contextual similarity is defined as the similarity between these sense vectors:

$$Sim_{context}(t, t') = CosSim(\hat{s}, \hat{s}\,')$$

3.2 Sentence Similarity Features

In the experiments reported on below, a range of different ways to represent sentences were tested. Sentence similarity was generally calculated by the average of the maximum similarity between pairs of terms from both sentences, respectively. The different ways of representing the data in combination with some sentence similarity measure will here be referred to as similarity *features*.

1. `MSRI-TermCentroid`:
   In each sentence, each term is represented as the sum of its sense vectors. This is similar to having one context vector, as in classical RI, but due to the sense-frequency filtering, potentially "noisy" senses are not included.

2. `MSRI-TermMaxSense`:
   For each bipartite term pair in the two sentences, their sense-pairs with maximum cosine similarity are used, one sense per term.

3. `MSRI-TermInContext`:
   A $5 + 5$ window around each (target) term is used as context for selecting one sense of the term. A window vector is calculated by summing the index vectors of the other terms in the window (i.e., except for the target term itself). The sense of the target term which is most similar to the window vector is used as the representation of the term.

4. `MSRI-TermHASenses`:
   Calculating similarity between two terms is done by applying the Hungarian Algorithm to all their bipartite sense pairs.

5. `RI-TermAvg`:
   Classical Random Indexing — each term is represented as a single context vector.

6. `RI-TermHA`:
   Similarity between two sentences is calculated by applying the Hungarian Algorithm to the context vectors of each constituent term.

The parameters were selected based on a combination of surveying previous work on RI (e.g., Sokolov (2012)), and by analysing how sense counts evolved during training. For MSRI, we used a similarity threshold of 0.2, a vector dimensionality of 800, a non-zero count of 6, and a window size of $5 + 5$. Sense vectors resulting from less than 50 observations were removed. For classical RI, we used the same parameters as for MSRI (except for a similarity threshold).

4 Experimental Setup

In order to explore the potential of the MSRI model and the textual similarity measures proposed here, experiments were carried out on data from the Semantic Textual Similarity (STS) shared task (Agirre et al., 2012; Agirre et al., 2013).

Given a pair of sentences, systems participating in this task shall compute how semantically similar the two sentences are, returning a similarity score between zero (completely unrelated) and five (completely semantically equivalent). Gold standard scores are obtained by averaging multiple scores obtained from human annotators. System performance is then evaluated using the Pearson product-moment correlation coefficient (ρ) between the system scores and the human scores.

The goal of the experiments reported here was not to build a competitive STS system, but rather to investigate whether MSRI can outperform classical Random Indexing on a concrete task such as computing textual similarity, as well as to identify which similarity measures and meaning representations appear to be most suitable for such a task. The system is therefore quite rudimentary: a simple linear regression model is fitted on the training data, using a single sentence similarity measure as input and the similarity score as the dependent variable. The implementations of RI and MSRI are based on JavaSDM (Hassel, 2004).

As data for training random indexing models, we used the CLEF 2004–2008 English corpus, consisting of approximately 130M words of newspaper articles (Peters et al., 2004). All text was tokenized and lemmatized using the TreeTagger for English (Schmid, 1994). Stopwords were removed using a customized version of the stoplist provided by the Lucene project (Apache, 2005).

Data for fitting and evaluating the linear regression models came from the STS development and test data, consisting of sentence pairs with a gold standard similarity score. The STS 2012 development data stems from the Microsoft Research Paraphrase corpus (`MSRpar`, 750 pairs), the Microsoft Research Video Description corpus (`MSvid`, 750 pairs), and statistical machine translation output based on the Europarl corpus (`SMTeuroparl`, 734 pairs). Test data for STS 2012 consists of more data from the same sources: `MSRpar` (750 pairs), `MSRvid` (750 pairs) and `SMTeuroparl` (459 pairs). In addition, different test data comes from translation data in the news domain (`SMTnews`, 399 pairs) and ontology mappings between OntoNotes and WordNet (`OnWN`, 750 pairs). When testing on the STS 2012 data, we used the corresponding development data from the same domain for training, except for `OnWN` where we used all development data combined.

The development data for STS 2013 consisted of all development and test data from STS 2012 combined, whereas test data comprised machine translation output (`SMT`, 750 pairs), ontology mappings both between WordNet and OntoNotes (`OnWN`, 561 pairs) and between WordNet and FrameNet (`FNWN`, 189 pairs), as well as news article headlines (`HeadLine`, 750 pairs). For simplicity, all development data combined were used for fitting the linear regression model, even though careful matching of development and test data sets may improve performance.

5 Results and Discussion

Table 2 shows Pearson correlation scores per feature on the STS 2012 test data using simple linear regression. The most useful features for each data set are marked in bold. For reference, the scores of the best performing STS systems for each data set are also shown, as well as baseline scores obtained with a simple normalized token overlap measure.

There is large variation in correlation scores, ranging from 0.77 down to 0.27. Part of this variation is due to the different nature of the data sets. For example, sentence similarity in the SMT domain seems harder to predict than in the video domain. Yet there is no single measure that obtains the highest score on all data sets. There is also no consistent difference in performance between the RI and MSRI measures, which seem to yield about equal scores on average. The `MSRI-TermInContext` measure has the lowest score on average, suggesting that word sense disambiguation in context is not beneficial in its current implementation.

The corresponding results on the STS 2013 test data are shown in Table 3. The same observations as for the STS 2012 data set can be made: again there was no consistent difference between the RI and MSRI features, and no single best measure.

All in all, these results do not provide any evidence that MSRI improves on standard RI for this particular task (sentence semantic similarity). Multi-sense distributional models have, however, been found to outperform single-sense models on other tasks. For example, Reisinger and Mooney (2010) report that multi-sense models significantly increase the correlation with human similarity judgements. Other multi-prototype distributional models may yield better results than their single-prototype counterparts on the STS task.

Features:	MSRpar	MSRvid	SMTeuroparl	SMTnews	OnWN	Mean
Best systems	0.73	0.88	0.57	0.61	0.71	0.70
Baseline	0.43	0.30	0.45	0.39	0.59	0.43
RI-TermAvg	0.44	0.71	**0.50**	**0.42**	0.65	**0.54**
RI-TermHA	0.41	0.72	0.44	0.35	0.56	0.49
MSRI-TermCentroid	**0.45**	0.73	**0.50**	0.33	0.64	0.53
MSRI-TermHASenses	0.40	**0.77**	0.47	0.39	**0.68**	**0.54**
MSRI-TermInContext	0.33	0.55	0.36	0.27	0.42	0.38
MSRI-TermMaxSense	0.44	0.71	**0.50**	0.32	0.64	0.52

Table 2: Pearson correlation scores per feature on STS 2012 test data using simple linear regression

Feature	Headlines	SMT	FNWN	OnWN	Mean
Best systems	0.78	0.40	0.58	0.84	0.65
Baseline	0.54	0.29	0.21	0.28	0.33
RI-TermAvg	0.60	**0.37**	0.21	0.52	0.42
RI-TermHA	**0.65**	0.36	0.27	0.52	0.45
MSRI-TermCentroid	0.60	0.35	**0.37**	0.45	0.44
MSRI-TermHASenses	0.63	0.35	0.33	**0.54**	**0.46**
MSRI-TermInContext	0.20	0.29	0.19	0.36	0.26
MSRI-TermMaxSense	0.58	0.35	0.31	0.45	0.42

Table 3: Pearson correlation scores per feature on STS 2013 test data using simple linear regression

Notably, the more advanced features used in our experiment, such as `MSRI-TermInContext`, gave very clearly inferior results when compared to `MSRI-TermHASenses`. This suggests that more research on MSRI is needed to understand how both training and retrieval can be fully utilized and optimized.

6 Conclusion and Future Work

The paper introduced a new method called Multi-Sense Random Indexing (MSRI), which is based on Random Indexing and performs on-the-fly clustering, as an efficient way to construct multi-prototype distributional models for word similarity. A number of alternative measures for word similarity were proposed, both context-dependent and context-independent, including new measures based on optimal alignment of word senses using the Hungarian algorithm. An extrinsic evaluation was carried out by applying the resulting models to the Semantic Textual Similarity task. Initial experimental results did not show a systematic difference between single-prototype and multi-prototype models in this task.

There are many questions left for future work. One of them is how the number of senses per word evolves during training and how the distribution of senses in the final model looks like. So far we only know that on average the number of senses keeps growing with more training material, currently resulting in about 5 senses per word at the end of training (after removing senses with frequency below the sense-frequency threshold). It is worth noting that this depends heavily on the similarity threshold for merging senses, as well as on the weighting schema used.

In addition there are a number of model parameters that have so far only been manually tuned on the development data, such as window size, number of non-zeros, vector dimensionality, and the sense frequency filtering threshold. A systematic exploration of the parameter space is clearly desirable. Another thing that would be worth looking into, is how to compose sentence vectors and document vectors from the multi-sense vector space in a proper way, focusing on how to pick the right senses and how to weight these. It would also be interesting to explore the possibilities for combining the MSRI method with the Reflective Random Indexing method by Cohen et al. (2010) in an attempt to model higher order co-occurrence relations on sense level.

The fact that the induced dynamic word senses do not necessarily correspond to human-created senses makes evaluation in traditional word sense disambiguation tasks difficult. However, correlation to human word similarity judgement may provide a way of intrinsic evaluation of the models (Reisinger and Mooney, 2010). The *Usim* bench mark data look promising for evaluation of word similarity in context (Erk et al., 2013).

It is also worth exploring ways to optimise the algorithm, as this has not been the focus of our work so far. This would also allow faster training and experimentation on larger text corpora, such as Wikipedia. In addition to the JavaSDM package (Hassel, 2004), Lucene (Apache, 2005) with the Semantic Vectors package (Widdows and Ferraro, 2008) would be an alternative framework for implementing the proposed MSRI algorithm.

Acknowledgements

This work was partly supported by the Research Council of Norway through the EviCare project (NFR project no. 193022) and by the European Community's Seventh Framework Programme (FP7/20072013) under grant agreement nr. 248307 (PRESEMT). Part of this work has been briefly described in our contribution to the STS shared task (Marsi et al., 2013).

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pages 385–393, Montreal, Canada, June. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pages 32–43, Atlanta, Georgia, June. Association for Computational Linguistics.

Apache. 2005. Apache Lucene open source package. `http://lucene.apache.org/`.

Trevor Cohen, Roger Schvaneveldt, and Dominic Widdows. 2010. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, 43(2):240–256, April.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172, Cambridge, Massachusetts, October. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):501–544.

Martin Hassel. 2004. JavaSDM package. `http://www.nada.kth.se/~xmartin/java/`. School of Computer Science and Communication; Royal Institute of Technology (KTH); Stockholm, Sweden.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Jeju Island, Korea. Association for Computational Linguistics.

Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, page 1036, Philadelphia, Pennsylvania. Erlbaum.

Adam Kilgarriff. 2000. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Harold W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Edwin Lughofer. 2008. Extensions of vector quantization for incremental clustering. *Pattern Recognition*, 41(3):995–1011, March.

Erwin Marsi, Hans Moen, Lars Bungum, Gleb Sizov, Björn Gambäck, and André Lynum. 2013. NTNU-CORE: Combining strong features for semantic similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pages 66–73, Atlanta, Georgia, June. Association for Computational Linguistics.

Roberto Navigli. 2009. Word Sense Disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69.

Carol Peters, Paul Clough, Julio Gonzalo, Gareth J.F. Jones, Michael Kluck, and Bernardo Magnini, editors. 2004. *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, volume 3491 of *Lecture Notes in Computer Science*. Springer-Verlag, Bath, England.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117, Los Angeles, California, June.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the 1st International Conference on New Methods in Natural Language Processing*, pages 44–49, University of Manchester Institute of Science and Technology, Manchester, England, September.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, March.

Artem Sokolov. 2012. LIMSI: learning semantic similarity by selecting random word subsets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pages 543–546, Montreal, Canada, June. Association for Computational Linguistics.

Tim Van de Cruys, Thierry Poibeau, and Anna Korhonen. 2011. Latent vector weighting for word meaning in context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1012–1022, Edinburgh, Scotland, July. Association for Computational Linguistics.

Dominic Widdows and Kathleen Ferraro. 2008. Semantic vectors: a scalable open source package and online technology management application. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 1183–1190, Marrakech, Morocco.

Paper C

Care episode retrieval: distributional semantic models for information retrieval in the clinical domain

Moen, Hans; Ginter, Filip; Marsi, Erwin; Peltonen, Laura-Maria;
Salakoski, Tapio, and Salanterä, Sanna

BMC Medical Informatics and Decision Making, 15(Suppl 2):S2, 2015.

Moen *et al. BMC Medical Informatics and Decision Making* 2015, **15**(Suppl 2):S2
http://www.biomedcentral.com/1472-6947/15/S2/S2

BMC
Medical Informatics & Decision Making

PROCEEDINGS Open Access

Care episode retrieval: distributional semantic models for information retrieval in the clinical domain

Hans Moen[1,2,3*], Filip Ginter[2†], Erwin Marsi[1†], Laura-Maria Peltonen[3,4], Tapio Salakoski[2,5], Sanna Salanterä[3,4]

From Louhi 2014: The Fifth International Workshop on Health Text Mining and Information Analysis
Gothenburg, Sweden. 27 April 2014

Abstract

Patients' health related information is stored in *electronic health records* (EHRs) by health service providers. These records include sequential documentation of care episodes in the form of clinical notes. EHRs are used throughout the health care sector by professionals, administrators and patients, primarily for clinical purposes, but also for secondary purposes such as decision support and research. The vast amounts of information in EHR systems complicate information management and increase the risk of information overload. Therefore, clinicians and researchers need new tools to manage the information stored in the EHRs. A common use case is, given a - possibly unfinished - care episode, to retrieve the most similar care episodes among the records. This paper presents several methods for information retrieval, focusing on care episode retrieval, based on textual similarity, where similarity is measured through domain-specific modelling of the distributional semantics of words. Models include variants of *random indexing* and the semantic neural network model *word2vec*. Two novel methods are introduced that utilize the ICD-10 codes attached to care episodes to better induce domain-specificity in the semantic model. We report on experimental evaluation of care episode retrieval that circumvents the lack of human judgements regarding episode relevance. Results suggest that several of the methods proposed outperform a state-of-the art search engine (Lucene) on the retrieval task.

Introduction

The development, adoption and implementation of health information technology, e.g. *electronic health record* (EHR) systems, is a strategic focus of health policies globally [1-4] and the amount of electronically documented health information is increasing exponentially as health records are becoming more and more computerised. The vast amounts of computerised health information complicate information management and increase the risk of information overload. At the same time, it creates opportunities for technological solutions to support health related and clinical decision making. For instance, the use of *natural language processing* (NLP) methods to facilitate researchers in discovering new knowledge to improve health and care.

EHRs are used throughout the health care sector by professionals, administrators and patients, primarily for clinical purposes, but also for secondary purposes such as decision support and research [5]. EHRs include structured and unstructured data, and they consist of a sequential collection of a patients health related information e.g. health history, allergies, medications, laboratory results and radiology images. Also, the different stages of a patient's clinical care are documented in the EHR as *clinical care notes*, which mainly consist of free text. A sequence of individual clinical care notes form a *care episode*, which is concluded by a discharge summary, as illustrated in Figure 1.

Information retrieval (IR) aims at retrieving and ranking documents from a large collection based on the information related needs of a user expressed in a

* Correspondence: hans.moen@idi.ntnu.no
† Contributed equally
[1]Department of Computer and Information Science, Norwegian University of Science and Technology, Sem Saelands vei 9, 7491 Trondheim, Norway
Full list of author information is available at the end of the article

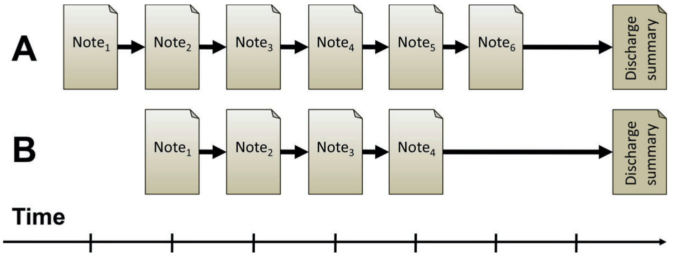


Figure 1 Illustration of care episode retrieval. The two care episodes (A and B) are composed of a number of individual clinical notes and a single discharge summary. Given an ongoing care episode (minus the discharge summary), the task is to retrieve other, similar care episodes.

search query [6]. IR has become a crucial technology for many organisations that deal with vast amounts of partly structured and unstructured (free text) data stored in electronic format, including hospitals and other health care providers. IR is an essential part of the clinical practice and clinicians, i.e. nurses and physicians search on the Internet for information, typically health literature, to solve clinical problems and for professional development [7]. Such online IR systems are associated with substantial improvements in clinicians decision making concerning clinical and health related problems [8,9]. To date, as the information in the EHRs is increasing, clinicians need new tools to manage the information. Therefore, IR from EHRs in general is a common and important task that, among other things, can support *Evidence-Based Practice* (EBP) through finding relevant care episodes and gathering sufficient evidence.

This paper focuses on the particular task of retrieving care episodes that are most similar to the sequence of clinical notes for a given patient, which we will call *care episode retrieval*. In conventional IR, the query typically consists of several keywords or a short phrase, while the retrievable units are typically documents. In contrast, in this work on care episode retrieval, the queries consist of the clinical notes contained in a care episode. The final discharge summaries for each care episode are assumed to be unavailable for constructing a query at retrieval time.

We envision a number of different use cases for a care episode retrieval system. Firstly, it could facilitate clinicians in care related decision making. For example, given a patient that is being treated in a hospital, an involved clinician may want to find previous patients that are similar in terms of their health history, symptoms or received treatments. Additional inputs from the clinician would enable the system to give more weight to keywords of particular interest within the care episodes, which would further be emphasized in the semantic similarity calculation during IR. This may support the clinician's care-related decision making when seeing what similar patients have received in terms of medication and treatment, what related issues such as bi-conditions or risks occurred, how other clinicians have described certain aspects, what clinical practice guidelines have been utilized, and so on. This relates to the principle of reasoning by analogy as used in textual case-based reasoning [10]. Secondly, when combined with systems for automatic summarization and trend detection, it could help health care managers to optimally allocate human resources with almost real time information concerning the overall situation on the unit for a specific follow-up period. Such a system could for example support managerial decision making with statistical information concerning care trends on the unit, adverse events and infections. Thirdly, it could facilitate knowledge discovery and research to improve care (cf. EBP). For instance, it could enable researchers to map or cluster similar care episodes to find common symptoms or conditions. In sum, care episode retrieval methods/systems hold large potential to improve documentation and care quality.

IR in the sense of searching text documents is closely related to NLP and is often considered a subfield of NLP. For example, stemming or lemmatization, in order to increase the likelihood of matches between terms in the query and a document, is a typical NLP task. From the perspective of NLP, care episode retrieval - and IR from EHRs in general - is a challenging task. It differs from general-purpose web search in that the vocabulary, the information needs and the queries of clinicians are highly specialised [11]. Clinical notes contain highly domain-specific terminology [12-14] and generic text processing resources are therefore often suboptimal or inadequate [15]. At the same time, development of dedicated clinical NLP tools and resources is often difficult and costly. For example, popular data-driven approaches to NLP are based on supervised learning, which requires substantial amounts of tailored training data, typically built through manual annotation by annotators who need both linguistic and clinical knowledge. Additionally, variations in the language and terminology used in sub-domains within and across health care organisations greatly limit the scope of applicability of such training data [12]. Moreover, resources are typically language-specific: most tools for processing English clinical text are of no use for our work on Finnish clinical text.

Recent work has shown that *distributional models of semantics*, induced in an unsupervised manner from large corpora of clinical and/or medical text, are well suited as a resource-light approach to capturing and representing domain-specific terminology [16-19]. The theoretical foundation for these models is the *distributional hypothesis* [20], stating that words with similar distributions in language - in the sense that they co-occur with overlapping sets of words - tend to have similar meanings. These models avoid most of the aforementioned problems associated with NLP resources. They do not involve the costly manual encoding of linguistic or clinical/medical knowledge by experts as required in knowledge-based approaches, nor do they involve equally costly investments in large-scale manual annotation and corpus construction as required for supervised learning. Instead, they can be constructed for any language or domain, as long as a reasonable amount of raw text in electronic format is available.

The work reported here investigates to what extent distributional models of semantics, built from a corpus of clinical text in a fully unsupervised manner, can be used to conduct care episode retrieval. The purpose of this study is to explore how a set of different distributional models perform in care episode retrieval, and also to determine how care episode similarity is best calculated. The models explored include several variants of *random indexing* and *word2vec*, methods which will be described in more detail in the 'Methods' section. These models allow us to compute the similarity between words, which in turn forms the basis for measuring similarity between texts such as individual clinical notes or larger care episodes. Several methods for computing textual similarity are proposed and experimentally tested in the task of care episode retrieval - being the main contribution of this paper.

It has been argued that clinical NLP should leverage existing knowledge resources such as knowledge bases about medications, treatments, diseases, symptoms and care plans, despite these not having been explicitly built for the purpose of clinical NLP [21]. Along these lines, a novel approach is presented here that utilizes the 10th revision of the International Classification of Diseases (ICD-10) [22] - a standardised tool of diagnostic codes for classifying diseases, labelled as meta-information to care episodes by clinicians - to better induce domain-specificity in the semantic model. Experimental results suggest that such models outperform a state-of-the art search engine (Lucene) on the task of care episode retrieval. Results also indicate that performance gain is achieved by most models when we utilize a list of health terms (cf. a health metathesaurus) for boosting term weights.

Apart from issues related to clinical terminology, another problem in care episode retrieval is the lack of benchmark data, such as the relevance scores produced by human judges commonly used for evaluation of IR systems. Although collections of care episodes may be available, producing gold standard similarity scores required for evaluation is costly. Another contribution of this paper is the proposal of evaluation procedures that circumvent the lack of human judgements regarding episode similarity. Two evaluation setups are used, one relying on ICD-10 codes attached to care episodes, and the other relying on textual semantic similarity between discharge summaries belonging to care episodes. Neither discharge summaries nor ICD-10 codes are used for constructing a query at retrieval time. This includes that sentences mentioning ICD-10 codes in free text are excluded from the query episodes. Despite our focus on the specific task of care episode retrieval, we hypothesize that the methods and models proposed here have the potential to increase performance of IR on clinical text in general.

This article extends earlier work published in Moen et al. [23]. New content includes the evaluation of various methods and setups on 40 instead of 20 query episodes, the introduction and evaluation of a new semantic model (W2V-ICD), and alternative ways of calculating care episode similarities.

The structure of this article is as follows. In the next section, 'Related work', we describe some related work. In the 'Task' section we describe in more detail the task of care episode retrieval, followed by a description of the data set of care episodes in the 'Data' section. The 'Methods' section presents six different distributional semantic models as well as two baselines. The 'Results' section reports the results of two experimental evaluations. The final two sections, 'Discussion' and 'Conclusion', are dedicated to discussion and conclusions respectively.

Related work

With the issues of information overload in hospitals and the general need for research and improvements in clinical care, several IR systems have been developed specifically for health records and clinical text. Examples are the *Electronic Medical Record Search Engine* (EMERSE) [24], the *StarTracker* [25], the *Queriable Patient Inference Dossier* (QPID) [26], the *MorphoSaurus* [27], and the *CISearch* [28]. These software are used by clinicians and researchers to seek information from EHRs. Other IR systems used in multiple domains, including the health domain, is the open source search engine, or framework, *Apache Lucene* (Lucene) [29] and the Terrier search engine [30]. Some research has been published in relation to the use of these systems in the clinical domain [11,26,28,31-34]. However, research evaluating the effect of these tools and their impact on care and patient outcomes seems to be scarce. In this work Lucene is used as a baseline.

