NTNU – Trondheim
Norwegian University of
Science and Technology

Named Entity Recognition in the Climate Change domain

An examination of NER systems for
climatological knowledge discovery

Sean William Holloway

Master of Science in Computer Science
Submission date: July 2015
Supervisor: Pinar Öztürk, IDI
Co-supervisor: Erwin Marsi, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science

NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Named Entity Recognition in the Climate Change domain

An examination of NER systems
for climatological knowledge discovery

by

Sean William Holloway

A thesis submitted in partiall fulfillment of the requirements for the
degree of Master of Computer Science (Civil Engineer)

in the
Faculty of Information Technology, Mathematics and Electrical Engineering
Department of Computer and Information Science

July 2015

Abstract

Climate change and its underlying processes is a complex and dynamic system that spans multiple fields of science. The relationships and interactions of this world altering phenomenon are difficult to analyse and do not always lend themselves to verification through experimentation. In light of this, science turns to an analysis of the building blocks of these interactions and attempts to build a larger model of what is thought to happen, in the abstract. While this is difficult enough inside a single field of science, climate change involves an unknown number of fields and the sheer amount of data is overwhelming. A system which is able to automatically identify the building blocks of these processes from scientific papers, and can then attempt to identify previously unknown connections, would give scientists clues for experimentation and thought. The creation of this system is the task of the Ocean-Certain initiative. In this paper, we begin the process of realization of this new system by examining the problem space and identifying goals that will make the task more concrete. Further, an analysis of existing Named Entity Recognition systems will be performed to find those which are viable for the Ocean-Certain project. Finally we look at chosen system improvements and give an outline as to how these systems could be improved in future endeavours.

Acknowledgements

I would first like to thank my advisors Pinar Öztürk and Erwin Marsi for their patience and assistance in creating this work, your help is and will always be, truly appreciated. Assistance from Ocean-Certain member Murat V. Ardelan, and NTNU IDI Professor Rune Sætre in giving field specific guidance, along with Sindre Næss for all our collaborations.

Special thanks go out to my partner Pernille for all her understanding and love during this project and thesis writing. Your support and unfailing willingness to help in any way possible made all the difference. This accomplishment would not be nearly as great without you.

Last but certainly not least, I would like to thank my family who raised me and allowed me to explore my love for knowledge, science, and computers. It was you who bought me my first science kits, and never thought twice when I spent hours locked in my room creating simple circuits. Your encouragement and understanding allowed me to travel to far off places, and gave me the fortitude to reach for my goals. Thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter contains the motivations behind this work as well as the Ocean-Certain project as a whole. Included is a description of the research goals used as objectives in this work (Section 1.2). Finally the structure of this thesis (Section 1.3) is given.

1.1 Motivation

As science advances, scientists have naturally split into specializations in order to study and understand the complex interactions within a field. It takes many years and large amounts of work to build up a sufficient pool of background knowledge, in part due to the specialized knowledge and training required. This specialization allows a scientist to work on the cutting edge in their field and perform experiments to drive progress. But what happens when a new field arises which crosses several different domains? How would scientists go about identifying questions to experiment on and how would they access the necessary knowledge to drive such experiments?

Typically a consortium of scientists will work together, using each other to fill in the different areas of expertise. While this approach works in many projects there are varied problems which can arise. A project may not have the funds or connections available to employ all scientists necessary to adequately cover an entire problem task. During the lifetime of a project it may enter into a new field in which no current member has the requisite knowledge. Challenges of logistics and scheduling can impact a team's performance and project outcome. As in any project the resources allotted limit the inclusion of personnel and specialists.

One way scientists fill in knowledge gaps is to rely on research articles published by other members of the scientific community. This method can also introduce new problems, such as identification of articles that are relevant, fact checking, and cross-validation. The amount of literature on any one subject can be enormous and an extreme amount of time can be used trying to find data which is applicable to a specific experiment.

The motivation behind Ocean-Certain is to take a new domain of research, climate change, and attempt to alleviate the problem of overwhelming amounts of cross-field data. To do so efficiently, existing tools will be found and evaluated by this study for their viability in the creation of a larger system by future Ocean-Certain(OC) projects.

1.2 Research Goals

1 - Identify system requirements

The first goal of this study is to identify the set of requirements for Named Entity Recognition(NER) systems in order to find suitable candidates for evaluation. This requires research into the types of information necessary in order to build a system that can understand the processes within climate change research. Together with desired output, other project requirements will be identified (such as availability, data-handling capability, etc.) which further refine the system specifications. These requirements will determine which systems are acceptable for subsequent formal evaluation against a manually annotated set of data.

2 - Find feasible Named Entity Recognition systems

Once a set of specifications is identified, a search for candidate NER systems will be performed. This search will be based off of survey papers, research articles, web searches, competitions, related field practices, and previous experience of past and current systems. In order to cover as much ground as possible, it will not be required that each system fulfill the system specifications precisely.

3 - Create "Gold Standard" annotated corpus

One-hundred article abstracts will be taken from a larger set of ten-thousand articles the Ocean-Certain project was given access to by the Nature Publishing Group[1]. These abstracts will be manually annotated with required system information to form a standard corpus against which candidate NER systems will be evaluated. This annotation task will conform to a set of rules and guidelines iteratively formed from experience gained while annotating. This process will include input from Natural Language experts, field experts, and computer science experts in order to create a robust and consistent rule set which suits both the OC project's needs and conforms to this specific task.

4 - Select a sub-set of feasible NER systems

An evaluation of candidate systems will take place in several stages in order to select those which show the most promise and allow for this work's completion on schedule. First, a system test will be performed in order to confirm that each system is working

[1] http://www.nature.com/

as expected and is sufficiently up-to-date to be formally evaluated. Next, a rundown of system specifications and capabilities will be discussed in order to identify a set of systems that will undergo more thorough evaluation and error analysis.

5 - Performance evaluation and identification of system weaknesses

After a suitable set of systems for each variable type is identified, a formal evaluation against the manually annotated gold standard will be performed. Each system's performance will be evaluated via the metrics of Precision, Recall, and F-measure.

Finally, an exploration of system weaknesses will be completed by examining the sets of false positive and false negative results returned. This data will be used to identify additional weaknesses in the annotation corpus, the annotation rules themselves, common system errors, and outliers such as spelling mistakes or the difficulty of complicated formulations.

1.3 Thesis Structure

This paper will detail the work done in the field of Named Entity Recognition as it applies to the Ocean-Certain Work Package One European Union project. From the explanation of Ocean-Certain's goals, a set of requirements will be outlined in order to direct the search for feasible NER systems (Research Goal 1). Next, a search will be performed to find as many NER systems that fully or partially fulfill these requirements (Research Goal 2). A "Gold Standard" corpus will then be created in order to evaluate how selected NER systems perform versus Ocean-Certain's needs (Research Goal 3). A subset of the existing NER systems will then be selected based upon their capabilities, along with additional considerations such as trainability, extensibility, and licensing practices (Research Goal 4). Finally, each selected system will be evaluated for performance against the Gold Standard corpus (Research Goal 5). Further, this paper will begin exploratory work on Named Entity Recognition system improvements and will include a discussion on further enhancements.

Chapter two will cover background information for the most important concepts in this paper as well as work on related projects both in- and outside the Ocean-Certain group. First a description of Ocean-Certain Work Package One and its goals, including

information about data mining and literature based discovery is given. The reason for Ocean-Certain's choice of focusing on the biological pump in its initial stages is then explained. Next this work will clarify Named Entity Recognition, what it is, and some of the challenges encountered while working on this study. Finally an examination of Named Entity Recognition methods utilized in other fields of study is outlined, as these provided a starting point and basis for some of the methods employed here.

Chapter three starts with an explanation of the role Named Entity Recognition systems play in various scenarios. This culminates in a description of why NER systems are necessary and evaluated in this study. Part two details the identification of Named Entity Recognition systems and describes additional considerations used in the selection of these systems for formal evaluation. Next is a description of the experimental data set utilised to evaluate the chosen NER systems. Finally an explanation of the evaluation metrics Precision, Recall, and F-measure, along with error types and additional metrics used is given.

In Chapter four this paper examines systems that were eventually selected for each entity class and includes a discussion about each. This discussion will include background information and the reasons for their selection. Further this chapter covers the experimental corpus and the guidelines used in its creation.

Chapter five will contain the results of this study's formal evaluation and an examination of the strengths and weaknesses found for each system. When multiple evaluation runs were necessary to gain a deeper understanding of how the system performs, each version implemented and tested will be covered. At the end of each entity class is a summary diagram and table for a comparative overview of performance. Finally an exploratory improvement of one selected system will be detailed and the outcome of these improvements will be discussed.

Finally, this paper will conclude with an overview of this paper in Chapter 6.

Due to space considerations only system error output will be included in the appendix in order to facilitate discussion. The full set of tables from each system test, along with their evaluation on each abstract, is available upon request.

Chapter 2

Background and Related Work

This chapter is a detailed examination of the background concepts surrounding this work and the Ocean-Certain project. Section 2.1.1 will first cover the Ocean-Certain project and its overarching goals, along with the specifics of Work Package One. A discussion of climate science and the biological pump, found in Section 2.1.1 and 2.1.3, is included to give an understanding of the complexity of these systems as well as the necessity of text mining/literature based discovery (See Section 2.1.2 for definitions). Next this chapter will cover Named Entity Recognition from origins to current techniques and practices. A look at some of the challenges facing this information extraction task that were encountered during this work will then be given. Section 2.1.5 covers the connection of variables together to form facts, which the results of this project will support. This section includes work currently in progress by other members of the project team as well as previous projects completed in the initial phases of Ocean-Certain Work Package One. Finally Section 2.2 will give an overview of work done in related fields which gave a starting point for the practices used in this study.

2.1 Key Concepts

2.1.1 Ocean-Certain

Ocean-Certain(OC), formally OCEAN-CERTAIN - "Ocean Food-web Patrol – Climate Effects: Reducing Targeted Uncertainties with an Interactive Network", is an initiative by the European Union to understand and support climate change research, starting by focusing on the Biological Pump (see Section 2.1.3). This initiative will foster a deeper understanding of current climate change research, and will assist both scientists and policy makers in making decisions about future efforts by providing knowledge. To do this, the project will use existing knowledge in the form of scientific papers and databases to generate an overview of the field as well as attempt to fill in missing gaps. By creating an interactive network of knowledge, the system will be able to identify areas of new inquiry and provide a pointer for further research by scientists. Please refer to the OC website[1] for more information. Ocean-Certain Work Package One(OCWP1) is the first stage of the Ocean-Certain initiative that will begin building a background knowledge database from scientific articles and other sources. This task will be completed using Literature-Based Knowledge Discovery and Text Mining.

2.1.2 Literature-Based Knowledge Discovery and Text Mining

OCWP1 will begin work on a system that is able to identify new knowledge from gathered data by constructing **facts** which will later be used to produce a **hypothesis**. A diagram of this process can be seen in Figure 2.1 created by OCWP1 project member Erwin Marsi[2].

On the left side are a collection of facts from scientific papers and other data, which are currently held to be relatively true. It is not the purpose nor goal of the Ocean-Certain project to determine the validity of these facts, rather it relies on the scientific community to determine their truth. Together with a database of background knowledge, this new system will be able to generate new hypotheses through *inference*[3], a conclusion that is reached based upon the factual evidence and a degree of reason. This is a simplified

[1] http://oceancertain.eu/what-is-ocean-certain/
[2] http://www.ntnu.edu/employees/emarsi
[3] https://en.wikipedia.org/wiki/Inference

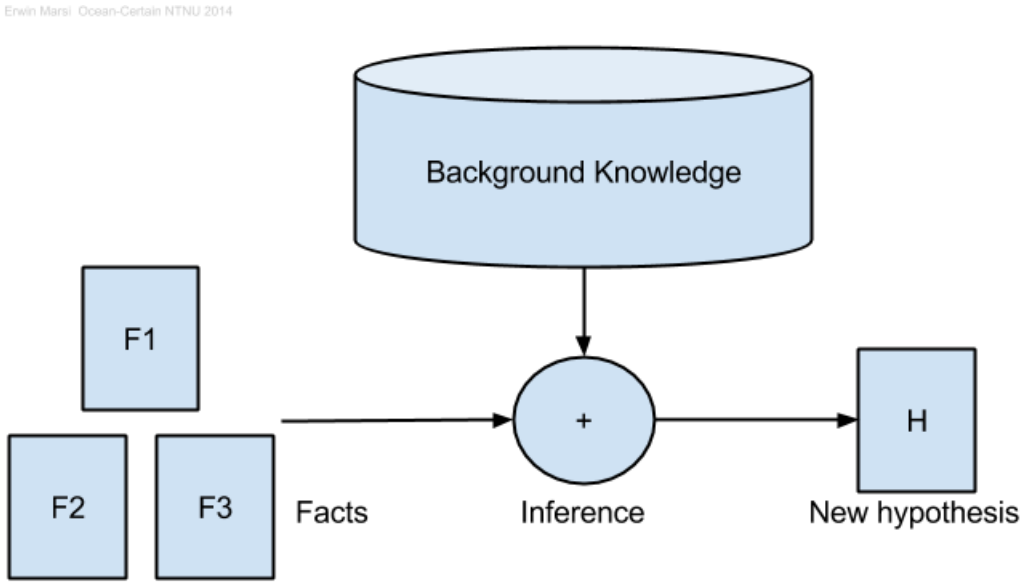

FIGURE 2.1: Literature Based Discovery concept

example of what is known as *Literature-Based Knowledge Discovery*(LBKD), named for its dependence on facts generated from scientific writings. LBKD was pioneered by Don Swanson(Swanson, 1986) to find connections in biomedical literature. While the methods may be different, research into this type of discovery was continued(Mack and Hehenberger, 2002) and is still ongoing(Holzinger and Jurisica, 2014).

There are a few important details to note in this concept, as these explain the goals and assumptions it is built upon. To begin with the system is reliant on facts and background knowledge in order to have the evidence necessary to infer a new hypothesis. In order to generate the variables for these facts a technique known as *Text Mining* is employed. Text mining in this context is the automatic processing of scientific writings using Named Entity Recognition (See Section 2.1.4) in order to identify and extract data(Tan, 1999) that can be used to build facts. Results of Named Entity Recognition are in the form of annotated sentences, where specific **entities** are labeled with their appropriate descriptor as in Figure 2.5. The process of combining these entities into facts is the focus of a parallel project by Erwin Marsi and its description can be found in Section 2.1.5. Text mining is also important in constructing background knowledge through the processing of *ontologies* (connected data) and other sources.

Finally the validity of the plausible hypothesis is not to be verified by this system or the OC project. It is a pointer for scientists and researchers to look at and confirm or deny.

2.1.3 Climate Change and the Biological Pump

The field of climate science is a relatively new field which has garnered a lot of interest from both the scientific community and the public at large. Governments around the world have been seeking information on climate change over the past few decades(Houghton and Callander, 1992; Change, 2007) in order to understand its ramifications. In the last few years more and more research is being devoted to this field, as its importance is fully realized(Pouchard and Noy, 2013; Marsi and Ardelan, 2014). This field involves many different areas of science including; oceanography, meteorology, geochemistry and biology. Even from this very short list, a wide range of expertise is needed to even begin research.

While climate science on a global scale can involve an incredible amount of processes, it was agreed upon that to start, a smaller area of focus was needed. Through examination of scientific articles together with input from scientists from many fields, the Biological Pump was chosen to be the ingress into this extensive field.

The Biological Pump

The Biological Pump is an important climatological process where carbon dioxide (CO_2) enters the world's oceans in several forms(Ducklow and Buesseler, 2001). This process is important for reducing the overall amount of CO_2 in the atmosphere, thus reducing the severity of the greenhouse effect together with overall temperature. This process is driven by several inter-connected systems, and is also fragile(Orr, 2005). Changes in temperature, atmospheric composition, CO_2 density, and more can have wide ranging effects on the effectiveness of the Biological Pump. A simplified diagram of this process can be seen in Figure 2.2.

In the first stage known as the soft-tissue pump, phytoplankton use CO_2 and other elements during photosynthesis to make the structures they need to live (carbohydrates, lipids, and proteins). This "fixes" CO_2 from the air into the soft tissues of the phytoplankton, seen in the upper left quadrant of Figure 2.2. Organisms then feed on the plankton and are in turn fed on by higher species, in a complex relation known as the *Food Web*[4]. Next, particulates from plankton and the organisms in the food web slowly sink to the bottom of the ocean and become a part of the sea bed in a process known as

[4]https://en.wikipedia.org/wiki/Food_web

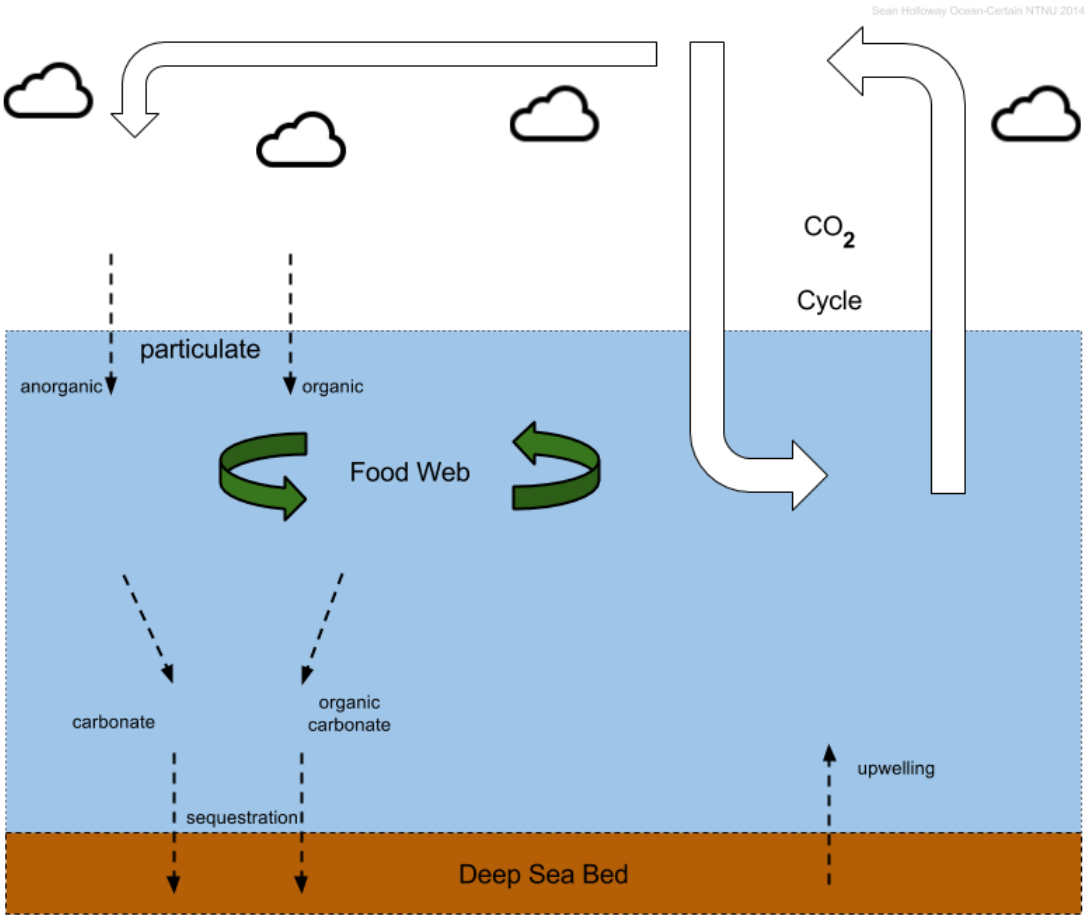

FIGURE 2.2: A simplified biological pump

sequestration[5]. Once there, the particulates containing CO_2 can remain for thousands of years, and this is how the majority of CO_2 is stored(CL, 2006).

The biological pump was chosen as it involves several fields of scientific study, including chemistry, biology, and meteorology. It involves several different processes working together e.g. the soft-tissue pump, the biological food web, and CO_2 sequestration, to form a greater whole. Thus it was seen a miniature version of the greater processes Ocean-Certain will try to analyse and understand through Literature-Based Knowledge Discovery.