One challenge related to clinical NLP is the limited availability of such data, mainly due to its sensitivity. Thus, many IR/search solutions that are in use in various EHR systems today are often off-the-shelf generic IR tools, or unique to the corresponding EHR systems. In other words, the underlying IR methods are seldom subject to research on clinical IR. However, in recent years through shared tasks such as the TREC Medical Records track [35,36] and the ShARe/CLEF eHealth Evaluation Lab [37], clinical data is becoming increasingly accessible to a broader audience of researchers, thus research on clinical NLP and IR has gained some momentum. Existing work on IR for health records relies to a large extent on performing some type of query expansion, and possibly some bootstrapping, through the use of tailored information sources, or corpus-driven statistical methods. Limsopatham et al. [38] attempts to improve IR on health records by inferring implicit domain knowledge, mainly done through query expansion that relies on tailored domain-specific resources and information from other high-ranked documents. Zhu and Carterette [39,40] performs query expansion mainly through the use of more generic resources, including ICD-9, Medical Subject Headings (MeSH) and Wikipedia. They also explore the use of a negation detection tool for information exclusion (ConText [41]).

Distributional semantic models have enjoyed a steady popularity for quite some time, and have for instance recently gained a lot of interest with the introduction of the word2vec method by Mikolov et al. [42]. Such methods for inducing models of distributional semantics, in an unsupervised and language independent fashion, have shown to perform well at a range of NLP tasks, including more generic IR [43,6,44-47] and clinical IR for health records, see participants of the TREC Medical Records track [35,36]. Noteworthy, Koopman et al. [17] performs a comparison of several distributional models at clinical IR, including models built using the methods random indexing (RI) [48] and latent semantic analysis (LSA) [49]. There is no doubt that further research and evaluation of such methods would contribute to potential improvements in NLP, IR and information management in the clinical domain. One method that lacks proper evaluation in this domain is word2vec.

A promising direction in clinical NLP have been demonstrated through methods/systems that utilize various external knowledge resources, other than just the actual words in the query and target, either for performing query expansion [40], or in the textual similarity metric [50]. This is some of the underlying inspiration for a novel method contribution in this paper, one that relies on exploiting ICD-10 codes that has been labelled the care episodes. However, instead of using these for direct query expansion, they are utilized in the actual training phase of the semantic models.

Existing work on clinical IR that we are aware of relies on evaluation data where the queries are short search phrases. This differs from the task presented here, where the query is an entire care episode.

Diagnosis and treatment codes, such as ICD codes, are often applied at the end of the patient stay, or even after discharged from the hospital. Automatic labeling of care episodes with ICD codes has been the subject of a number of studies, e.g. [51,52]. Arguably this task is somewhat related to our task as far as the use of ICD codes for evaluation is concerned.

Task

The task addressed in this study is retrieval of care episodes that are similar to each other in terms of their clinical free text. The purpose is to explore how a set of different distributional models perform in care episode retrieval, and also to determine how care episode similarity is best calculated. In contrast to the normal IR setting, where the search query is derived from a text stating the user's information need, here the query is based on another care episode, which we refer to as the *query episode*. As the query episode may document ongoing treatment, and thus lack a discharge summary and ICD-10 code, neither of these information sources can be relied upon for constructing the query. The task is therefore to retrieve the most similar care episodes using only the information contained in the free text of the clinical notes in the query episode. An overview showing the steps in our experimental setup is illustrated in Figure 2.

Evaluation of retrieval results generally requires an assessment of their relevancy to the query. To perform automatic evaluation, a *gold standard* is needed, which specifies the relevant documents from the collection for each query. It is common to produce such a gold standard through (semi-) manual work, relying on multiple human experts to select or rank documents according to their relevancy to a given query. Hence, obtaining such judgements is typically costly and time-consuming. Moreover, for the care episode retrieval task, the manual work would require experts in the clinical domain.

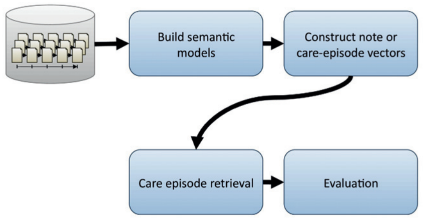


Figure 2 Experimental setup overview. Figure shows an overview of the various steps in our experimental setup.

In relation to this study, with the help of an expert in the clinical domain, we tried to assess the feasibility of creating such a gold standard for the care episode retrieval task. What we found was that assessing whether or not two care episodes are similar, strictly based on the information in both texts, was a difficult task with a lot of room for (individual) interpretation, especially for the top-ranked care episodes. By looking at the top-10 care episodes retrieved by the various semantic models and Lucene for a particular query episode, we found almost all of them had many overlapping clinical features with the query episode, even if they did not share the same primary ICD-10 code. In many cases they shared ICD-10 codes, but not necessarily the primary ones. Also, even though many patients were hospitalized due to similar reasons and/or backgrounds, this did not necessarily mean that they were treated in response to the exact same diagnosis, given the same treatments, or received those treatments in the same order. We estimate the two most time-consuming sub-tasks to be (1) creating explicit and unambiguous guidelines for the human evaluators, possibly unique ones for each query episode; (2) performing the evaluation for the required number of care episodes (average being 357 care episodes for each of the 40 queries when looking at the top 100 retrieved care episodes per query for each model/system). In addition, it is important to have enough human evaluators evaluating the same data to be able to verify that inter-annotator agreement is of a sufficiently high level. We concluded that the effort required for creating such a gold standard was simply too much for the resources we had access to.

As we did not have access to the required resources for creating the evaluation set manually, we opted for an alternative approach. Two different evaluation strategies were used in an attempt to approximate human relevance judgements. The first evaluation method is based on the assumption that a retrieved episode is relevant if its ICD-10 code is identical to that of the query episode. The second method assumes that a retrieved episode is relevant if its discharge summary is semantically similar to that of the query episode. In this setting, crucially, discharge summaries or ICD-10 codes are not used in either query construction or episode retrieval. Both evaluation methods are further detailed in the sections 'Experiment 1: ICD-10 code identity' and 'Experiment 2: Discharge summary overlap' respectively.

Data

The data set used in this study consists of the electronic health records from patients with any type of heart related problem that were admitted to one particular university hospital in Finland between the years 2005-2009. Of these, the clinical notes written by physicians are used (i.e. we did not use the corresponding nursing notes). An assent for the research was obtained from the Ethics Committee of the Hospital District (17.2.2009 §67) and permission to conduct the research was obtained from the Medical Director of the Hospital District (2/2009). The total data set consists of 66884 care episodes, which amounts to 398040 notes and 64 million words in total. Words here refer to terms identified through the lemmatization, except terms being pure numbers. This full set was used for training of the semantic models. To reduce the computational demands of experimentation, a subset was used for evaluation purposes, comprising 26530 care episodes, amounting to 155562 notes and 25.7 million words in total.

Notes are mostly unstructured, consisting of Finnish clinical free text.

The care episodes have been manually labeled according to ICD-10. Codes are normally applied at the end of the patients' hospital stay, or even after the patient has been discharged from the hospital. Care episodes have commonly one primary ICD-10 code attached and optionally a number of additionally secondary codes. As extraction of the potential secondary ICD-10 codes is non-trivial, we have in this study only used the primary ICD-10 codes - used for training two of the semantic models and for evaluation purposes in Experiment 1.

ICD-10 codes have an internal structure that reflects the classification system ranging from broad categories down to fine-grained subjects. For example, the first character (J) of the code J21.1 signals that it belongs to the broad category *Diseases of the respiratory system*. The next two digits (21) classify the subject as belonging to the subcategory *Acute bronchiolitis*. Finally, the last digit after the dot (1) means that it belongs to the sub-subclass *Acute bronchiolitis due to human metapneumovirus*. There are 356 unique "primary" ICD-10 codes in the evaluation data set.

Methods

Semantic models

A crucial part in retrieving similar care episodes is having a good similarity measure. Here similarity between care episodes is measured as the semantic similarity between the words they contain (see section 'Compute care episode similarity'). Semantic similarity between words is in turn measured through the use of distributional semantic models. In this way, no explicit query expansion step is performed, but potentially indirect word matches are found through the various models. Several model variants are tested, utilizing different techniques and parameters for building them. The models trained and tested in this paper are: (1) classic random indexing with a sliding window using term index vectors and term context vectors (RI-Word); (2) random indexing with index vectors for clinical notes (RI-Note); (3) random indexing with index vectors for ICD-10 codes (RI-ICD); (4) a version of random indexing where only the term index vectors are used (RI-Index); (5) a semantic neural network model, using *word2vec* (W2V) to build word context vectors (W2V); and (6) a W2V version of the RI-ICD method, using a modified version of W2V for training (W2V-ICD).

RI-Word

Random indexing (RI) [48] is a method for building a (pre) compressed vector space model with a fixed dimensionality, done in an incremental fashion. RI involves the following two steps: First, instead of allocating one dimension in the multidimensional vector space to a single word, each word is assigned an "index vector" as its unique signature in the vector space. Index vectors are generated vectors consisting of mostly zeros together with a randomly distributed set of several 1's and -1's, uniquely distributed for each unique word; the second step is to induce "context vectors" for each word. A context vector represents the *contextual meaning* of a word. This is done using a sliding window of a fixed size to traverse a training corpus, inducing context vectors for the center/target word of the sliding window by summing the index vectors of the neighbouring words in the window. An example illustrating how RI-Word works is shown in Figure 3.

As the dimensionality of the index vectors is fixed, the dimensionality of the vector space will not grow beyond the size $W \times Dim$, where W is the number of unique words in the vocabulary, and Dim being the pre-selected dimensionality to use for the index vectors. As a result, RI models are significantly smaller than plain vector space models, making them a lot less computationally expensive. Additionally, the method is fully incremental (additional training data can be added at any given time without having to retrain the existing model), easy to parallelize, and scalable, meaning that it is fast and can be trained on large amounts of text in an on-line fashion.

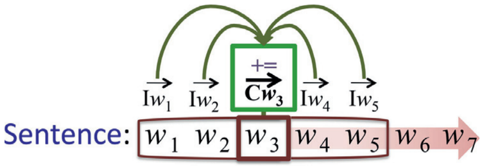


Figure 3 Training the RI-Word model. A sliding window with a size of five words is moved over the text, word by word. The context vector $\overrightarrow{Cw_3}$ for the word in the center of the window w_3 is updated by adding the index vectors of the other words within the window, i.e. $\overrightarrow{Iw_1}$, $\overrightarrow{Iw_2}$, $\overrightarrow{Iw_4}$ and $\overrightarrow{Iw_5}$. As a result, context vector $\overrightarrow{Cw_3}$ records the fact that w_3 co-occurs with word w_1, w_2, w_4 and w_5. The training process continues with moving the sliding window one position to the right and repeating the addition operation for context vector $\overrightarrow{Cw_4}$, and so on until the end of the training text is reached.

RI-Note

Contrary to sliding window approach used in RI-Word, a RI model built with *note index vectors* first assigns unique index vectors to every clinical note in the training corpus. In the training phase, each word in a note gets the corresponding note index vector added to its context vector. See Figure 4 for an illustration of how RI-Note works.

RI-ICD

Based on the principle of RI with note index vectors, we here explore a novel method for constructing a vector space model by exploiting the ICD-10 code classification done by clinicians. Instead of using note index vectors, we now use *ICD-code index vectors*. First, a unique index vector is assigned to each chapter and sub-chapter in the ICD-10 taxonomy. This means assigning a unique index vector to each "node" in the ICD-10 taxonomy, as illustrated in Figure 5. For each clinical note in the training corpus, the index vector of the their primary ICD-10 code is added to all words within it. In addition, all the index vectors for the ICD-codes higher in the taxonomy are added, each weighted according to their position in the hierarchy. A weight of 1 is given to the full code, while the weight is halved for each step upwards in the hierarchy. The motivation for the latter is to capture a certain degree of similarity between codes that share an initial path in the taxonomy. As a result, this similarity gets encoded in the resulting model. As an example, illustrated in Figure 5: for a clinical note labelled with the code J21.1, we add the following index vectors to the context vectors of all its constituting words: $\overrightarrow{I_J} \times 0.125$, $\overrightarrow{I_{J2}} \times 0.25$, $\overrightarrow{I_{J21}} \times 0.5$ and $\overrightarrow{I_{J21.1}} \times 1.0$.

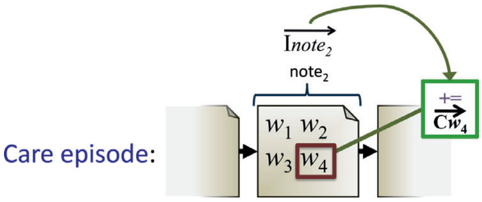


Figure 4 Training the RI-Note model. Word w_4's context vector, $\overrightarrow{Cw_{4'}}$ is updated by adding the index vector $\overrightarrow{Inote_2}$ of the note it is part of. The same update is applied to all other words in the note. As a result, context vectors for words co-occurring in the same note become more similar.

The underlying hypothesis for building a model in this way is that it may capture relations between words in a way that better reflects the clinical domain, compared with the other domain-independent methods for modelling.

RI-Index

As an alternative to using context vectors for words, we simply only use their index vectors in place of context vectors, therefore not modelling their "contextual meaning". When constructing note or care episode vectors directly from word index vectors (see the 'Compute care episode similarity' section), the resulting vectors represent a compressed version of a TF*IDF matrix, which again is similar to Lucene.

W2V

Recently, a novel method for inducing vector space models was introduced by Mikolov et al. [42], stemming from the research in deep learning and neural network language models. While the overall objective of learning a continuous vector space representation for each word based on its textual context remains, the underlying algorithms are substantially different from traditional methods such as LSA and RI. The model is based on a somewhat simplified neural network with as many input nodes as there are vocabulary items, a hidden linear projection layer with as many nodes as is the desired dimensionality of the vector space, and finally a hierarchical soft-max output layer. Every node in the hidden projection layer calculates a linear combination of the values of the input layer nodes (0 or 1), as its own value. The nodes of the output layer, in turn, calculate a linear combination of the hidden layer node outputs, which is subsequently passed through a specific non-linear function. The network is trained one input-output example pair at a time, and for each pair the difference between the expected and the actual output of the network is calculated. The linear combination weights in the network are subsequently adjusted to decrease the error using the *back-propagation* procedure. This procedure is repeated for all training data pairs, often in several passes over the entire training dataset, until the network converges and the error does not decrease any further. The application of neural networks specifically in word prediction tasks is presented, for example, by Bengio et al. [53].

The main distinguishing features specific to the W2V model are the linear (as opposed to the traditionally non-linear) hidden layer, and the usage of the efficient hierarchical soft-max output layer, which allows for a substantial decrease in the number of output nodes that need to be considered for the back-propagation. Combined, these two techniques allow the network to be efficiently trained on billions of tokens worth of input text. There are two distinct regimes in which the network is trained, or in other words, two ways to define the task on which the network is trained. In the *skip-gram* architecture, the network is trained given a focus word to predict a nearby word. I.e. a sliding window of typically ±10 tokens wide is slid across the text with the focus word at its center and each word within the window is in turn considered a prediction target. The focus word - context word pairs then constitute the word pairs used to train the network. The single input node corresponding to the focus word is activated while setting all other input layer nodes to zero (also referred to as *one hot* representation), and the error in the output layer prediction of the context word is back-propagated through the network. It is important to note that the output layer predictions are only necessary to train the network and we are not interested in them otherwise. To understand on intuitive level why the network learns efficient representations, consider the two-step process of the prediction: first, the input layer is used to activate the hidden, representation layer and second, the hidden layer is used to activate the output layer and predict the context word. To maximize the performance on this task, the network is thus forced to assign similar hidden layer representations to words which tend to have similar contexts. Since these representations form the W2V model, distributionally similar words are given similar vector representations. An alternative training regime is the *BoW* (bag of words) architecture. In this architecture, all words from the context are used at once to activate the respective nodes in the input layer, and the focus word is the prediction target. In a sense, it is the reverse of the skip-gram architecture. The main advantage of the BoW regime is its speed, because only a single update of the network is necessary per each context, unlike in the skip-gram architecture, where as many updates are performed as there are words in the context. Regardless of the training regime, the vector space representation of a word is the weight vector from its corresponding input node to the hidden layer. As

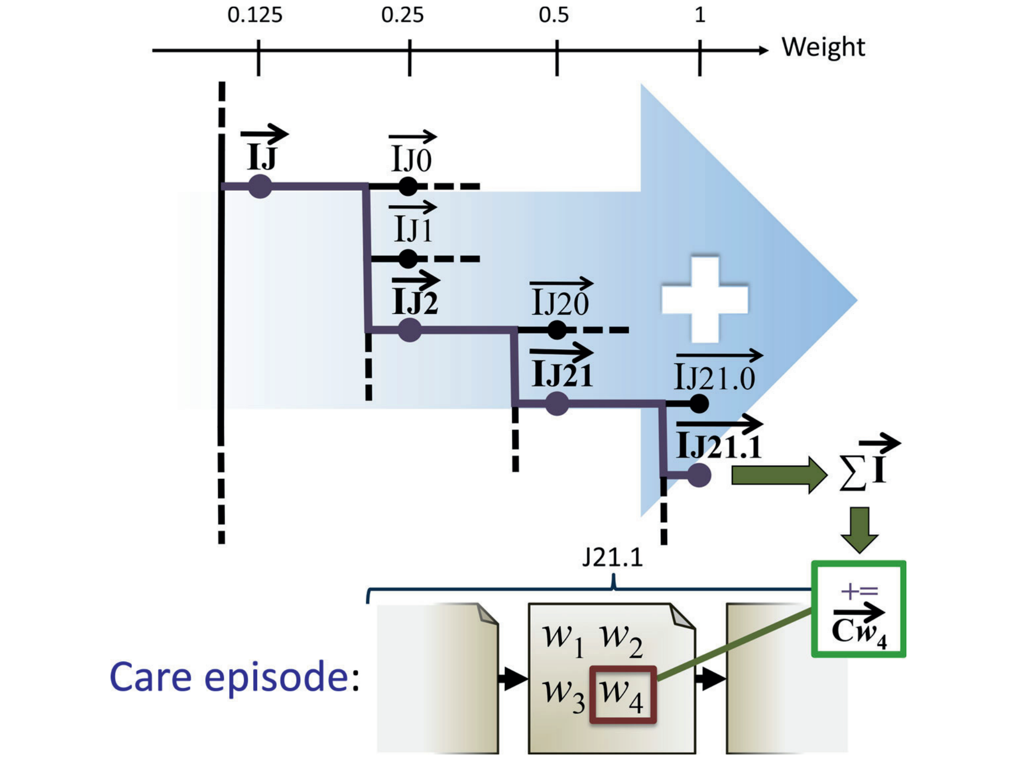


Figure 5 Training the RI-ICD model. Word w_4 occurs in a note that is part of a care episode labeled with the ICD-10 code J21.1. Its context vector $\overrightarrow{Cw_4}$ is therefore updated by adding the index vector for the code J21.1. This context vector is constructed from the weighted sum of index vectors of its parts: $\sum \overrightarrow{I} = (0.125 \times \overrightarrow{I_J}) + (0.25 \times \overrightarrow{I_{J2}}) + (0.5 \times \overrightarrow{I_{J21}}) + (1.0 \times \overrightarrow{I_{J21.1}})$. As a result, w_4's context vector becomes more strongly associated with the code J21.1 and - to a lesser degree - with all superclasses of J21.1 in the ICD-10 taxonomy. The same update is applied to the context vectors of all other words in care episodes labeled as J21.1.

mentioned previously, the hidden-to-output layer weights are discarded after training. See Figure 6 for an example illustrating how model training with W2V is carried out.

One of the main practical advantages of the W2V method lies in its scalability, allowing the training on billions of words of input text in the matter of several hours, setting it apart from the majority of other methods of distributional semantics. Additionally, the W2V method has been shown to produce representations that preserves important linguistic regularities [54]; as elaborated by Levy and Goldberg [55].

W2V-ICD

As will be shown, the RI-ICD method offers a notable advantage over the standard RI in the care episode retrieval task. We therefore introduce a novel variant of the W2V algorithm which implements the same insights as the RI-ICD method. As the starting point serves the fact that only the input-to-hidden layer weights define the final vector space representation. Therefore, as long as we preserve the input and hidden layers as in the original W2V architecture, i.e. a single input node for every word and a hidden layer with as many nodes as is the dimensionality of the representation, we are free to modify the prediction task of the network. In this case, we will use the ICD-10 codes for the corresponding clinical notes as the prediction target, training the network to predict the ICD-10 code of the note given a word from it. Following a similar intuition as for the skip-gram architecture, in order to maximize the performance on the task, the network will assign similar representation to words which occur in notes with the

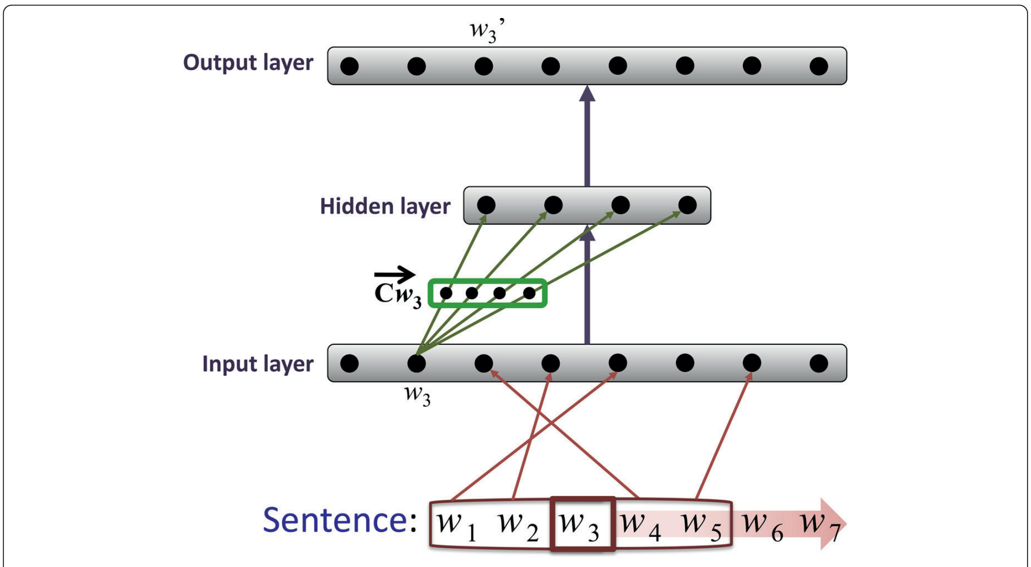


Figure 6 Training the W2V BoW model. A sliding window with the size of five words is moved over the text, word by word. The input layer nodes of the network corresponding to the words in the context window of the word w_3 are activated. The error in the output layer prediction and the expected prediction for the focus word w_3 is back-propagated through the network. When the training is completed, the context vector $\overrightarrow{Cw_3}$ constitutes the set of weights connecting the input layer node for w_3 and the hidden layer.

same ICD-10 codes. This objective clearly mirrors the motivation for the RI-ICD method. As in RI-ICD, we make use of the hierarchical structure of the ICD-10 codes, as illustrated in Figure 5, whereby not only the full ICD-10 code, but also its structural parts constitute training targets for the network. For each note, the network is thus trained on all pairs of a word from the note on the input layer, and a structural segment of the ICD-10 code as the prediction target. We use the ICD-10 code segments and their frequencies to define a vocabulary, whereupon we induce the hierarchical softmax layer in exactly the same manner as in the standard W2V algorithm. We implement the exact same weighting as in the RI-ICD method, giving ICD-10 code segments a weight which decreases as the generality of the segment increases. We then use these weights to scale the update gradient propagated through the network. See Figure 7 for an example how this training is done.

Compute care episode similarity

After having computed a semantic model, or six in this case, together with the baselines, the next step is to calculate care episode similarities for the retrieval task. Multiple ways of calculating care episode similarities exist.

We explore two overall approaches: One where each care episode is viewed as a single document, with all corresponding notes concatenated (SingleSim); Another where each care episode is viewed as a set of individual notes. For the latter, care episode similarity between two care episodes is calculated from pairwise note similarities and aggregating into a single similarity score. This again can be done in multiple ways. Three approaches are explored here (AvgSim, HASim and NWSim).

SingleSim: Single care episode vectors

Here we compute a separate vector for each care episode by summing the word vectors for all words in the care episode. The resulting vector is divided by the total number of words in the episode to normalize for differences in length between care episodes. Similarity between care episodes is then determined by computing the cosine similarity between their vectors.

AvgSim: Average note vector similarity

Each individual note within a care episode gets its own note (document) vector by summing the word vectors for all words in the note. In order to compute the similarity between two episodes, we take the average over the exhaustive pairwise similarities between their notes. That is, for every note in the first care episode, we compute its similarity to every note in the second care episode, and

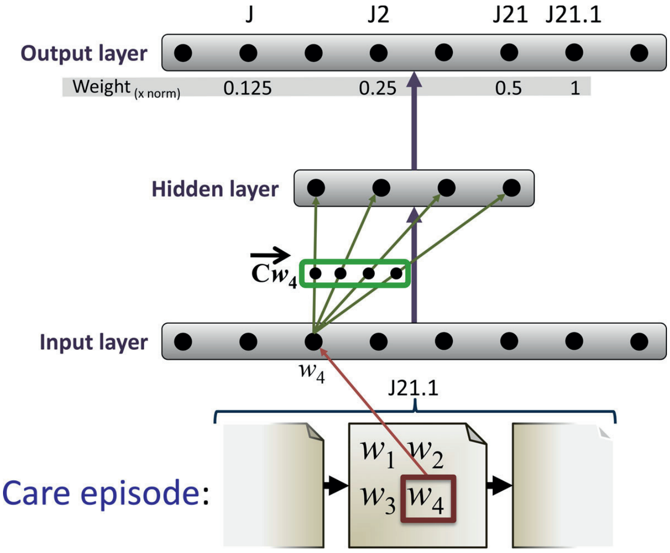


Figure 7 Training the W2V-ICD model. Word w_4 occurs in a note that is part of a care episode labeled with the ICD-10 code J21.1. The input node corresponding to w_4 is activated and the error between the output layer prediction and the expected prediction for J21.1 is back-propagated through the network. Same procedure is repeated for J21, with the update scaled by 0.5, and J2 scaled by 0.25, and finally J, scaled by 0.125. When the training is completed, the context vector $\overrightarrow{Cw_4}$ is formed by the weights connecting the input node corresponding to w_4 and the hidden layer of the network.

then take the average over all these pairwise similarities. Similarity between notes is again measured by the cosine similarity between their vectors.

HASim: Hungarian Algorithm for pairing of note vectors

For two care episodes, a note-to-note similarity matrix is calculated, populated with pairwise note vector similarities. By applying the *Hungarian Algorithm* [56], we compute the optimal pairing of each note in one episode to exactly one other note in the other episode, maximizing the sum of similarities. The final similarity between two care episodes is this sum of their paired notes similarities, multiplied by two, and divided by the total number of notes in the two care episodes (Equation 1). See Figure 8 for an example showing how the notes are paired using the Hungarian Algorithm.

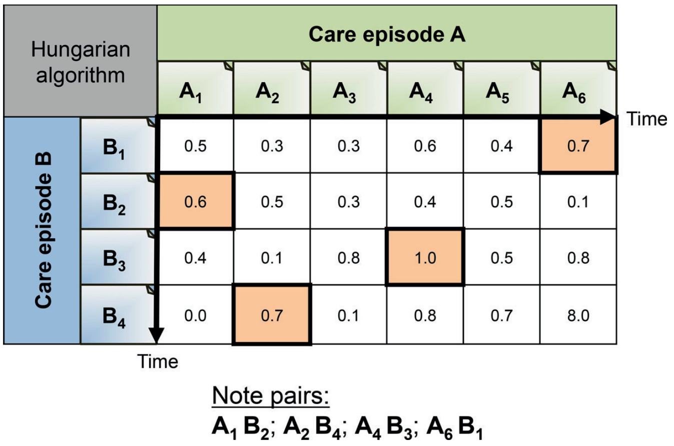
$$Sim(A, B) = \frac{2 \times \sum CosSim(\overrightarrow{A_i}, \overrightarrow{B_j})}{A.noteCount + B.noteCount}$$

NWSim: Needleman-Wunsch algorithm for sequence alignment of note vectors

Here we utilize a sequence alignment algorithm called *Needleman-Wunsch* [57] to align two episodes by their most similar notes. A note in one care episode can be aligned with zero or one notes in the other care episode. See Figure 9 for an example showing how the notes are aligned using the Needleman-Wunsch algorithm. The difference with the Hungarian Algorithm is that the temporal order of the notes is preserved. In other words, crossing alignment are not allowed. This reflects the intuition that care episodes sharing treatments in the same order are more likely to be similar than episodes with the same treatments in a different temporal order. We found that using the overall score produced by the Needleman-Wunsch algorithm for care episode similarity did not give any good results at this task. Instead, similarity between two care episodes is calculated from pairwise note vector similarities for the aligned notes. Final care episode similarity scores are obtained by using Equation 1.

Word vector weighting

The word vectors used in calculating care episode similarities, as described in section 'Compute care episode similarity', are all normalized and weighted before they are used. Common to all is that they are first normalized and multiplied by their Inverse Document Frequency (IDF) weight [58]. Such weighting is done in an attempt to weight words by their overall relevancy compared to the other words on corpus level. It essentially gives more weight to words occurring in few documents

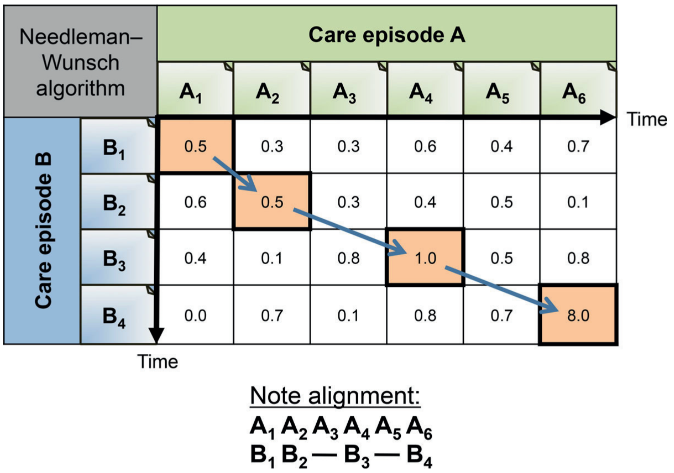


Hungarian algorithm	A_1	A_2	A_3	A_4	A_5	A_6
B_1	0.5	0.3	0.3	0.6	0.4	0.7
B_2	0.6	0.5	0.3	0.4	0.5	0.1
B_3	0.4	0.1	0.8	1.0	0.5	0.8
B_4	0.0	0.7	0.1	0.8	0.7	8.0

Note pairs:

$A_1 B_2$; $A_2 B_4$; $A_4 B_3$; $A_6 B_1$

Figure 8 Hungarian algorithm for note pairing. Figure showing an example of how the Hungarian algorithm would find the optimal clinical note pairs for care episode A and B.



Needleman–Wunsch algorithm	A_1	A_2	A_3	A_4	A_5	A_6
B_1	0.5	0.3	0.3	0.6	0.4	0.7
B_2	0.6	0.5	0.3	0.4	0.5	0.1
B_3	0.4	0.1	0.8	1.0	0.5	0.8
B_4	0.0	0.7	0.1	0.8	0.7	8.0

Note alignment:

$A_1 A_2 A_3 A_4 A_5 A_6$
$B_1 B_2 — B_3 — B_4$

Figure 9 Needleman-Wunsch algorithm for note alignment. Figure showing an example of how the Needleman-Wunsch algorithm would align clinical note pairs for care episode A and B.

(notes in our case) while giving less weight to those occurring in many documents. We refer to this weighting method as IDFWeight.

As a part of the experiment reported here, we aim to improve upon the domain-independent IDF weighting. For this, we boost the weight of words with clinical relevancy. This is accomplished by doubling the IDF weight of words occurring in a Finish health metathesaurus [59], which contains terms extracted from: vocabularies and classifications from FinMeSH; ICD-10; ICPC-2 (International Classification of Primary Care); the ATC-classification (generic drug names by WHO); the NOMESCO classification for surgical procedures; the Finnish vocabulary on nursing. This weighting method will be referred to as IDF*MetathesaurusWeight. Each of the approaches to calculating care episode similarity, with the models described in section ‘Semantic models’, are tested both with and without such metathesaurus-based re-weighting of word vectors.

Baselines

Two baselines were used in this study. The first one is random retrieval of care episodes, which can be expected to give very low scores and serves merely as a sanity check. The second one is Apache Lucene [29], a state-of-the-art search engine based on look-up of similar documents through a reverse index and relevance ranking based on a TF*IDF-weighted vector space model. Care episodes and underlying notes were indexed using Lucene. Similar to the other models/methods, all of the free text in the query episode, excluding the discharge summary and any sentence mentioning ICD-10 codes, served as the query string provided to Lucene. Being a state-of-the-art IR system, the scores achieved by Lucene in these experiments should indicate the difficulty of the task.

Results

Experiment 1: ICD-10 code identity

As explained in the ‘Task’ section, we lack a gold standard data set for care episode retrieval, consisting of relevant documents per query according to judgements by human experts. Therefore the relevance of retrieved episodes is estimated using a proxy. In this experimental setup, evaluation is based on the assumption that a retrieved episode is relevant if its ICD-10 code is identical to that of the query episode. It should be stressed that ICD-10 codes, i.e. possible free-text sentences that mention an ICD-10 code, are not included in the queries when conducting the experiment.

In the experiment we strove to have a setup that was as equal as possible for all models and systems, both in terms of text pre-processing and in terms of the target model dimensionality when inducing the vector space models. The clinical notes are split into sentences, tokenized, and lemmatized using a Constraint-Grammar based morphological analyzer and tagger extended with clinical vocabulary [60]. After stop words were removed [61], the total training corpus contained 39 million words (minus the query episodes), while the evaluation subset contained 18.5 million words. The vocabulary consisted of 0.6 million unique words.

In total, 40 care episodes were randomly selected to serve as the query episodes during testing, with the requirement that each had different ICD-10 codes and consisted of a minimum of six clinical notes. The average number of words per query episode is 796. The number of correct episodes per query episode varies between 9 and 1654. The total is 18440 episodes with an average length of 461 words per episode. When conducting the experiment all care episodes were retrieved for each of the 40 query episodes.

The RI- and W2V-based models have all a predefined dimensionality of 800. For the RI-based models, 4 non-zeros were used in the index vectors. For the RI-Word model, a narrow context window was employed (5 left + 5 right), weighting index vectors according to their distance to the target word ($weight_i = 2^{1-dist_{it}}$). In addition, the index vectors were shifted once left or right depending on what side of the target word they were located, similar to *direction vectors* as described in Sahlgren et al. [62]. These parameters for RI were chosen based on previous work on semantic textual similarity [63]. Also a much larger window of 20+20 was tested, but without noteworthy improvements.

The W2V-based models are trained using the BoW architecture and otherwise default parameters. The W2V-ICD model is trained with 10 iterations with a progressively decreasing learning rate, starting from 0.04. The utilized software was: Apache Lucene (version 4.2.0) [29]; The word2vec tool [64], for training the W2V model; A modified W2V implementation from the gensim library [65], for training of the W2V-ICD-based models; JavaSDM package [66], which served as the basis for the RI-based methods. Evaluation scores were calculated using the *TREC eval* tool [67].

As we have two different word vector weighting methods, and four different ways to calculate care episode similarities, a total of eight test runs was conducted:

- IDFWeight & SingleSim (Table 1).
- IDFWeight & AvgSim (Table 2).
- IDFWeight & HASim (Table 3).
- IDFWeight & NWSim (Table 4).
- IDF*MetathesaurusWeight & SingleSim (Table 5).
- IDF*MetathesaurusWeight & AvgSim (Table 6).
- IDF*MetathesaurusWeight & HASim (Table 7).
- IDF*MetathesaurusWeight & NWSim (Table 8).

Performance on care episode retrieval was assessed using three different evaluation measures:

- *Precision at 10* (P@10) denotes the precision among the top-10 results, in other words, the proportion of episodes with the same ICD-10 code as the query episode among the first 10 retrieved episodes. This probably best reflects the clinical usage scenario where a user is only prepared to check the highest ranked results, but is not willing to go through all results. P@10 scores reported are averages over 40 queries.
- R-precision (Rprec) is defined as the precision at the R-th position in the results, where R is the number of correct entries in the gold standard. This

Table 1 Experiment 1: Results from the IDFWeight & SingleSim setup.

IR model	MAP	P@10	Rprec
Lucene	0.1210	0.2800	0.1527
RI-Word	0.0915	0.2475	0.1300
RI-Note	0.1035	0.2850	0.1356
RI-ICD	0.2478	0.4250	0.2601
RI-Index	0.1171	0.3075	0.1555
W2V	0.1557	0.3000	0.1892
W2V-ICD	0.2666	0.3975	0.2874
Random	0.0178	0.0175	0.0172

Table 2 Experiment 1: Results from the IDFWeight & AvgSim setup.

IR model	MAP	P@10	Rprec
Lucene	0.0915	0.1564	0.0963
RI-Word	0.0317	0.0667	0.0465
RI-Note	0.0530	0.1308	0.0701
RI-ICD	0.1481	0.2256	0.1645
RI-Index	0.0599	0.1026	0.0654
W2V	0.1200	0.2128	0.1510
W2V-ICD	0.2357	0.3462	0.2499
Random	0.0178	0.0175	0.0172

Table 3 Experiment 1: Results from the IDFWeight & HASim setup.

IR model	MAP	P@10	Rprec
Lucene	0.1045	0.2385	0.1230
RI-Word	0.0319	0.1154	0.0456
RI-Note	0.0425	0.1487	0.0639
RI-ICD	0.0464	0.2333	0.0640
RI-Index	0.0840	0.2231	0.1112
W2V	0.0791	0.2513	0.1124
W2V-ICD	0.0917	0.2359	0.1311
Random	0.0178	0.0175	0.0172

Table 4 Experiment 1: Results from the IDFWeight & NWSim setup.

IR model	MAP	P@10	Rprec
Lucene	0.0812	0.2282	0.0938
RI-Word	0.0288	0.0795	0.0384
RI-Note	0.0358	0.0486	0.1000
RI-ICD	0.0407	0.1821	0.0560
RI-Index	0.0552	0.1923	0.0742
W2V	0.0647	0.1949	0.0954
W2V-ICD	0.0938	0.2410	0.1264
Random	0.0178	0.0175	0.0172

Table 5 Experiment 1: Results from the IDF*MetathesaurusWeight & SingleSim setup.

IR model	MAP	P@10	Rprec
Lucene	0.1210	0.2800	0.1527
RI-Word	0.0958	0.2600	0.1355
RI-Note	0.1161	0.2975	0.1501
RI-ICD	0.2372	0.4200	0.2541
RI-Index	0.1387	0.3100	0.1775
W2V	0.1619	0.3125	0.1968
W2V-ICD	0.2580	0.3850	0.2793
Random	0.0178	0.0175	0.0172

Table 6 Experiment 1: Results from the IDF*MetathesaurusWeight & AvgSim setup.

IR model	MAP	P@10	Rprec
Lucene	0.0915	0.1564	0.0963
RI-Word	0.0313	0.0641	0.0455
RI-Note	0.0559	0.1385	0.0741
RI-ICD	0.1453	0.2462	0.1632
RI-Index	0.0680	0.1000	0.0732
W2V	0.1280	0.2333	0.1592
W2V-ICD	0.2280	0.3410	0.2454
Random	0.0178	0.0175	0.0172

Table 7 Experiment 1: Results from the IDF*MetathesaurusWeight & HASim setup.

IR model	MAP	P@10	Rprec
Lucene	0.1045	0.2385	0.1230
RI-Word	0.0318	0.1128	0.0451
RI-Note	0.0430	0.1538	0.0631
RI-ICD	0.0452	0.2256	0.0627
RI-Index	0.0930	0.2385	0.1225
W2V	0.0814	0.2308	0.1176
W2V-ICD	0.0874	0.2359	0.1257
Random	0.0178	0.0175	0.0172

Table 8 Experiment 1: Results from the IDF*MetathesaurusWeight & NWSim setup.

IR model	MAP	P@10	Rprec
Lucene	0.0812	0.2282	0.0938
RI-Word	0.0288	0.0872	0.0379
RI-Note	0.0354	0.1179	0.0500
RI-ICD	0.0393	0.1821	0.0537
RI-Index	0.0601	0.2231	0.0810
W2V	0.0663	0.2051	0.0972
W2V-ICD	0.0890	0.2333	0.1196
Random	0.0178	0.0175	0.0172

indicates the proportion of the top-R retrieved episodes with the same ICD-10 code as the query episode, where R is the number of episodes with the same ICD-10 code in the whole collection. Our Rprec scores are averages over 40 queries.

- Mean average precision (MAP) is defined as the mean of the average precision over all (40) queries. For each query, the average is the precision value obtained for the top k documents, each time a relevant doc is retrieved. This is probably the most commonly used evaluation measure in IR. Moreover, it is very sensitive to ranking, so systems that rank the most similar episodes first receive higher MAP scores.

For the models using normal IDF weighting of word vectors (IDFWeight) the MAP, P@10 and Rprec results from each model, baselines, and the different ways to calculate care episode similarities, are shown in Tables 1, 2, 3, and 4. More precisely, results using IDFWeight and SingleSim are shown in Table 1. Table 2 shows the results from IDFWeight and AvgSim. Table 3 shows the results from IDFWeight and HASim. Table 4 shows the results from IDFWeight and NWSim. Best overall results among these are achieved with the setup SingleSim. Here, model W2V-ICD achieves highest MAP and Rprec scores, closely followed by RI-ICD. RI- ICD achieves the best P@10 scores. For the other setups, where each care episode is viewed as a collection of notes, shown in Tables 2, 3 and 4, the AvgSim approach to calculating care episode similarities achieves highest scores for most models. The exceptions are Lucene and RI-Index (and P@10 scores for RI-Word), which achieve noteworthy better scores with the HASim approach. No models achieve best scores with the NWSim approach. On average, W2V, W2V-ICD and RI-ICD outperforms Lucene. The other models either achieve scores that are comparable to Lucene, or lower. The latter is especially the case for the AvgSim, HASim and NWSim. Lucene and RI-Index seem to somewhat follow each other in terms of performance, which is as expected, given the similarities in how they are trained.

For the models using IDF weighting and double weight to words matching those in a health metathesaurus (IDF*MetathesaurusWeight), results are shown in Tables 5, 6, 7, and 8. The same trends are seen here with regards to scoring, where all models achieve best scores with SingleSim. No performance is gained in viewing each care episode as a collection of multiple individual notes.

When comparing the differences between IDFWeight (Tables 1, 2, 3, and 4) with IDF*MetathesaurusWeight (Tables 5, 6, 7, and 8), most setups and models achieve increased scores with IDF*MetathesaurusWeight. This is however not the case for the two models relying on ICD-10 codes for training, namely RI-ICD and W2V- ICD.

Experiment 2: Discharge summary overlap

This experiment uses a different evaluation method in which the semantic similarity between discharge summaries is used as a proxy for relevance. It assumes that a retrieved episode is relevant if its discharge summary is semantically similar to that of the query episode. It should be emphasized that discharge summaries are not used in either query construction or episode retrieval. Using the discharge summaries of the query episodes, the top 100 care episodes with the most similar discharge summary were selected. This procedure was repeated for each model - i.e. the six different semantic models and Lucene - resulting in seven different test sets, each consisting of 40 query episodes with their corresponding 100 most similar collection episodes. The top 100 was used rather than a threshold on the similarity score, because otherwise seven different thresholds would have to be chosen.

Subsequently a 7-by-7 experimental design was followed where each retrieval method, or model, was tested against each test set. At retrieval time, for each query episode, the system retrieves and ranks 1000 care episodes. It can be expected that when identical methods are used for retrieval and test set construction, the resulting bias gives rise to relatively high scores. In contrast, averaging over the scores for all seven construction methods is expected to be a less biased estimator of performance. The way these average scores are calculated is exemplified in Table 9 for MAP scores. This is done in the same way for the other scores (Retrieved count and P@10), but not shown for reasons of space. The resulting average scores for each care episode similarity calculation approach, over the various models, are shown as follows: Retrieved counts in Figure 10, MAP in Figure 11, and P@10 are shown in Figure 12.

The same models/systems and their parameters were used here as in Experiment 1. The Random baseline achieved the following average scores: Retrieved count = 151, MAP = 0.0004, P@10 = 0.0022.

For the results reported in Figures 10, 11 and 12, IDFWeight word weighting was used for generating both the result sets and the evaluation sets, however we also tried using the IDF*MetathesaurusWeight weighting approach when generating the result sets. When evaluated on the evaluation sets generated with IDFWeight weighting, we observed that the results for each model were typically slightly better compared to the result sets generated with IDFWeight weighting for the following models: RI-Word, RI-Note, RI-Index and W2V (average score gain +3.39%). And similar to Experiment 1, this

Table 9 Experiment 2: MAP scores for different IR models (rows) when using different models for measuring discharge summary similarity (columns).

Test set	Lucene	RI-Word	RI-Note	RI-ICD	RI-Index	W2V	W2V-ICD	Average	Rank
IR model									
Lucene	0.084	0.046	0.041	0.050	0.030	0.062	0.071	**0.055**	4
RI-Word	0.041	0.049	0.029	0.036	0.016	0.048	0.051	**0.039**	7
RI-Note	0.048	0.041	0.063	0.061	0.024	0.050	0.074	**0.052**	5
RI-ICD	0.059	0.036	0.054	0.149	0.033	0.058	0.124	**0.073**	2
RI-Index	0.063	0.033	0.044	0.048	0.043	0.052	0.065	**0.050**	6
W2V	0.075	0.051	0.052	0.079	0.035	0.094	0.106	**0.070**	3
W2V-ICD	0.089	0.053	0.070	0.150	0.046	0.094	0.187	**0.098**	1

This table also illustrates the general approach to how the average scores are calculated for the results graphs for Experiment 2.

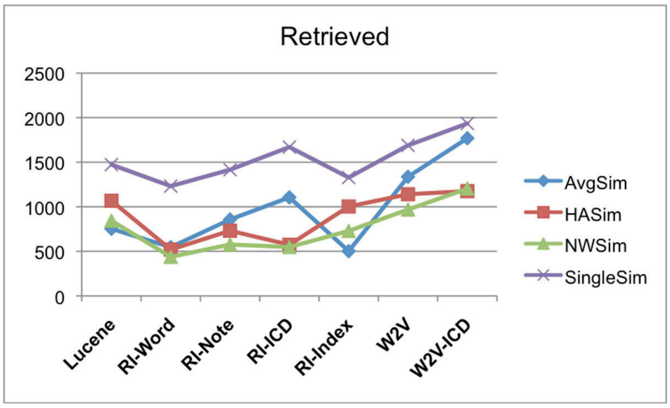


Figure 10 IDFWeight-Eval - IDFWeight-Results - Retrieved counts. Average number of correctly retrieved care episodes (max 4000) for different similarity measures using the various IR models. For creating the evaluation set the IDFWeight weighting was used, and also the retrieval was done using the IDFWeight weighting.

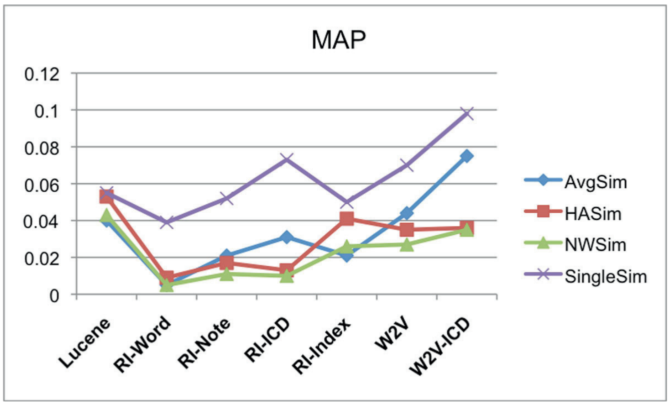


Figure 11 IDFWeight-Eval - IDFWeight-Results - MAP. Average MAP scores for different similarity measures using the various IR models. For creating the evaluation set the IDFWeight weighting was used, and also the retrieval was done using the IDFWeight weighting.