[5]https://en.wikipedia.org/wiki/Carbon_sequestration

2.1.4 Named Entity Recognition

Named Entity Recognition(NER) is a sub-task of larger information extraction tasks that arose from the MUC convention initiated and financed by DARPA[6] (Defense Advanced Research Projects Agency). The primary task of NER is to identify specific types of information within data, and tags that information with meta data, e.g. the type of information it belongs to(Nadeau and Sekine, 2007). For example if we wished to find the names of any person in a report, an NER system would analyse the text and tag any name it found with "Person".

The Message Understanding Conference(MUC) began with defense related tasks involving military reports(Grishman and Sundheim, 1996). DARPA assembled teams of researchers to compete against one another to attain the best performance possible. Teams would find ways to fill in data on events, commonly; agents involved, what the cause was, time, date, and consequences of the event. A need for formal evaluation of these results gave rise to some of the performance measuring practices still used today, namely Precision, Recall, and the F-measure (See Section 3.2.4).

Eventually the MUC conference moved onto evaluation of more civil sources like journalistic articles. At MUC-3 news reports would be the new medium to be analysed, looking for terrorist activities in Latin America(Grishman and Sundheim, 1996). This shift in mediums would introduce new problems into the field as the structure of information would be more varied and could include more languages. Up until the last MUC conference, MUC-7, the problem definition would change many times from corporate information to airplane crashes(Elaine Marsh, 1998).

Over the lifetime of the MUC conference it was noted that many of the tasks given were based around finding specific information from resources. Early research(Coates-Stephens, 1992; Thielen, 1995) found that the problem could be generalized into finding proper names in articles. Eventually the first three named entities, Person, Location, and Organization were identified as primary sources and called the "enamex" entities. Later other important units would be found and classified. Time and date would be termed "timex", numbers and percentages as "numex". These entities would be marked

[6] http://www.darpa.mil/

with meta data tags that identify the group of information the entity belongs to. These tags became known as *annotations*[7].

In Figure 2.3 an example NER system pipeline can be seen, split into sub-tasks for the extraction of named entities from textual data into a database. For this example, articles, a form of textual document, are given to the NER system for processing. In **Format Analysis** the system attempts to decipher what type of documents were given, in order to better understand what information is important and what can be removed(Maynard and Wilks, 2001). A typical situation would be if given an HTML document the NER system would be able to strip away formatting and documents tags, while retaining the text meant to be read by a user (which is typically the part to be processed). In some cases only a single form of document is meant to be processed by the system, in which case this step may be skipped.

The **Tokenizer/POS tagger** is responsible for segmenting the text and assigning *Part-of-Speech* tags to each segment (or *token*). First the system will attempt to detect and separate sentences based upon a set of rules and pattern recognition. This task has challenges of ambiguity as sentence endings can be either relatively unambiguous (? or !) or more difficult with a simple period (.) as these can be used in abbreviations like "Mr. Anderson" or numbers (4.2%). A common method to overcome this challenge is whenever the system detects a period it uses a binary classifier (yes/no) to decide if the sentence is ending or not based upon rules[8]. Once sentence splitting is completed the system will move on to segmentation, where each sentence is split into appropriate segments e.g. words, punctuation, and numbers. This task attempts to detect a word and normalize or stem it, as words can have multiple forms (run, ran, running) and these need to be understood to be talking about a single action. The methods and practices to solve this task are too extensive to cover in this paper but typically involve word form dictionaries and lexicons, more information can be found in resources such as the Stanford NLP Group[9]. More advanced systems could be asked to recognize a range of additional concepts such as chemical formulae, mathematical expressions, or diagrams in addition to simple words. Next the system attempts to assign a part-of-speech tag to each token, in order to gain context and decipher how it fits into the larger structure. The most common case would be marking verbs and nouns. If you wanted the system

[7] https://en.wikipedia.org/wiki/Annotation
[8] http://www.nactem.ac.uk/dtc/dtcTsuruoka.pdf
[9] http://nlp.stanford.edu/

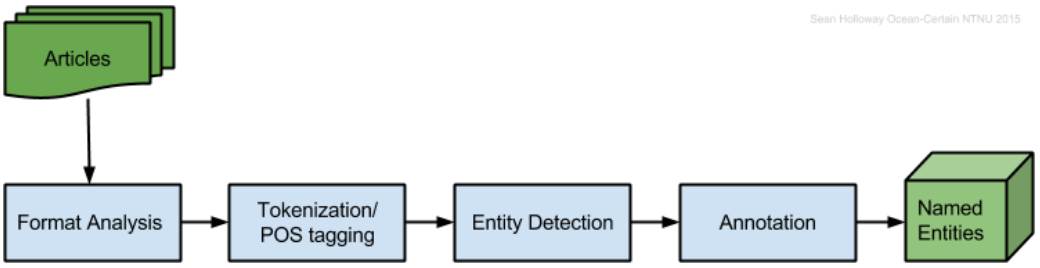

FIGURE 2.3: Example Named Entity Recognition Pipeline

to find "Person" then it could be more confident marking a token that was a noun as a person rather than a verb. The difficulty here is that many words can belong to several parts-of-speech. The word "well" can be used as an adverb (Writing is going well), a noun (The well is dry), or an adjective (We are doing well).

This brings us to **Entity Detection** in which system looks at each token and attempts to identify them as a Named Entity or not. While there are many different kinds of identifiers, from hand-crafted rule sets to Hidden Markov Models, a gazetteer system is common. Gazetteering is the use of crafted dictionaries or word lists that a system will compare against each token to see if it is a viable candidate. These crafted dictionaries allow for a very precise identification system, but are not very flexible in handling new data. It is common for a dictionary entry to contain the exact word the system is to find, as well as name variants. For example "John Smith" may also allow for "J. Smith", "John S.", or "John", depending on the degree of matching required. Once an entity is ready to be classified the system moves on to **Annotation** in which it tags the entity with the appropriate meta data. When complete the system will return either a list of these annotated entities or the entire document with the annotations included in-line.

As the importance and practicality of the research done for the MUC conferences became known the practice of Named Entity Recognition became more mainstream. New applications of NER were imagined and applied in different scenarios for many reasons. From science, to news, to corporate information handling the methods and abilities of NER systems would be driven forward in order to organize a new world. With the explosion of the Internet a vast amount of data became more available, however the unstructured and varied nature of that information would present new challenges for those trying to analyse it. The ability to answer basic questions from text(Scaria and Clark; Berant et al., 2014) could open new doors in data processing and comprehension, such as the summary of knowledge in scientific articles. Social media would become an important

source of information for many parties, and the extraction of information from these mediums would be increasingly important. A corporation that was interested in how it is perceived by users could use Named Entity Recognition to find tweets(Alan Ritter and Etzioni, 2011) or blog posts that talk about their corporation. News channels could comb through any number of sources in an attempt to find and report on new stories in the quickest and most efficient manner possible.(Ekbal and Bandyopadhyay, 2008; Miller and Stone, 1999) The applications of a system that can process huge amounts of data at a rapid pace is only limited by the thinking of people who are willing to use it.

Three main methods of entity recognition are now commonly thought of as standard. Rule-based is a strain where pattern matching and heuristics are leveraged to find relevant information, however these systems are often very strict in the information they find. ProMiner is an example system which uses a rule-based system for entity detection.(Hanisch and Fluck, 2005) This method is based on a sequence of rules that follows a priority system to identify entities but is constrained by problems of data validity (no false information), rule consistency (rules do not conflict), and rule independence (no redundant rules), among others. The formation of this rule-set is very time consuming and only typically used when processing known documents. An application of this could be in the health sector for processing patient data sheets where it is already known what type of information should be on each line.

Gazetteer systems use lexicons, dictionaries, and sets of pre-assembled information to find entities that are sought after, but again this knowledge needs to be predetermined. For example if a user wished to find any document that involves any "City" in the United States a list of such cities would be compiled and the system would perform a direct search of any entity found to see if it was on the list. While this may seem straight forward, it can be difficult to compile such a list and is very time consuming. Data validity is a problem due to real-world changes, in the example given new cities may be built, or old ones disappear. A city may lose its status and become a town, or the city could undergo a name change. Name variation is also a challenge, a "Person" could be referred to as a range of names as in the examples given under the NER pipeline explanation.

The most contemporary of the these methods are machine learning approaches which use stochastic models and/or evolving weight calculations to identify candidate entities.

Hidden Markov Models(Zhou and Su, 2002), Conditional Random Fields(Settles, 2004), and Maximum Entropy(Borthwick, 1999) are examples of these types of systems. While not dependent on a carefully crafted rule-set or dictionary, these methods depend on a large amount of training data in order to adapt and update their system to match a problem task. This training data and verification does require a significant amount of resources and time to complete, however the benefit is a lower cost of maintenance versus a rule-based system. A degree of human supervision is typically necessary in order to resolve ambiguity and to refine a machine learning system, especially if transitioning to a new problem area. A machine learning system will often have a poorer performance than e.g. a rule-based system, but the ability to incorporate new data and the applications towards ambiguous data hold the appeal of researchers.

Systems can also be comprised of one or more of these approaches in combination, to handle the different challenges of each stage in the NER pipeline. Under the discussion of the NER pipeline these different approaches are outlined to give a short introduction to the varied methods used in solving these tasks. In this work a range of NER system types are evaluated, from Chemspot's CRF and pattern based matcher to the gazetteering system of Linnaeus2. It is important to understand these different approaches in order to more accurately define and understand the strengths and weaknesses of each system. With this knowledge it is possible to impart information as to how each system can be utilized or modified to fit the needs of the Ocean-Certain project.

Challenges facing NER

With the extensive applicability of Named Entity Recognition in different scenarios, along with the evolving nature of information and its containers, there are many challenges facing NER today. In order to limit the scope of challenges faced system developers will often make an assortment of assumptions and limitations on what an NER system is asked to do. Often it is assumed that an NER system will not be able to work on an undefined amount of data types, they are typically crafted to handle a single type such as text. Further, an NER system may only handle one or more formats of that data type, whether it be web pages (HTML) or plain text.

The system designers will often choose a set of entity types the system will be able to recognize, depending on the domain it is intended for. An example would be in news text, Person/Organization/Location is common, while in bio medicine Gene/Protein

names are the norm. Other limitations such as language, structure, context domain, or types of identifiable information may be imposed. While a perfect system would be the ideal, the degree of performance necessary to be useful can also be an assumption that is made.

Next this paper will examine some of the challenges encountered in this work, in order to gain an understanding of their structure and composition, as well as the impact they can have in finding suitable NER systems for the larger OC project. Here the term "language" is used as a representation of information based upon context, from normal words to chemical formulae to numerical data.

Use of informal language

It is typical for anyone who is writing anything to introduce their own style of language, taken from their upbringing, experiences, and interests. This can be evident in the use of words, phrases, terms, or expressions used to convey their ideas. People working within an area of expertise will often form their own informal set of terms that are communally used and accepted by their peers, as these will be understood from that common context. Often this problem will arise when extracting information from less formal sources such as text messages, tweets, or blog posts(Alan Ritter and Etzioni, 2011; Liu and Zhou., 2011).

While this is a smaller problem in scientific articles, as they are written to be read and assessed by the wider scientific community, our project necessitates the examination of articles from many fields of study. Cross-field variances in concept definition and understanding, common term usage, data structuring, and contextual starting point can lead to mis-identification of entities by an NER system.

In "A System for Adaptive Information Extraction from Highly Informal Text"(i Alemany and Carrascosa, 2011) this problem is examined. Research into using current language processing tools together with string procedures, machine learning and custom processing identified ways to handle such problems in short text messages, classified ads, or tweets. It is noted that while some aspects of the problem could be solved by standard tokenization and chunking, additional resources would be necessary to provide acceptable performance on tasks like semantic tagging.

Information chunking

Information chunks are sets of words that make up a larger entity, such as "United Emirates" or "carbon dioxide". In some forms the individual words are not entities by themselves, but in the latter example "carbon" could also be taken as an entity alone. Naturally a system that identifies "carbon dioxide" as a single entity is optimal but this context sentitivy is not simple to implement.

This problem area is one that is faced by many NER systems and has a thorough background of research, the paper "Named Entity Recognition using an HMM-based Chunk Tagger"(Zhou and Su, 2002) uses Hidden Markov Models for information chunking. Compilation methods using the strengths of different systems together has been attempted as described in "ETL Ensembles for Chunking, NER and SRL"(dos Santos and Fernandes, 2010), which use bagging and random subspaces to perform chunking.

While these methods may be available the implementation of such could become a very resource consuming task. In addition information chunking is often an embedded part of an NER system, and therefore difficult to replace or modify. This makes a well functioning information chunker an important part of our choice in NER system.

Applicability of different domain methods

A different type of challenge arises when one uses methods that are used in one domain towards another. As seen in previous sections, variances in grammar can have a large impact on performance. Studies have seen as much as a forty percent reduction in performance when testing a system on the intended structured corpus, versus an unstructured one(Ciaramita and Altun., 2005). Other differences in domain such as common grammatical structures, naming conventions, specificity of entities, and more can affect the tailoring of NER systems.

The Ocean-Certain goal of using current NER systems towards a similar, yet different field could yield some of these problems. The hope is that by combining different NER systems, using cross-validation and tailored resources, these individual systems can be adapted to suit the project needs.

2.1.5 Connecting variables into facts

As seen in Figure 2.2 (discussed in Section 2.1.3) there are many **processes** which need to be identified by the Ocean-Certain system, and are necessary to understand the Biological Pump as a whole. Each process is the combination of many sub-processes which can each be further reduced to simpler and simpler forms. A process here is defined as one variable having a direct relation to another variable, causing a change, however for clarity larger macro-processes will also be referred to as a process. These changes are typically distinct and definable such as increases or decreases, though these are very simple examples and by no means a complete list.

To begin with three key factors were identified that would be imperative to building a model of these processes;

Chemical compounds were a natural choice as carbon dioxide (CO_2), along with many other elements, are involved in the Biological Pump (Section 2.1.3). Without a knowledge of how and where these chemical compounds act and relate, an understanding of the Biological Pump is not possible.

Marine species were also a natural choice as these are the biological agents that perform the CO_2 fixations. Further, the food web also plays a large part in the movement of carbon inside this system.

The third key factor, location, was decided upon after deliberation for its contextual value. From surface waters (euphotic layer), to the deep sea bed, location information can give a lot of information on what type of process is taking place. This information will eventually allow the Ocean-Certain system to have increased confidence that it understands the processes it is trying to identify. If the system can understand that a paragraph is talking about the euphotic layer, it can be more confident that phytoplankton will probably be involved in photosynthesis.

In addition these variables have been studied in entity identification research previously. This meant that it should be possible to find existing systems that would fulfill Ocean-Certain's needs without having to develop the systems themselves. As well as saving time and resources, this would allow the Ocean-Certain team to focus efforts into improving existing systems where necessary.

Each of these three variables are referred to as **entity types** within the domain of Named Entity Extraction. With these variables identified the system can then proceed in attempting to connect them together with relations.

Entity Identification

Much like Swanson(Swanson, 1986) did in the mid-eighties to identify his hypothesis on fish oil, and later simulated by others(Cameron and Rindflesch, 2013), this approach attempts to identify variables and events which make up the building blocks of processes. An event is an identifier used to denote a change of status of something, such as an increase or decrease. A variable is an actor such as a species or chemical which is affected by an event.

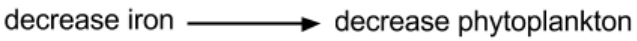

FIGURE 2.4: Example process

In Figure 2.4 a very simple process is illustrated, both "iron" and "phytoplankton" are variables connected by a relation. The events "decreases" together with their variables shows the process, namely that a decrease in iron leads to a decrease in phytoplankton.

The Ocean-Certain project's approach attempts to identify these processes automatically from thousands of scientific articles. Using Named Entity Recognition (NER) together with machine learning, Ocean-Certain aims to be able to span multiple scientific fields and any amount of data to create a database of such events.

To begin with, a system which is able to identify the variables which are deemed intrinsic to these processes needs to be created. With the entity types chosen, work can begin on finding NER systems capable of identifying them in text.

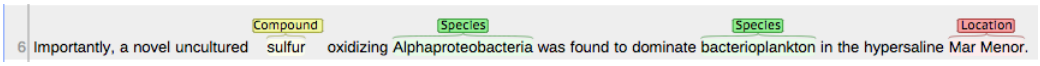

FIGURE 2.5: Example annotated sentence

Figure 2.5 is an example of the Ocean-Certain goal for existing NER systems. While this example is from a manually annotated document, a system which is able to do this automatically is the objective. In this example, NER systems are able to identify all three of the target entity types (chemical compounds, species, and locations) embedded in a sentence from an article abstract. The reason for researching and developing a

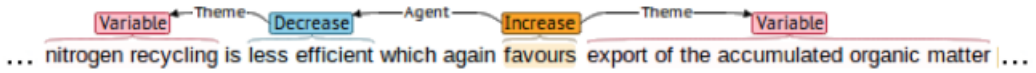

FIGURE 2.6: Theme-agent relationship

system capable of this automatic annotation is the generation of facts as discussed in Section 2.1.2. Finding NER systems capable of extracting these target entity types will be the first step in creating the larger Ocean-Certain system.

Fact Creation

Fact creation is the task of creating process structures that the OC system will later use in the creation of chains of processes. Previous work by Elias Aamot and Erwin Marsi(Marsi and Ardelan, 2014) researched the creation of tools to connect variables together with processes of change. This is a top-down approach to the problem which gave insight into the types information that would be necessary. A snippet of their program output can be seen in in Figure 2.6.

The basis of *this* study is the ability to find and annotate entities within scientific articles that these processes will be built upon. This work is a bottom-up approach which begins with finding these anchor terms to be connected together in future studies. At its conclusion the imagined OC project system will work in these steps;

1. Define set of scientific articles to be analysed.
2. Identify and annotate key variables. (This study)
3. Expand key variables to include necessary information to create events.
4. Identify processes of change.
5. Connect events together to form the identifies processes.
6. Connect processes together in order to identify chains of processes.

The entities selected in Section 2.1.5, along with the additional requirements of NER systems as explained above, form the set system requirements that will be used to find NER systems.

2.2 Other Related Work

2.2.1 NER in Bio-medicine

The field of bio-medicine has seen recent advancements in NER development that made it a good starting point for this study, and the Ocean-Certain project. Often the applications of these technologies focus on the relationship between genes and proteins. Their interactions connected with diseases and chemistry make up the foundation of pharmaceutical innovation, a field which has a strong motivation and resource pool to fund research. From the popular Watson computer being adapted for health-care, "IBM's Watson Gets Its First Piece Of Business In Healthcare"(Upbin, 2013) to the extraction of events in "Event-based Information Extraction for the biomedical domain: the Caderige project"(Alphonse, 2004) there are many comparisons to be drawn.

"Extracting synonymous gene and protein terms from biological literature"(Yu and Agichtein, 2003) gives information on tasks that involve identifying synonyms of the same entity, even from researchers within the same field. They examine four different methods to solving this problem and evaluate the different methods over a large biomedical corpus. Comparatively these methods could assist in developing a way to identify cross-domain synonyms and structures of information which are similar. As chemicals are an early focus "Information extraction technologies for the life science industry" makes special mention of challenges related to recognizing these entities.

For the larger extraction problems, such as finding contextual information in relation to events, we see studies such as "Open Information Extraction from Biomedical Literature Using Predicate-Argument Structure Patterns"(Nguyen and Tojo, 2013). This study attempts to extract any type of relation or fact from biomedical literature. "Learning Recursive Patterns for Biomedical Information Extraction"(Berardi and Malerba, 2007) uses inductive logic, recursion, and pattern recognition to try and find dependencies between entities, a context sensitive task.

From this short list it can be seen that the field of bio-medicine has a large volume of research applicable towards the Ocean-Certain project. However the fact that the research focuses on a different domain means adaptation and ingenuity will be necessary to be successful.

Chapter 3

NER in the Climate Change domain

The first section of this chapter is a discussion on the necessity of Named Entity Recognition in text mining scenarios, literature-based knowledge discovery, the Ocean-Certain project, and this study. Section 3.2 begins with the initial guidelines for finding viable NER systems along with criteria used to reject candidates for formal evaluation. Next, this chapter details the choices made for the creation of an experimental data corpus including size, scope, and content. Section 3.2.3 examines the guidelines and tools used in creating the manually annotated "Gold Standard" corpus that will represent a perfect system output to be measured against. The evaluation metrics and standards are covered in Section 3.2.4 along with an explanation of the relationship between them, together with their mathematical equations. Finally additional considerations that impacted the choice of NER systems falling outside the scope of the previous sections are discussed.