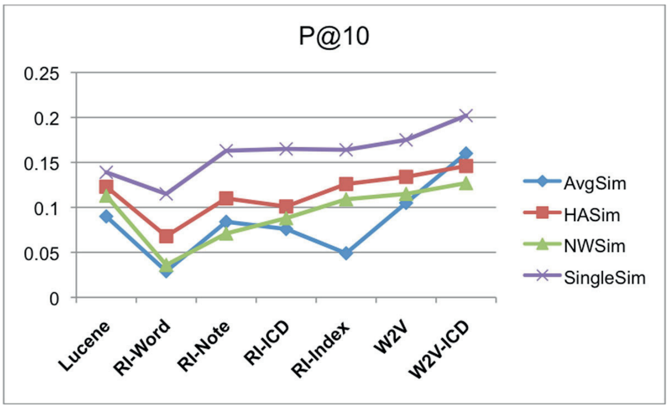


Figure 12 IDFWeight-Eval - IDFWeight-Results - P@10. Average P@10 scores for different similarity measures using the various IR models. For creating the evaluation set the IDFWeight weighting was used, and also the retrieval was done using the IDFWeight weighting.

was not the case for the RI-ICD and W2V-ICD models (average score gain −1.83%).

Discussion

The goal of the experiments was to determine which distributional semantic model work best for care episode retrieval, and what the best way of calculating care episode similarity is. The experimental results show that several models outperform Lucene. This suggests that distributional semantic models contribute positively to calculating document/note similarities in the clinical domain, compared with straight forward word matching (cf. RI-Index and Lucene). Both W2V and RI-Word utilize a narrow contextual sliding window during training. The scores indicate that W2V induces a model that, among these two, is better suited for IR with the approach taken here. RI-Word did perform relatively bad, and there are reasons to believe that performance gains can be achieved through adjusting and/or optimizing the utilized weighting (cf. TF*IDF), vector normalization, and model training parameters [68,69].

The relatively good performance of the RI-ICD and W2V-ICD models suggests that exploiting structured or encoded information in building semantic models for doing clinical NLP is a promising direction that calls for further investigation. This applies to clinical NLP as well as other domains and NLP tasks. This approach concurs with the arguments in favor of reuse of existing information sources in Friedman et al. [21]. On the one hand, it may not be surprising that these models perform best in Experiment 1, given that both modelling/training and evaluation method here rely on ICD-10 codes. On the other hand, being able to accurately retrieve care episodes with similar ICD-10 codes evidently has practical value from a clinical perspective. With the evaluation used in Experiment 1, one could argue that the best performance would be achieved by a dedicated ICD-10 classification system. However, in an IR context a labeling of each care episode by a small number of ICD- 10 codes does not provide sufficient information to allow full (relative) similarity rankings of the care episodes. One would thus not be able to retrieve e.g. the top 10 most similar care episodes to a query episode in a ranked (descending) order.

Additional support for the ICD-10 code based models comes from a different evaluation strategy that makes use of the discharge summaries associated with each care episode. This method is based on the idea that similar care episodes are likely to have similar discharge summaries. Therefore an approximation of the gold standard for a query can be obtained from the top-n episodes in the collection with a summary most similar to that of the query. Notice that, same as for the ICD-10 codes, the discharge summary is not used for constructing the query. However, this approach has some drawbacks. For example, similarity between discharge summaries must be measured using the same distributional models as used in retrieval, introducing a certain amount circularity. There is also no principled way to determine the value of *n* when taking the top-n results. Yet, when using this evaluation method - which does not rely on ICD-10 codes - the ICD-based models still perform best (cf. results reported in [23]), suggesting that their good performance is not only due to the use of ICD-10 codes for evaluation purposes.

Further, the results indicates, for most models whose word vectors are built from word distribution statistics, performance gains when heightened weight is given to words matching those in a health metathesaurus. Such a list of health terms is something that can easily be obtained in most languages. The fact that RI-ICD and W2V-ICD did not benefit from such re-weighting of word vectors can be explained through how these models are trained, and that the "statistical correct" semantic meanings of words, especially in relation to the ICD-10 codes, is already induced through the training phase.

All models performed best when care episodes were viewed as atomic documents (SingleSim). Thus there seems to be no performance gain in taking the internal structure of each care episode into account, i.e., the individual clinical notes. One possible reason for this being the case would be that each note on their own, compared to all text in a full care episode, do not contain enough information to be properly discriminative for this task.

In our data a single care episode can potentially span across several hospital wards. A better correlation between the similarity measures is to be expected when using care episodes originating from a single ward. Also, taking into consideration all ICD-10 codes for care episodes - not only the primary one - could potentially improve discrimination among care episodes. This could be useful for extending the RI-ICD and W2V-ICD models by training them on the secondary ICD-10 codes as well.

Input to the models for training was limited to the free text in the clinical notes, with the exception of the use of ICD-10 codes in the RI-ICD and W2V-ICD models. Additional sources of information could, and probably should, be utilized in an actual care episode retrieval system deployed in a hospital. A prime candidate is the structured and coded information commonly found in EHR systems. Examples are patient diagnosis and treatment codes, lab test values, dates, wards visited, medications, care episode span, previous diagnosis, age, sex, classified events, and so on. Some of these may belong to an ontology or thesaurus with a certain internal structure that could be exploited, such as SNOMED CT [70] and UMLS [71] (for languages where this can be applied). Other potential sources include user- supplied keywords or information units/concepts that have been extracted from, or matched against, free text using some type of information extraction tool such as MetaMap [72]. Such structured information can be used directly for IR, or indirectly through training of models as demonstrated in the current work. One potential issue with the use of structured information is that it may be incomplete or missing, giving rise to the problem of "missing values".

Conclusion

This paper proposes the new task of *care episode retrieval* as a special instance of information retrieval in the clinical domain. It was argued that the specialized clinical language use calls for dedicated NLP resources, which are mostly lacking - especially for languages other than English - and costly to build. Distributional models of semantics, built from a collection of raw clinical text in a fully unsu- pervised manner, were proposed as a resource-lean alternative. Several variants of *random indexing* and *word2vec* were proposed and experimentally tested. Also several approaches to calculating the overall care episode similarity on the basis of their word similarities were explored.

As manually constructing a gold standard is costly, two new evaluation strategies are introduced. One relies on the ICD-10 codes attached to care episodes, the other relies on discharge summaries. Two innovative distributional models were presented - RI-ICD and W2V-ICD - which leverage the ICD-10 codes to enhance domain- specific word similarity. These models also proved to yield best performance, out- performing a state-of-the-art search engine (Lucene). Taking the internal structure of care episodes into account, including attempts at optimal pairing or temporal alignment of individual clinical notes, did not yield any improvements.

The work presented here suggests that training a representation to associate additional knowledge to that obtained from the free text alone, such as structured domain information, is beneficial to the computation of semantic similarity. We have demonstrated how ICD-10 codes can be used indirectly for care episode retrieval, and we hypothesize that the utilized methods may also perform well when applied to more generic IR tasks within the clinical domain. Other types (structured) information units and concepts should also be explored in future work. Also, it is likely that a similar approach can be used for IR and NLP in other domains.

Our evaluation, as well as that in most of the related work, is based on pure retrieval performance. Future work on IR in the clinical domain should arguably focus more on evaluating IR-systems in terms of support for care and patient outcomes.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Hans Moen (HM) was responsible for coordinating the work and was involved in all parts of it, and had the main responsibility for the manuscript. HM had the responsibility of the overall design of the study and for carrying out the experiments. HM initiated the idea of using ICD codes for training models. HM also implemented most of the utilized code.
Filip Ginter (FG) contributed to the experimental setup, writing and reviewing of the manuscript. FG did the work related to training of the W2V model. FG also initiated the idea of translating the RI-ICD principles into W2V, i.e. the W2V-ICD method, including implementing and training it.
Erwin Marsi (EM) contributed to the design of the RI models, experimental setup, data analysis, and to writing and reviewing of the manuscript. EM also contributed with supervision thoughout the work.

Laura-Maria Murtola (LM) contributed to the writing and reviewing of the manuscript. LM did also contribute with clincal domain knowledge in most aspects of the work, and provided access to the utilized metathesaurus.
Tapio Salakoski (TS) contributed through reviewing of the manuscript and supervision. TS also provided access to the utilized clinical data.
Sanna Salanterä (SS) contributed through reviewing of the manuscript and by providing access to the utilized clinical data. SS also contributed with supervision and ideas leading up to the ICD-based models.

Acknowledgements

Parts of the work reported here are based on an earlier study reported in the workshop paper [23]. This study was partly supported by the Research Council of Norway through the EviCare project (NFR project no. 193022), the Turku University Hospital (EVO 2014), and the Academy of Finland (project no. 140323). The study is a part of the research projects of the Ikitik consortium (http://www.ikitik.fi). We would like to thank Juho Heimonen for assisting us in pre-processing the data. We would also like to thank the reviewers for their comments and suggestions.

Declarations

Publication costs for this article were funded by the corresponding author's institution[1].
This article has been published as part of *BMC Medical Informatics and Decision Making* Volume 15 Supplement 2, 2015: Proceedings of Louhi 2014: The Fifth International Workshop on Health Text Mining and Information Analysis. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcmedinformdecismak/supplements/15/S2.

Authors' details

[1]Department of Computer and Information Science, Norwegian University of Science and Technology, Sem Saelands vei 9, 7491 Trondheim, Norway. [2]Department of Information Technology, University of Turku, Joukahaisenkatu 3-5, 20520 Turku, Finland. [3]Department of Nursing Science, University of Turku, Lemminkäisenkatu 1, 20520 Turku, Finland. [4]Turku University Hospital, Kiinamyllynkatu 4-8, 20521 Turku, Finland. [5]Turku Centre for Computer Science (TUCS), Joukahaisenkatu 3-5, 20520 Turku, Finland.

Published: 15 June 2015

References

1. **European Commission: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: eHealth Action Plan 2012-2020 - Innovative healthcare for the 21st century.** 2012 [http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012DC0736&from=EN].
2. Blumenthal D, Tavenner M: **The "meaningful use" regulation for electronic health records.** *New England Journal of Medicine* 2010, **363(6)**:501-504.
3. Jha AK: **Meaningful use of electronic health records: the road ahead.** *JAMA* 2010, **304(15)**:1709-1710.
4. Bartlett C, Boehncke K, Haikerwal M: **E-health: enabler for australia's health reform.** Melbourne: Booz & Company, Melbourne; 2008.
5. Häyrinen K, Saranto K, Nykänen P: **Definition, structure, content, use and impacts of electronic health records: a review of the research literature.** *Int J Med Inform* 2008, **77(5)**:291-304.
6. Manning CD, Raghavan P, Schütze H: In *Introduction to Information Retrieval. Volume 1.* Cambridge university press Cambridge, Cambridge, UK; 2008:1-18.
7. Younger P: **Internet-based information-seeking behaviour amongst doctors and nurses: a short review of the literature.** *Health Information & Libraries Journal* 2010, **27(1)**:2-10.
8. Westbrook JI, Coiera EW, Gosling AS: **Do online information retrieval systems help experienced clinicians answer clinical questions?** *Journal of the American Medical Informatics Association* 2005, **12(3)**:315-321.
9. Westbrook JI, Gosling AS, Coiera EW: **The impact of an online evidence system on confidence in decision making in a controlled setting.** *Medical Decision Making* 2005, **25(2)**:178-185.
10. Lenz M, Hübner A, Kunze M: **Textual cbr.** *Case-based Reasoning Technology* Springer, New York, USA; 1998, 115-137.
11. Yang L, Mei Q, Zheng K, Hanauer DA: **Query log analysis of an electronic health record search engine.** *AMIA Annual Symposium Proceedings* 2011, **2011**:915-924.
12. Rector AL: **Clinical terminology: why is it so hard?** *Methods Inf Med* 1999, **38(4-5)**:239-252.
13. Friedman C, Kra P, Rzhetsky A: **Two biomedical sublanguages: a description based on the theories of Zellig Harris.** *J Biomed Inform* 2002, **35(4)**:222-235.
14. Allvin H, Carlsson E, Dalianis H, Danielsson-Ojala R, Daudaravičius V, Hassel M, *et al*: **Characteristics and analysis of Finnish and Swedish clinical intensive care nursing narratives.** *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents* Association for Computational Linguistics; 2010, 53-60.
15. Shatkay H: **Hairpins in bookstacks: information retrieval from biomedical text.** *Briefings in Bioinformatics* 2005, **6(3)**:222-238.
16. Pedersen T, Pakhomov SV, Patwardhan S, Chute CG: **Measures of semantic similarity and relatedness in the biomedical domain.** *Journal of Biomedical Informatics* 2007, **40(3)**:288-299.
17. Koopman B, Zuccon G, Bruza P, Sitbon L, Lawley M: **An evaluation of corpus-driven measures of medical concept similarity for information retrieval.** *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* ACM; 2012, 2439-2442.
18. Henriksson A, Moen H, Skeppstedt M, Daudaravi V, Duneld M, *et al*: **Synonym extraction and abbreviation expansion with ensembles of semantic spaces.** *J Biomed Semantics* 2014, **5(1)**:6.
19. Cohen T, Widdows D: **Empirical distributional semantics: Methods and biomedical applications.** *Journal of Biomedical Informatics* 2009, **42(2)**:390-405.
20. Harris ZS: **Distributional structure.** *Word* 1954, **10**:146-162.
21. Friedman C, Rindflesch TC, Corn M: **Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the national library of medicine.** *J Biomed Inform* 2013, **46(5)**:765-773.
22. **World Health Organization and others: International classification of diseases (icd).** 2013.
23. Moen H, Marsi E, Ginter F, Murtola LM, Salakoski T, Salanterä S: **Care episode retrieval.** *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL* Association for Computational Linguistics, Gothenburg, Sweden; 2014, 116-124.
24. Hanauer DA: **EMERSE: the electronic medical record search engine.** In *AMIA Annual Symposium Proceedings. Volume 2006.* American Medical Informatics Association; 2006:941.
25. Gregg W, Jirjis J, Lorenzi NM, Giuse D: **Startracker: an integrated, web-based clinical search engine.** *AMIA Annual Symposium Proceedings* American Medical Informatics Association; 2003, 855.
26. Campbell EJ, Krishnaraj A, Harris M, Saini S, Richter JM: **Automated before-procedure electronic health record screening to assess appropriateness for GI endoscopy and sedation.** *Gastrointestinal Endoscopy* 2012, **76(4)**:786-792.
27. Markó K, Schulz S, Hahn U: **Morphosaurus-design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain.** *Methods of Information in Medicine* 2005, **44(4)**:537-545.
28. Natarajan K, Stein D, Jain S, Elhadad N: **An analysis of clinical queries in an electronic health record search utility.** *Int J Med Inform* 2010, **79(7)**:515-522.
29. Cutting D: **Apache Lucene open source package.** 1999 [http://lucene.apache.org/].
30. Ounis I, Amati G, Plachouras V, He B, Macdonald C, Lioma C: **Terrier: A high performance and scalable information retrieval platform.** *Proceedings of the OSIR Workshop* Citeseer; 2006, 18-25.
31. Zheng K, Mei Q, Hanauer DA: **Collaborative search in electronic health records.** *Journal of the American Medical Informatics Association* 2011, **18(3)**:282-291.
32. Alkasab TK, Harris MA, Zalis ME, Dreyer KJ, Rosenthal DI: **A case tracking system with electronic medical record integration to automate outcome tracking for radiologists.** *J Digit Imaging* 2010, **23(6)**:658-665.
33. Spat S, Cadonna B, Rakovac I, Gütl C, Leitner H, Stark G, Beck P: **Enhanced information retrieval from narrative german-language clinical text documents using automated document classification.** *Stud Health Technol Inform* 2008, **136**:473-478.
34. Schulz S, Daumke P, Fischer P, Müller M: **Evaluation of a document search engine in a clinical department system.** *AMIA Annual Symposium Proceedings* American Medical Informatics Association; 2008, 647-651.
35. Voorhees E, Tong R: **Overview of the TREC 2011 medical records track.** *The Twentieth Text REtrieval Conference Proceedings TREC* 2011.

36. Hersh WR, Voorhees EM: **Overview of the TREC 2012 medical records track.** *The Twenty-first Text REtrieval Conference Proceedings TREC* 2012.
37. Suominen H, Salanterä S, Velupillai S, Chapman WW, Savova G, Elhadad N, *et al*: **Overview of the ShARe/CLEF eHealth evaluation lab 2013.** *Information Access Evaluation Multilinguality, Multimodality, and Visualization* 2013, 212-231.
38. Limsopatham N, Macdonald C, Ounis I: **Inferring conceptual relationships to improve medical records search.** *OAIR '13 Proceedings of the 10th Conference on Open Research Areas in Information Retrieval* 2013, 1-8.
39. Zhu D, Carterette B: **Combining multi-level evidence for medical record retrieval.** *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing* 2012, 49-56.
40. Zhu D, Carterette B: **Improving health records search using multiple query expansion collections.** *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference On* 2012, 1-7.
41. Harkema H, Dowling JN, Thornblade T, Chapman WW: **Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports.** *J Biomed Inform* 2009, **42(5)**:839-851.
42. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J: **Distributed representations of words and phrases and their compositionality.** *Advances in Neural Information Processing Systems* 2013, **26**:3111-3119.
43. Berry MW, Dumais ST, O'Brien GW: **Using linear algebra for intelligent information retrieval.** *SIAM Review* 1995, **37(4)**:573-595.
44. Ruiz M, Eliasmith C, López A: **Exploring the Use of Random Indexing for Retrieving Information.** *Statistical Language and Speech Processing - First International Conference* 2013.
45. Carrillo M, Villatoro-Tello E, Lopez-Lopez A, Eliasmith C, Montes-y-Gomez M, Villasenor-Pineda L: **Representing context information for document retrieval.** *Flexible Query Answering Systems* 2009, **5822(4)**:239-250.
46. Vasuki V, Cohen T: **Reflective random indexing for semi-automatic indexing of the biomedical literature.** *J Biomed Inform* 2010, **43(5)**:694-700.
47. Le QV, Mikolov T: **Distributed representations of sentences and documents.** *Proceedings of The 31st International Conference on Machine Learning* 2014, 1188-1196.
48. Kanerva P, Kristofersson J, Holst A: **Random Indexing of Text Samples for Latent Semantic Analysis.** *Proceedings of 22nd Annual Conference of the Cognitive Science Society* 2000, 103-106.
49. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R: **Indexing by latent semantic analysis.** *Journal of the American Society for Information Science* 1990, **41(6)**:391-407.
50. Symonds M, Zuccon G, Koopman B, Bruza P, Nguyen A: **Semantic judgement of medical concepts: Combining syntagmatic and paradigmatic information with the tensor encoding model.** *Australasian Language Technology Association Workshop* 2012, 15.
51. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR: **A systematic literature review of automated clinical coding and classification systems.** *Journal of the American Medical Informatics Association* 2010, **17(6)**:646-651.
52. Henriksson A, Hassel M, Kvist M: **Diagnosis code assignment support using random indexing of patient records-a qualitative feasibility study.** *AIME'11 Proceedings of the 13th conference on Artificial intelligence in medicine* 2011, 348-352.
53. Bengio Y, Ducharme R, Vincent P, Janvin C: **A neural probabilistic language model.** *Journal of Machine Learning Research* 2003, **3**:1137-1155.
54. Mikolov T, Yih Wt, Zweig G: **Linguistic regularities in continuous space word representations.** *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 2013, 746-751.
55. Levy O, Goldberg Y: **Linguistic regularities in sparse and explicit word representations.** *Proceedings of Eighteenth Conference on Computational Natural Language Learning (CoNNL-2014)* 2014.
56. Kuhn HW: **The Hungarian method for the assignment problem.** *Naval Research Logistics Quarterly* 1995, **2**:83-97.
57. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *Journal of Molecular Biology* 1970, **48(3)**:443-453.
58. Jones KS: **A statistical interpretation of term specificity and its application in retrieval.** *Journal of Documentation* 1972, **28(1)**:11-21.
59. **FinMeSH - the solid ground for a health metathesaurus in Finland.** [http://www.kaypahoito.fi/documents/10184/18136/Posteri+FinMeSH.pdf], Poster presented at European Association of Health Information and Libraries (EAHIL) Workshop, September 12–15, Krakow, Poland (2007).
60. Karlsson F: **Constraint Grammar: a Language-independent System for Parsing Unrestricted Text.** Mouton de Gruyter, Berlin and New York; 1995.
61. Hyppänen H: **Finnish stopword list.** 2007 [http://www.nettiapina.fi/finnish-stopword-list/].
62. Sahlgren M, Holst A, Kanerva P: **Permutations as a means to encode order in word space.** *Proceedings of the Annual Meeting of the Cognitive Science Society* 2008.
63. Moen H, Marsi E, Gambäck B: **Towards dynamic word sense discrimination with random indexing.** *Proceedings of the Workshop on Continuous Vector Space Models and Their Compositionality* 2013, 83-90.
64. Mikolov T: **Word2vec tool.** 2013 [https://code.google.com/p/word2vec/].
65. Řhůřek R, Sojka P: **Software Framework for Topic Modelling with Large Corpora.** *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* 2010, 45-50.
66. **JavaSDM package.** 2004 [http://www.nada.kth.se/~xmartin/java/].
67. **National Institute of Standards and Technology: TREC eval tool.** 2006 [http://trec.nist.gov/trec_eval/].
68. Henriksson A, Hassel M: **Optimizing the dimensionality of clinical term spaces for improved diagnosis coding support.** *Proceedings of the Louhi Workshop on Health Document Text Mining and Information Analysis* 2013, 1-6.
69. Moen H, Marsi E: **Cross-lingual random indexing for information retrieval.** *Statistical Language and Speech Processing* 2013, 164-175.
70. **International Health Terminology Standards Development Organisation: Supporting Different Languages.** [http://www.ihtsdo.org/snomed-ct/snomed-ct0/different-languages/].
71. **Unified Medical Language System.** [http://www.nlm.nih.gov/research/umls/].
72. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC: **Medex: a medication information extraction system for clinical narratives.** *Journal of the American Medical Informatics Association* 2010, **17(1)**:19-24.

doi:10.1186/1472-6947-15-S2-S2
Cite this article as: Moen *et al.*: Care episode retrieval: distributional semantic models for information retrieval in the clinical domain. *BMC Medical Informatics and Decision Making* 2015 **15**(Suppl 2):S2.

On evaluation of automatically generated clinical discharge summaries

Moen, Hans; Heimonen, Juho; Murtola, Laura-Maria; Airola, Antti; Pahikkala, Tapio; Terävä, Virpi; Danielsson-Ojala, Riitta; Salakoski, Tapio, and Salanterä, Sanna

In *Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI 2014)*, pages 101–114, Trondheim, Norway, 2014. CEUR Workshop Proceedings.

Paper D

On Evaluation of Automatically Generated Clinical Discharge Summaries

Hans Moen[1], Juho Heimonen[2,5], Laura-Maria Murtola[3,4], Antti Airola[2],
Tapio Pahikkala[2,5], Virpi Terävä[3,4], Riitta Danielsson-Ojala[3,4],
Tapio Salakoski[2,5], and Sanna Salanterä[3,4]

[1] Department of Computer and Information Science,
Norwegian University of Science and Technology, Norway
[2] Department of Information Technology, University of Turku, Finland
[3] Department of Nursing Science, University of Turku, Finland
[4] Turku University Hospital, Finland
[5] Turku Centre for Computer Science, Finland
hans.moen@idi.ntnu.no
{juaheim,lmemur,ajairo,aatapa,vmater,rkdaoj}@utu.fi
{tapio.salakoski,sansala}@utu.fi

Abstract. Proper evaluation is crucial for developing high-quality computerized text summarization systems. In the clinical domain, the specialized information needs of the clinicians complicates the task of evaluating automatically produced clinical text summaries. In this paper we present and compare the results from both manual and automatic evaluation of computer-generated summaries. These are composed of sentence extracts from the free text in clinical daily notes – corresponding to individual care episodes, written by physicians concerning patient care. The purpose of this study is primarily to find out if there is a correlation between the conducted automatic evaluation and the manual evaluation. We analyze which of the automatic evaluation metrics correlates the most with the scores from the manual evaluation. The manual evaluation is performed by domain experts who follow an evaluation tool that we developed as a part of this study. As a result, we hope to get some insight into the reliability of the selected approach to automatic evaluation. Ultimately this study can help us in assessing the reliability of this evaluation approach, so that we can further develop the underlying summarization system. The evaluation results seem promising in that the ranking order of the various summarization methods, ranked by all the automatic evaluation metrics, correspond well with that of the manual evaluation. These preliminary results also indicate that the utilized automatic evaluation setup can be used as an automated and reliable way to rank clinical summarization methods internally in terms of their performance.

Keywords: Summarization Evaluation, Text Summarization, Clinical Text Processing, NLP

1 Introduction

With the large amount of information generated in health care organisations today, information overload is becoming an increasing problem for clinicians [1,2]. Much of the information that is generated in relation to care is stored in electronic health record (EHR) systems. The majority of this is free text – stored as clinical notes – written on a daily basis by clinicians about care of individual patients. The rest of the information contained in EHRs is mainly images and structured information, such as medication, coded information and lab values. Towards tackling the problems of information overload, there is a need for (EHR) systems that are able to automatically generate an overview, or summary, of the information in these health records - this applies to both free text and structured information. Such systems would enable clinicians to spend more time treating the patients, and less time reading up on information about the patients. However, in the process of developing such summarization systems, quick and reliable evaluation is crucial.

A typical situation where information overload is frequently encountered is when the attending physician is producing the discharge summary at the end of a care episode. Discharge summaries are an important part of the communication between different professionals providing the health care services and their aim to ensure the continuity of a patients care. However, there are challenges with these discharge summaries as they are often produced late, and the information they contain tend to be insufficient. For example, one study showed that discharge summaries exchanged between the hospital and the primary care physicians is often lacking information, such as diagnostic test results (lacking in 33-63%), treatment progression (lacking in 7-22%), medications (lacking in 2-40%), test results (lacking in 65%), counseling (lacking in 90-92%) and follow-up proposals (lacking in 2-43%) [3]. One reason for this is that, during discharge summary writing process, the physicians tend to simply not have the time to read everything that has been documented in the clinical daily notes. Another reason is the difficulty of identifying the most important information to include in the discharge summary.

Computer-assisted discharge summaries and standardized templates are measures for improving the transfer time and the quality of discharge information between the hospital and the primary care physicians [3]. Furthermore, computer-assisted discharge summary writing using automatic text summarization could improve the timeliness and quality of discharge summaries further. Another more general user scenario where text summarization would be useful is when clinicians need to get an overview of the documented content in a care episode, in particular in critical situations when this information is needed without delay.

Automatic summarization of clinical information is a challenging task because of the different data types, the domain specificity of the language, and

In: E.A.A. Jaatun, E. Brooks, K.E. Berntsen, H. Gilstad, M. G. Jaatun (eds.): Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI 2014), Trondheim, Norway, 19-MAY-2014, published at http://ceur-ws.org

the special information needs of the clinicians [4]. Producing a comprehensive overview of the structured information is a rather trivial task [5]. However, that is not the case for the clinical notes and the free text they contain. Previously, Liu et al. [6] applied the automated text summarization methods of the MEAD system [7] to Finnish intensive care nursing narratives. In this work the produced summaries were automatically evaluated against corresponding discharge reports. The authors found that some of the considered methods outperformed the random baseline method, however, the authors noted that the results were overall quite disappointing, and that further work was needed in order to develop reliable evaluation methods for the task.

We have developed an extraction based text summarization system that attempts to automatically produce a textual summary of the free text contained in all the clinical (daily) notes related to a – possibly ongoing – care episode, written by physicians. *The focus of this paper is not on how the summarization system works, but rather on how to evaluate the summaries it produces.* In our ongoing work towards developing this system, we have so far seven different *summarization methods* to evaluate, including a `Random` method and an `Oracle` method. The latter method representing an upper bound for the automatic evaluation score. Having a way to quickly and automatically evaluate the summaries that these methods produce is critical during method development phase. Thus the focus of this paper is how to perform such automated evaluation in a reliable and cost-effective way.

Automatic evaluation of an extraction based summary is typically done through having a gold standard, or "gold summary", for comparison. A gold summary is typically an extraction based summary produced by human experts [8]. Then one measures the textual overlap, or similarity, between a targeted summary and the corresponding gold summary, using some metric for this purpose. However, we do not have such manually tailored gold summaries available. Instead we explore the use of the original physician-made discharge summaries for evaluation purposes as a means of overcoming this problem. These discharge summaries contain sentence extracts, and possibly slightly rewritten sentences, from the clinical notes. They also typically contain information that has never been documented earlier in the corresponding care episode, which makes them possibly suboptimal for the task of automatic evaluation.

To explore whether this approach to automatic evaluation is viable, we have also conducted *manual evaluation* of a set of summaries, and then compared this to the results from the automatic evaluation. A possible correlation between how the manual and automatic evaluation ranks the summarization methods would mean that further automatic evaluation with this approach can be considered somewhat reliable. In this study, automatic evaluation is mainly performed using the *ROUGE evaluation package* [9]. The manual evaluation was done by domain experts who followed an evaluation scheme/tool that we developed for this purpose. Figure 1 illustrates the evaluation experiment.

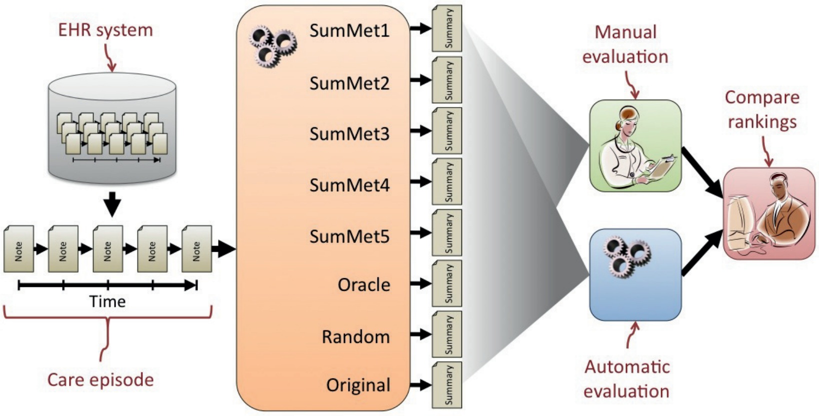


Fig. 1. Illustration of the evaluation experiment.

2 Data

The data set used in this study contained the electronic health records of approximately 26,000 patients admitted to a Finnish university hospital between the years 2005–2009 with any type of cardiac problem. An ethical statement (17.2.2009 §67) and the organisational permission (2/2009) from the hospital district was obtained before collection of this data set.

To produce data suited for automatic summarization, discharge summaries were extracted and associated to the daily notes they summarize. There were two types of discharge summaries: internal (written when the patient is moved to another ward and summarizing the time spent on the given ward) and final (written when the patient leaves the hospital and summarizing the whole stay). Note that a final discharge also summarizes any internal summaries written during the stay.

All notes and discharge summaries were lemmatized at the sentence level using the morphological analyser FinTWOL [10] and the disambiguator FinCG [11] by Lingsoft, Inc.[6]. Stopwords were also removed[7]. The preprocessed corpus contained 66,884 unique care episodes with 39 million words from a vocabulary of 0.6 million unique terms.

The full corpus was utilized in deriving statistics about the language for some of the summarization methods. For the summarization and evaluation experiment, the corpus was narrowed down to the care episodes having I25 (Chronic ischaemic heart disease; including its sub-codes) as the primary ICD-10 code and consisting of at least 8 clinical notes, including the discharge summary. The

[6] http://www.lingsoft.fi

[7] http://www.nettiapina.fi/finnish-stopword-list/

latter condition justifies the use of text summarization. The data were then split into the *training* and *test* sets, containing 159 and 156 care episodes, for the parameter optimization and evaluation of summarization methods, respectively.

3 Text Summarization

All summarization methods used in this study are based on *extraction-based* multi-document summarization. This means that each summary consist of a subset of the content in the original sentences, found in the various clinical notes that the summary is produced from [12]. This can be seen as a specialized type of multi-document summarization since each document, or clinical note, belong to the same patient, and together constitute a connected sequence. In the presented evaluation, seven different summarization methods are used, including `Random` and `Oracle`, resulting in seven different summaries per care episode. The original physician made discharge summary, `Original`, which is used as the *gold summary* for automatic evaluation, is also included in the manual evaluation. For the automatic evaluation, this gold summary is viewed as the perfect summary, thus having a perfect F-score (see Section 4.1). As stated earlier, the focus of this paper is on evaluating the text summaries produced by a summarization system. Thus the description of the utilized summarization system and the various methods used will be explained in more detail in a forthcoming extended version of this work. However, the two trivial control methods, `Random` and `Oracle`, deserves some explanation here.

`Random` This is the baseline method, which works by composing a summary through simply selecting sentences randomly from the various clinical notes. This method should give some indication of the difficultly level of the summarization task at hand.

`Oracle` This is a control-method that has access to the gold summary during the summarization process. It basically tries to optimize the ROUGE-N2 F-scores for the generated summary according to the gold summary, using a greedy search strategy. This summarization method can naturally not be used in a real user scenario, but it represents the upper limit for what is possible to achive in score for an extraction based summary, or summarization method, when using ROUGE for evaluation.

The other summarization methods are here referred to as `SumMet1`, `SumMet2`, `SumMet3`, `SumMet4` and `SumMet5`.

For each individual care episode, the length of the corresponding gold summary served as the length limit for all the seven generated summaries. This was mainly done so that a sensible automatic evaluation score (F-score, see Section 4.1) could be calculated. In a more realistic user scenario, a fixed length could be used, or e.g. a length that is based on how many clinical notes the summaries are generated from. Each computer generated summary is run through a post-processing step, where each sentence are sorted according to when they were written.

4 Evaluation Experiment

We conducted and compared two types of evaluation, *automatic evaluation* and *manual evaluation* in order to evaluate the different summarization methods. The purpose of this study is primarily to find out if there is a correlation between the conducted automatic evaluation and the manual evaluation. This will further reveal which of the automatic evaluation metrics that correlates the most with the scores from the manual evaluation. As a result, we would get some insight into the reliability of the selected approach to automatic evaluation. Ultimately this study can help us in assessing the reliability of this evaluation approach, so that we can further develop the underlying summarization system.

In the automatic evaluation we calculated the F-score for the overlap between the generated summaries and the corresponding gold summaries using the ROUGE evaluation package. As gold summaries we used the original discharge summary written by a physician. This gold summary is thus considered to be the optimal summary[8], so we assume that it to always have an F-score of 1.0.

The conducted evaluation can be classified as so called *intrinsic evaluation.* This means that the summaries are evaluated independently of how they potentially affect some external task [8].

4.1 Automatic Evaluation

ROUGE metrics, provided by the ROUGE evaluation package [9] (see e.g. [13]), are widely used as automatic performance measures in the text summarization literature. To limit the number of evaluations, we selected four common variants:

- **ROUGE-N1**. Unigram co-occurrence statistics.
- **ROUGE-N2**. Bigram co-occurrence statistics.
- **ROUGE-L**. Longest common sub-sequence co-occurrence statistics.
- **ROUGE-SU4**. Skip-bigram and unigram co-occurrence statistics.

These metrics are all based on finding word *n*-gram co-occurrences, or overlaps, between a) one or more *reference summaries*, i.e. gold summary, and b) the *candidate summary* to be evaluated.

Each metric counts the number of overlapping units (the counting method of which depends on the variant) and uses that information to calculate recall (R), precision (P), and F-score (F). The recall is the ratio of overlapping units to the total number of units in the reference while the precision is the ratio of overlapping units to the total number of units in the candidate. The former

[8] This is of course not always the truth from a clinical perspective, but we leave that to another discussion.

describes how well the candidate covers the reference and the latter describes the quality of the candidate. The F-score is then calculated as

$$F = \frac{2PR}{P + R} \ , \tag{1}$$

The evaluations were performed with the lemmatized texts with common stopwords and numbers removed.

4.2 Manual Evaluation

The manual evaluation was conducted independently by three domain experts. Each did a blinded evaluation of the eight summary types (seven machine generated ones plus the original discharge summary) for five care episodes. Hence, the total sum of evaluated summaries per evaluator was 40. All summaries were evaluated with a 30-item schema, or *evaluation tool* (see Table 1). This tool was constructed based on the content criteria guideline for medical discharge summaries, used in the region where the data was collected. So each item correspond to a criteria in this guideline. In this way, items were designed to evaluate the medical content of the summaries from the perspective of discharge summary writing. When evaluating a summary, each of these items were evaluated on a 4-class scale from -1 to 2, where, -1 = not relevant, 0 = not included, 1 = partially included and 2 = fully included. The evaluators also had all the corresponding clinical notes at their disposal when performing the evaluation.

The items in our tool are somewhat comparable to the evaluation criteria used in an earlier study of evaluating neonate's discharge summaries, where the computer generated discharge summaries using lists of diagnoses linked to ICD-codes [14]. However, the data summarized in the aforementioned work is mainly structured and pre-classified data, thus the summarization methods or performance is not comparable to our work.

The evaluators experienced the manual evaluations, following the 30-item tool, to be very difficult and extremely time consuming. This was mainly due to the evaluation tool, i.e. its items, being very detailed and required a lot of clinical judgement. Therefore, for this study, only five patient care episodes and their corresponding summaries, generated by the summarization system, were evaluated by all three evaluators. This number of evaluated summaries are too small for generalization of the results, but this should still give some indication of the quality of the various summarization methods in the summarization system. The 30 items in the manual evaluation tool are presented in Table 1.

4.3 Evaluation Statistics

In order to test whether the differences in the automatic evaluation scores between the different summarization methods were statistically significant, we performed the Wilcoxon signed-rank test [15] for each evaluation measure, and each pair of methods at significance level $p = 0.05$. We also calculated the p-values

Table 1. Items evaluated in the manual evaluation.

Evaluation criteria
Care period
Care place
Events (diagnoses/procedure codes) of care episode
Procedures of care episode
Long-term diagnoses
Reason for admission
Sender
Current diseases, which have impact on care solutions
Effects of current diseases, which have impact on care solutions
Current diseases, which have impact on medical treatment solutions
Effects of current diseases, which impact on medical treatment solutions
Course of the disease
Test results in chronological order with reasons
Test results in chronological order with consequences
Procedures in chronological order with reasons
Procedures in chronological order with consequences
Conclusions
Assessment of the future
Status of the disease at the end of the treatment period
Description of patient education
Ability to work
Medical certificates (including mention of content and duration)
Prepared or requested medical statements
A continued care plan
Home medication
Follow-up instructions
Indications for re-admission
Agreed follow-up treatments in the hospital district
Other disease, symptom or problem that requires further assessment
Information of responsible party for follow-up treatment

for manual evaluation. These were obtained with the paired Wilcoxon test computed over the 30 mean content criteria scores of the 30-item evaluation tool (see Table1). The mean scores were calculated by averaging the manually entered scores of the three evaluators and five care episodes. The -1 values indicating non-relevance were treated as missing values (i.e. they were ignored when calculating the averages). Also here a significance level of $p = 0.05$ was used. The agreement between the three evaluators was investigated by calculating the intraclass correlation coefficient (ICC) for all manually evaluated summaries using the two-way-mixed model.

To identify which of the automatic evaluation metrics that best follows the manual evaluation, Pearson product-moment correlation coefficient (PPMCC) and Spearman's rank correlation coefficient (Spearman's rho) [16] were calculated between the normalized manual evaluation scores and each of the automatic evaluation scores (from Table 2).

5 Evaluation Results

The results from the automatic and the manual evaluations are presented in Table 2. The scores from the automatic evaluation are calculated from the average F-scores from the 156 test care episodes, while the results from the manual eval-

Table 2. Evaluation results, each column are ranked internally by score.

Rank	**ROUGE-N1** F-score	**ROUGE-N2** F-score	**ROUGE-L** F-score	**ROUGE-SU4** F-score	**Manual** (norm$_{max}$)
1	Original 1.0000	Original 1.0000	Original 1.0000	Original 1.0000	Original 1.0000
2	Oracle 0.7964	Oracle 0.7073	Oracle 0.7916	Oracle 0.6850	SumMet2 0.6738
3	SumMet2 0.6700	SumMet2 0.5922	SumMet2 0.6849	SumMet2 0.5841	Oracle 0.6616
4	SumMet5 0.5957	SumMet5 0.4838	SumMet5 0.5902	SumMet5 0.4723	SumMet5 0.6419
5	SumMet1 0.4785	SumMet1 0.3293	SumMet1 0.4717	SumMet1 0.3115	SumMet3 0.5326
6	SumMet4 0.3790	SumMet4 0.2363	SumMet4 0.3725	SumMet4 0.2297	SumMet1 0.5167
7	Random 0.3781	Random 0.2094	Random 0.3695	SumMet3 0.2013	Random 0.5161
8	SumMet3 0.3582	SumMet3 0.2041	SumMet3 0.3521	Random 0.2001	SumMet4 0.5016

uation are the average scores from a subset of five care episodes (also included in the automatic evaluation), all evaluated by three domain experts. The latter scores have all been normalized by dividing the scores of the highest ranking method. This was done in an attempt to scale these scores to the F-scores from the automatic evaluation.

All automatic metrics and the manual evaluation agreed in terms of what summarization method belongs to the top three, and the bottom four. When calculating significance levels for the automatic evaluations for the five highest ranked methods, the differences were always significant. However, in several cases the differences between the three lowest ranked methods, those being SumMet4, SumMet3 and Random, were not statistically significant. These results are in agreement with the fact that all the evaluation measures agreed on which the five best performing methods were, whereas the three worst methods are equally bad, all performing on a level that does not significantly differ from the approach of picking the sentences for the summary randomly.

For the manual evaluation, the original discharge summaries scored significantly higher than any of the generated summaries. Furthermore, the summaries produced by the Oracle, SumMet2 and SumMet5 methods were evaluated to be significantly better than those produced by the four other methods. Thus, the manual evaluation divided the automated summarization methods into two distinct groups, one group that produced seemingly meaningful summaries, and the other that did not work significantly better than the Random method. The divi-

[8] The original discharge summary is of course not a product of any of the summarization methods.

Table 3. PPMCC and Spearman's rho results, indicating how the scoring by the automatic evaluation metrics correlates with the normalized manual evaluation scores.

Evaluation metric	PPMCC (p-values)	Spearman's rho (p-values)
ROUGE-N1	0.9293 (0.00083)	0.8095 (0.01490)
ROUGE-N2	0.9435 (0.00043)	0.8095 (0.01490)
ROUGE-L	0.9291 (0.00084)	0.8095 (0.01490)
ROUGE-SU4	**0.9510** (0.00028)	**0.8571** (0.00653)

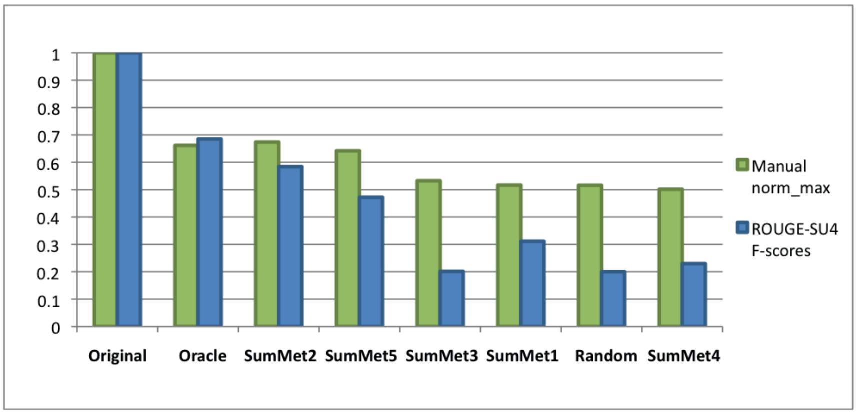


Fig. 2. Graph showing the evarage manual scores (norm$_{max}$), calculated from five care episodes (evaluated by three domain experts), and average F-scores by ROUGE-SU4, calculated from 156 care episodes. The order, from left to right, is sorted according to descending manual scores.

sion closely agrees with that of the automated evaluations, the difference being that in the manual evaluation also `SumMet1` ended up in the bottom group of badly performing summarization methods.

In Table 3 are the PPMCC and Spearman's rho results, indicating how each automatic evaluation metric correlates with the manual evaluation scores. Spearmans rho is a rank-correlation measure, so it does not find any difference between most of the measures, since they rank the methods in exactly the same order (except ROUGE-SU4, which has a single rank difference compared to others). In contrast, PPMCC measures the linear dependence taking into account magnitudes of the scores in addition to the ranks, so it observes some extra differences between the measures. This shows that ROUGE-SU4 has the best correlation compared to the manual evaluation. Figure 2 illustrates the normalized manual evaluation scores with the ROUGE-SU4 F-scores.

6 Discussion

All automatic evaluation metrics and the manual evaluation agreed that the top three automatic summarization methods significantly outperform the `Random` method. These methods are `SumMet2`, `SumMet5` and `Oracle`. Thus we can with a certain confidence assume that this reflects the actual performance of the utilized summarization methods. `Oracle` is of course not a proper method, given that it is "cheating", but it is a good indicator for what the upper performance limit it.

The reliability of the manual evaluation is naturally rather weak, given that only five care episodes were evaluated by the three evaluators. The findings of the manual evaluation are not generalizable due to the small number of care episodes evaluated. Therefore, more manual evaluation is needed to confirm these findings.

On a more general level, the results indicate that the utilized approach – using the original discharge summaries as gold summaries – is a seemingly viable approach. This also means that the same evaluation framework can potentially be transferred to clinical text in other countries and languages who follow a similar hospital practice as in Finland.

The manual evaluation results show that the various summarization methods are less discriminated in terms of scores when compared to the automatic evaluation scores. We believe that this is partly to blame for the small evaluation set these scores are based on, and also because of the evaluation tool that was utilized. For these reasons we are still looking into ways to improve the manual evaluation tool before we conduct further manual evaluation. It is interesting to see that `SumMet2` is considered to outperform the `Oracle` method, according to the manual evaluators.

6.1 Lessons learned from the manual evaluation

The agreement between the evaluators in the manual evaluation was calculated with the 40 different summaries evaluated by each of the three evaluators. The ICC value for the absolute agreement was 0,68 (95% CI 0,247-0,853, p<0,001). There is no definite limit in the literature on how to interpret ICC values, but there are guidelines that suggest that values below 0.4 are poor, values from 0.4 to 0.59 are fair, values from 0.6 to 0.74 are good and values from 0.75 to 1.0 are excellent in terms of the level of interrater agreement [17]. The agreement between the evaluators in the manual evaluation was good, based on these suggested limits. This means that there were some differences between the evaluations conducted by the three evaluators, which indicates that the criteria used in the 30-item manual evaluation tool allowed this variance, and therefore, the tool with its items need further development. Another aspect is that the evaluators would need more training concerning the use of the criteria and possibly more strict guidelines.

Furthermore, the evaluators reported that the manual evaluation was difficult and very time consuming, due to the numerous and detailed items in the manual evaluation tool. They also reported that the judgement process necessary when evaluating the summaries was too demanding. It became obvious that several of the items in the evaluation tool were too specifically targeting structured information. This means information that is already identified and classified in the health record system, which does not need to be present in the unstructured free text from where the summaries are generated. Examples are 'Care period', 'Care place' and 'Sender'. In the future, a shorter tool, i.e. less items, with stricter criteria and more detailed guidelines for the evaluators is needed. One important property of such a new tool would be, when used by the human evaluators, that good and bad summaries (i.e. summarization methods) are properly discriminated in terms of scoring.

7 Conclusion and Future Work

In this paper we have presented the results from automatic and manual evaluation of seven different methods for automatically generating clinical text summaries. The summary documents was composed from the free text of the clinical daily notes written by physicians related to patient care.