3.1 Role of NER in text mining scenarios

As seen previously the role of Named Entity Recognition is the ability to automate the process of finding specific kinds of information within data. This automation allows for the processing of large amounts of data quickly, in order to perform various tasks in order to accomplish a goal as described in Section 2.1.4. While the term "data" can refer to many things e.g. text, sound, or picture, the most common form applied to NER is text. This study involves the evaluation of NER systems in order to find existing systems that can process scientific text, and to determine their suitability for the larger OC project. It is not the goal of this work to develop a new system, but to understand the strengths and weaknesses of these existing systems in order to advise further work on systems that may possess the capabilities to handle basic text mining.

Text mining[1], described in Section 2.1.2, is the processing of text to identify requested information using techniques of pattern recognition. The ability to classify a document, find the semantic meaning of a statement, or determine the cause of an action can all be seen as a goal of text mining. Named Entity Recognition forms the basis of text mining by identifying the different parts of a text as belonging to a specific set entities, or not. To classify a document we may wish the NER system to find the names of people, organizations, or locations in order to determine what the document is talking about.

3.1.1 NER in Literature-Based Knowledge Discovery(LBKD)

Pioneered by Don R. Swanson(Swanson, 1986) in 1986, LBKD is the examination of existing knowledge from scientific articles in order to find new relations. Normally, new knowledge is formed by experiments done in laboratories, but in this approach analysis of current knowledge is iteratively combined in order to generate undiscovered relations. To form his previously undiscovered hypothesis of fish oil helping migraine headaches, known as the *Reynaud-Fish Oil Hypothesis*, Swanson used statistics to analyse current facts and look for patterns, e.g. the statistically high data points where one term is together in many publications with another. In a basic rendition of his scenario, two facts were found;

[1] https://en.wikipedia.org/wiki/Text_mining

(1) Migraines are often caused by low magnesium levels.

(2) Magnesium levels in the body are elevated by taking fish oil.

The combination of these implies that by taking fish oil, it is possible to reduce a migraine. While seemingly simple it is the process of determining all the facts from large databases of publications that is the problem. Named Entity Recognition will replace the statistical analysis by automatically finding the variables these facts are based upon (discussed in the next section).

3.1.2 Text Mining and LBKD in Ocean-Certain

In the case of the Ocean-Certain project, Text Mining is to be combined with Literature-Based Knowledge Discovery to process scientific articles related to climate change. This processing has the goal of finding chosen entities in order to build up larger relations that will be used to form facts. To begin, the entities **Chemical Substance**, **Biological Species**, and **Location** were chosen to be the targets of existing NER systems.

Named Entity Recognition has the goal of finding what the Ocean-Certain project calles *agents of change* within scientific articles. By finding these agents, it is possible to understand the inner-workings of large processes. By an *agent of change* it is meant any variable which causes the change in another variable, as seen in 2.4. Here we have two variables, "iron" and "phytoplankton", which have the relation that a decrease in iron leads to a decrease in phytoplankton. This event is known as the "Iron Hypothesis"[2] and could be postulated by facts found in scientific articles. To begin finding such an event a Named Entity Recognition system would be tasked with finding the element "iron", as well as the biological species "phytoplankton".

3.1.3 NER systems in this study

For this study we focus on NER systems that are able to identify the three chosen entities of chemical compounds, species, and locations within scientific articles. The use of existing systems will shorten the development of of the OC project and allow the team members to focus on new systems or improvements/adaptations of the chosen

[2]https://en.wikipedia.org/wiki/Iron_fertilization

existing systems for their problem scenario. The evaluated systems are developed for different domains and scenarios than Ocean-Certain's target domain, climate change, but by breaking down the OC task into sub-tasks, it is believed systems can be found to solve some of these individual problems without the necessary resources required to develop systems from scratch.

An example sentence together with the goal annotations can be seen in Figure 2.5, where the chosen entities are tagged with their respective information type. Section 3.2.1 describes the initial requirements for candidate NER systems.

3.2 Evaluation Methodology

3.2.1 Pre-selection of NER systems

The initial requirements for a feasible NER system included being able to identify the types of information required. The system should optimally be run locally but online NER systems were also considered. A system which had an open license for use would also be required, as Ocean-Certain does not want to spend an extensive amount of time in legal negotiations. Access to the source code would allow OC to extend or adjust the system as necessary, instead of being a black box where only pre- or post-processing steps were available to improve performance. Systems which were more recent, or were more recently updated, were also given a higher priority. While not crucial, it was thought that advances in NER methods a system used would lead to better results under formal evaluation.

Many systems were found that had additional problems that negated their inclusion for formal performance review. This included being severely out of date or no longer supported by any form of body. Many systems that were previously available became pay-for or were bought up by bigger corporations that were not open to their use. Some systems required the use of outside programs or other software which was not available, or required an extensive knowledge of unknown languages/systems. This set of requirements forms the basis for pre-selection of NER systems as detailed in Research Goal 1.

3.2.2 The experimental data set

There are many guidelines to creating an experimental data set, some of which were not feasible due to time constraints. While automation of evaluation was a possibility for some systems, others required manual calculation which proved to be a large task. As there was only one assessor several limitations were accepted. It is also noteworthy that this evaluation is only a precursor to larger tests in an attempt to narrow down our choice of systems.

First, only twenty-five of the original one-hundred abstracts were used for this evaluation. While not extensive by any means it was felt that this gave a good indication of what a system could do and the performance level we would be working with. It also allowed us to conduct more analysis for each system giving a deeper understanding of a systems capabilities and limitations, rather than a shallow overview.

There was no attempt to ensure the existence of at least one mention in every abstract. Although this can lead to a skewed statistical result it was felt that because our analysis was sufficiently deep, these numbers would be understood in the context of a systems limitations. In addition, with an understanding of the problem area we are working in it is a certainty that this situation will arise when the system is put into practice. In essence this made our data set more true to our actual data set than it would otherwise be.

3.2.3 Manually annotated "Gold Standard"

Manual annotation is a very domain knowledge intensive task and is the result of many iterations. To assist in this an internal discussion page on the projects wiki was created to assemble a set of rules and guidelines. These rules were added to based upon experiences while annotating the one-hundred abstract corpus. Later these rules were discussed and resolved in meetings with all members of the project team. This evaluation was run using the first iteration rule-set and annotated corpus.

Annotation was done with the use of the Brat Rapid Annotation[3] tool. This tool had been used in previous work and was a good fit for this study. Brat is available both to be

[3] http://brat.nlplab.org/

run on a web server for cooperative annotation and locally in the absence of web server resources. A simple and intuitive graphical user interface allows for speedy annotation while still maintaining an overview for precision. The system allows for multiple sets of annotations over the same database, which allows for an iterative approach to annotation and a history of past sets. Finally the annotations themselves are stored separately which allows for automated evaluation which this experiment took advantage of in several cases.

This Gold Standard represents the "goal" in mind and would be the perfect result set of an NER system.

3.2.4 Evaluation of NER systems

System evaluation is performed by inputting the abstract corpus into each system and comparing the results given versus the gold standard. To be given a perfect score a system would have to produce exactly the same identifications as in the gold standard, with no missing or additional results.

A graphical representation of this process can be seen in Figure 3.1.

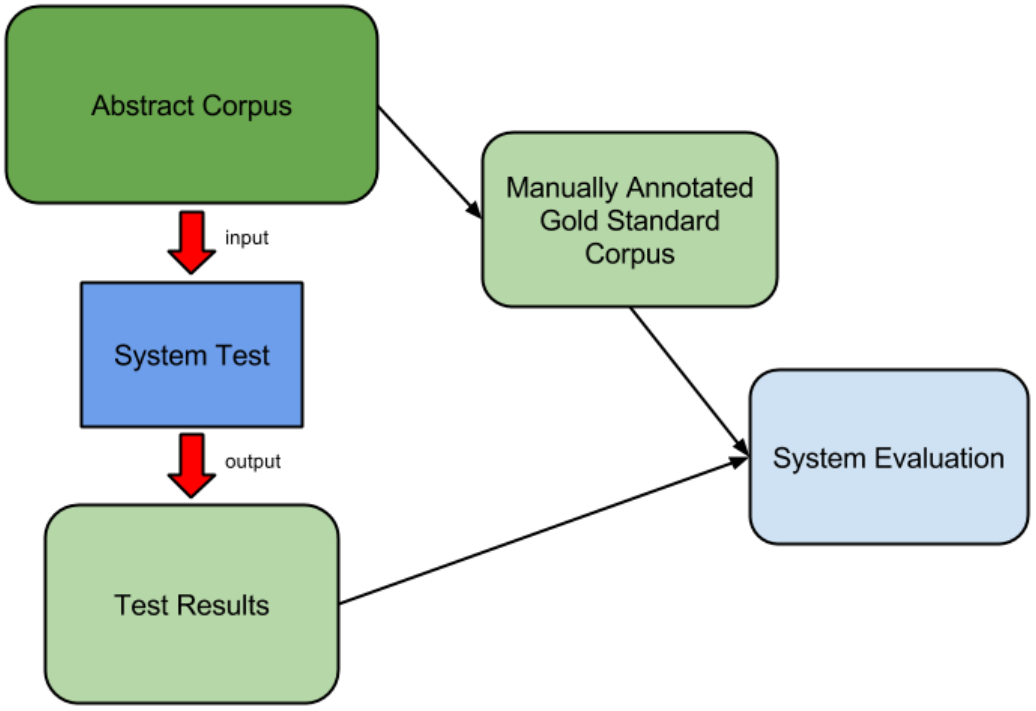

FIGURE 3.1: A simplified biological pump

There are many types of evaluation statistics that can be generated for an NER system. In the interest of focus, three important measures of performance were chosen with the addition of two types of system error.

Two of those measures, Precision and Recall, form the basis of much deeper analysis and are a good measure of performance(Powers, 2011). F-measure is used as an overall statistical performance measure which, while shallow, gives an overview of how well a system does. The identification of Type I and Type II errors (See Section 3.2.4) is a commonly used analysis to identify system weaknesses.

Precision, Recall and F-measure

Precision and Recall are two sides of performance evaluation which measure how well a system is able to identify correct terms, and how well it is able to identify as many terms as possible. Put another way "Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity."(Wikipedia, 2013)

Often the set of terms that we wish the system to identify is called the set of relevant elements. The relation of these terms can be seen in Figure 3.2 from Wikipedia(Wikipedia, 2013).

To calculate precision and recall three variables from the examination of system returned results versus the gold standard are needed. True positives are entities the tested system returned as a result and were also annotated in the gold standard. False positives were entities the tested system returned as a result which were not in the gold standard. False negatives are entities that the gold standard annotated as a valid entity but the tested system did not return. True negatives are uninteresting as this is the set of results not annotated which the system also did not return as a valid result. See below under "Identifying Type I and Type II errors" for a more detailed explanation.

From these variables precision and recall can be calculated, as seen in Equation 3.1 and 3.2 respectively(Wikipedia, 2013);

$$Precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{3.1}$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \tag{3.2}$$

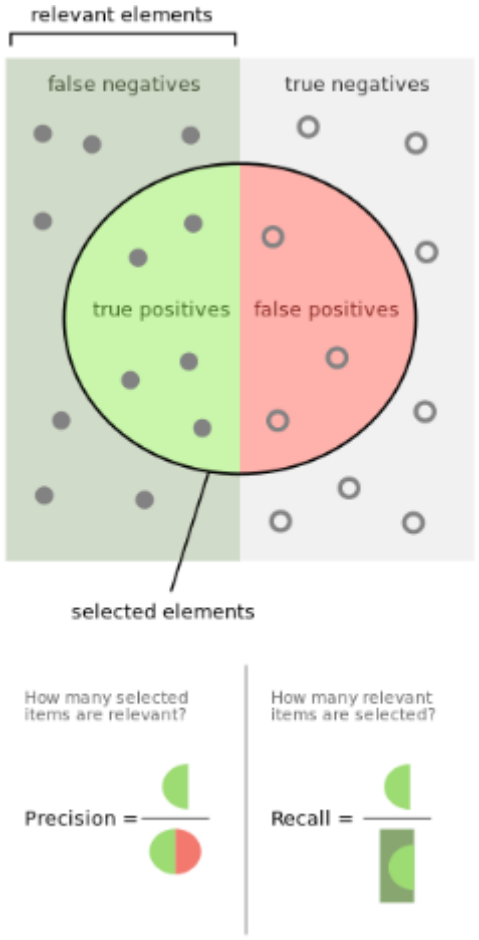

FIGURE 3.2: Precision - Recall relation

F-measure is the harmonic mean of precision and recall, typically used as an overall evaluation score. See Equation 3.3. As both precision and recall can be seen as the ratio of hits versus misses, or a *rate*, the harmonic mean gives an average of these. It is notable that given a precision and recall of zero, F-measure calculators will return an F-measure score of 1 or 100%. This is equitable to recall and precision being set to 100% given a divisor of zero and was discussed with OC project team members. As this is technically correct it was allowed in this work, as the frequency of this happening should be relatively small, but is taken under consideration during evaluation.

$$F - measure = \ 2 \ * \ \frac{Precision * Recall}{Precision + Recall} \tag{3.3}$$

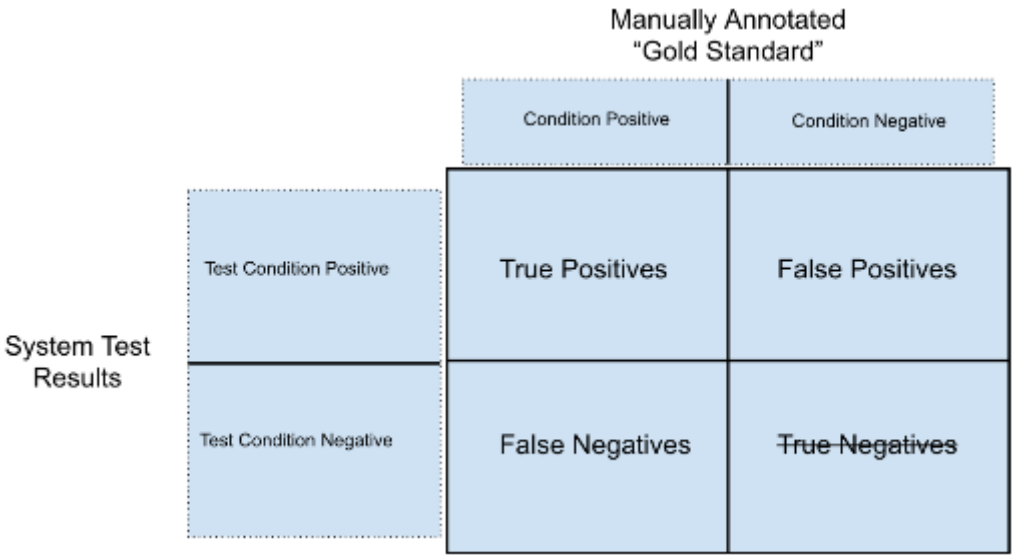

FIGURE 3.3: Type I and Type II error relation

Identifying Type I and Type II errors

The identification of Type I (*false positive*) and Type II (*false negative*) errors gives a sense of the cases when the system is not performing as expected. This requires examination of both the gold standard data set and system returned results against one another. A Type I error occurs when the evaluated system returns a result which are not marked in the gold standard. These represent "false hits" and negatively impact the precision score of the system, and the reason they appear is varied. Sources of Type I errors could represent a failure of the gold standard to correctly annotate all instances of an entity type, in this study it was found that systems which were to find chemical compounds would annotate "water", while the gold standard did not. Another source of Type I errors is ambiguous words that the system is unsure about, often when the word is capitalized but the system cannot resolve what the word is. Capitalized words are regarded by most systems as important, and they will often be annotated regardless.

Type II errors are when the system has not annotated a term which is annotated in the gold standard. This can be seen as "misses" by the system. Any time a system has a Type II error this affects recall, or the coverage of terms expected to be annotated. Again, the sources of these errors can vary, from a system unable to handle the scope of data envisioned, to a problem with term identification or segmentation. A common Type II error encountered during this study's evaluation was the inability to correctly identify terms that were separated by a white space.

Conversely, a term which is annotated by both the evaluated system and the gold standard is a *true positive* or "hit". True negatives, when the system does not annotate a term and we do not wish it to, are not considered by this evaluation scheme. The relation of these terms is visualized in Figure 3.3 to give a sense of how they fit together.

Additional considerations

Some performance errors are not taken into account in a strict precision/recall/f-measure evaluation but have an impact on system performance. A system that covers more entity types than was begun with in the gold standard is such a consideration. Where possible this study will comment on and refine a system to more closely match the intentions, e.g. the improvements to Oscar3. At times the capabilities and flexibility of a system may result in the selection of that system over ones that have a superior performance, exemplified in Linnaeus2. Throughout this evaluation these sources of error are considered and covered under the discussion section for each system in Chapter 5.

Chapter 4

Experiment

Chapter four is an extensive discussion of the choices made in regards to which NER systems to formally evaluate (Research Goal 4) from the original pool of discovered systems (Research Goal 2). Each entity type is considered individually in Section 4.1, including entity-specific requirements and reported performance of the NER systems on various external corpora. Further is a description of the experimental data set used in evaluation (Section 4.2) and the formal guidelines used in creating the manually annotated gold standard (Section 4.3)(Research Goal 3).

4.1 System selection

4.1.1 Chemical systems

Four systems were originally considered for this evaluation; CheNER(Usié and Valencia, 2014), Chemspot 2.0(Rocktäschel and Leser, 2012), Oscar3(Corbett and Murray-Rust, 2006), and Oscar4(Jessop and Murray-Rust, 2011).

Upon initial testing CheNER was not able to annotate data through the use of the Graphical User Interface. The system simply returned no results from several different abstracts which were chosen to contain different types of chemical compounds and had chemicals represented in different forms. Examination of this error did not reveal any cause but subsequently CheNER was dropped from formal evaluation.

Chemspot 2.0 was chosen as the system was relatively new, last updated in 2014, and was accompanied by a review paper, in addition to being under a Common Public License. From the Chemspot website[1] the system is described as;

"a set of tools for named entity recognition and classification of chemicals in natural language texts, including trivial names, abbreviations, molecular formulas and IUPAC entities. Since the different classes of relevant entities have rather different naming characteristics, ChemSpot uses a combined approach of employing a Conditional Random Field and a dictionary, as well as pattern-based recognition, a classifier model and several methods for consolidating all annotations. ChemSpot also performs named entity normalization by assigning identifiers from numerous chemical databases. It achieves an F1 measure of 79.0% on the SCAI corpus."

Several factors are noteworthy here, mainly that the system handles different types of chemical information such as names and formulae. The hybrid approach of using Conditional Random Fields, dictionaries, and pattern based recognition is an approach which is fast becoming an industry standard. Chemspot 2.0 also includes span data alongside the system annotated entities, which helps in being precise with evaluation scores. Automatic recognition of ChEBI identifiers and InCHI formulae are secondary, but useful, pieces of information that are noted as being useful in later stages. These

[1] https://www.informatik.huberlin.de/de/forschung/gebiete/wbi/resources/chemspot/chemspot

factors, together with the F1 measure on the SCAI corpus fueled our interest in this system.

Oscar3 was chosen as a shallower Named Entity Recognizer with a focus on chemical entities. The advantage of this system is that it can automatically annotate found entities with chemical structures from the ChEBI ontological database. It also provides more information alongside annotated hits such as confidence level, ontological ID numbers, InChI structures, and language position information. As this was a shallower NER system more false positive hits were expected, but pre- or post-processing was thought to alleviate many of these errors. Oscar3 does not include any span data, however the results are given in-line with the article text and in XML format. With sufficient automation and parsing this span data would easily be retrievable where necessary.

Oscar4 is the updated version of Oscar3 and was considered in order to evaluate more recent trends in NER systems. This system however is still in the early stages of development and consists mainly of a core library toolset with an API. As formal evaluation and use would require the creation of a wrapper program it was eventually set aside for consideration at a later date. This was due to time constraints and the amount of work it would take to create such a program simply for evaluation.

In the end it was decided to select both Chemspot and Oscar 3 for formal evaluation. This would allow a more focused chemical NER system versus a shallower one, and both seemed to be a good fit for the Ocean-Certain requirements. Since Oscar3 is a more general NER system, it would naturally return a result set that would have an abundance of errors. To alleviate this problem it was decided that evaluation of the result set would happen in three stages; first the full result set would be analysed, then the result set of just returned elements, finally the result set that consisted of elements and any hit above a 90% confidence level. This would allow for the evaluation of Oscar3 in conditions that were closer to Chemspot, as well as give an idea of what other capabilities Oscar3 possessed.

4.1.2 Biological species systems

The three systems selected for testing were Linnaeus 2.0(Gerner and Bergman, 2010), SPECIES(Pafilis and Jensen, 2013), and OrganismTagger(Naderi and Witte, 2011) which all showed a reasonable aptitude for species identification.

Linnaeus 2.0 was released in 2011 and is a general purpose dictionary based matcher. This system was interesting as it is very flexible both in terms of input and output. Greater control over the system is given to users by way of having direct access to the dictionary used for matching. There is also an option to run Linnaeus as a server instead of locally, which allows for load balancing and the processing of a very large amount of documents in a timely manner. With this scalability in mind, Linnaeus 2.0 was a good match for both the current and future needs of Ocean-Certain. Original testing of Linnaeus 2.0 was done using the internal general term matching dictionary which contains the top ten thousand mentions from Medline.

SPECIES is a system developed after Linnaeus and uses, in part, its naming conventions for binomials but also includes abbreviations, common names, and acronyms. Using the NCBI Taxonomy it is also reported to be able to handle misspellings which can confuse entity identification. The developers describe a command-line tool that is significantly faster than Linnaeus. An interesting note was that SPECIES was developed on an abstract-based corpus and it was thought this could be useful if Ocean-Certain decided to pre-process new documents based upon their abstracts.

OrganismTagger uses a pipeline based system called GATE which was developed for users to be able to build their own system using modules. This modularity would give a large amount of flexibility to the system and its modular nature would be easier to keep updated or fix as need be. OrganismTagger is a hybrid system which uses both rule-matching and machine learning to achieve entity recognition and the performance of such systems was desirable. Other considerations are the ability to normalize scientific names as well as strain and common name detection. The system's performance on the OT (precision 95%, recall 94%) and Linnaeus-100 (99% and 97%) corpus are both very intriguing. There was interest in seeing how the OrganismTagger system would perform in a different problem space.

4.1.3 Location systems

While location identification is one of the older entity extraction tasks it was equally challenging finding systems to match the Ocean-Certain specifications. For this project CoreNLP(Manning and McClosky, 2014), OpenNLP(Baldridge, 2005), and IllinoisNE Tagger(Ratinov and Roth, 2009) were selected for evaluation.

CoreNLP is from Stanford Natural Language Processing Group as a model based classifier with a long continuous development history. This system uses CRF classification and is able to use custom built models for entity identification. As one of the leading systems in use in entity extraction it was a natural choice for consideration. In addition the project team has previous experience with CoreNLP which is a valuable asset.

Apache OpenNLP is a Maven based system for common NLP tasks and is a command line toolkit. This project is a collaboration from volunteers and the Apache Software Foundation, developed with the idea of being a base system for larger systems to be built upon. Using perceptron and maximum entropy machine learning techniques OpenNLP is a pipeline system which can use trained models to identify entities. It is possible to test each system function using the command line but can also be accessed through an API for full system runs.

Finally IllinoisNE Tagger is a gazetteer system the can either identify enamex entities or a larger 18-label entity set. The gazetteer is based upon Wikipedia, word class models, and non-local features to create a diverse entity tagger system. When considering systems for formal evaluation this was considered a strength due to the varied nature of our target entities.

4.2 Experimental data

Initial testing for the chemical systems will consist of twenty-five article abstracts from the larger ten-thousand corpus data set. The evaluations will all use the same twenty-five abstracts in order to evaluate their performance against each other. This initial testing is to refine and streamline the evaluation process, as well as to begin refinements of the Oscar3 system. Refinement was necessary as Oscar3 is a shallow NER system, meaning it is designed to cover a lot of entity types and thus would not perform well against our

evaluation scheme out of the box. After the initial testing run the abstract corpus will be expanded to one hundred abstracts with the same guidelines. This would allow for more data points in order to give a better overview of the tested systems capabilities.

These article abstracts were chosen for their focus around the biological pump and/or relevance to climate change research. However, these articles were chosen at random within those guidelines in order to avoid bias. As explained in Section 3.2.3, these abstracts are not screened to always include at least one mention of an entity. This decision means recall and precision scores can be skewed (as 0/0 is a 100% score) but as this would affect every system equally, was disregarded. These articles covered several different fields of study including microbiology, geoscience, nature, and microbial ecology.

4.3 Manually annotated "Gold Standard"

To start, a basic annotation set was created to be the basis for subsequent annotation runs and to begin discussion with OC project team members over specific outlying examples. The initial annotating found the three chosen entity types (chemical compounds, biological species, and locations) with the intention to find only specific entities, no general names or types. As there is no one definitive biological species taxonomy, the NCBI Taxonomy[2] was used to check if a suspected word was a biological species or not. Although this limits the annotations to only entities found within this taxonomy, it is considered a comprehensive one and used extensively. Modifiers such as "human-induced" or "iron-binding" would not be annotated at first, in order to asses the capability of systems to capture these more complex entities but not negatively impact systems performance who cannot. Further, a modifier can change what type of information is given in a very complex way, "water-column" is describing a *location* and not the compound *water*. "Low-oxygen" is a term describing a detail of an environment, and can be seen as an event. Given the large amount of work that would be required to analyse and adapt to such changes, together with the relative sparseness of these types of words, exclusion was determined to be the best path. The rule-set for the first annotation run is summarised here and was based off the GENIA Corpus Annotation Guidelines[3] as well as the CRAFT Corpus Annotation Guidelines[4].

[2] http://www.ncbi.nlm.nih.gov/taxonomy
[3] http://www.nactem.ac.uk/genia/
[4] http://bionlp-corpora.sourceforge.net/CRAFT/

Annotation Run 01 - Only specific compounds, geographic locations, species. No modifiers.

- Species names that are shortened will be annotated. Ex. C. Watsonii, P. Globosa.
- Complex compound formulae will not be annotated. Ex. $C_8H_{10}N_4O_2$
- A mention that contains a less specific entity as a subunit, annotate the larger entity only. Ex. carbon dioxide as "carbon dioxide", not "carbon".
- Earth will be annotated.
- Human will be annotated.
- Species genus, phylum, or class groups will not be annotated. Ex. Synechococcus, Prochlorococcus
- Compound acronyms will be annotated. Ex. DMS (dimethylsulphide), DMSP (dimethylsulphidepropionate)

The second run of annotation includes an expansion of the biological species annotations to include genus/phylum/class names. This was done to facilitate and evaluate the improvements of Linnaeus2 (Section 5.4) as a more general species NER system. Two other refinements were included, "sugar" would henceforth be annotated as a chemical compound, as would "water". While "coral" was annotated by many chemical systems, it was chosen to not be included as this would cause confusion between the chemical meaning and species.

In the final iteration, general species names such as "phytoplankton" and "human" would be annotated. Modifiers connected to relevant entities not annotated in the first run would now annotate the entity only. For example "human-induced" would now annotate "human". Continuing with the improvement of Linnaeus2, bacterial genus/phylum/class would be annotated (e.g. Actinobacteria, Alphaproteobacteria) in order to assess the coverage of the dictionary modifications for the system.

Chapter 5

Results and Discussion

This chapter contains the results from each system evaluation and a discussion detailing the strengths and weaknesses found (Research Goal 5). Where necessary it will detail improvements and adjustments taken under evaluation in order to provide a firm grasp of each systems' capabilities. In each evaluation the system weaknesses are outlined, with a more detailed examination of Type I and Type II errors given. At the end of each section is a summary showing average metric scores and a graphical presentation in order to understand how the systems performed against each other within each entity type. In Section 5.4 this paper examines the improvements made to the Linnaeus2 system through expansion and improvement of the dictionary used for entity recognition via gazetteering.

5.1 Chemical systems

Initial testing of Chemspot 2.0 and Oscar3 on an experimental abstract corpus consisting of twenty-five articles revealed the extent of Oscar3's shallow NER implementation. After consideration it was decided that the coverage from this system could possibly be utilised and post-processing techniques would be implemented to asses how the system performed as a purely chemical NER tool. Testing of this implementation was done by filtering out different variations of the full Oscar3 output and evaluated versus previous scores (Figure 5.1. The goal of this filtering was to keep as much recall as possible while improving Oscar3's precision. Results of both Chemspot 2.0 and Oscar3 on the twenty-five abstract corpus (along with the different variations of Oscar3) is discussed in Section 5.1.1. Finally both systems are tested on the full one-hundred abstract corpus and will be discussed in Section 5.1.3.

5.1.1 25 abstract analysis

5.1.1.1 Analysis of Chemspot 2.0

Results of this evaluation can be found in Appendix A, Table A.1.

Chemspot 2.0 was able to pick out chemicals with an average amount of precision (67.1%) and a reasonably high degree of recall (77.6%). The F-measure score of 71.9% is close to the reports performance on the SCAI Corpus (79.0%)(Rocktäschel and Leser, 2012) with the reduced performance being attributable to being in a more natural context and being evaluated against a smaller amount of documents.

Both common chemical names and their formulas were often identified, even in the cases when being connected to a non-entity such as "Low-oxygen". Compound names were also annotated fairly consistently even with variations in spelling.

Overall this system performed to expectations, however it was unable to identify compounds which were split by white spaces. This would be a major shortcoming for the Ocean-Certain system, and unless it is possible compensate with pre- or post-processing, this would disqualify Chemspot 2.0 from use.

The most common false positive (Type I) error was the labeling of the entity "water". Further discussion is warranted on whether "water" should be labeled as an entity, or whether this result should be removed by post-processing but not counted against the system. This is because it is a compound, yet such a general and widely used one that the information benefits of taking this entity further are marginal at best. For the same reasons "sugar" will also need to be discussed and its inclusion into the next round of annotation decided.

System Type II errors showed a large weakness in Chemspot as the system is not able to identify compounds which are divided by a white space. This can be seen in the missed entities "carbon dioxide" (articles 10, 15, 18, 20, 21), "nitrous oxide" (article 17), and "hydrogen sulphide" (article 19). It seems Chemspot 2.0 simply annotates the individual chemical instead of analysing surrounding words for larger compound names. While this error could possibly be alleviated by post-processing steps, it would require a large amount of time and resources to go back through each annotation result and find out if it was possibly a part of a larger entity. This is a problem of information chunking as described in Chapter 2 Section 2.1.4.

There are several additional considerations as the first annotation run did not annotate compounds or chemicals that were connected to non-entities by a hyphen, e.g. "low-oxygen". Without previous knowledge of system capabilities this was decided in order to simplify evaluation as much as possible and give as wide an array of systems a chance to compete. The entity "DMSP" (dimethylsulphide propionate) is also understandably an entity that an NER system would commonly miss as this acronym could stand for a wide range of things. Pre-processing to add this acronym to a pattern based dictionary, or post-processing for the identification of specific acronyms such as these may ultimately be the best way to catch such examples.

5.1.1.2 Analysis of Oscar3 version 1

Results of this evaluation can be found in Appendix A, Table A.2.

As expected the Oscar3 system identified a large amount of entities, from chemicals and compounds to acronyms and additional information. While this amount of results gave a low precision score of 31.1%, it naturally gave a higher recall of 92.4%. With an

overall F-measure of 38.7% it is easy to see that this system, by itself and in this form, would not be very suitable for Ocean-Certain. However, leveraging the coverage of the system could eventually allow for cross-validation of other systems or use as a contextual information generator.

The large number of false positive results is immediately apparent but not unexpected. Most commonly "marine", "nutrient", "water", and "DOC" are annotated in error. Many examples of capitalized names being annotated are seen ("Actinobacteria", "Flavobacteria", "Porites") which are not useful in this context but could possibly be used to cross-check for species later. One of the major errors of the Oscar3 system is the annotation of "In" (articles 2, 4, 16, 23) as an element. This seems to occur when "In" is at the start of a paragraph and could present some difficulty to verify later as an actual element in the text. Further, "P" is being annotated as phosphorus (article 22) which arises when a formal name is shortened, as in "P. Globosa". As with the Chemspot system "water" is an error which will possibly be annotated at a later stage, however for consistency it is considered an error here.

Very few results were missed, only four from the entire data set. One is a misspelling ("dimethyl sulphide") which is not expected for a system to pick up. Two compound names separated by a white space ("nitrous oxide" in article 17, "hydrogen sulphide" in article 19) were also missed.

5.1.1.3 Analysis of Oscar3 version 2

Results of this evaluation can be found in Appendix A, Table A.3.

This result set is only annotations Oscar3 marked as "Elements" or having compound formulae in the form of InCHI[1] or SMILES[2].

To assess if Oscar3 could perform in a more specific context versus Chemspot this analysis was made on a subset of the total results returned. Trading some recall (84.9%) in order to elevate the precision (65.5%) score. With an overall F-measure of 66.2% Oscar3 does not perform as well as Chemspot but still has a higher recall rate. However the weaknesses leading to Oscar3's low precision score can nearly be halved by the inclusion of "water" in the gold standard.

[1] http://www.iupac.org/home/publications/eresources/inchi.html
[2] https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

Of the false positive, Oscar3 annotated a lot of instances of "water" and "waters" (articles 2, 3, 4, 8, 10, 12, 13, 15, 16, 18, 22). Twenty-one errors are "water" which would significantly change Oscar3s precision score and overall F-measure. As "In" and "P" were annotated as elements these are still present as false positives.

False negatives increased as a result of this output filtering, notably missing some compounds such as "dimethylsulphide". It is interesting to see that "dimethylsulfide" is still correctly annotated as an entity and this may be a result of dictionary bias or common-name mistakes. Another error that is introduced is the system missing compound short names, "CO2" (articles 12, 20, 24) in some instances. "DMS" and "DMSP" are also missed as these are acronyms that may not be technically equated with their longer forms.

These results are closer to what is expected of an NER system and it could possibly be used together with either pre- and post-processing or crossed with results from the entire result set.

5.1.1.4 Analysis of Oscar3 version 3

Results of this evaluation can be found in Appendix A, Table A.4.

This result set is annotations Oscar3 marked as elements, compound formulae in the form of InCHI or SMILES, plus any annotation that had a confidence level of 90% or higher.

This set was an evaluation of how confident Oscar3 was annotating entities which were not specifically connected to an element or compound. Precision was only improved by 0.1% (65.6%) while recall recovered a more significant amount (90.8%) versus the elements-only result set of 85.9%. The overall F-measure recovered significantly as well to 69.9% which is reasonably close to Chemspot's score of 71.9%.

The inclusion of high confidence annotations introduced a few new false positive results versus the elements only set. "Rhodophyta" and "Chlorophyta" from article 23 were re-introduced along with "anoxygenic" in article 16.

This subset was able to recover "dimethylsulphide" as a result from article 3 as well as "Dimethylsulfoniopropionate" from article 16. The system still misses "CO2" from quite

a few articless, recovering some instances but not all. "DMSP" is still not recovered by the system with this subset as anticipated, being a very unofficial acronym.

5.1.2 Chemical Compound Evaluation Comparison (25 abstract)

Figure 5.1 and the accompanying data table (Table 5.1) shows a comparison of evaluation scores from the refinement of the Oscar3 system over the twenty-five abstract corpus using Chemspot as a guideline.

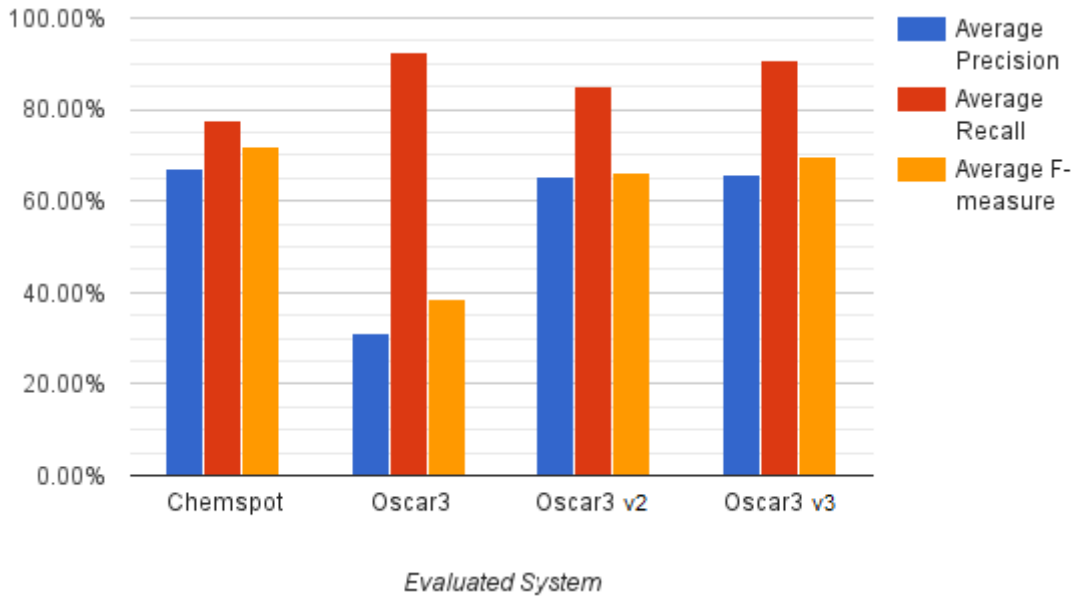

FIGURE 5.1: Chemical NER Evaluation Summary (25 abstract)

System Name	Avg Precision	Avg Recall	Avg F-measure
Chemspot 2.0	67.1%	77.6%	71.9%
Oscar3	31.1%	92.4%	38.7%
Oscar3 v2	65.5%	84.9%	66.2%
Oscar3 v3	65.6%	90.8%	69.9%

TABLE 5.1: Chemical NER Evaluation Data (25 abstracts)

5.1.3 100 abstract analysis

5.1.3.1 Analysis of Chemspot 2.0

Chemspot's performance over the one-hundred abstract corpus was comparable its performance on the twenty-five abstract corpus, with a precision of 67.7%, recall of 76.3%, and F-measure of 65.6%. These results show an ability to identify chemical compounds in a large array of instances, however the shortcomings while applied to this studies' scenario merits doubt as to Chemspot's viability. An examination of system errors reveals a tendency to annotate capitalized terms even with no corresponding ontological connection, along with abbreviations that have no discernible connection to chemical elements. Further, the system misses large numbers of entities that occur over two or more words, as seen in the twenty-five abstract evaluation.

Some errors given by Chemspot are attributable to choices made when developing the annotation guidelines and can therefore be ignored. The false positive result "water" is a common example of this, but at only 33/263 instances does not require specific handling to avoid skewing results versus other systems. Similarly, the ability of the system to find entities connected to modifiers, e.g. "low-oxygen" or "water-column" in articles 10 and 18 respectively, is a choice by the annotators to ignore as explained in Chapter 4 Section 4.3.

Elements of the system's pattern recognition can be seen in the annotation of abbreviations previously connected to a chemical compound. The false positive "OMZ" (Oxygen Minimum Zone) found in article 17, "BGE" (Butyl Glycidyl Ether) in article 58, show how the system can connect these concepts together. However the exact conditions necessary for this process need to be examined as it was not successful in every case. Chemspot did not annotate "DMSP" (Dimethylsulfoniopropionate) seen in articles 16 or 75, while "DMS" (Dimethylsulfide) in article 16 was. While not a goal of this evaluation, it should be noted the ability of a system to perform this kind of connection for future studies as this capability may be leveraged in other systems.

An oddity noticed in Chemspot is the identification of "Trichodesmium" in articles 32 and 78. This study could not find any connection to a chemical compound and research found that the identified word is a biological species. The term "Peninsula" was also often misidentified (articles 26 and 33) and may indicate either attempts to annotate

words that may be misspelled, or an over weighing of capitalized terms in decision making.