Seven automatic summarization methods were evaluated. For the automatic evaluation the corresponding original discharge summaries were used as gold summaries for doing the automatic evaluation. Among these summarization methods were the control-methods `Random` and `Oracle`. Four ROUGE metrics were used for the automatic evaluation, ROUGE-N1, ROUGE-N2, ROUGE-L and ROUGE-SU4.

The evaluation results seem promising in that the ranking order of the various summarization methods, ranked by all the automatic evaluation metrics, correspond well with that of the manual evaluation. These preliminary results indicates that the utilized automatic evaluation setup can be used as an automated and reliable way to rank clinical summarization methods internally in terms of their performance.

More manual evaluation, on a larger sample of care episodes, is needed to confirm the findings in this study. In this context, more research is needed to make a manual evaluation tool that better discriminates good from bad summaries, as well as being easier to use by evaluators. This preliminary work provided us good insight and ideas about how to further develop the manual evaluation tool, suited for a larger-scale manual evaluation.

As future work, we also plan to conduct so called *extrinsic evaluation* of the summarization methods, meaning that the various summaries produced by the system are evaluated in terms of their impact on clinical work.

8 Acknowledgements

This study was financially supported partly by the Research Council of Norway through the EviCare project (NFR project no. 193022), the Turku University Hospital (EVO 2014), and the Academy of Finland (project no. 140323). The

study is a part of the research projects of the Ikitik consortium (`http://www.ikitik.fi`).

References

1. Hall, A., Walton, G.: Information overload within the health care system: a literature review. Health Information & Libraries Journal **21**(2) (2004) 102–108
2. Van Vleck, T.T., Stein, D.M., Stetson, P.D., Johnson, S.B.: Assessing data relevance for automated generation of a clinical summary. In: AMIA Annual Symposium Proceedings. Volume 2007., American Medical Informatics Association (2007) 761
3. Kripalani, S., LeFevre, F., Phillips, C.O., Williams, M.V., Basaviah, P., Baker, D.W.: Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. Jama **297**(8) (2007) 831–841
4. Feblowitz, J.C., Wright, A., Singh, H., Samal, L., Sittig, D.F.: Summarization of clinical information: A conceptual model. Journal of biomedical informatics **44**(4) (2011) 688–699
5. Roque, F.S., Slaughter, L., Tkatšenko, A.: A comparison of several key information visualization systems for secondary use of electronic health record content. In: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, Association for Computational Linguistics (2010) 76–83
6. Liu, S.: Experiences and reflections on text summarization tools. International Journal of Computational Intelligence Systems **2**(3) (2009) 202218
7. Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In: Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization. Volume 4 of NAACL-ANLP-AutoSum '00., Association for Computational Linguistics (2000) 21–30
8. Afantenos, S., Karkaletsis, V., Stamatopoulos, P.: Summarization from medical documents: a survey. Artificial intelligence in medicine **33**(2) (2005) 157–177
9. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In Marie-Francine Moens, S.S., ed.: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain, Association for Computational Linguistics (July 2004) 74–81
10. Koskenniemi, K.: Two-level model for morphological analysis. In Bundy, A., ed.: Proceedings of the 8th International Joint Conference on Artificial Intelligence. Karlsruhe, FRG, August 1983, William Kaufmann (1983) 683–685
11. Karlsson, F.: Constraint grammar as a framework for parsing running text. In: Proceedings of the 13th Conference on Computational Linguistics - Volume 3. COLING '90, Stroudsburg, PA, USA, Association for Computational Linguistics (1990) 168–173
12. Nenkova, A., McKeown, K.: Automatic summarization. Foundations and Trends in Information Retrieval **5**(23) (2011) 103–233
13. Dang, H.T., Owczarzak, K.: Overview of the tac 2008 update summarization task. In: Proceedings of text analysis conference. (2008) 1–16
14. Lissauer, T., Paterson, C., Simons, A., Beard, R.: Evaluation of computer generated neonatal discharge summaries. Archives of disease in childhood **66**(4 Spec No) (1991) 433–436

15. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics **1** (1945) 80–83
16. Lehman, A.: JMP for basic univariate and multivariate statistics: a step-by-step guide. SAS Institute (2005)
17. Cicchetti, D.V.: Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological assessment **6**(4) (1994) 284

Comparison of automatic summarisation methods for clinical free text notes

Moen, Hans; Peltonen, Laura-Maria; Heimonen, Juho; Airola, Antti; Pahikkala, Tapio; Salakoski, Tapio, and Salanterä, Sanna

Artificial Intelligence in Medicine, 67:25–37, 2016.

Paper E

Artificial Intelligence in Medicine 67 (2016) 25–37

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Comparison of automatic summarisation methods for clinical free text notes

Hans Moen[a,b,c,*], Laura-Maria Peltonen[c,d], Juho Heimonen[b,e], Antti Airola[b], Tapio Pahikkala[b,e], Tapio Salakoski[b,e], Sanna Salanterä[c,d]

[a] Department of Computer and Information Science, Norwegian University of Science and Technology, Sem Saelands vei 9, 7491 Trondheim, Norway
[b] Department of Information Technology, University of Turku, Joukahaisenkatu 3–5, 20520 Turku, Finland
[c] Department of Nursing Science, University of Turku, Lemminkäisenkatu 1, 20520 Turku, Finland
[d] Turku University Hospital, Kiinamyllynkatu 4–8, 20521 Turku, Finland
[e] Turku Centre for Computer Science (TUCS), Joukahaisenkatu 3–5, 20520 Turku, Finland

ARTICLE INFO

Article history:
Received 4 May 2015
Received in revised form
14 December 2015
Accepted 5 January 2016

Keywords:
Automatic text summarisation
Summarisation evaluation
Distributional semantics
Word space models
Clinical text processing
Electronic health records

ABSTRACT

Objective: A major source of information available in electronic health record (EHR) systems are the clinical free text notes documenting patient care. Managing this information is time-consuming for clinicians. Automatic text summarisation could assist clinicians in obtaining an overview of the free text information in ongoing care episodes, as well as in writing final discharge summaries. We present a study of automated text summarisation of clinical notes. It looks to identify which methods are best suited for this task and whether it is possible to automatically evaluate the quality differences of summaries produced by different methods in an efficient and reliable way.
Methods and materials: The study is based on material consisting of 66,884 care episodes from EHRs of heart patients admitted to a university hospital in Finland between 2005 and 2009. We present novel extractive text summarisation methods for summarising the free text content of care episodes. Most of these methods rely on word space models constructed using distributional semantic modelling. The summarisation effectiveness is evaluated using an experimental automatic evaluation approach incorporating well-known ROUGE measures. We also developed a manual evaluation scheme to perform a meta-evaluation on the ROUGE measures to see if they reflect the opinions of health care professionals.
Results: The agreement between the human evaluators is good (ICC = 0.74, $p < 0.001$), demonstrating the stability of the proposed manual evaluation method. Furthermore, the correlation between the manual and automated evaluations are high (> 0.90 Spearman's rho). Three of the presented summarisation methods ('Composite', 'Case-Based' and 'Translate') significantly outperform the other methods for all ROUGE measures ($p < 0.05$, Wilcoxon signed-rank test and Bonferroni correction).
Conclusion: The results indicate the feasibility of the automated summarisation of care episodes. Moreover, the high correlation between manual and automated evaluations suggests that the less labour-intensive automated evaluations can be used as a proxy for human evaluations when developing summarisation methods. This is of significant practical value for summarisation method development, because manual evaluation cannot be afforded for every variation of the summarisation methods. Instead, one can resort to automatic evaluation during the method development process.

1. Introduction

1.1. Background

Information overload in the health sector is becoming an increasing problem for clinicians [1,2]. They have to read masses of text (such as clinical notes, guidelines and scientific literature) to satisfy their information needs. Lack of time and resources to do this properly causes problems such as errors, frustration, inefficiency and communication failures [3].

* Corresponding author at: Department of Computer and Information Science, Norwegian University of Science and Technology, Sem Saelands vei 9, 7491 Trondheim, Norway. Tel.: +47 97502647.
E-mail addresses: hans.moen@idi.ntnu.no (H. Moen), lmemur@utu.fi (L.-M. Peltonen), juaheim@utu.fi (J. Heimonen), ajairo@utu.fi (A. Airola), aatapa@utu.fi (T. Pahikkala), tapio.salakoski@utu.fi (T. Salakoski), sansala@utu.fi (S. Salanterä).

http://dx.doi.org/10.1016/j.artmed.2016.01.003

The contents of electronic health record (EHR) systems are largely composed of clinical notes (or clinical narratives) in the form of unstructured and unclassified text. The clinical notes written during a single care episode, i.e. a stay in a hospital, can be quite voluminous, especially for patients suffering from more complex and long-term health problems. Knowing the medical history of a patient is vital for a clinician, but scanning through clinical notes consumes precious time that could be better spent treating the patient.

Automatic summarisation of the free text content in care episodes could assist clinicians in at least two ways. First, it could provide an (indicative) overview of the documentation of a care episode. Together with structured data (such as laboratory test results, images, diagnostic codes and personal information) it could help clinicians to familiarise themselves with the content of the care episode and the patient's problems, which is particularly useful if the information is needed urgently. Second, it may help in writing a discharge summary of a care episode. Discharge summaries are crucial in communication between different health care providers and they are needed to ensure continuity of care. However, there are a number of challenges with them, ranging from being produced late to having insufficient information. For example, Kripalani et al. [4] showed that discharge summaries exchanged between hospitals and primary care physicians are often lacking some of the expected information, such as that related to treatment progression, counselling and follow-up proposals. Computer-assisted discharge summaries and standardised templates are measures for improving transfer time and the quality of discharge information between hospitals and primary care physicians [4]. The utilisation of automatic text summarisation could improve the timeliness and quality of discharge summaries even further.

Central to this work is the focus on *resource-lean*[1] and *language-independent* methods. Such methods are important for languages such as Finnish, for which no major manually constructed lexical resources suited for the comprehensive semantic analysis of clinical text are available.

1.2. Related work

This study focuses on the extraction-based summarisation approach, in which the summary is generated by selecting a subset of sentences from the relevant text. This approach is viable because a sizeable portion of a clinical text summary is created by copying or deriving information from clinical notes [2,5–7]. See [8,9] for example, for more information on extraction-based text summarisation.

A central issue in extraction-based summarisation is how to determine what the most relevant content to be included in a summary is. Common techniques of extraction-based summarisation include topic-based sentence extraction [10,11], where the relevance of a sentence is computed with respect to one or more topics of interest; and centrality-based sentence extraction [12,13], where the sentences that are the most (strongly) associated with others are selected on the assumption that they constitute the best coverage of the documents. In order to avoid including redundant information, it is common to apply the maximal marginal relevance criterion [10] or similar techniques that take sentence overlap into account. Purely statistical (data-driven) approaches to text summarisation are often referred to as 'knowledge poor', whereas those using knowledge resources are considered 'knowledge rich'. The latter could, for example, include the use of an ontology that models medical and clinical concepts as well as their relationships.

In their recent review, Mishra et al. [14] indicated that there is a growing interest in knowledge-rich approaches in the biomedical domain, coinciding with the increased availability of comprehensive lexical resources, such as the WordNet ontology [15] and the UMLS compendium [16] (including SNOMED-CT [17], ICD [18] and MeSH [19]). There are several language tools that rely on these resources, such as MetaMap [20], cTAKES [21] and SemRep [22]. Other commonly used resource types include, for example, annotated corpora designed for machine learning (ML) algorithms (see e.g. [6,23]). However, one disadvantage to approaches that rely on manually constructed resources is that they are often not applicable across domains or languages [24,25]. WordNet and UMLS (SNOMED-CT, in particular), for example, are only available in a few languages. The cost of adapting existing resources to new languages, domains or tasks, or constructing new resources, is often high.

The use of *distributional semantic methods* represents a resource-light approach to capturing terminology in clinical texts [26–31]. These methods rely on the *distributional hypothesis* [32] for constructing *distributional semantic models* from word co-occurrence statistics in an unsupervised manner, typically using a very large corpus of unannotated text. The aim is to model similarities, or relatedness, between linguistic items (e.g. words) in a way that reflects their relative *semantic meaning*. Distributional semantic models representing word-level semantic similarity are often referred to as *word space models* (WSMs). In a WSM, a *word context vector* is created for each unique word in the underlying corpus. Further, each context vector represents a point in the 'word space' and their internal distances reflect their semantic similarities. Similarities between context vectors are then calculated to quantify the semantic similarity as a numeric value (for example, using the *cosine similarity function*). Popular techniques and frameworks for constructing WSMs include latent semantic analysis (LSA) [33], random indexing (RI) [34] and Word2vec (W2V) [35]. The domain-specificity of the corpus used for constructing the model has been shown to be important for the usefulness of semantic similarities to the intended task [27]. Distributional semantic models in various forms have been extensively used in text summarisation, e.g. [9,13,36].

To the best of our knowledge, the task of automatically generating textual summaries from clinical notes has been pursued by relatively few researchers, which is also evident in recent reviews and related works, for example [14,24]. We have identified several pieces of work focusing on the task of automatically generating textual summaries from unstructured clinical notes. Liu [37] used the MEAD summarisation toolkit. Van Vleck et al. [2] performed structured interviews to identify and classify phrases that clinicians considered relevant to explaining a patient's history. Meng et al. [6] used an annotated training corpus together with tailored semantic patterns to determine what information should be repeated in a new clinical note or summary. Velupillai and Kvist [23] focused on recognising diagnostic statements in clinical text, learning from an annotated training corpus, and classifying these based on the level of certainty they have in them. Extracted diagnostic statements are then used to produce a text summary. Others have worked on more conceptual models for understanding and supporting the generation of information summaries in the clinical domain [38,39].

The evaluation of computer-generated summaries is typically performed by comparing the generated summary with a *gold standard* (or reference summary), which represents the ideal manually constructed summary or summaries. The *ROUGE*[2] *evaluation*

[1] We are striving towards using as little manual labour as possible.

[2] ROUGE is short for *recall oriented understudy for gisting evaluation*.

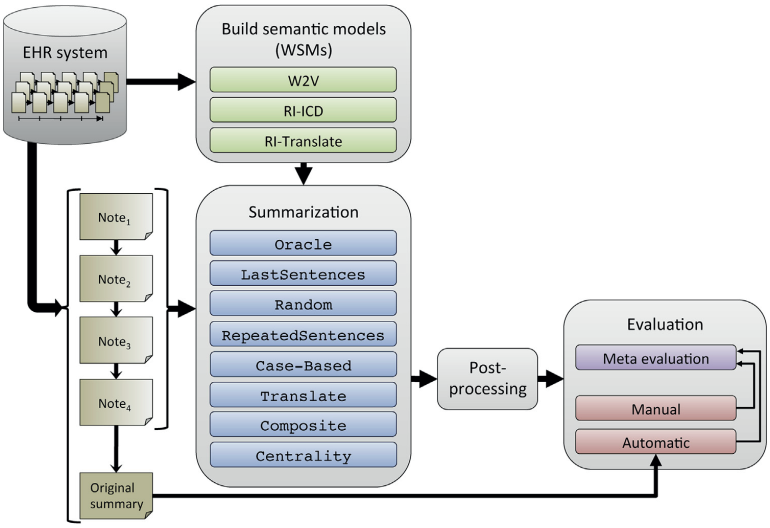


Fig. 1. Experimental set-up. The figure shows how the experiment was conducted.

package [40] has become a de facto evaluation metric in text summarisation. The required gold standard summaries are costly to create given the manual work required. This is particularly the case in specialised domains where domain experts are required. Lissauer et al. [3] evaluated computer-generated discharge summaries from neonate reports by manually comparing them to dictated summaries, as well as analysing them to see whether they contained the required information according to guidelines. Liu [37] performed automatic evaluation of computer-generated summaries of clinical nursing notes by using the original discharge reports as gold standard summaries. Moen et al. [41] applied both manual and automatic evaluation to the assessment of the reliability of automatic evaluation; the manual evaluation was performed by domain experts and the automatic evaluation was performed by using ROUGE to calculate the similarity between the computer-generated summaries and the original discharge summaries (produced by clinicians).

1.3. Objectives

The main contributions of this study can be summarised as follows:

- Proposal and implementation of four novel automatic summarisation methods designed for summarising the free text in care episodes;
- Proposal and implementation of a methodology for conducting the manual evaluation of automatically generated care episode summaries;
- Empirical analysis of automatic evaluation measures through comparison with manual evaluations;
- Performance assessment of the four novel automatic summarisation methods along with four baseline methods.

The data used in this study is Finnish clinical text, but since the applied methods are language-independent, the contributions should also be relevant to other languages. The overall set-up is illustrated in Fig. 1.

2. Material and methods

2.1. Data

The data set used in this study consists of EHRs from patients with any type of heart-related problem that were admitted to a single university hospital in Finland between 2005 and 2009. Of these, the clinical notes written by physicians on the various wards that the patients visited were used. However, notes written by nurses were not included. Fig. 2 shows an example of a clinical note.

Ethical approval for the research was obtained from the ethics committee of the hospital district (17.2.2009 §67) and permission to conduct the research was obtained from the medical director of the hospital district (2/2009). The total set consisted of 66,884 care episodes, which amounts to 398,040 notes and 64 million words[3] in total. This full set, minus 308 care episodes reserved for optimisation and evaluation (see below), were used for constructing the WSMs (see Section 2.2).

The notes are mostly unstructured, consisting of clinical free text in Finnish. Various subheadings do occur in the clinical notes, but these are not standardised, structured, or uniquely recognised in our corpus. Thus, we treat these in the same way as the rest of the text. Some of the sentences are, according to the EHR system, considered to be metadata — such as names of the authors, dates, wards and so on. We treat the free text sentences and the meta-text sentences as two separate text types, so these are not mixed in the sense that they cannot belong to the same 'sentence topic' clusters, which are described in Section 2.3.

Each care episode has been manually labelled with ICD-10 codes by clinicians as a part of the original care process. These are normally applied at the end of the patient's hospital stay, or after they are discharged from hospital. Care episodes commonly have one primary ICD-10 code attached to them, and a number of optional secondary codes. In this study the primary ICD-10 code is used in constructing the RI-ICD WSM, as described in Section 2.2.

[3] A word is defined as tokens containing at least one letter.

English translation:
61-years old female with Crohn's disease. Attended cycling event in Salo, flu prirorlry. Arfter cycling, experienced breathing difficulties and went to the emergency department and elevated herart enzymes and incompensation were found. Was admitted to the ICU for care of incompensation and pneumonia. In UKG 2.6. ef 30%. In coronary angiography, significrant stenoses in RCA, LCX and mrarin. Trordray, elective quadrurple bypass LITA-LAD, Ao-LOM-LPL and Ao-RBD, in which goord flow. Pre.op. the posterior wall of the left ventricle and the septum contract lamely, ef about 35%, mitral valve 1-2/4 leak. Aortic cross-clamp time 1 h 32 min. Post.op. ef over 40%. On basis of the UKG-finding pre.op. Simdax-infusion was initiated. On arrival to ICU, haemodynamics was stable, norepinephrine administered. Cardiac index 3,2. Warming-up and weaning in ventilator.

Finnish original:
61-vuotias nainen jolla Crohnin tauti. Salossa ollessaan osallistunut pyöräilytapahtumaan, edelträvrästi flunssaa. Pyöräilyn jrälkeen hengitysvaikeuksien takia TYKSin ensiapuun ja todettu sydränentsyymit kohonneiksi ja inkompensaatiota. Otettu teho-osastolle inkompensaation ja pneumonian hoitoon. UKG:ssa 2.6. ef 30%. Koronaariangiossa merkitträvrät stenoosit RCA:ssa, LCX:ssa ja prärärungossa. Tränrärän elektiivinen neljrän suonen ohitus LITA-LAD, Ao-LOM-LPL ja Ao-RBD, joihin hyvrät virtaukset. Pre.op. vasemman kammion takaseinrä ja septum supistuvat vaisusti, ef noin 35%, mitraaliläpässä 1-2/4 vuoto. Aortan sulkuaika 1 t 32 min. Post.op. ef yli 40%. UKG-löydöksen perusteella potilaalle aloitettu jo pre.op. Simdax-infuusio. Teho-osastolle saapuessa hemodynamiikka stabiilia, noradrenaliini menossa. Cardiac index 3,2. Lämmitys ja vieroitus respiraattorissa.

Fig. 2. Example of a clinical note. This is a fake case originally created in Finnish by domain experts, then translated into English. Common misspellings are included intentionally.

In the presented experiments, we restrict our evaluation to the care episodes which have the primary ICD code I25 — chronic ischemic heart disease, including subcodes (I25.0, I25.1, etc.). As a further restriction, to justify the use of text summarisation, we consider only care episodes consisting of seven or more clinical notes written by physician. In order to guarantee that the methods are tested on independent test data that is not used for developing the summarisation models, the 308 care episodes are split into two subsets:

- A *summarisation optimisation set*, consisting of 152 care episodes, used for optimising parameters related to the summarisation methods.
- A *summarisation evaluation set*, consisting of 156 care episodes, used for evaluation in the conducted experiments. This is further split into two subsets of 20 and 136 care episodes, the former subset being evaluated in Experiment 1 and the latter in Experiment 2.

This splitting is performed according to the year in which the care episodes were carried out.

2.2. Word space models used for sentence similarity and summarisation

We use a method based on the RI technique, utilising the ICD-10 codes attached to care episodes (RI-ICD) to construct a WSM for the purpose of calculating (semantic) similarity between care episodes. We also use the RI technique in constructing a 'cross-text' translation model (RI-Translate), and the W2V method is used to construct a WSM for the purpose of calculating sentence-to-sentence similarities, as well as sentence-to-document[4] similarities. The cosine similarity metric is used to calculate vector similarities.

Random indexing and RI-ICD

RI [34] is a technique for constructing a (pre-)compressed WSM with a fixed dimensionality, done in an incremental fashion. This is achieved by initiating *index vectors* for each unique word in the corpus. An index vector is a vector of a fixed dimensionality, containing mainly zeros along with a small number of randomly assigned non-zeros, typically 1 or −1. During training, context vectors for words are constructed by adding index vectors to them. In this way, the dimensionality of the context vectors remains constant. In this work we use a version of RI where context features are based on the ICD-10 code classifications of care episodes. We have called this RI-ICD,[5] previously introduced in [42].

RI-translate

Another RI-based method used here is one intended for cross-lingual translation purposes, described in [43]. We refer to it as RI-Translate. The method constructs a bilingual WSM that connects words in the *source language* (SL) to their translated counterparts in the *target language* (TL). In practice, we operate with two models, one for SL and one for TL, where both belong to the same semantic space in that they are constructed with the same set of *index vectors*. For training, pre-aligned translation pairs (in this case, aligned sentences) connecting the SL to the TL are used as training instances. The training takes place as follows: for each translation pair (SL–TL), a unique index vector is generated and added to the corresponding context vectors for words in the SL and TL models. This will result

[4] Documents are in this case the clinical notes.

[5] A vector dimensionality of 800 was used, and the number of non-zeros for the index vectors was set to four.

Table 1
Top ten most similar words according the W2V-based WSM for the query words 'pain' and 'foot', together with the corresponding cosine similarity scores. The words have been translated from Finnish to English.

Pain	(kipu)	*cossim*	Foot	(jalka)	*cossim*
Pain sensation	(kiputuntemus)	0.5097	Lower limb	(alaraaja)	0.5905
Ache	(särky)	0.4835	Ankle	(nilkka)	0.3731
Pain symptom	(kipuoire)	0.4173	Limb	(raaja)	0.3454
Chest pain	(rintakipu)	0.4042	Shin	(sääri)	0.3405
Dull pain	(jomotus)	0.4000	Peripheral	(periferisia)	0.3112
Backpain	(selkäkipu)	0.3953	Callus	(känsä)	0.3059
Pain seizure/attack	(kipukohtaus)	0.3904	Top of the foot	(jalkapöytä)	0.2909
Pain status	(kiputila)	0.3685	Upper limb	(yläraaja)	0.2879
Abdominal pain	(vatsakipu)	0.3653	Peripheral	(perifer)	0.2875
Discomfort	(vaiva)	0.3614	In lower limb	(alaraajassa)	0.2707

in high cosine similarity between words in the SL model and the TL model that have often occurred in the same translation pairs.