Twenty false positives come from the apparent inability of the system to connect larger entities together. Continuing on from the twenty-five abstract list we see examples in article 29 with "calcium carbonate" and article 55 "nitrous oxide". This is the biggest failing of the Chemspot system, and while post-processing techniques may help to alleviate the issue it is not believed to be viable in this situation. This is because Chemspot does not identify the later parts of these entities in any way, e.g. "carbonate" or "oxide", meaning an attempt to connect them together would require a full run by another NER system to even begin connecting these entities together.

In cases where the system missed annotations we see one very large trend besides the exclusion of multi-word entities discussed above. A large amount of these Type II errors, one-hundred and fourteen of two-hundred errors (57 %) are the non-identification of simple chemical element acronyms. These simple acronyms are often "C" (Carbon), "N" (Nitrogen), and "P" (Phosphorus) as seen in articles; 32, 34, 35, 39, and more. It is possible the ambiguity of these acronyms causes the system to not have enough confidence in classifying them, however with the density of such acronyms in an article the pattern recognizer should be able to pick these up. In addition, Chemspot touts the cross validation of several types of classifiers (CRF, pattern-based, dictionary etc. discussed in Section 4.1.1) which should be able to properly handle this identification.

With the double issue of missing multi-word entities and lacking the sensitivity to correctly identify simple acronyms, Chemspot 2.0 is not a good fit for the purposes of the larger Ocean-Certain project. The system does show some aptitude for finding larger acronyms which are connected to chemical compounds, and this capability could be studied and improved upon in further work.

5.1.3.2 Analysis of Oscar3 version 3

Oscar3 v3 was able to perform comparably or better than Chemspot in all areas, and though this is the result of filtering Oscar3's output, it did not require advanced system changes or complex analysis to do so. The output filtering extracts output only where entities are annotated as Elements, have an InCHI/SMILES compound formula, or an identification confidence of 90% or greater. With these modifications, the precision attained was 67.4%, recall 86.6%, and the F-measure 67.2%.

As with Chemspot, "water" was annotated by the system but counted as an error to allow for comparison between systems. In future work this entity may either be an allowed entity type or be filtered out of the output.

Oscar3 does show a tendency to annotate words which are technically element acronyms, such as "In" (Indium) or "At" (Astatine), when they are capitalized at the beginning of a sentence. This can be seen in a large amount of articles and comprises 36/199 (18%) of the false positives reported by the system. Analysis of the system output does reveal that while these errors are marked as elements, they often have a low confidence score which could be used to filter them out, however this is not a comprehensive solution. A better alternative to rectify these mistakes would be the use of pattern recognition to detect when these occur at the beginning of a sentence, as in normal language you would not start such an acronym.

A second trend in false positive results is the annotation of biological species which are capitalized in the text. We see this in article 23 ("Rhodophyta", "Chlorophyta"), several articles with "Trichodesmium" (32, 78, 95), and article 37 with "Daphnia". While the cause of this is unknown, all of these annotations had a greater than 90% confidence level. In the case of "Trichodesmium" the ending -esmium may be close enough to common element endings with variations of -ium to warrant annotation by the system. As the larger OC system will also be looking for biological species, it may be worthwhile to have a cross-reference check with the other NER systems in order to decide which entity type these belong to.

In opposition of Chemspot, the Oscar3 system shows a weakness in annotating slightly more complex chemical compound acronyms e.g "O2", "CO2", "N2", while it does well identifying simpler acronyms. Oscar3 does come with the option to modify their string

matching algorithms however, so this should be a relatively simple task to solve. It can also be noted that the Oscar4 system will be further along in development by the time of writing and may be more adequate in identifying these missed acronyms.

The coverage of the Oscar3 system lends itself to heavy modification and the extraction of results as necessary by users, which sets it apart in terms of viability for the OC project. Analysis of the weaknesses of this system show more room for improvement, with the implementation of such being relatively inexpensive in terms of time and resources. These factors make Oscar3 an appealing choice for chemical entity recognition and is recommended by this study.

5.1.4 Chemical Compound Evaluation Comparison

Figure 5.2 and the accompanying data table (Table 5.2) shows a comparison of evaluation scores from the examination of chemical compound NER systems examined over the full one-hundred abstract corpus in this work.

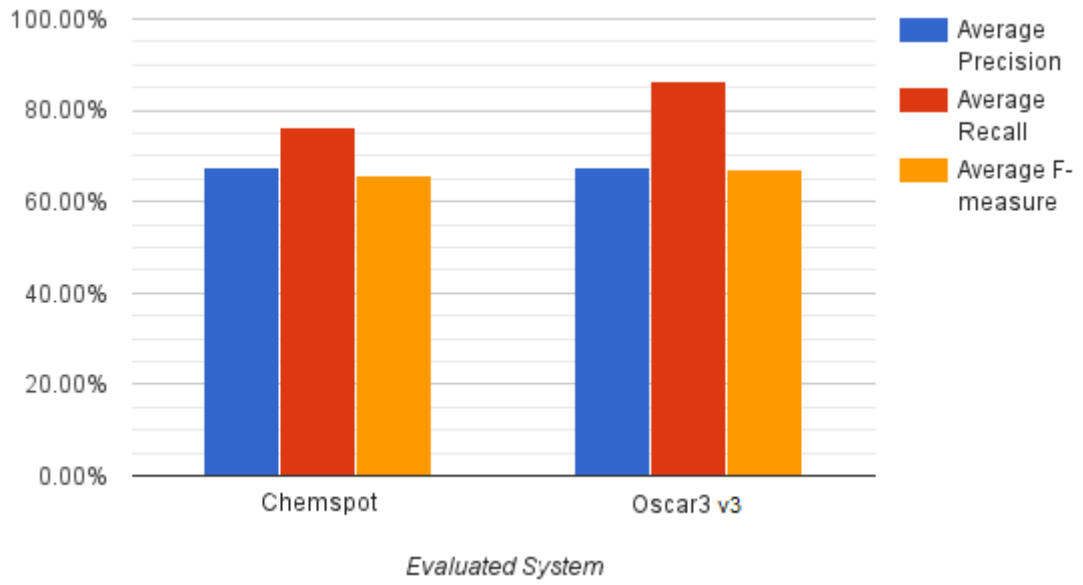

FIGURE 5.2: Chemical NER Evaluation Summary (100 abstracts)

System Name	Avg Precision	Avg Recall	Avg F-measure
Chemspot 2.0	67.7%	76.3%	65.6%
Oscar3 v3	67.4%	86.6%	67.2%

TABLE 5.2: Chemical NER Evaluation Data (100 abstracts

5.2 Biological species systems

5.2.1 Analysis of SPECIES

The SPECIES NER system was able to identify all the entities annotated in our gold standard corpus, with few false positives. With a recall of 100.0% and a precision of 97.7% this system was able to find nearly everything this study required. However it should be noted that this system is based on the NCBI taxonomy, which provides more context for these scores as the gold standard is also based upon NCBI. In addition, biological species that matched our annotation guidelines were few over the one-hundred abstract corpus, only 64 entities found versus chemical compounds of which there were 693.

Of the false hits reported by the system, four are the result of the annotation guidelines not allowing for modified entities such as "human-induced" in article 21 and "gamme-Proteobacterium" in article 34. The comprehensiveness of the NCBI taxonomy and the SPECIES system allowed for the inclusion of acronyms associated with species that are not normally considered, e.g. viruses. Article 35 discusses a virus "PpV" which is specific to the P. pouchetii bacteria and was thus annotated by SPECIES.

While this system, on the surface, seems more than adequate for Ocean-Certain it should be judged in the context given. Upon further examination of SPECIES it is easy to gain access to the individual files comprising the dictionary look-up terms and is extensible by a user should they wish. However the fact that the system does not use any sort of link to the NCBI taxonomy in their species list is something that should be considered when choosing to do so.

5.2.2 Analysis of OrganismTagger

OrganismTagger(OT) is another NER system based around the NCBI taxonomy, but is implemented via the GATE[3] development platform. GATE provides a modular, pipeline system for programs to be built and executed upon, provided they can build programs compatible with the necessary modules they wish to use. This adds another level of complexity and knowledge needed for developers not familiar with the system, but could

[3] https://gate.ac.uk/

be used if given the time or the existence of previous experience. OrganismTagger uses a hybrid detection system, combining rule-based NER together with machine learning techniques to perform its task. The connection to the NCBI database gives the system an edge in this evaluation, as the annotation gold standard was also based off of this same database, but the performance of OT is still high nonetheless. With a recall just below SPECIES at 99.0%, OT has a lower precision of 88.0% giving it an overall F-measure of 89.0%.

The system's false positive errors are nearly all (17 of 20) instances of the NCBI database containing very general names for certain biological species. As we can see in articles 17, 30, 36, 57, 61, 80 and 86 the entity "mum" is annotated, because it is a common name for "Chrysanthemum". Similarly "Canary" in article 14 is connected to the common canary. These results could be removed from the dictionary OT uses, but this may be a more difficult task given that you would have to modify the GATE module itself to do so. The two other instances of false positives, "marine diatoms" and "marine bacterium" are also common names given under specific species in NCBI. The one instance of "human" annotated from article 28 is the result of the annotation gold standard not accepting "human-induced".

A single article, number 86, was responsible for OrganismTagger not receiving a 100% score for recall in that it missed the entity "Richelia intracellularis" and its shortened name "R. intracellularis" for unknown reasons. It is possible that this was an omission from the dictionary used by OT in error.

The OrganismTagger did quite well on this task even with the low count of possible entities to annotate, but its dependence on the GATE system warrants skepticism towards its inclusion in the larger Ocean-Certain system. Unless the entire OC project decides to use GATE, it would be technically difficult to streamline this system's use and output together with other systems that do not use GATE. In addition, the output formats of GATE do not lend themselves towards easy parsing and can be difficult to handle automatically. GATE's reliance on a graphical user interface also makes it infeasible to design automated systems around it, given the lack of command line tools. It should be noted however that this study did not have previous experience with GATE and it is possible that these issues are surmountable.

5.2.3 Analysis of Linnaeus 2.0

Initial testing of this system uses the internal dictionary provided by the developers, which is composed of the top ten-thousand most frequently mentioned species found in MEDLINE. The MEDLINE[4] database which is a collection of biomedical literature from the U.S. National Library of Medicine. Linnaeus is a gazetteering system using a set of dictionaries to positively identify entities and as such is an easy system to modify. In Section 5.4 this study researches the expansion and adaptation of the Linnaeus dictionary to improve performance while expanding the task domain.

Linnaeus was able to perform at a relatively equal level with SPECIES and OrganismTagger, while using a manually crafted gazetteer dictionary created from MEDLINE articles. This system scored a precision of 85.8%, recall of 95.7%, and overall F-measure of 84.3% which are good scores when applied to a different domain. Again, the low amount of gold standard annotations should be taken into consideration when assessing these scores.

As with OrganismTagger the Linnaeus system annotates common names of species, "mum" and "canary" with the addition of "lake" (a common name for a shovel-nosed lake frog) in articles 16, 30, 37, 59, and 61. This pushes the recall score slightly lower than the previous two systems but should not be regarded as an additional hurdle. A few other general names are also annotated, again similarly to previous systems, with "marine diatoms" and "marine bacterium" marked as well as "bacteriophage" in article 22.

The Linnaeus system misses a total of six species names, "Phaeocystis pouchetti" in article 35, "Methylobacterium oryzae", "Methylosulfonomonas methylovora", "Hyphomicrobium sp" in article 45, "Trichodesmium tenue" in article 78, and "Richelia intracellularis" in article 86. These names are possibly missed due to the internal dictionary of Linnaeus only taking the top 99% of species names from Medline, and could be corrected by a simple expansion of the dictionary. The original assumption that OrganismTagger left out "R. intracellularis" in error may be incorrect given that Linnaeus also did not mark this entity. It is possible that this species name is not commonly accepted enough by the scientific community to warrant its inclusion in a formal dictionary. Consultation

[4] http://www.nlm.nih.gov/bsd/pmresources.html

with biologists may be required for further refinement of dictionaries and guidelines created as to the level of inclusion such a system should allow for non-recognized species names.

Several factors besides performance should be considered when evaluating Linnaeus for inclusion in the OC project system that set it apart from SPECIES and OrganismTagger. The Linnaeus system is able to handle multiple format types (MEDLINE, PMC, BMC, OTMI, text, etc.) for processing which could be expanded to include formats as necessary by the project team. It is flexible in the types of outputs available (XML, HTML, TSV, database) which allows for more options when developing a streamlined system. An important consideration is that Linnaeus can be run as a server, allowing for load-balancing and and multiple clients to be utilizing this resource simultaneously. As the Ocean-Certain system is likely to be running a larger amount of processing tasks and data sources, this capability may become crucial to building a well functioning system.

5.2.4 Species Evaluation Comparison

Figure 5.3 and the accompanying data table (Table 5.3) shows a comparison of evaluation scores from the examination of biological species NER systems examined in this work.

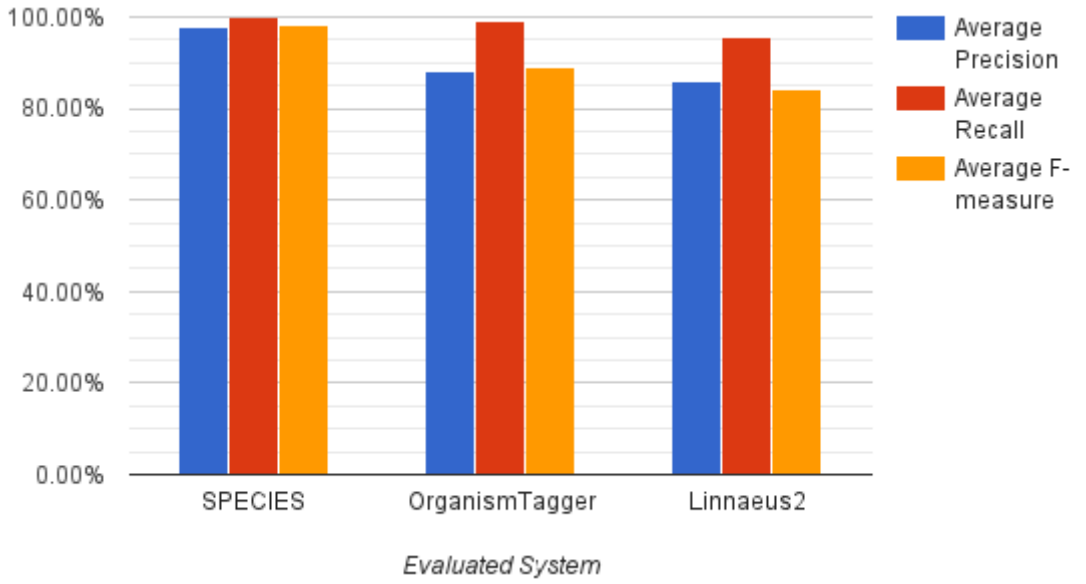

FIGURE 5.3: Biological Species NER Evaluation Summary (100 abstracts)

System Name	Avg Precision	Avg Recall	Avg F-measure
SPECIES	97.7%	100.0%	98.2%
OrganismTagger	88.0%	99.0%	89.0%
Linnaeus2	85.8%	95.7%	84.3%

TABLE 5.3: Biological Species NER Evaluation Data

5.3 Location systems

5.3.1 Analysis of CoreNLP

CoreNLP is the Named Entity Recognizer of the Stanford Natural Language Processing Group's processing tool and is created to identify the enamex entity types Person, Organization, and Location. Output is in an XML format with each entity type marked making for easy extraction of locations for this evaluation. The OC project team has a lot of experience with the use of CoreNLP and it is used in the system created in Work Package One for connecting variables into facts (See Section 2.1.5). This system uses a CRF classifier (explained in Section 2.1.4) and has the capability of users training their own models in order to adapt the system to a different domain task.

Preliminary analysis recognized a problem in parsing this output as the performance of CoreNLP was much lower than expected, with a precision of only 52.0%, and recall of 49.6%. Upon analysis it was found that when a location name occurred over several words, such as "San Marco", the system would correctly identify the different parts but there was no easily definable connection between the two. By recognizing that these entities naturally followed one another in the output, an adaptation was taken to merge these different parts together to form a whole entity. This system change lead to a second analysis of CoreNLP and is discussed in the next section.

5.3.2 Analysis of CoreNLP version 2

With the changes in output CoreNLP performs with a more reasonable degree of accuracy and coverage. Achieving a precision of 78.7%, recall of 74.0% with over F-measure of 71.7% the system model and output parsing still needs a degree of work. Without a more extensive pattern-matching system in place to find entities that should be connected together, the output evaluation is susceptible to entities that span across sentences as is the case in article 16. In the text "Antarctic lake" and "Organic Lake" are in different sentences, but the makeup of the XML output has no markers to show that this is the case. The same error can be seen in two other cases where a location is at the end of a sentence and the capitalized word at the start of the next is assumed to be

part of the same entity. These cases are article 39 with "Baltic Sea Photosynthesis" and article 48 with "Sargasso Sea Bacterioplankton".

A main weakness of the CoreNLP system is the importance it places on capitalized words, which leads to the annotation of many entities that are not one of its main annotation types. This problem can be identified as a result of using a system on a different domain than intended, as discussed in Challenges facing NER (Section 2.1.4). In a conventional text such as a news story, the use of capitalized words would be nearly exclusively used for enamex types, while in scientific texts a wider range of information and proper names are used. A total of forty-two false positive errors across fifteen articles are because of this, seen for example in article 6 with "Alphaproteobacteria", article 33 with "Flavobacteria", "Alphaproteobacteria" and "Gammaproteobacteria", among others. As the CoreNLP system is model based, the training and refinement of this system by building a new model is possible, but expensive to realize.

Examination of the false negative errors by the CoreNLP system do not show any exact patterns as to why the system chooses to annotate some mentions of entities but not others. In article 56 the system correctly identifies three instances of "Artic" but misses two others with no discernible reason. It may simply be a result of the CRF classifier model becoming confused, or the model may be applied to a scenario which is too different. In either case a larger example set and evaluation would need to be performed in order to update and adapt CoreNLP for use in the OC project.

We see from article 42 where CoreNLP identifies "Wadden Sea" versus the gold standard "German Wadden Sea", as well as "Northern Sweden" versus "Sweden" in article 59, that exactly what defines a location needs to be considered further. While a dictionary of places would be beneficial for this type of disambiguation, the creation of such is difficult and carries with it a large maintenance requirement or it can quickly become outdated. Whether to adapt a system to the researchers needs or to adjust the guidelines of what is expected of the system is left up to future work.

5.3.3 Analysis of OpenNLP

OpenNLP was an outlier in this evaluation given that it had a very high precision score (93.0%) but had a low recall rate (57.0%). The combined F-measure of 63.4% was the worst of the three evaluated systems if the first version of CoreNLP is not considered. This was the only system to use Maximum Entropy Models (See Section 2.1.4) and Perceptron based machine learning, and is built as tool which uses specific models to find specific entity types. These models are easily changed and can be updated or adapted with a reasonable amount of time invested. The developers of OpenNLP have included extensive documentation on every stage of their pipeline, which is useful for other projects attempting to refine the system behaviour.

A review of the false positive errors given by OpenNLP shows errors in segmentation and pattern recognition when trying to tokenize words. Specifically parentheses seem to confuse the system, seen in article 56 with "(8.15" and article 84 with "Mediterranean Sea)" being annotated. Another pattern observed is the tendency to switch between annotating locations with additional parameters such as "North" (article 59), or not. This can also be seen in article 34 where "South Pacific Ocean" and "Western South Pacific Ocean" are annotated simply as "Pacific Ocean". These errors represent a small inability to deal with more complex name structures and require refinement of the tokenizer to solve.

This inability may also be represented in the amount of false negatives given, as they are often complicated structures. However the model used for this evaluation is likely basic and not equipped to deal with the full range of location names used in the evaluation abstracts. More obscure locations such as "Faroe-Shetland Channel" (article 53), "Prydz Bay" (article 61), and "Bedford Basin" (article 98) may indicate a simpler model. However even more basic examples of locations that would likely be included are also missed, including common country names e.g. "Spain" in article 6, "Chile" in article 17, and "Canada" in article 92. Given these errors an extensive model would need to be constructed if OpenNLP was chosen to be in the Ocean-Certain project system.

While the high precision is a quality that would be beneficial to Ocean-Certain, a high recall rate is likely more important. With a high recall rate, pre- and post- processing methods can refine output and improve precision without the system itself needing substantial work.

5.3.4 Analysis of Illinois NET

IllinoisNE Tagger is a system based on gazetteer dictionaries taken from Wikipedia and uses word class models for the examination of unlabeled text. It has the ability to label entities of four different types (person/organization/location/misc) or and expanded eighteen entity type set. The system performed at an even level with other systems in this class with a precision of 77.4%, recall of 72.5% and F-measure of 80.5%.

One of the main sources of false positive errors is the annotation of capitalized words which are not locations. This pattern may reveal a tendency to assign too much importance to capitalized words found in sentences. Examples are seen in eight separate articles, notably article 13 with "Gamma-", "Delta-" and, "Actinobacteria", article 59 with "Cyanobacteria", "Bacteroidetes", "Flavobacteria", and "Mimivirus" among others. A second source of Type I error is the annotation of acronyms that could be referring to a location, but taken in context are not. Article 14 has "PA" and "PE" annotated incorrectly, article 45 has "NT", and article 88 has "BLPR" in multiple instances. Further examples of the system's weakness in context and segmentation ability is the incorrect segmentation of location names which span multiple words. This problem is one of information chunking, and is exemplified in article 34 where IllinoisNE annotated "Western" and "Western South" apart from the their main location name "Pacific Ocean".

While some of the system's Type II errors can be connected to a weakness of context sensitivity or information chunking as explained previously, it also seems to lack an extensive location name dictionary. In article 6 both "Mar Menor" and "Albufera" are not annotated despite having Wikipedia articles specifically about these locations. Additionally, many locations are mis-identified as organizations or miscellanious by IllinoisNE despite having an obvious connection to a location when taken in context. This issue of context sensitivity is admittedly difficult, and may require and extensive modification of the system's dictionary to improve.

5.3.5 Location Evaluation Comparison

Figure 5.4 and the accompanying data table (Table 5.4) shows a comparison of evaluation scores from the examination of location NER systems examined in this work.

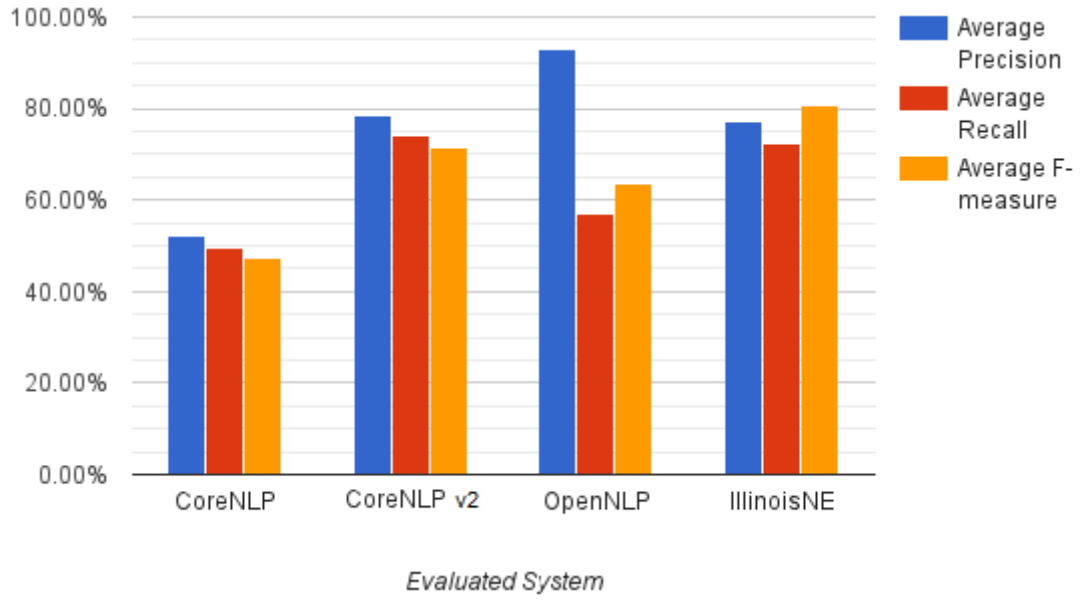

FIGURE 5.4: Location NER Evaluation Summary (100 abstracts)

System Name	Avg Precision	Avg Recall	Avg F-measure
CoreNLP	52.0%	49.6%	47.5%
CoreNLP v2	78.7%	74.0%	71.7%
OpenNLP	93.0%	57.0%	63.4%
Illinois NET	77.4%	72.5%	80.5%

TABLE 5.4: Location NER Evaluation Summary Data

5.4 Linnaeus2 Dictionary Improvements

Over the course of this work it was observed that the corpus articles would commonly speak about biological species in different ways. These include referring to species groups (e.g. Genus/Phylum/Class), and using very specific names for bacteria (strains). This information was deemed sufficiently important to the larger Ocean-Certain project to warrant examination. After consideration, Linnaeus was chosen for experimental dictionary expansion for several reasons;

1. Linnaeus is primarily a gazetteer system and is dictionary based.
2. The included dictionary stems from MEDLINE entries, which were not available to this project.
3. Current entries are already given with taxonomic connection information.
4. The simplicity of the dictionaries meant extension was not overly resource consuming.

In addition, it was possible to extract the additional species information that was necessary from the NCBI Taxonomy. NCBI has structured their entity groups in a logical way that made automatic parsing of the entries possible.

The goal of this experiment was to add more comprehensive species information to Linnaeus' dictionary and asses its performance versus an expanded Gold Standard Annotation. This would allow the Ocean-Certain project to estimate what future improvements and refinement might cost in terms of resources. With the target Gold Standard being significantly more complex, performance is not required to be superior to the previous configuration. The results of this dictionary extension can be seen in Figure 5.5 and each step is explained below. The segment marked "Initial test" shows the performance score of Linnaeus from the initial evaluation completed earlier in this paper.

Master's student Sindre Næss was tasked with extracting all Genus/Phylum/Class information from the NCBI Taxonomy, as well as bacterial names together with their strains. His previous work with ontologies and the NCBI Taxonomy allowed for the rapid compilation of this data and is described in his paper "Generalization of Named Entities Using Linked Data"[cit].

The first set of dictionary expansion information supplied was a list of Genus/Phylum/-Class names for any species found in the NCBI Taxonomy. This list was added to the full external dictionary supplied by the Linnaeus developers. A preliminary test to verify that the new information was understood and in use by Linnaeus was done, and is seen in the performance graph (Figure 5.5) marked **Expanded Dictionary**. With a reported recall of 100.0%, the system was now identifying more varied species information from the original Gold Standard. The resulting drop in precision to 32.0% is a repercussion of the system finding entities that were not currently annotated manually. This verification allowed for the continuation of this experiment by expanding the set of target species to be annotated.

While the original Gold Standard annotation included only specific species names (See Section 4.3), it was now necessary expand the rule-set to include species groups and strains. Any species group such as "Alphaproteobacteria" (a class of bacteria), or "LKM11" (an environmental clade) was included. After completion of this annotation run another evaluation of Linnaeus was taken marked as **Gold Standard expansion**. When complete, the performance scores for Linnaeus were promising with precision back up to 73.4%, though recall did drop to 81.2%. This recall drop is attributable to our Gold Standard annotation now requiring the annotation of bacteria strains which our new dictionary was not yet expanded to include.

Under **Dictionary Improvement (Bacteria)** the dictionary was again expanded to include more specific bacteria names and strains. This is reflected in the (somewhat marginal) improvement of precision and recall to 75.3% and 84.6% respectively. While now performing near the level of similar systems like SPECIES and OrganismTagger, an analysis of system errors showed that name variations for species were being missed. A final improvement of the dictionary was to take the various added species names, and generate their shortened names and variations using a Python script. For example "Crocosphaera watsonii" would now also have the entries "C. watsonii" and "Corcosphaera sp" added to allow the system to identify these variations as the same thing.

This final improvement gave another small increase in performance to a precision of 77.2% and a recall score of 91.0%. These scores are comparable to systems built solely around the NCBI Taxonomy and over a much more intricate set of entities. Further improvements are possible with the inclusion of additional taxonomical databases into

the dictionary, and a refinement of existing name variations. The assistance of biologists to determine a set of rules for name variation as well as allowed species names could also be beneficial, as this specialized knowledge was not available at the time of writing.

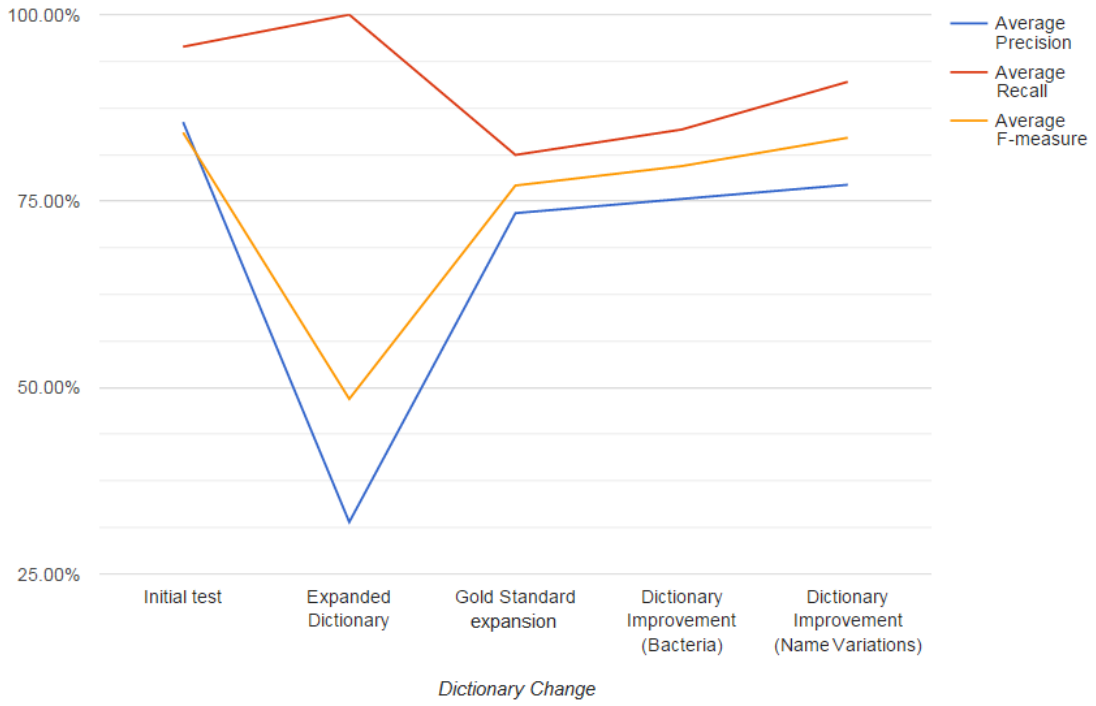

FIGURE 5.5: Linnaeus 2 Dictionary Improvement Summary

	Avg Precision	Avg Recall	Avg F-measure
Initial test	85.6%	95.7%	84.2%
Expanded Dictionary	32.0%	100.0%	48.5%
Gold Standard expansion	73.4%	81.2%	77.1%
Dictionary Improvement 1	75.3%	84.6%	79.7%
Dictionary Improvement 2	77.2%	91.0%	83.5%

TABLE 5.5: Linneaus 2 Dictionary Improvement Data

Chapter 6

Conclusion

The aim of this work has been the examination and evaluation of existing Named Entity Recognition systems for specific entity types. This evaluation is to determine the strengths and weaknesses of each system, in order to ascertain their viability for the Ocean-Certain initiative. A set of research goals was drafted in Chapter 1 in order to clarify and direct the purpose of this study while underway. In Chapter 2, background information on key concepts was provided to give readers the knowledge required to understand the purpose of Ocean-Certain, as well as the terminology and methods used in Named Entity Recognition. Further, the process of connecting variables into facts was described in order to demonstrate the importance of finding suitable NER systems to fit Ocean-Certain's needs. Chapter 3 outlined the role of Named Entity Recognition in different scenarios, along with the methodology used to evaluate the different systems covered. The reasoning behind the selection of certain NER systems for formal evaluation along with the details of the "Gold Standard" annotation set to be evaluated against was given in Chapter 4. Finally the results of each system test, along with an examination of the strengths and weaknesses revealed, was discussed in Chapter 5. An objective opinion about the viability of each system, based upon the results and the necessary work to improve performance, was also given. This study then went further by expanding the capabilities of the Linnaeus 2.0 NER system in Section 5.4.

Appendix A

System Evaluation Error Tables

In order to keep the size of this paper to a reasonable level, only the system error tables will be included in the Appendix. These error tables list the false positive and false negative errors of each system by abstract. The error results for the twenty-five abstract tests are contained in a single table. To improve readability, each one-hundred error table is broken up into 4 segments. These error tables are labeled in the following format;

<system name> Errors <number> of 4

Summary evaluation results of precision, recall, and F-measure can be found for each system in Chapter 5, at the end of their respective sections.

The full results tables for each system, showing the above metrics for each individual abstract, will be included in a .zip file and is available upon request.

The full output tables, which show the system returned annotations versus the gold standard over every article will be included in a .zip file with this thesis and is available upon request.

TABLE A.1: Chemspot 2.0 Errors (25 abstracts)

Abstract	False Positive (test:yes, gold standard:no)	False Negative (test:no, gold standard: yes)
00001#10.1038#nmicro1749		
00002#10.1038#ngeo1765	vitamins	Phosphorus
00003#10.1038#ngeo1910	water	sulphur
00004#10.1038#nmicro1180	water, water	
00005#10.1038#nmicro1751		
00006#10.1038#srep00490		
00007#10.1038#nmicro956		
00008#10.1038#nature04161		
00009#10.1038#ngeo964	iron-binding	
00010#10.1038#ngeo1916	Low-oxygen	carbon dioxide, Carbon dioxide
00011#10.1038#nmicro1750		
00012#10.1038#nmicro2386	water, Coral, water	
00013#10.1038#ismej.2011.12	leucine	
00014#10.1038#ismej.2010.33	dimethyl	
00015#10.1038#35037500	water, CO	carbon dioxide, carbon dioxide, dimethyl sulphide
00016#10.1038#ismej.2013.69	oxygen, OMZ, 16S	DMSP, DMSP, DMSP, DMSP
00017#10.1038#ismej.2013.144	water-column	nitrous oxide
00018#10.1038#nature13123		carbon dioxide
00019#10.1038#nmicro1745		hydrogen sulphide
00020#10.1038#ismej.2010.79		carbon dioxide, carbon dioxide
00021#10.1038#ncomms3934	13C-, 13C	carbon dioxide
00022#10.1038#ismej.2013.135	Coral, sugar, sugars, 16S	
00023#10.1038#ismej.2012.161	Ace, 'C-Ace', O2-H2S, water, C-Ace	
00024#10.1038#ismej.2010.28	high-CO2	
00025#10.1038#nature07235		

TABLE A.2: Oscar3 v1 Errors (25 abstracts)

Abstract	False Positive (test:yes, gold standard:no)	False Negative (test:no, gold standard: yes)
00001#10.1038#nmicro1749	marine, protein, marine	
00002#10.1038#ngeo1765	nutrient, nutrient, marine, waters, nutrient, nutriens, nutrients, In, nutrient, vitamins, marine, nutrient, nutrient, marine, nutrient	
00003#10.1038#ngeo1910	water, emission	
00004#10.1038#nmicro1180	water, oxic, marine, Lysis, groups, water, In	
00005#10.1038#nmicro1751	taxa, nutrient, size, taxa	
00006#10.1038#srep00490	marine, cyanobacteria, groups, LD12	
00007#10.1038#nmicro956	food, PCD, proteins	
00008#10.1038#nature04161	water, brine, food	
00009#10.1038#ngeo964	iron, iron	
00010#10.1038#ngeo1916	role, marine, nutrients, nutrient, fossil fuel, waters, oxygen, nutrient	
00011#10.1038#nmicro1750	Marine, marine	
00012#10.1038#nmicro2386	process, waters	
00013#10.1038#ismej.2011.12	water, nutrient, waters, food, nutrients, waters, role, DOC, structure, DOC, waters, l-1, DOC, l-1, l-1, DOC, l-1, addition, structure, all, Gamma, Delta, Actinobacteria, SAR11, SAR116, Flavobacteria, DOC, structure, water, DOC, structure, water	
00014#10.1038#ismej.2010.33	structure, structure, cyclonic, FF, nucleic acid, leucine, structure, 0-, PA, PHA, leucine, FF, FF, 200-, PA, PHA, FF, structure, FF, epi	
00015#10.1038#35037500	A, algal, waters, dimethyl, waters, waters, fate, algal, water, mass	dimethyl sulphide
00016#10.1038#ismej.2013.69	marine, water, size, Marinobacter, anoxygenic, oxidation, oxidation, marine, nutrient, In, oxydation, light	
00017#10.1038#ismej.2013.144	size, oxygen, Marine, oxygen, OMZ, size, 16S, rRNA, size, OMZ, size, OMZ, size, size, OMZ, antibiotic, element, protein, marine, nitric, oxide, reduction, size, denitrification, role, attachment, OMZ	nitrous oxide
00018#10.1038#nature13123	water	
00019#10.1038#nmicro1745	famine, food, transport, water	hydrogen sulphide
00020#10.1038#ismej.2010.79	marine, marine, Marine	
00021#10.1038#ncomms3934	structure, A, mass, fertilizers, nutrient	
00022#10.1038#ismej.2013.135	algal, role, elements, marine, lysis, algal, algal, water, gamma, alphaproteobacterial, mass spectrometry, 13C, 15N, P, lysis, P, lysis, mass spectrometry, P, role, algal macroalgal, algal, structure, Porites, Ochrophyta, Rhodophyta, Cholorophyta, DOC, DCNSs, DOC, DCNS, 16S, rRNA, macroalgal, DCNS, Ochrophyta, Rhodophyta, algal, DCNS, DOC, algal, DCNS, structure, taxa, Erythrobacteraceae, In, macroalgal, algal, DCNS	
00023#10.1038#ismej.2012.161	GSB, Chlorobiaceae, An, GSB, C-Ace, O2-H2S, water, DNA sequence, nine, 19-fold,	
00024#10.1038#ismej.2010.28	C-Ace, A, protein, light, C-Ace, GSB, bacteriochlorophylls, GSB	
00025#10.1038#nature07235	nutrient, role, elements, A, addition, fate, food, mineral, nutrients, addition, mineral, nutrients, mineral, nutrient, mineral, nutrients	CO2

TABLE A.3: Oscar3 v2 Errors (25 abstracts)

Abstract	False Positive (test:yes, gold standard:no)	False Negative (test:no, gold standard: yes)
00001#10.1038#nmicro1749		
00002#10.1038#ngeo1765	waters, In	
00003#10.1038#ngeo1910	water	dimethylsulphide, dimethylsulphide, dimethylsulphide, dimethylsulphide, dimethylsulphide
00004#10.1038#nmicro1180	water, water, In	
00005#10.1038#nmicro1751		
00006#10.1038#srep00490		
00007#10.1038#nmicro956		
00008#10.1038#nature04161	water	
00009#10.1038#ngeo964		iron, iron, iron
00010#10.1038#ngeo1916	waters, oxygen	
00011#10.1038#nmicro1750		
00012#10.1038#nmicro2386	waters	CO2, CO2
00013#10.1038#ismej.2011.12	water, water, waters, waters, water, water	
00014#10.1038#ismej.2010.33	leucine, leucine	
00015#10.1038#35037500	waters, waters, waters, water	dimethyl sulphide
00016#10.1038#ismej.2013.69	water, In	CO, DMS, Dimethylsulfoniopropionate, DMSP, DMSP, DMSP, DMSP, DMS
00017#10.1038#ismej.2013.144	oxygen, oxygen	nitrous oxide
00018#10.1038#nature13123	water	
00019#10.1038#nmicro1745	hydrogen, water	hydrogen sulphide
00020#10.1038#ismej.2010.79		CO2, CO2
00021#10.1038#ncomms3934		
00022#10.1038#ismej.2013.135	water, P, P, P	
00023#10.1038#ismej.2012.161	In	sugar, sugars, galactose, fucose
00024#10.1038#ismej.2010.28		CO2, CO2
00025#10.1038#nature07235		

TABLE A.4: Oscar3 v3 Errors (25 abstracts)