When querying the system, the context vector(s) corresponding to the query in the SL model is used as the query. This query vector is then matched against the units in the TL model, using cosine similarity, in order to find the most likely translation(s). This method is used for summarisation purposes in the `Translate` method, as described below.

Word2vec

Word2vec [35] is a framework for constructing WSMs using a neural network. In this work we utilise the W2V *CBOW* architecture.[6] Table 1 shows an example of how the W2V-based model captures semantic similarity relations. We use this model in the various summarisation methods for computing sentence-to-sentence similarities and sentence-to-document similarities.

Composing sentence and document vectors

Sentence context vectors are composed by normalising and summing (pointwise summation) the constituent word context vectors weighted by their sentence term frequency multiplied by their global inverted sentence frequency (TF*ISF). A similar approach is used for constructing context vectors representing clinical notes, but here weighting is based on term frequency multiplied with global inverted document frequency (TF*IDF) [44], where each clinical note is considered as a document.

2.3. Summarisation methods

We evaluated eight different summarisation methods. `Oracle` is an (unrealistic) reference method that has access to the true/original discharge summary when selecting sentences to extract, providing an upper boundary to how well an extractive summarisation method can work for our data. `LastSentences` and `Random` are simple reference methods that a successful summarisation method should be able to outperform. `Centrality` is a standard baseline approach that is commonly used in the field, and the remaining four methods, called `RepeatedSentences`, `Case-Based`, `Translate` and `Composite`, are proposed methods developed specifically with the clinical domain in mind.

For each care episode, the length of the summary generated by each summarisation method is set to have a fixed size equal to the word count of the accompanying discharge summary (i.e. the 'gold standard summary' for the care episode). Sentences are iteratively extracted from the clinical notes until the total word count becomes equal to or just exceeds the word count of the discharge summary. Therefore, generated summaries can have a word count equal to the discharge summary, or exceed this limit by a subset of the words in the last extracted sentence. This way of dynamically selecting the summarisation length is mainly done to enable the calculation of the automatic evaluation scores (*F*-score), described in Section 2.4.2, which assumes equal length of the target summary (the summary being evaluated) and the gold standard summary.

In the summarisation methods `RepeatedSentences`, `Case-Based`, `Translate` and `Composite`, a type of *topic clustering* is used to perform redundancy reduction. We found that each sentence is typically informative, self-sustaining in information content, and independent of other sentences within a single note. All sentences are first clustered into unlabelled *sentence topics* in an unsupervised way using the W2V model. A cosine similarity threshold λ, optimised on the *summarisation optimisation set*, is used for determining whether two sentences can be considered similar or not — whether or not they belong to the same topic based on their cosine similarity. The underlying approach is somewhat comparable to how similar paragraphs are detected and merged in McKeown et al. [45] with the aim of reducing sentence redundancy. Since we know when each sentence was written, and if we assume that we are able to cluster sentences that discuss the same topic across clinical notes (e.g. the state of a patient's pain), we can also assume that the latest written sentence belonging to a topic is the most representative of the latest information concerning that topic. Therefore, we allow the latest written sentence belonging to each topic cluster to be the representative sentence. In an attempt to most effectively model sentence topic clusters, the clustering approach is done as follows: first, we assume that all sentences in the first clinical note of a care episode belong to different topics (see Note$_1$ in Fig. 3 for an illustration). Then we iterate through the care episode, from the first to the last written note, and assign each sentence to either existing topics ($cos\ sim \geq \lambda$) or new topics ($cos\ sim < \lambda$) based on their cosine similarities in relation to λ. In the utilised similarity comparison, the latest added sentence of a topic always represents that topic. A sentence can only belong to one topic, so if a sentence is similar to two or more topics, i.e. $cos\ sim > \lambda$, the sentence is assigned to the most similar topic.[7]

Original discharge summary

The original discharge summary is a text written by a clinician, typically a physician, to summarise a care episode. These summaries are thus written at the end of each care episode, and often contain extracts from the accompanying clinical notes. They also typically contain a certain amount of as-yet undocumented information which focuses on follow-up treatment, and are meant for the receiving ward (if any) or the primary care sector.

[6] For W2V a window size of 5 + 5 and a dimensionality of 800 was used.

[7] In the unlikely event that the cosine similarities between a sentence and two or more topics are the same, the sentence is assigned to a random topic among these.

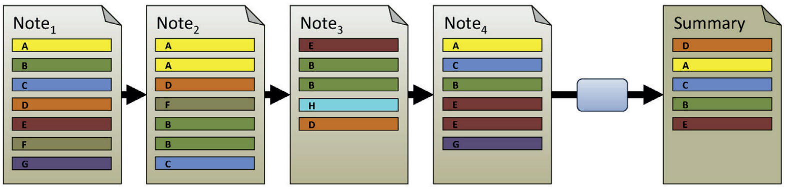


Fig. 3. Summarisation method RepeatedSentences. The example illustrates how summaries are produced by sentence topic clustering and topic scoring. The highest scoring topics from highest to lowest are B, E, A, C, D, G, F and H. In the generated summary displayed here, the topics are sorted by the post-processing step, and the three lowest scoring topics, G, F and H, are excluded.

In this work, the original discharge summary serves as the gold standard summary for its accompanying care episode in the automatic evaluation approach that is used (see Section 2.4.2). In addition, some of the summarisation methods use these in their underlying training (`Translate`) or in the summarisation phase (`Case-Based`). Naturally, for a care episode that is to be summarised, the accompanying original discharge summary will not be available to the summarisation system in a realistic scenario.

Summarisation method: `Oracle`

This is a control method that has access to the original discharge summary during the summarisation process. It optimises the ROUGE-N2 *F*-scores (see Section 2.4.2) for the generated summary according to the gold summary, using a greedy search strategy. That is, the method extracts sentences one by one from the clinical notes until it reaches the length threshold, always picking sentences that result in the highest possible ROUGE-N2 score. This method is cheating, since it has access to the original discharge summary in the summarisation process. Still, it represents the upper limit for what is achievable in terms of ROUGE-N2 scores for an extraction-based summary.

Summarisation method: `LastSentences`

The latest written clinical note at any point should supposedly represent the current state of the patient. By selecting the latest information found in the last or latest written information, one can intuitively assume that this information is important in a (discharge) summary. In this method, the summary is simply constructed from the *N* last written sentences during the care episode, where *N* is the number of sentences needed to reach the length threshold. Intuitively, this represents a strong baseline.

Summarisation method: `Random`

This baseline method constructs summaries by randomly selecting sentences from the care episode until the length threshold is reached. It provides a lower boundary to the performance, which any meaningful summarisation approach should aim to significantly outperform.

Summarisation method: `RepeatedSentences`

Meng et al. [6] argue that information being repeated across clinical notes is an indicator of its relevance with respect to inclusion in subsequent notes in the sequence. The use of time features is also explored by Lim et al. [46] in the task of multi-document summarisation of news article documents. The underlying hypothesis for the `RepeatedSentences` method is that information that is repeated in multiple clinical notes throughout a care episode, with the emphasis on when it was written, is the most important information to include in a summary. Features from the initial sentence topic clustering step are used for scoring. A topic is assigned a score based on the sum of the *order* of when, in the care episode, its underlying sentences were written. For example, if a topic contains sentences from clinical note numbers 3, 5 and 6 (numbered relative to the dates they were written), the topic score becomes 14. The *N* highest scoring sentence topics (i.e. their representative sentences) are included in the final summary. The `RepeatedSentences` summarisation method is illustrated in Fig. 3.

Summarisation method: `Case-Based`

`Case-Based`, or 'case-based summarisation', is here an analogy to *case-based reasoning* (CBR) [47] which performs a type of *textual* case-based reasoning (TCBR) [48]. CBR involves retrieving existing or older 'cases' with similar content as the target problem, and then reusing the solution of the retrieved case (or cases) to solve the target problem. In a similar manner, this principle is applied here in text summarisation. The underlying hypothesis is that patients with (the most) similar care episodes (according to the documented text in their clinical notes) have similar content in their discharge summaries. The sentences from these discharge summaries are then treated as the central 'topics' for what to include in the summary. This is in line with *evidence-based practice* (EBP) in that relevant care episodes are identified and the information found there is relied upon as 'evidence' for what should be included in the summary.

Given a target care episode that is to be summarised, we first *retrieve* the top five most similar care episodes using information retrieval on care episode level (i.e. 'care episode retrieval'). For this the RI-ICD method is used (explained in [42], Section 4.1). Then we *reuse* these by iterating through each sentence in their discharge summaries. The representative/last sentences from each sentence topic in the target care episode (as described earlier) is then scored by their cosine similarity to each of these using the W2V model. Out of these, the *N* highest scoring sentences are included in the generated summary. Fig. 4 illustrates this using a modification of the 'CBR cycle' from [47].[8]

Summarisation method: `Translate`

Here we use the RI-Translate method, as explained in Section 2.2, for the purpose of text summarisation. Instead of translation between languages, it is used for 'cross-text-type translation' — translating from the text in clinical notes (care episodes without discharge summaries) to the text found in the discharge summaries, while limiting the translation candidates (i.e. sentences) to also come from the sentence topics in the clinical notes. The aim is thus to construct a type of translation system that can map sentences in clinical notes to the most probable sentences to be found in an accompanying discharge summary, based on translation statistics learnt from a large clinical corpus.

First, a translation, or cross-text-type WSM is constructed using the RI-Translate method. Here the source language (SL) consists of

[8] Fig. 4 also contains the steps *revise* and *retain*, but these are outside the scope of this work.

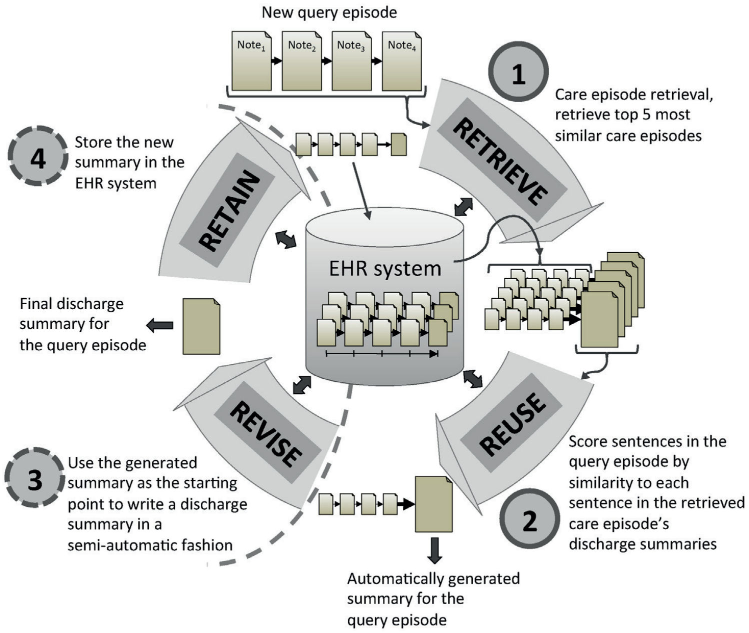

New query episode
Note1
Note2
Note3
Note4
1
Care episode retrieval, retrieve top 5 most similar care episodes
RETRIEVE
4
Store the new summary in the EHR system
RETAIN
EHR system
Final discharge summary for the query episode
REVISE
REUSE
3
Use the generated summary as the starting point to write a discharge summary in a semi-automatic fashion
2
Score sentences in the query episode by similarity to each sentence in the retrieved care episode's discharge summaries
Automatically generated summary for the query episode

Fig. 4. The `Case-Based` summarisation method illustrated as a 'CBR cycle', based on the CBR cycle introduced by Aamodt and Plaza [47]. The left side of the dashed line is not utilised in this work, but illustrates how the full CBR cycle can be used in a hospital setting.

the text in the clinical notes, while the discharge summaries constitute the text target language (TL). Training instances are rather coarse, as each care episode represents a single training instance. More precisely, for each care episode, the context vectors of the words in the underlying clinical notes (SL) and those in its accompanying discharge summary (TL) have a unique index vector added to them.

When summarising a care episode, each sentence (in the corresponding clinical notes, pre-clustered into sentence topics) has two sentence vectors constructed, one using word context vectors from the SL model, and the other using the TL model. Then, each sentence vector built with the SL model is iteratively used to query the system. Sentences represented by the TL model are then ranked by their overall *max* cosine similarity scores to these queries, and the top *N* sentences are included in the final summary.

Summarisation method: `Composite`

In this composite method, the sentence-scoring features from the methods `RepeatedSentences`, `Case-Based` and `Translate` are combined. We found that the best automatic evaluation scores (*F*-scores) from the *summarisation optimisation set* were achieved when the scores by `Case-Based` and `Translate` were kept as their initial cosine scores, while for `RepeatedSentences`, the sentence scores were first normalised by dividing on the *max* scoring sentence. This normalisation converts the scores to be within the same range as the cosine-based scores, ranging from 0 to 1. These three feature scores are then simply totalled to create the final feature score for each sentence. Finally the top *N* sentences are selected for the final summary.

Summarisation method: `Centrality`

The centrality (or centroid) principle is the most commonly relied on summarisation technique for many generic text types and domains. It is based on ranking sentences by how representative they are of the central information of the text that is to be summarised. In existing work, a range of methods have been used to compute sentence centrality in extraction-based summarisation. The *PageRank* method [49] has been extensively used for this purpose as a graph-based approach. We decided to base our implementation on the method presented in [13], which relies on a graph representation together with a WSM (RI). To construct the WSM, we used W2V instead of RI because preliminary testing indicated that this model performed better. Here, weighted PageRank for text is used, referred to as 'TextRank' [50]. Edges between nodes, i.e. between sentences, are weighted according to the precalculated sentence similarity using W2V. Each sentence also has an initial score similar to the cosine similarity between the sentence and the corresponding clinical note, represented as sentence vectors and document vectors, respectively. In addition, to adapt this approach to multiple documents, i.e. multiple clinical notes, we have extended this method with a sentence centrality ranking that works on multiple notes, in a similar way to how it is done in [51]. This is done by multiplying edge weights by one of two preset constants. Constant $\imath$ is multiplied with intra-note edge weights, i.e. edges going between sentences within the same clinical note; and inter-note edges are multiplied with the constant ϵ.[9]

2.3.1. Post-processing

Simple post-processing is applied to each summary for the purpose of rearranging the sentence order. Sentences are sorted according to the date they were written (i.e. using the date of the clinical note they belong to). Internal ranking between sentences

[9] In the experiment we used a PageRank α value of 0.90, $\imath$ was 0.3, while ϵ was 1.0.

from the same date is carried out according to internal sentence order. If two sentences from two different notes have the same date stamp, ranking is performed according to their chronological note IDs. Meta-sentences (described in Section 2.1) are placed first and rearranged internally.

2.4. Experiment and evaluation

The following two experiments were conducted:

- **Experiment 1**: The first experiment focuses on determining the reliability of the automatic evaluation. This is done by comparing how the manual and automatic evaluations (four ROUGE measures) correlate in terms of the relative rankings of the eight summarisation methods. Here, 20 care episodes (a subset of the 156 care episodes in the *summarisation evaluation set*) are evaluated both manually and automatically. Spearman's rank correlation coefficient (Spearman's rho) [52] is calculated between the average manual evaluation scores and the average scores for each of the automatic evaluation metrics for each summarisation method.
- **Experiment 2**: In the second experiment, the summarisation methods are tested on a larger evaluation set of 136 care episodes (the 156 care episodes in the *summarisation evaluation set* minus the 20 used in Experiment 1). The evaluation is performed in an automated manner using four ROUGE measures. The aim is primarily to determine which summarisation method produces the best summaries. In order to test whether the different scores achieved by the different summarisation methods were statistically significant, we performed the Wilcoxon signed-rank test [53] based on the scores from each ROUGE measure, for each summarisation method pair.

In both experiments, we use the same eight summarisation methods described to construct summaries for each care episode. The utilised WSMs are first constructed using the full corpus described in Section 2.1, minus the optimisation and evaluation sets mentioned.

A preliminary version of our evaluation set-up has been described in [41].[10] Our comparison of manual and automatic evaluation is similar to the analysis conducted by Chin-Yew Lin in [40] on English newswire data when introducing the ROUGE measures.

One goal is to see if our automatic evaluation set-up is reliable, given the uncertainties related to using the original discharge summaries as gold standard summaries. This is done by independently analysing whether or not there is a correlation between how human evaluators rank the performance of the summarisation methods and how automatic evaluation metrics rank these same summarisation methods. Furthermore, we aim to reliably establish which of the tested summarisation methods (and underlying features) perform best.

2.4.1. Manual evaluation

The manual evaluation is conducted by three domain experts in the clinical field: two physicians and one nurse, all professionals in specialised care and each with over five years' experience of working with patients suffering from heart-related health problems.

Table 2
Scheme used for the manual evaluation.

Evaluation criteria	Rating
Sender	yes = 1, no = 0
Reason for admission	yes = 1, no = 0
Long-term diagnosis	yes = 1, no = 0
Procedures (e.g. operation, angioplasty)	yes = 1, no = 0
Tests (e.g. lab-test, X-ray, EKG)	yes = 1, no = 0
Medication	yes = 1, no = 0
Health status at discharge	yes = 1, no = 0
Plans for the future	yes = 1, no = 0
Readability: how good is the flow of the text?	0.0–1.0, 0.0 = bad to 1.0 = excellent
Readability: how good is the content of the summary?	0.0–1.0, 0.0 = bad to 1.0 = excellent

A pre-study focusing on the same type of manual evaluation was conducted in [41]. In this pre-study, a 30-item evaluation scheme (or tool) for manual evaluation was developed based on the hospital districts' guidelines for writing discharge summaries. It used a 4-point scale ranging from −1 to 2, where, −1 = not relevant, 0 = not included, 1 = partially included and 2 = fully included. However, using this scheme turned out to be difficult and extremely time-consuming. One reason for this is that quite a few of the items were somewhat overlapping and very fine-grained, like 'conclusions', 'assessment of the future' and 'status of the disease at the end of the treatment period'. Other items were rarely documented by clinicians (physicians) in the clinical notes written during an ongoing care episode, such as 'status of the disease at the end of the treatment period', 'ability to work' and 'continued care plan'. In addition, a couple of the items were redundant as they concern what we refer to as structured information in the EHR system, such as 'care place' and 'care period', and there is little value in trying to extract this from the text. Therefore, this manual evaluation scheme was further developed to a more simplified version with only ten criteria items.[11] Eight of the ten criteria were rated dichotomically 'yes' or 'no'. These criteria items concern the contents of the discharge summary, where 'yes' means that the summary includes content related to the criteria. Moving from a 4-point scale to a 2-point scale was done to simplify the evaluation further. The two remaining criteria concern the readability of the summary and were rated on a scale of 0.0–1.0, where 0.0 was poor and 1.0 excellent. The scheme used in the manual evaluation is shown in Table 2. Information about what type of note each sentence belongs to, and when it was written, was presented as metadata for the manual evaluators.

Each evaluator evaluated the same 20 care episodes, with eight summaries per care episode. The inter-rater agreement between the three evaluators was calculated with the intraclass correlation coefficient (ICC) [54] with a two-way mixed model using IBM SPSS Statistics version 22. Based on the existing literature, we found no fixed limit regarding the interpretation of ICC values; one suggestion is that values below 0.4 are *poor*, values from 0.4 to 0.59 are *fair*, values from 0.6 to 0.74 are *good*, and values from 0.75 to 1.0 are *excellent* [55]. The inter-rater agreement between the evaluators in this study was *good* (ICC = 0.744, 95% CI 0.722–0.766, $p < 0.001$).

Given the quite concrete evaluation criteria in Table 2, one could intuitively assume that the best summarisation approach would be to focus on extracting those exact ten criteria items. As a result, we experimented with one summarisation method that aimed to do just that. However, this performed poorly in both manual and automatic evaluation. The main reason for this is that we do not have

[10] The *F*-scores from the automatic evaluation are on average noticeably lower in this study than those reported in [41]. This is primarily because here we excluded a specific type of note from all care episodes, a type of summary for the patients, which is often written at the same time as the final discharge summary, and their contents tend to be very similar; sometimes identical. In addition, some of the methods used in this experiment are new or different.

[11] A pilot test was conducted in the process of developing the manual evaluation scheme.

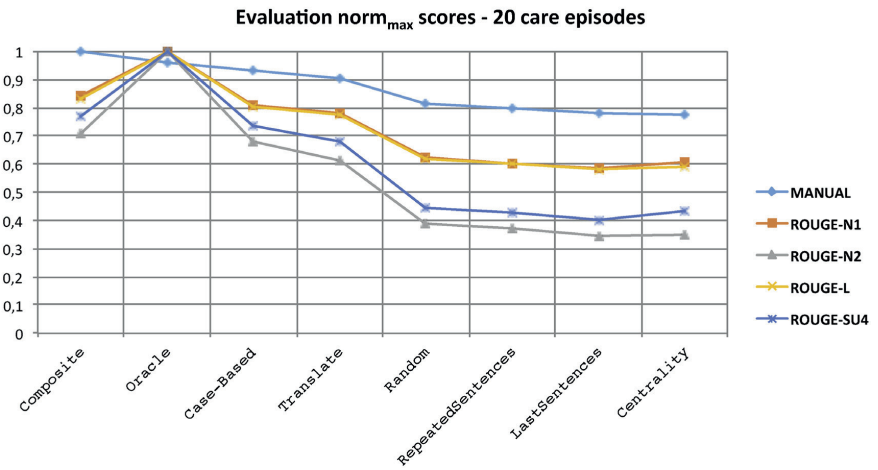


Fig. 5. Graph illustrating the trend for how the automatic evaluation metrics correlate with the manual evaluation of summaries from 20 care episodes. All evaluation entries have been normalised by dividing the scores by their *max* scores. The summarisation methods are arranged according to the manual evaluation scores, and the lines visualise how ROUGE measures follow the trend of the manual evaluations.

any good way of mapping the criteria descriptions to the content in the clinical notes. For example, there is no straightforward way of mapping 'long-term diagnosis' to a sentence not explicitly containing these exact or similar words. A sentence mentioning long-term diagnosis could be: 'the patient has been suffering from high blood pressure for the last four years.'

2.4.2. *Automatic evaluation*

Automated evaluation of summaries generated from a care episode is performed by using the accompanying original discharge summary as a gold standard. This exploratory approach circumvents the need for manually constructing such a gold standard.

The *ROUGE evaluation toolkit* [40] contains multiple *n*-gram-based evaluation metrics that are commonly used for automatic summarisation scoring, such as in the document understanding conferences (DUC) and the text analysis conferences (TAC) [56]. ROUGE basically works by calculating the *n*-gram overlap between a *target summary* (the summary that is to be evaluated), and one or more *gold standard summaries*. The outputs from these metrics are precision, recall and *F*-score, reflecting the overlap between the target and gold standard summaries. The average *F*-scores are what we report here. As there are several metrics to choose from, we use the following[12]:

- ROUGE-N1 unigram co-occurrence statistics.
- ROUGE-N2 bigram co-occurrence statistics.
- ROUGE-L longest common sub-sequence co-occurrence statistics.
- ROUGE-SU4 skip-bigram and unigram co-occurrence statistics.

3. Results

3.1. *Results for Experiment 1*

To visualise how the evaluations correlate, we have plotted the scores from the manual and automatic evaluations in a graph, shown in Fig. 5.

Table 3
Spearman's rho results, indicating how the automatic evaluation metrics correlate with the manual evaluation scores over 20 care episodes.

Evaluation metric	Spearman's rho (*p*-values)
ROUGE-N1	0.9048 (0.00201)
ROUGE-N2	0.9524 (0.00026)
ROUGE-L	0.9524 (0.00026)
ROUGE-SU4	0.9048 (0.00201)

The correlations between manual and automatic evaluations were calculated using Spearman's rho. The results are shown in Table 3. Based on the statistical analysis and *p*-values in Table 3, the four ROUGE measures have a high correlation with the manual evaluations.

3.2. *Results for Experiment 2*

The results from the automatic evaluation of 136 care episodes are shown in Table 4. The *r* columns show the internal ranking of each summarisation method for each evaluation measure. A more illustrative representation is shown in Fig. 6.

We calculated significance levels using the Wilcoxon signed-rank test, with $p < 0.05$ (with Bonferroni correction for multiple hypothesis testing). Based on the p-values the methods could be divided into three groups. First, `Oracle` significantly outperformed all the other methods against all of the ROUGE measures (highest *p*-value: $2.12 \cdot 10^{-22}$ ROUGE-N1 `Oracle` vs. `Translate`). Second, `Composite`, `Case-Based` and `Translate` significantly outperformed the methods in the third group — `RepeatedSentences`, `LastSentences`, `Centrality` and `Random` — against all ROUGE measures (highest *p*-value $3.74 \cdot 10^{-4}$ ROUGE-N2 `Translate` vs. `LastSentences`). In this third group, no method significantly differed from the `Random` method in terms of at least one ROUGE measure. The *p*-values for all comparisons are included in the supplementary materials.

Based on the analysis we can divide the methods (not counting the `Oracle` method) into two groups: `Composite`, `Case-Based` and `Translate` are successful at producing summaries that outperform the simple baseline methods in all comparisons, whereas the

[12] We found the listed ROUGE metrics to be the most commonly used metrics in the literature.

Table 4
F-scores from the automatic evaluation of 136 care episodes. The order of the summarisation methods is the same as in Fig. 5.

Sum. method	ROUGE-N1	*r*	ROUGE-N2	*r*	ROUGE-L	*r*	ROUGE-SU4	*r*
`Composite`	0.3820	2	0.1849	2	0.3678	2	0.1865	2
`Oracle`	0.4819	1	0.2865	1	0.4683	1	0.2694	1
`Case-Based`	0.3634	4	0.1741	3	0.3497	4	0.1764	3
`Translate`	0.3703	3	0.1661	4	0.3551	3	0.1720	4
`Random`	0.3043	7	0.1177	7	0.2949	7	0.1241	7
`RepeatedSentences`	0.3301	5	0.1408	5	0.3196	5	0.1463	5
`LastSentences`	0.3287	6	0.1398	6	0.3184	6	0.1462	6
`Centrality`	0.2862	8	0.1027	8	0.2743	8	0.1151	8

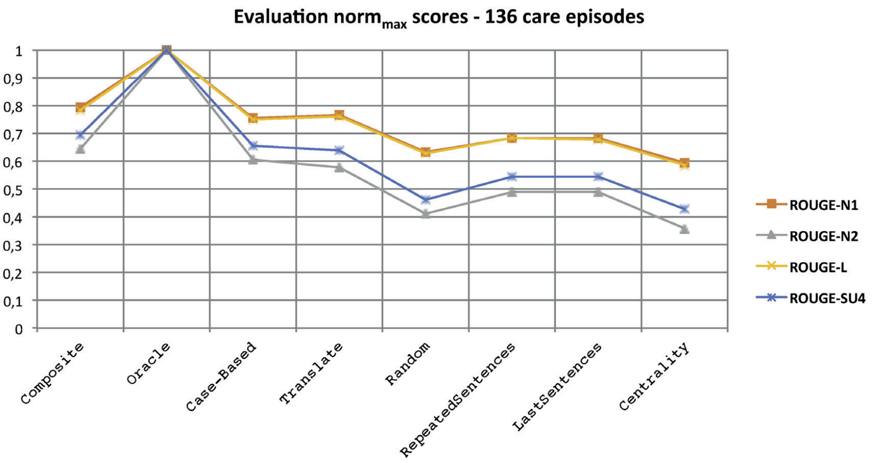


Fig. 6. Graph illustrating how the various summarisation methods perform against a set of 136 care episodes. All evaluation entries have been normalised by dividing scores by their *max* scores. The order of the summarisation methods is the same as in Fig. 5; the lines highlight the similar trends for the ROUGE measures over all the summarisation methods.

`Centrality` and `RepeatedSentences` methods fall in the same group with the simple baseline approaches.

Without the Bonferroni correction, the significantly differing groups would be as follows:

1. `Oracle`
2. `Composite`
3. `Case-Based`, `Translate`
4. `RepeatedSentences`, `LastSentences`
4. `Centrality`, `Random`

4. Discussion

In this work we consider a variety of resource-lean and language-independent summarisation methods for clinical text. These methods circumvent the need for tailored language resources and tools. The proposed summarisation methods utilise WSMs constructed from word co-occurrence statistics in a large corpus of clinical text (see Section 2.1). This enables us to capture various semantic similarity relations in the clinical text in an automatic, data-driven way. The aim is not to construct perfect summaries that can fully replace individual clinical notes or completely automate the process of producing discharge summaries, for example. Rather, this work is a step towards exploring ways of automatically constructing indicative clinical text summaries by relying on purely statistical features for determining a sentence's significance.

We introduce a scheme that domain experts can use to manually compare the relative quality of different automatically produced summaries (and the underlying summarisation methods). The proposed scheme consists of a 10-item questionnaire measuring the expert's opinion of the readability of the summary, and whether it has relevant content. The scheme has been developed based on experiences from our preliminary study on evaluating clinical summarisation methods [41], resulting in a more streamlined tool that is easier to use consistently.

However, such manual evaluation requires human input and is thus impractical to use during summarisation method development, where rapid feedback is required when testing different method variations. Therefore, we also use the ROUGE toolkit for performing automated evaluation. We also seek to establish whether the automated ROUGE-based evaluation can be used in place of human evaluation in the context of clinical summarisation. This meta-evaluation is performed through rank correlation coefficient analysis between the manual and automated evaluation. Finally, we aim to establish which summarisation method performs best in the task of clinical summarisation.

The results from Experiment 1 show that there is a correlation between how the manual and automatic evaluations rank the different summarisation methods. This indicates that using an automated ROUGE-based evaluation set-up is feasible. Further, it shows that the automatic evaluation scores, with the applied evaluation set-up, are reliable for determining what summarisation method performs best. The observation that the manual evaluators preferred the `Composite` method to the `Oracle` method indicates that the greedy search strategy, based on the original discharge summary, does not necessarily produce the best possible extraction-based (discharge) summary.

The results from Experiment 2 show that the methods `Composite`, `Case-Based` and `Translate` all work better than the basic baseline methods ($p < 0.05$) (not taking into consideration the `Oracle` method), whereas the `Centrality` baseline fails to outperform even the `Random` baseline with this data. `Composite`, which consists of combined features from `RepeatedSentences`, `Case-Based` and `Translate` consistently has the highest ROUGE performances. However, the difference between these and the next best methods is not statistically significant against all ROUGE measures following the Bonferroni correction.

When producing the summaries, the `Composite` method combines the following basic principles:

- The importance of a sentence depends on how many times the same or similar information has been mentioned throughout a care episode (`RepeatedSentences`).
- By looking at discharge summaries of other similar care episodes, one can assess the importance of a sentence based on whether or not the same or similar information has been written in these summaries (`Case-Based`).
- If, using a WSM-based translation system, a sentence (its vector representation) can be 'translated' into a vector representation that is similar to how this same sentence would look in the translated word space, it should be considered for inclusion in the final summary (`Translate`).
- Clustering sentences into topics that span across clinical notes in a care episode allows for the removal of redundancy.

`Centrality` is evaluated as being one of the lowest-scoring summarisation methods. Given its broad usage in text summarisation for other domains, this deserves a closer analysis. We asked the evaluators to comment on the structure and content of the summaries that this method produced using open-ended questions. The three questions were:

1. What important information is missing from the summary?
2. What information in the summary is unnecessary?
3. How logical is the structure of the summary?

The following sums up what they wrote based on the analysis of five summaries:

- *Disorganised structure of text, confusing, illogical order or structure.*
- *The end is missing.*
- *Cannot get an overall view of patients' care episode.*
- *Important information is missing.*
- *Information is diffuse and fragmented.*
- *Sentences are not connected.*
- *Too many details about unimportant stuff.*

This seems to indicate that the most 'central' information, independent of when it was written, is not a good indicator of the information that clinicians want to have in the discharge summary. This method did not include sentence topic clustering, which was used in several of the other methods. This further supports the importance of such topic clustering despite the relatively poor performance of `RepeatedSentences`. In future work, other variations and implementations of centrality-based methods should be evaluated, e.g. through the use of the MEAD system [57], similarly to how it is done in [37].

`LastSentences` performs relatively poorly in comparison to many of the other summarisation methods. This is an interesting observation in that it suggests that reading only the latest written information or note(s) is suboptimal when the task is to write a discharge summary. It also suggests that there are reasons to believe that it *is* beneficial for clinicians to use text summarisation systems in their work, e.g. to assist in highlighting relevant information documented earlier in a care episode.

Even with our rather coarse-grained manual evaluation, when applied to a limited number of care episodes, a high correlation is seen with the automatic evaluation. Hence, this automatic evaluation approach can be used to rank the different summarisation methods in order of effectiveness. And since such manual evaluation is not affordable every time a summarisation method has been modified, or when a new method is developed, it should be possible to resort to this automatic evaluation during the method development process.

This study raises questions about the usability, reliability and usefulness of such (imperfect) automatic summarisation systems, particularly when used at the point of care. This is difficult to assess based on the utilised evaluation approach and scores achieved here. The question is: what does it actually mean to have a system that is able to generate textual summaries containing parts of or all (i.e. perfect evaluation score) the content one would expect to find in a manually-created discharge summary? One answer is that it would likely provide a good starting point for a clinician who is about to write the actual discharge summary. It is also likely that the same automatically generated summary would provide an indicative overview of the information having been documented during the corresponding care episode, from a clinician's perspective. However, patient safety issues must be considered before this kind of system is taken into practice. On the one hand, it is important that the most relevant information needed for safe care provision is assured in automatically generated summaries. On the other hand, as long as clinicians treat the generated summaries as an indicative summary, this could be a helpful feature in EHR systems, particularly in situations where time is of the essence. Future research including more user-centred evaluation is required to answer this question in more detail.

A weakness of this study is the validity of the evaluation. The utilised manual evaluation scheme is quite coarse-grained in that it contains ten criteria items, and the rating is done on the level of 'yes' or 'no'. However, in the previously mentioned pre-study [41], a 30-item evaluation scheme was tested, using a four-point scale, but was found to be too detailed and time-consuming to use.

The automatic evaluation is performed using the original discharge summary as a gold standard summary, despite the fact that these discharge summaries are not themselves produced in a purely extractive way. This is reflected in the fact that the ROUGE scores are arguably quite low compared to scores reported in various other studies on text summarisation (see e.g. [40]). The scores achieved by `Oracle` indicate the maximum ROUGE-N2 scores achievable with an extractive-based summarisation system for our data. However, it is encouraging to see that there is a correlation in terms of relative goodness between manual and automatic evaluation, both here and in the pre-study [41], which is promising for future work in this direction.

An alternative evaluation approach would be to manually develop gold standard summaries in a purely extractive way for a set of care episodes, replacing the original discharge summaries as gold standard summaries in the automatic evaluation. This approach was not pursued here, as it is more resource-intensive, but it would possibly give us more reliable results. Another approach would be to use the summarisation system in a (simulated) clinical setting with clinicians as users. Such an evaluation approach is referred to as *extrinsic evaluation*, and could potentially shed light on the impact on documentation speed and quality, as well as on health care quality and patient outcomes. This type of evaluation could also potentially provide directions for future work on improving the summarisation system.

Currently it is difficult to assess the usefulness and potential impact that this type of summarisation system could have in a real clinical setting. On the one hand, it could be a convenient tool for clinicians in terms of providing an indicative textual overview of ongoing care episodes, for example. On the other hand, the possible imperfection of the information presented in the generated summaries must be considered in relation to potential patient safety issues. Future work should focus more on *extrinsic evaluation* by evaluating how the use of automatic text summarisation systems in a clinical setting will impact on documentation speed and quality, as well as health care quality and patient outcomes. Here we believe that a more user-guided summarisation system is needed, enabling real-time incremental summary generation, similar to the methods proposed in [58]. This would mean that the computer-generated summary, or the sentences that it suggests for inclusion in the final summary, are calculated based on analysing what content the user has already written in (or imported into) the summary.

5. Conclusion

This work on the automated summarisation of free text in care episodes introduces and evaluates both a framework for evaluating summarisation methods, as well as novel methods for performing the summarisation. Most of the presented summarisation methods rely on statistical information derived from a large corpus of clinical text, this includes various WSMs. The best performing summarisation methods, according to the applied evaluation, are `Composite`, `Case-Based` and `Translate`. The ROUGE-based evaluation measures are shown to correlate highly with the manual evaluation in terms of relative ranking. Based on these results, we believe that the explored sentence features, especially those in the `Composite` method, provide useful directions on how to approach this summarisation task in a resource-lean fashion. Further studies are needed to assess the applicability of such methods in real-world clinical settings.

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgements

This study was partly supported by the Research Council of Norway through the EviCare project (project no. 193022), Turku University Hospital (EVO 2014), and the Academy of Finland (project no. 140323). The study is part of the research projects of the IKITIK consortium (http://www.ikitik.fi). We would like to thank the manual evaluators for their contributions. We would also like to thank Filip Ginter for assisting us in the work on Word2vec. Finally, we would like to thank the reviewers for their insightful comments. This paper has been proofread by Lingsoft Language Services Oy, and this was financed by the Department of Computer and Information Science, Norwegian University of Science and Technology.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.artmed.2016.01.003.

References

[1] Hall A, Walton G. Information overload within the health care system: a literature review. Health Inf Libr J 2004;21(2):102–8, http://dx.doi.org/10.1111/j.1471-1842.2004.00506.x.

[2] Van Vleck TT, Stein DM, Stetson PD, Johnson SB. Assessing data relevance for automated generation of a clinical summary. In: Teich JM, Suermondt J, Hripcsak G, editors. AMIA annual symposium proceedings. 2007. p. 761–5.

[3] Lissauer T, Paterson C, Simons A, Beard R. Evaluation of computer generated neonatal discharge summaries. Arch Dis Child 1991;66(4 Spec No.): 433–6.

[4] Kripalani S, LeFevre F, Phillips CO, Williams MV, Basaviah P, Baker DW. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. J Am Med Assoc 2007;297(8):831–41.

[5] Sørby ID, Nytrø Ø. Does the electronic patient record support the discharge process? A study on physicians' use of clinical information systems during discharge of patients with coronary heart disease. Health Inf Manag J 2005;34(4):112–9.

[6] Meng F, Taira RK, Bui AA, Kangarloo H, Churchill BM. Automatic generation of repeated patient information for tailoring clinical notes. Int J Med Inform 2005;74(7–8):663–73.

[7] Wrenn JO, Stein DM, Bakken S, Stetson PD. Quantifying clinical narrative redundancy in an electronic health record. J Am Med Inform Assoc 2010;17(1):49–53, http://dx.doi.org/10.1197/jamia.M3390.

[8] Afantenos S, Karkaletsis V, Stamatopoulos P. Summarization from medical documents: a survey. Artif Intell Med 2005;33(2):157–77.

[9] Nenkova A, McKeown K. Automatic Summarization. Found Trends Inf Retr 2011;5(2–3):103–233, http://dx.doi.org/10.1561/1500000015.

[10] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Croft WB, Moffat A, van Rijsbergen CJ, Wilkinson R, Zobel J, editors. Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval. 1998. p. 335–6.

[11] Goldstein J, Mittal V, Carbonell J, Kantrowitz M. Multi-document summarization by sentence extraction. In: Proceedings of the 2000 NAACL-ANLP workshop on automatic summarization – volume 4, NAACL-ANLP-AutoSum'00. 2000. p. 40–8, http://dx.doi.org/10.3115/1117575.1117580.

[12] Patil K, Brazdil P. SumGraph: text summarization using centrality in the pathfinder network. Int J Comput Sci Inf Syst 2007;2(1):18–32.

[13] Chatterjee N, Mohan S. Extraction-based single-document summarization using random indexing. In: Proceedings of the 19th IEEE international conference on tools with artificial intelligence – volume 02, ICTAI'07. 2007. p. 448–55, http://dx.doi.org/10.1109/ICTAI.2007.28.

[14] Mishra R, Bian J, Fiszman M, Weir CR, Jonnalagadda S, Mostafa J, et al. Text summarization in the biomedical domain: a systematic review of recent research. J Biomed Inform 2014;52:457–67, http://dx.doi.org/10.1016/j.jbi.2014.06.009.

[15] Miller GA. WordNet: a lexical database for English. Commun ACM 1995;38(11):39–41, http://dx.doi.org/10.1145/219717.219748.

[16] Unified medical language system [cited 10th August 2015]. http://www.nlm.nih.gov/research/umls.

[17] International Health Terminology Standards Development Organisation: supporting different languages [cited 10th August 2015]. http://www.ihtsdo.org/snomed-ct/snomed-ct0/different-languages.

[18] World Health Organization, International Classification of Diseases (ICD).

[19] U.S. National Library of Medicine, MeSH (Medical Subject Headings) [cited 10th August 2015]. http://www.ncbi.nlm.nih.gov/mesh.

[20] Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229–36.

[21] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507–13.

[22] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003;36(6):462–77.

[23] Velupillai S, Kvist M. Fine-grained certainty level annotations used for coarser-grained e-health scenarios. In: Gelbukh A, editor. Computational linguistics and intelligent text processing, vol. 7182 of lecture notes in computer science. Berlin/Heidelberg: Springer; 2012. p. 450–61, http://dx.doi.org/10.1007/978-3-642-28601-8_38.

[24] Kvist M, Skeppstedt M, Velupillai S, Dalianis H. Modeling human comprehension of Swedish medical records for intelligent access and summarization systems – future vision, a physician's perspective. In: Fensli R, Dale J, editors. 9th Scandinavian conference on health informatics. 2011.

[25] Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009;42(5):760–72.

[26] Pedersen T, Pakhomov SV, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007;40(3):288–99.

[27] Koopman B, Zuccon G, Bruza P, Sitbon L, Lawley M. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In: Chen X, Lebanon G, Wang H, Zaki MJ, editors. 21st ACM international conference on information and knowledge management, CIKM'12. 2012. p. 2439–42, http://dx.doi.org/10.1145/2396761.2398661.

[28] Henriksson A, Moen H, Skeppstedt M, Daudaravi V, Duneld M. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. J Biomed Semant 2014;5(1):6.

[29] Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. J Biomed Inform 2009;42(2):390–405.

[30] Cohen R, Aviram I, Elhadad M, Elhadad N. Redundancy-aware topic modeling for patient record notes. PLOS ONE 2014;9(2), http://dx.doi.org/10.1371/journal.pone.0114677.

[31] Vine LD, Zuccon G, Koopman B, Sitbon L, Pruza P. Medical semantic similarity with a neural language model. In: Li J, Wang XS, Garofalakis MN, Soboroff I, Suel T, Wang M, editors. Proceedings of the 23rd ACM international conference on conference on information and knowledge management, CIKM 2014. 2014. p. 1819–22, http://dx.doi.org/10.1145/2661829.2661974.

[32] Harris ZS. Distributional structure. Word 1954;10:146–62.

[33] Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci 1990;41(6):391–407.

[34] Kanerva P, Kristofersson J, Holst A. Random Indexing of text samples for latent semantic analysis. In: Proceedings of 22nd annual conference of the Cognitive Science Society. 2000. p. 1036.

[35] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K, editors. Advances in Neural Information Processing Systems 26. Neural Information Processing Systems Foundation; 2013. p. 3111–9.

[36] Luhn HP. The automatic creation of literature abstracts. IBM J Res Dev 1958;2(2):159–65.

[37] Liu S. Experiences and reflections on text summarization tools. Int J Comput Intell Syst 2009;2(3):202–18.

[38] Sarkar K, Nasipuri M, Ghose S. Using machine learning for medical document summarization. Int J Database Theory Appl 2011;4:31–49.

[39] Abulkhair M, ALHarbi N, Fahad A, Omair S, ALHosaini H, AlAffari F. Intelligent integration of discharge summary: a formative model. In: Al-Dabass D, Uthayopas P, Sa-nguanpong S, Niramitranon J, editors. 4th international conference on intelligent systems modelling & simulation. IEEE; 2013. p. 99–104.

[40] Lin C-Y. Rouge: a package for automatic evaluation of summaries. In: Marie-Francine Moens SS, editor. Text summarization branches out: proceedings of the ACL-04 workshop. 2004. p. 74–81.

[41] Moen H, Heimonen J, Murtola L-M, Airola A, Pahikkala T, Terävä V, et al. On evaluation of automatically generated clinical discharge summaries. In: Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI). 2014. p. 101–14.

[42] Moen H, Marsi E, Ginter F, Murtola L-M, Salakoski T, Salanterä S. Care episode retrieval. In: Velupillai S, Duneld M, Kvist M, Dalianis H, Skeppstedt M, Henriksson A, editors. Proceedings of the 5th international workshop on health text mining and information analysis (Louhi)@ EACL. 2014. p. 116–24.

[43] Karlgren J, Sahlgren M, Järvinen T, Cöster R. Dynamic lexica for query translation. In: Peters C, Clough P, Gonzalo J, Jones G, Kluck M, Magnini B, editors. Multilingual information access for text, speech and images, vol. 3491 of lecture notes in computer science. Berlin/Heidelberg: Springer; 2005. p. 150–5, http://dx.doi.org/10.1007/11519645_15.

[44] Jones K. A statistical interpretation of term specificity and its application in retrieval. J Doc 1972;28(1):11–21.

[45] McKeown K, Klavans J, Hatzivassiloglou V, Barzilay R, Eskin E. Towards multidocument summarization by reformulation: progress and prospects. In: Hendler J, Subramanian D, editors. Proceedings of the sixteenth national conference on artificial intelligence and eleventh conference on innovative applications of artificial intelligence. 1999. p. 453–60.

[46] Lim J-M, Kang I-S, Bae J-H, Lee J-H. Sentence extraction using time features in multi-document summarization. In: Myaeng S, Zhou M, Wong K-F, Zhang H-J, editors. Information retrieval technology, vol. 3411 of lecture notes in computer science. Berlin/Heidelberg: Springer; 2005. p. 82–93, http://dx.doi.org/10.1007/978-3-540-31871-2_8.

[47] Aamodt A, Plaza E. Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Commun 1994;7(1):39–59.

[48] Lenz M, Hübner A, Kunze M. Textual CBR. In: Lenz M, Burkhard H-D, Bartsch-Spörl B, Wess S, editors. Case-based reasoning technology, vol. 1400 of lecture notes in computer science. Berlin/Heidelberg: Springer; 1998. p. 115–37, http://dx.doi.org/10.1007/3-540-69351-3_5.

[49] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Comput Netw ISDN Syst 1998;30(1):107–17, http://dx.doi.org/10.1016/S0169-7552(98)00110-X.

[50] Mihalcea R. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In: Proceedings of the ACL 2004 on interactive poster and demonstration sessions, ACLdemo'04. 2004.

[51] Wan X, Yang J. Improved affinity graph based multi-document summarization. In: Proceedings of the human language technology conference of the NAACL, companion volume: short papers, NAACL-Short'06. 2006. p. 181–4.

[52] Lehman A. JMP for basic univariate and multivariate statistics: a step-by-step guide. Cary, NC, USA: SAS Institute; 2005.

[53] Wilcoxon F. Individual comparisons by ranking methods. Biometrics 1945;1:80–3, http://dx.doi.org/10.2307/3001968.

[54] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420–8.

[55] Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6(4):284–90.

[56] Dang HT, Owczarzak K. Overview of the TAC 2008 update summarization task. In: Proceedings of text analysis conference 2008 workshop – notebook papers and results. 2008. p. 1–16.

[57] Radev DR, Jing H, Budzikowska M. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: Proceedings of the 2000 NAACL-ANLP workshop on automatic summarization, vol. 4 of NAACL-ANLP-AutoSum'00. 2000. p. 21–30, http://dx.doi.org/10.3115/1117575.1117578.

[58] Sankarasubramaniam Y, Ramanathan K, Ghosh S. Text summarization using Wikipedia. Inf Process Manag 2014;50(3):443–61.