Abstract	False Positive (test:yes, gold standard:no)	False Negative (test:no, gold standard: yes)
00001#10.1038#nmicro1749		
00002#10.1038#ngeo1765	waters, In	
00003#10.1038#ngeo1910		
00004#10.1038#nmicro1180	water, water, In	
00005#10.1038#nmicro1751		
00006#10.1038#srep00490		
00007#10.1038#nmicro956		
00008#10.1038#nature04161	water	
00009#10.1038#ngeo964	iron, iron	
00010#10.1038#ngeo1916	waters, oxygen	
00011#10.1038#nmicro1750		
00012#10.1038#nmicro2386	waters	CO2, CO2
00013#10.1038#ismej.2011.12	water, water, waters, waters, water, water	
00014#10.1038#ismej.2010.33	leucine, leucine	
00015#10.1038#350375OO	waters, waters, waters, water	dimethyl sulphide
00016#10.1038#ismej.2013.69	anoxygenic, In	CO, DMS, DMSP, DMSP, DMSP, DMSP
00017#10.1038#ismej.2013.144		nitrous oxide
00018#10.1038#nature13123	oxygen, oxygen	
00019#10.1038#nmicro1745	water	hydrogen sulphide
00020#10.1038#ismej.2010.79		
00021#10.1038#ncomms3934		
00022#10.1038#ismej.2013.135	water, P, P, P	
00023#10.1038#ismej.2012.161	Rhodophyta, Cholorphyta, DOC, Rhodophyta, DOC, In	sugars, galactose, fucose
00024#10.1038#ismej.2010.28	sulfur, sulfur, water	
00025#10.1038#nature07235		CO2, CO2

TABLE A.5: Chemspot 2.0 Errors 1 of 4

Abstract	False Positive (test:yes, gold standard:no)	False Negative (test:no, gold standard: yes)
00001#10.1038#nmicro1749		
00002#10.1038#ngeo1765	vitamins	
00003#10.1038#ngeo1910	water	sulphur
00004#10.1038#nmicro1180	water, water	
00005#10.1038#nmicro1751		
00006#10.1038#srep00490		
00007#10.1038#nmicro956		
00008#10.1038#nature04161		
00009#10.1038#ngeo964	iron-binding	iron
00010#10.1038#ngeo1916	carbon, Carbon, Low-oxygen	carbon dioxide, Carbon dioxide
00011#10.1038#nmicro1750		
00012#10.1038#nmicro2386		carbon
00013#10.1038#ismej.2011.12	water, Coral, water	carbon
00014#10.1038#ismej.2010.33		leucine
00015#10.1038#35037500	dimethyl, water	carbon dioxide, carbon dioxide, dimethyl sulphide
00016#10.1038#ismej.2013.69	water	DMSP, DMSP, DMSP, DMSP, DMS, DMSP
00017#10.1038#ismej.2013.144	OMZ, 16S, OMZ, OMZ, OMZ, nitrous, OMZ	nitrous oxide, oxygen
00018#10.1038#nature13123	water-column	carbon dioxide
00019#10.1038#nmicro1745	hydrogen, water	hydrogen sulphide
00020#10.1038#ismej.2010.79	carbon	carbon dioxide, carbon dioxide
00021#10.1038#ncomms3934	nitrogen:phosphorus	carbon dioxide, phosphorus, nitrogen, carbon, phosphorus
00022#10.1038#ismej.2013.135	13C-, 13C	
00023#10.1038#ismej.2012.161	Coral, 16S, Coral	
00024#10.1038#ismej.2010.28	Ace, 'C-Ace', O2-H2S, water, Ace, C-Ace, oxygen, C-Ace, sulfate, sulfur	O2, H2S
00025#10.1038#nature07235	high-CO2	

TABLE A.6: Chemspot 2.0 Errors 2 of 4

00026#10.1038#ismej.2012.31	Peninsula, Peninsula	
00027#10.1038#ngeo1921	water	carbon dioxide
00028#10.1038#nature12857		carbon dioxide
00029#10.1038#35038000	calcium	carbon dioxide, carbon dioxide, calcium carbonate
00030#10.1038#ismej.2010.185	sulfur, Ace	
00031#10.1038#srep04661	ammonia, urease	
00032#10.1038#ismej.2013.121	dinitrogen-fixing, water, Trichodesmium	C, N, C, N, C, N, C, N
00033#10.1038#ismej.2012.28	Peninsula, Peninsula, ammonia, 3-hydroxypropionate/4-hydroxybutyrate, tricarboxylic, ammonia, nitrite	
00034#10.1038#ismej.2011.152	picocyanobacterium, sugars, diazotrophs	N, N
00035#10.1038#ismej.2008.125	PO43	C, N, P, C, N, P, N, P, N, P, N, P, N, P
00036#10.1038#ngeo1265	water, Water	15N
00037#10.1038#ismej.2014.39	water, water	carbon
00038#10.1038#nature04159		carbon dioxide
00039#10.1038#ismej.2011.20	water, water	N2, N2, N, O2, N2, N2, N2, C, C, N, O2
00040#10.1038#ismej.2009.60	chlorophyll	
00041#10.1038#ismej.2008.15		phosphate, P, P, P
00042#10.1038#ismej.2009.29	chlorophyll, NO3	NO3-
00043#10.1038#nature06267	carbon:nitrogen	carbon dioxide
00044#10.1038#nature01165		carbon
00045#10.1038#ismej.2012.130	Methylosulfonomonas, chlorophyll, chlorophyll	
00046#10.1038#ngeo234	water-column	
00047#10.1038#nature04158		
00048#10.1038#ismej.2008.117	water, 16S	
00049#10.1038#ismej.2013.199		
00050#10.1038#ismej.2009.94	dinoflagellates, D, DFL12T, vitamins, vitamin, oxygen, vitamin, DFL12T, vitamins, DFL12T, oxygen, D, DFL12T, D, DFL12T	

TABLE A.7: Chemspot 2.0 Errors 3 of 4

00051#10.1038#ismej.2007.70	ribulose-1, 5-bisphosphate, water	carbon
00052#10.1038#ismej.2010.211	vitamin, vitamin, B12, B12, B12, B12, B12, B12, B12, vitamin	carbon dioxide
00053#10.1038#ismej.2012.14	13C, water	carbon
00054#10.1038#ismej.2010.53	colony-water, carbon-fixation	Carbon, O2, O2, C, N2, N2, N2, O2, C, N, C, N
00055#10.1038#nmicro2439	nitrous	carbon dioxide, nitrous oxide
00056#10.1038#srep02339		carbonate
00057#10.1038#ismej.2011.201	amino	
00058#10.1038#ismej.2007.86	BGE, BGE, BGE, BGE, BGE, BGE, BGE	
00059#10.1038#ismej.2009.120	Carboxylic, amino, AAs, carbohydrates, CHs, CA, AA, CH, water, CA, AA, CH, carbon, acetate	Carboxylic acids, amino acids
00060#10.1038#ismej.2014.16		
00061#10.1038#ismej.2010.83		
00062#10.1038#nature08985	Revolution, carbon:nitrate, carbon:nitrate, carbon:nitrate	
00063#10.1038#ismej.2013.190		
00064#10.1038#ismej.2011.132		
00065#10.1038#ngeo2108	bottom-water	carbon
00066#10.1038#nature11229	iron, iron	carbon dioxide
00067#10.1038#ismej.2009.8	cyanobacterium, N, nucleotide, iron-stress, C	N2, nitrogenase
00068#10.1038#nature04160		
00069#10.1038#ismej.2012.44	low-oxygen, OMZ, [O2]=0, 13C:15N, amino, [O2]=0, [O2]=2-15, OMZ, low-oxygen	carbon, C, N, O2, O2, C, N, amino acids, O2
00070#10.1038#nclimate1507	Carbon, high-CO2	Carbon dioxide, CO2
00071#10.1038#ismej.2014.11	water, nitrous, ammonia, water, Water, Water, ammonia, water, chlorophyll	nitrous oxide
00072#10.1038#nature02437	low-chlorophyll, Alaska, silicic, silicic, geo, iron, iron	silicic acid, silicic acid
00073#10.1038#ismej.2010.91	chlorophyll, Chl, Chl, Pyrosequencing, 16S	
00074#10.1038#nclimate1989	DMSP-sulphur, DMSP-sulphur, 35S-DMSP, 35S, 35S-DMSP, 3H-leucine	CO2
00075#10.1038#ismej.2011.118		sulphur, DMSP

TABLE A.8: Chemspot 2.0 Errors 4 of 4

00076#10.1038#ismej.2011.50			nitrate, 15N
00077#10.1038#ngeo1757			C, N, P
00078#10.1038#ismej.2012.13	Trichodesmium, Trichodesmium, Trichodesmium, Trichodesmium		iron, N2, N2
00079#10.1038#ismej.2014.1	nucleotide, PCR-amplicons, metagenomes, water		N
00080#10.1038#ismej.2010.36	carbon		carbon dioxide
00081#10.1038#ismej.2013.216			carbon
00082#10.1038#ismej.2014.60			Ne, Ne
00083#10.1038#ismej.2014.71	diazotrophs		N2, N2, N2, N2, PO43, oxygen
00084#10.1038#ismej.2010.197	ammonium, urease, acetyl-CoA, 4-hydroxybutyryl, [14C]HCO3		C, C, HCO3, 14C
00085#10.1038#ismej.2009.105	estuarine, chlorophyll-alpha		N, ammonia, C, N
00086#10.1038#ismej.2008.56	N-fixing, R, R, NO3+NO2, nitrogen		NO3, NO2, N
00087#10.1038#ncomms2981			
00088#10.1038#ismej.2009.142	water, 16S, VIS4, water, VIS2, VIS3, VIS5, VIS6		
00089#10.1038#ncomms4271			
00090#10.1038#nmicro2831	water		
00091#10.1038#ismej.2011.214	water, virus-to-prokaryote, VPR		
00092#10.1038#ismej.2011.157	carbon		carbon dioxide
00093#10.1038#ismej.2011.182	N2-fixation, N2-fixation, chlorophyll, 15N-labeled		ammonium, N
00094#10.1038#ismej.2013.79	water, SUP05, water		
00095#10.1038#ismej.2007.5			phosphate, PO43-, N, PO43-
00096#10.1038#ngeo1830			carbon dioxide, carbon dioxide, C, C, C, C, C
00097#10.1038#17351	silicic		silicic acid
00098#10.1038#ismej.2013.234			methanol, carbon monoxide
00099#10.1038#ismej.2011.117	16S, water, chlorophyll		
00100#10.1038#srep01292			

TABLE A.9: Oscar3 v3 Errors 1 of 4

Abstract	False Positive (test:yes, gold standard:no)	False Negative (test:no, gold standard: yes)
00001#10.1038#nmicro1749		
00002#10.1038#ngeo1765	In	
00003#10.1038#ngeo1910	water	
00004#10.1038#nmicro1180	In, water, water	
00005#10.1038#nmicro1751		
00006#10.1038#srep00490		
00007#10.1038#nmicro956		
00008#10.1038#nature04161	water	
00009#10.1038#ngeo964	iron	
00010#10.1038#ngeo1916	oxygen	
00011#10.1038#nmicro1750		
00012#10.1038#nmicro2386		CO2, CO2
00013#10.1038#ismej.2011.12	water, waters, waters, water	
00014#10.1038#ismej.2010.33		
00015#10.1038#35037500	iron, water	dimethyl sulphide
00016#10.1038#ismej.2013.69	sulfur, In, water, anoxygenic	CO, DMS, DMSP, DMSP, DMSP, DMSP, DMS, DMSP
00017#10.1038#ismej.2013.144		nitrous oxide, nitric
00018#10.1038#nature13123	water	
00019#10.1038#nmicro1745	hydrogen, water	hydrogen sulphide
00020#10.1038#ismej.2010.79		
00021#10.1038#ncomms3934		
00022#10.1038#ismej.2013.135	P, P, P, water	
00023#10.1038#ismej.2012.161	In, Rhodophyta, Chlorophyta, DOC, Rhodophyta, DOC	sugars, galactose, fucose
00024#10.1038#ismej.2010.28	sulfur, sulfur, sulfate, oxygen, water	O2, H2S
00025#10.1038#nature07235		CO2

TABLE A.10: Oscar3 v3 Errors 2 of 4

00026#10.1038#ismej.2012.31	In, anoxygenic		
00027#10.1038#ngeo1921	In, water		
00028#10.1038#nature12857			
00029#10.1038#35038000			
00030#10.1038#ismej.2010.185	In, sulfur		
00031#10.1038#srep04661	ammonia		
00032#10.1038#ismej.2013.121	dinitrogen, water, Trichodesmium, Trichodesmium		
00033#10.1038#ismej.2012.28	In, In, ammonia, 3-hydroxypropionate, tricarbolyxic acid, ammonia, nitrite	taurine	
00034#10.1038#ismej.2011.152	At, N2, picocyanobacterium		
00035#10.1038#ismej.2008.125	lysis, lysis, lysis	NH4+, PO43-, NH4+	
00036#10.1038#ngeo1265	In, water	15N, 14N	
00037#10.1038#ismej.2014.39	Daphnia, Daphnia, DOC, Daphnia, water, water, DOC, Daphnia, Daphnia, DOC, Daphnia, DOC	leucine	
00038#10.1038#nature04159			
00039#10.1038#ismej.2011.20	N, water, water	O2, O2, O2, O2, N2, N2, N2, N2, NH4+, NH4+	
00040#10.1038#ismej.2009.60	water, chlorophyll		
00041#10.1038#ismej.2008.15			
00042#10.1038#ismej.2009.29	hydrolysis	NO3-	
00043#10.1038#nature06267			
00044#10.1038#nature01165			
00045#10.1038#ismej.2012.130	In, methanol		
00046#10.1038#ngeo234	In, phosphorus, nitrogen, phosphate, water		
00047#10.1038#nature04158	In		
00048#10.1038#ismej.2008.117	water		
00049#10.1038#ismej.2013.199			
00050#10.1038#ismej.2009.94	oxygen, oxygen	arginine	

TABLE A.11: Oscar3 v3 Errors 3 of 4

00051#10.1038#ismej.2007.70	ribulose-1,5-biphosphate	CO2, CO2, CO2, CO2, CO2, CO2
00052#10.1038#ismej.2010.211	At, carbon	CO2, CO2,
00053#10.1038#ismej.2012.14	water	
00054#10.1038#ismej.2010.53	carbon, In, cyanobacterium, water, ammonium, water	O2, O2, O2, O2, C, N2, N2, N2, N2, NH4+, O2
00055#10.1038#nmicro2439		nitrous oxide
00056#10.1038#srep02339		CO2, CO2
00057#10.1038#ismej.2011.201	water	glutamine
00058#10.1038#ismej.2007.86	carbon, DOC, DOC	
00059#10.1038#ismej.2009.120	water	carbohydrates, acetate, Carboxylic acids, amino acids
00060#10.1038#ismej.2014.16	P, waters	
00061#10.1038#ismej.2010.83	In, Chlorophyta, Chlorophyta, Chlorophyta	
00062#10.1038#nature08985	At,	
00063#10.1038#ismej.2013.190		
00064#10.1038#ismej.2011.132		
00065#10.1038#ngeo2108	water, water	
00066#10.1038#nature11229	iron, iron, water	
00067#10.1038#ismej.2009.8	In, C, C, C, C, iron, cyanobacterium	N2, nitrogenase
00068#10.1038#nature04160		
00069#10.1038#ismej.2012.44	oxygen, oxygen, In, At	C, O2, O2, O2, amino acids
00070#10.1038#nclimate1507	In	CO2, CO2, CO2, CO2, CO2, CO2
00071#10.1038#ismej.2014.11	water, ammonia, water, Water, Water, ammonia, water, chlorophyll	nitrous oxide
00072#10.1038#nature02437	iron, iron, iron, silica	
00073#10.1038#ismej.2010.91	aminopeptidase, Marinobacter	
00074#10.1038#nclimate1989	phosphate, In	CO2
00075#10.1038#ismej.2011.118	3H-leucine, cyanobacterial, DMSP-sulphur	sulphur, DMSP, DMSP, DMSP, sulphur, DMSP

TABLE A.12: Oscar3 v3 Errors 4 of 4

00076#10.1038#ismej.2011.50			15N
00077#10.1038#ngeo1757	C		
00078#10.1038#ismej.2012.13	Trichodesmium, cyanobacterial, Trichodesmium, Trichodesmium, Trichodesmium, Trichodesmium, Trichodesmium, Trichodesmium, Trichodesmium	N2, N2	
00079#10.1038#ismej.2014.1			
00080#10.1038#ismej.2010.36	cyanobacterial, Euk-B	CO2, CO2, CO2, CO2, CO2, CO2	
00081#10.1038#ismej.2013.216	In, Sulfitobacter	N2, N2, N2, N2, N2,	
00082#10.1038#ismej.2014.60	In		
00083#10.1038#ismej.2014.71	In, water, deoxygenation	NO2, PO43	
00084#10.1038#ismej.2010.197	As, ammonia, acetyl-CoA	CO2, HCO3, 14C	
00085#10.1038#ismej.2009.105		C	
00086#10.1038#ismej.2008.56	nitrogen, cyanobacterium, cyanobacterium, hterocyst	NO3, NO2, Si(OH)4	
00087#10.1038#ncomms2981			
00088#10.1038#ismej.2009.142	water, water		
00089#10.1038#ncomms4271			
00090#10.1038#nmicro2831	In, water		
00091#10.1038#ismej.2011.214	In, water, water		
00092#10.1038#ismej.2011.157		O2, O2	
00093#10.1038#ismej.2011.182		ammonium, leucine, N	
00094#10.1038#ismej.2013.79	In, water, water		
00095#10.1038#ismej.2007.5	In, cyanobacterial, Trichodesmium, cyanobacterial	PO43-, PO43-, PO43-	
00096#10.1038#ngeo1830		C, C, C, C, C	
00097#10.1038#17351		silica, CO2, silicic acid	
00098#10.1038#ismej.2013.234	In, No, water	sulfur	
00099#10.1038#ismej.2011.117	In, water, water, chlorophyll		
00100#10.1038#srep01292	In, In		

TABLE A.13: SPECIES Errors 1 of 4

Abstract	False Positive (test:yes, gold standard:no)	False Negative (test:no, gold standard: yes)
00001#10.1038#nmicro1749		
00002#10.1038#ngeo1765		
00003#10.1038#ngeo1910		
00004#10.1038#nmicro1180		
00005#10.1038#nmicro1751		
00006#10.1038#srep00490		
00007#10.1038#nmicro956		
00008#10.1038#nature04161		
00009#10.1038#ngeo964		
00010#10.1038#ngeo1916		
00011#10.1038#nmicro1750		
00012#10.1038#nmicro2386		
00013#10.1038#ismej.2011.12		
00014#10.1038#ismej.2010.33		
00015#10.1038#35037500		
00016#10.1038#ismej.2013.69		
00017#10.1038#ismej.2013.144		
00018#10.1038#nature13123		
00019#10.1038#nmicro1745		
00020#10.1038#ismej.2010.79		
00021#10.1038#ncomms3934	Human, human	
00022#10.1038#ismej.2013.135		
00023#10.1038#ismej.2012.161		
00024#10.1038#ismej.2010.28	bacteriophage	
00025#10.1038#nature07235		

TABLE A.14: SPECIES Errors 2 of 4

00026#10.1038#ismej.2012.31		
00027#10.1038#ngeo1921		
00028#10.1038#nature12857	human	
00029#10.1038#35038000		
00030#10.1038#ismej.2010.185		
00031#10.1038#srep04661		
00032#10.1038#ismej.2013.121		
00033#10.1038#ismej.2012.28		
00034#10.1038#ismej.2011.152	gamma-Proteobacterium	
00035#10.1038#ismej.2008.125	PpV, PpV, PpV, PpV	
00036#10.1038#ngeo1265		
00037#10.1038#ismej.2014.39		
00038#10.1038#nature04159		
00039#10.1038#ismej.2011.20		
00040#10.1038#ismej.2009.60		
00041#10.1038#ismej.2008.15		
00042#10.1038#ismej.2009.29		
00043#10.1038#nature06267		
00044#10.1038#nature01165		
00045#10.1038#ismej.2012.130		
00046#10.1038#ngeo234		
00047#10.1038#nature04158		
00048#10.1038#ismej.2008.117		
00049#10.1038#ismej.2013.199		
00050#10.1038#ismej.2009.94		

TABLE A.15: SPECIES Errors 3 of 4

00051#10.1038#ismej.2007.70	
00052#10.1038#ismej.2010.211	
00053#10.1038#ismej.2012.14	
00054#10.1038#ismej.2010.53	
00055#10.1038#nmicro2439	
00056#10.1038#srep02339	
00057#10.1038#ismej.2011.201	
00058#10.1038#ismej.2007.86	
00059#10.1038#ismej.2009.120	
00060#10.1038#ismej.2014.16	
00061#10.1038#ismej.2010.83	
00062#10.1038#nature08985	
00063#10.1038#ismej.2013.190	
00064#10.1038#ismej.2011.132	
00065#10.1038#ngeo2108	
00066#10.1038#nature11229	
00067#10.1038#ismej.2009.8	
00068#10.1038#nature04160	
00069#10.1038#ismej.2012.44	
00070#10.1038#nclimate1507	
00071#10.1038#ismej.2014.11	
00072#10.1038#nature02437	
00073#10.1038#ismej.2010.91	
00074#10.1038#nclimate1989	
00075#10.1038#ismej.2011.118	

TABLE A.16: SPECIES Errors 4 of 4

00076#10.1038#ismej.2011.50		
00077#10.1038#ngeo1757		
00078#10.1038#ismej.2012.13		
00079#10.1038#ismej.2014.1		
00080#10.1038#ismej.2010.36		
00081#10.1038#ismej.2013.216		
00082#10.1038#ismej.2014.60		
00083#10.1038#ismej.2014.71		
00084#10.1038#ismej.2010.197		
00085#10.1038#ismej.2009.105		
00086#10.1038#ismej.2008.56		
00087#10.1038#ncomms2981		
00088#10.1038#ismej.2009.142		
00089#10.1038#ncomms4271		
00090#10.1038#nmicro2831		
00091#10.1038#ismej.2011.214		
00092#10.1038#ismej.2011.157		
00093#10.1038#ismej.2011.182		
00094#10.1038#ismej.2013.79		
00095#10.1038#ismej.2007.5		
00096#10.1038#ngeo1830		
00097#10.1038#17351		
00098#10.1038#ismej.2013.234		
00099#10.1038#ismej.2011.117		
00100#10.1038#srep01292		

TABLE A.17: OrganismTagger Errors 1 of 4

Abstract	False Positive (test:yes, gold standard:no)	False Negative (test:no, gold standard: yes)
00001#10.1038#nmicro1749		
00002#10.1038#ngeo1765		
00003#10.1038#ngeo1910		
00004#10.1038#nmicro1180		
00005#10.1038#nmicro1751		
00006#10.1038#srep00490		
00007#10.1038#nmicro956		
00008#10.1038#nature04161		
00009#10.1038#ngeo964		
00010#10.1038#ngeo1916		
00011#10.1038#nmicro1750		
00012#10.1038#nmicro2386		
00013#10.1038#ismej.2011.12		
00014#10.1038#ismej.2010.33	Canary, Canary	
00015#10.1038#35037500		
00016#10.1038#ismej.2013.69		
00017#10.1038#ismej.2013.144	mum, mum, mum, mum	
00018#10.1038#nature13123		
00019#10.1038#nmicro1745		
00020#10.1038#ismej.2010.79		
00021#10.1038#ncomms3934		
00022#10.1038#ismej.2013.135		
00023#10.1038#ismej.2012.161		
00024#10.1038#ismej.2010.28		
00025#10.1038#nature07235		

TABLE A.18: OrganismTagger Errors 2 of 4

00026#10.1038#ismej.2012.31		
00027#10.1038#ngeo1921		
00028#10.1038#nature12857	human	
00029#10.1038#35038000		
00030#10.1038#ismej.2010.185	mum, mum	
00031#10.1038#srep04661		
00032#10.1038#ismej.2013.121		
00033#10.1038#ismej.2012.28		
00034#10.1038#ismej.2011.152		
00035#10.1038#ismej.2008.125		
00036#10.1038#ngeo1265	mum	
00037#10.1038#ismej.2014.39		
00038#10.1038#nature04159		
00039#10.1038#ismej.2011.20		
00040#10.1038#ismej.2009.60		
00041#10.1038#ismej.2008.15		
00042#10.1038#ismej.2009.29		
00043#10.1038#nature06267		
00044#10.1038#nature01165		
00045#10.1038#ismej.2012.130		
00046#10.1038#ngeo234		
00047#10.1038#nature04158		
00048#10.1038#ismej.2008.117		
00049#10.1038#ismej.2013.199		
00050#10.1038#ismej.2009.94		

TABLE A.19: OrganismTagger Errors 3 of 4

00051#10.1038#ismej.2007.70		
00052#10.1038#ismej.2010.211		
00053#10.1038#ismej.2012.14		
00054#10.1038#ismej.2010.53		
00055#10.1038#nmicro2439		
00056#10.1038#srep02339		
00057#10.1038#ismej.2011.201	mum, mum, mum	
00058#10.1038#ismej.2007.86		
00059#10.1038#ismej.2009.120		
00060#10.1038#ismej.2014.16		
00061#10.1038#ismej.2010.83	mum	
00062#10.1038#nature08985		
00063#10.1038#ismej.2013.190		
00064#10.1038#ismej.2011.132		
00065#10.1038#ngeo2108		
00066#10.1038#nature11229		
00067#10.1038#ismej.2009.8		
00068#10.1038#nature04160		
00069#10.1038#ismej.2012.44		
00070#10.1038#nclimate1507		
00071#10.1038#ismej.2014.11		
00072#10.1038#nature02437		
00073#10.1038#ismej.2010.91		
00074#10.1038#nclimate1989		
00075#10.1038#ismej.2011.118		

TABLE A.20: OrganismTagger Errors 4 of 4

00076#10.1038#ismej.2011.50		
00077#10.1038#ngeo1757		
00078#10.1038#ismej.2012.13		
00079#10.1038#ismej.2014.1		
00080#10.1038#ismej.2010.36	mum, mum, mum	
00081#10.1038#ismej.2013.216	marine bacterium	
00082#10.1038#ismej.2014.60		
00083#10.1038#ismej.2014.71		
00084#10.1038#ismej.2010.197		
00085#10.1038#ismej.2009.105		
00086#10.1038#ismej.2008.56	mum	Richelia intracellularis, Richelia intracellularis, R. intracellularis, R. intracellularis, R. intracellularis, R. intracellularis, R. intracellularis
00087#10.1038#ncomms2981		
00088#10.1038#ismej.2009.142		
00089#10.1038#ncomms4271		
00090#10.1038#nmicro2831		
00091#10.1038#ismej.2011.214		
00092#10.1038#ismej.2011.157		
00093#10.1038#ismej.2011.182		
00094#10.1038#ismej.2013.79		
00095#10.1038#ismej.2007.5		
00096#10.1038#ngeo1830		
00097#10.1038#17351	marine diatoms	
00098#10.1038#ismej.2013.234		
00099#10.1038#ismej.2011.117		
00100#10.1038#srep01292		

TABLE A.21: Linnaeus 2.0 Errors 1 of 4

Abstract	False Positive (test:yes, gold standard:no)	False Negative (test:no, gold standard: yes)
00001#10.1038#nmicro1749		
00002#10.1038#ngeo1765		
00003#10.1038#ngeo1910		
00004#10.1038#nmicro1180		
00005#10.1038#nmicro1751		
00006#10.1038#srep00490		
00007#10.1038#nmicro956		
00008#10.1038#nature04161		
00009#10.1038#ngeo964		
00010#10.1038#ngeo1916		
00011#10.1038#nmicro1750		
00012#10.1038#nmicro2386		
00013#10.1038#ismej.2011.12		
00014#10.1038#ismej.2010.33	Canary, Canary	
00015#10.1038#35037500		
00016#10.1038#ismej.2013.69	lake, Lake, lake, Lake, lake, Lake, lake	
00017#10.1038#ismej.2013.144	mum, mum, mum, mum	
00018#10.1038#nature13123		
00019#10.1038#nmicro1745		
00020#10.1038#ismej.2010.79		
00021#10.1038#ncomms3934	Human, human	
00022#10.1038#ismej.2013.135	bacteriophage	
00023#10.1038#ismej.2012.161		
00024#10.1038#ismej.2010.28	Lake, Lake	
00025#10.1038#nature07235		

TABLE A.22: Linnaeus 2.0 Errors 2 of 4

00026#10.1038#ismej.2012.31			
00027#10.1038#ngeo1921			
00028#10.1038#nature12857	human		
00029#10.1038#35038000			
00030#10.1038#ismej.2010.185	lake, lake, Lake, lake, mum, mum, lake, lake, lake, Lake		
00031#10.1038#srep04661			
00032#10.1038#ismej.2013.121			
00033#10.1038#ismej.2012.28			
00034#10.1038#ismej.2011.152			
00035#10.1038#ismej.2008.125		Phaeocystis pouchetii, Phaeocystis pouchetii, P. pouchetii, P. pouchetii, P. pouchetii, P. pouchetii, P. pouchetii	
00036#10.1038#ngeo1265	mum		
00037#10.1038#ismej.2014.39	lake, Lake		
00038#10.1038#nature04159			
00039#10.1038#ismej.2011.20			
00040#10.1038#ismej.2009.60			
00041#10.1038#ismej.2008.15			
00042#10.1038#ismej.2009.29			
00043#10.1038#nature06267			
00044#10.1038#nature01165			
00045#10.1038#ismej.2012.130		Methylobacterium oryzae, Methylosulfonomonas methylovora, Hyphomicrobium sp	
00046#10.1038#ngeo234			
00047#10.1038#nature04158			
00048#10.1038#ismej.2008.117			
00049#10.1038#ismej.2013.199			
00050#10.1038#ismej.2009.94			

TABLE A.23: Linnaeus 2.0 Errors 3 of 4

00051#10.1038#ismej.2007.70		
00052#10.1038#ismej.2010.211		
00053#10.1038#ismej.2012.14		
00054#10.1038#ismej.2010.53		
00055#10.1038#nmicro2439		
00056#10.1038#srep02339		
00057#10.1038#ismej.2011.201	mum, mum, mum	
00058#10.1038#ismej.2007.86		
00059#10.1038#ismej.2009.120	lake	
00060#10.1038#ismej.2014.16		
00061#10.1038#ismej.2010.83	mum, Lake, Lake	
00062#10.1038#nature08985		
00063#10.1038#ismej.2013.190		
00064#10.1038#ismej.2011.132		
00065#10.1038#ngeo2108		
00066#10.1038#nature11229		
00067#10.1038#ismej.2009.8		
00068#10.1038#nature04160		
00069#10.1038#ismej.2012.44		
00070#10.1038#nclimate1507		
00071#10.1038#ismej.2014.11		
00072#10.1038#nature02437		
00073#10.1038#ismej.2010.91		
00074#10.1038#nclimate1989		
00075#10.1038#ismej.2011.118		

TABLE A.24: Linnaeus 2.0 Errors 4 of 4

00076#10.1038#ismej.2011.50		
00077#10.1038#ngeo1757		
00078#10.1038#ismej.2012.13		Trichodesmium tenue
00079#10.1038#ismej.2014.1		
00080#10.1038#ismej.2010.36	mum, mum, mum	
00081#10.1038#ismej.2013.216	marine bacterium	
00082#10.1038#ismej.2014.60		
00083#10.1038#ismej.2014.71		
00084#10.1038#ismej.2010.197		
00085#10.1038#ismej.2009.105		
00086#10.1038#ismej.2008.56	mum	Rhelia intracellularis, Rhelia intracellularis, R. intracellularis, R. intracellularis, R. intracellularis, R. intracellularis, R. intracellularis
00087#10.1038#ncomms2981		
00088#10.1038#ismej.2009.142		
00089#10.1038#ncomms4271		
00090#10.1038#nmicro2831		
00091#10.1038#ismej.2011.214		
00092#10.1038#ismej.2011.157		
00093#10.1038#ismej.2011.182		
00094#10.1038#ismej.2013.79		
00095#10.1038#ismej.2007.5		
00096#10.1038#ngeo1830		
00097#10.1038#17351	marine diatoms	
00098#10.1038#ismej.2013.234		
00099#10.1038#ismej.2011.117		
00100#10.1038#srep01292		

TABLE A.25: CoreNLP v2 Errors 1 of 4

Abstract	False Positive (test:yes, gold standard:no)	False Negative (test:no, gold standard: yes)
00001#10.1038#nmicro1749		Earth's
00002#10.1038#ngeo1765		
00003#10.1038#ngeo1910	Arctic sea	Arctic, Arctic, Arctic, Arctic
00004#10.1038#nmicro1180		
00005#10.1038#nmicro1751		
00006#10.1038#srep00490	Alphaproteobacteria, Alphaproteobacteria	Mediterranean, Mar Menor, Mar Menor, Albufera
00007#10.1038#nmicro956	Earth	Earth's, Earth's
00008#10.1038#nature04161		Polar ocean
00009#10.1038#ngeo964		
00010#10.1038#ngeo1916		
00011#10.1038#nmicro1750		
00012#10.1038#nmicro2386		
00013#10.1038#ismej.2011.12	Gamma, Betaproteobacteria, Bacteriodetes, Actinobacteria, Flavobacteria	
00014#10.1038#ismej.2010.33		
00015#10.1038#35037500		
00016#10.1038#ismej.2013.69	Antarctic lake Organic Lake	Antarctic, Organic Lake, Organic Lake, Antarctic
00017#10.1038#ismej.2013.144		
00018#10.1038#nature13123		
00019#10.1038#nmicro1745		
00020#10.1038#ismej.2010.79		
00021#10.1038#ncomms3934	Earth	Earth's
00022#10.1038#ismej.2013.135		
00023#10.1038#ismej.2012.161	Ochrophyta, Chlorophyta, Rhodophyta, Alphaproteobacteria	French Polynesia
00024#10.1038#ismej.2010.28	Antarctic a Green	Antarctic a
00025#10.1038#nature07235		Arctic, Arctic

TABLE A.26: CoreNLP v2 Errors 2 of 4

00026#10.1038#ismej.2012.31		
00027#10.1038#ngeo1921		
00028#10.1038#nature12857		
00029#10.1038#35038000		
00030#10.1038#ismej.2010.185		Ace Lake
00031#10.1038#srep04661	Arctic Ocean Thaumarchaeota	Arctic Ocean, Polar Seas
00032#10.1038#ismej.2013.121		
00033#10.1038#ismej.2012.28	Flavobacteria, Alphaproteobacteria, Gammaproteobacteria, Alphaproteobacteria, Flavobacteria, Gammaproteobacteria, Flavobacteria	
00034#10.1038#ismej.2011.152		
00035#10.1038#ismej.2008.125		
00036#10.1038#ngeo1265	Sargasso Sea Phytoplankton	Sargasso Sea
00037#10.1038#ismej.2014.39	Daphnia galeata, Daphnia, Daphnia	
00038#10.1038#nature04159		
00039#10.1038#ismej.2011.20	Baltic Sea Photosynthesis	
00040#10.1038#ismej.2009.60		
00041#10.1038#ismej.2008.15		Sep Reservoir
00042#10.1038#ismej.2009.29	Wadden Sea	German Wadden Sea
00043#10.1038#nature06267		
00044#10.1038#nature01165		
00045#10.1038#ismej.2012.130		Atlantic Ocean, Mauritanian
00046#10.1038#ngeo234		
00047#10.1038#nature04158		
00048#10.1038#ismej.2008.117	Sargasso Sea Bacterioplankton, Bermuda Atlantic	Sargasso Sea, Bermuda, Atlantic
00049#10.1038#ismej.2013.199	San Pedro Ocean Time	San Pedro
00050#10.1038#ismej.2009.94		

TABLE A.27: CoreNLP v2 Errors 3 of 4

00051#10.1038#ismej.2007.70			
00052#10.1038#ismej.2010.211	Zn	Pacific	
00053#10.1038#ismej.2012.14		Earth's	
00054#10.1038#ismej.2010.53	Baltic Sea Carbon, Aphanizomenon	Baltic Sea	
00055#10.1038#nmicro2439			
00056#10.1038#srep02339		Polar Oceans, Arctic, Antarctic, Antarctic, Arctic, Polar oceans	
00057#10.1038#ismej.2011.201	Southern California Bight, Bacteroidetes, Flavobacteria, Mimivirus	Southern California	
00058#10.1038#ismej.2007.86			
00059#10.1038#ismej.2009.120	Northern Sweden	Sweden	
00060#10.1038#ismej.2014.16	Micromonas	Arctic Oceans, Arctic, Arctic	
00061#10.1038#ismej.2010.83	Haptophyta, Cercozoa, Perkinsozoa, Chlorophyta, Haptophyta	Bourget, Aydat, Pavin, Bourget	
00062#10.1038#nature08985			
00063#10.1038#ismej.2013.190	Antarctic McMurdo Dry Valley	Antarctic, Antarctic, McMurdo Dry Valley, Fryxell, Bonney	
00064#10.1038#ismej.2011.132			
00065#10.1038#ngeo2108	Snowball Earth	Earth	
00066#10.1038#nature11229			
00067#10.1038#ismej.2009.8			
00068#10.1038#nature04160			
00069#10.1038#ismej.2012.44			
00070#10.1038#nclimate1507			
00071#10.1038#ismej.2014.11			
00072#10.1038#nature02437			
00073#10.1038#ismej.2010.91			
00074#10.1038#nclimate1989			
00075#10.1038#ismej.2011.118			

TABLE A.28: CoreNLP v2 Errors 4 of 4

00076#10.1038#ismej.2011.50		English Channel
00077#10.1038#ngeo1757		
00078#10.1038#ismej.2012.13		Pacific Oceans
00079#10.1038#ismej.2014.1		North Pacific Subtropical Gyre
00080#10.1038#ismej.2010.36	Atlantic Ocean Global	Atlantic Ocean
00081#10.1038#ismej.2013.216		
00082#10.1038#ismej.2014.60		
00083#10.1038#ismej.2014.71		
00084#10.1038#ismej.2010.197	Crenarchaeota	
00085#10.1038#ismej.2009.105	Crenarchaeota	Great Barrier Reef
00086#10.1038#ismej.2008.56	Mediterranean Sea Biological	Mediterranean Sea, Levantine Basin
00087#10.1038#ncomms2981		
00088#10.1038#ismej.2009.142	Flavobacteria, Flavobacteria	
00089#10.1038#ncomms4271		
00090#10.1038#nmicro2831	Earth	Earth's
00091#10.1038#ismej.2011.214		
00092#10.1038#ismej.2011.157		
00093#10.1038#ismej.2011.182		South Pacific Gyre, South Pacific Gyre
00094#10.1038#ismej.2013.79	Oceanospirillales, Flavobacteria, Actinobacteria, Betaproteobacteria	
00095#10.1038#ismej.2007.5		Great Barrier Reef
00096#10.1038#ngeo1830		
00097#10.1038#17351		
00098#10.1038#ismej.2013.234		
00099#10.1038#ismej.2011.117	Bathycoccus, Micromonas	Atlantic
00100#10.1038#srep01292		

TABLE A.29: OpenNLP Errors 1 of 4

Abstract	False Positive (test:yes, gold standard:no)	False Negative (test:no, gold standard: yes)
00001#10.1038#nmicro1749		Earth's
00002#10.1038#ngeo1765		Southern Ocean, Pacific
00003#10.1038#ngeo1910		
00004#10.1038#nmicro1180		
00005#10.1038#nmicro1751		
00006#10.1038#srep00490		Mar Menor, Albufera de Valencia, Spain, Mar Menor, Albufera
00007#10.1038#nmicro956	Archaean	Earth's, Earth's
00008#10.1038#nature04161		Polar ocean
00009#10.1038#ngeo964		
00010#10.1038#ngeo1916		
00011#10.1038#nmicro1750		
00012#10.1038#nmicro2386		
00013#10.1038#ismej.2011.12		Paopao Bay, Moorea, French Polynesia
00014#10.1038#ismej.2010.33		Canary Islands
00015#10.1038#35037500		Southern Ocean, Southern Ocean
00016#10.1038#ismej.2013.69		Antarctic, Organic Lake, Vestfold Hills, Antarctica, Organic Lake, Organic Lake, Antarctic
00017#10.1038#ismej.2013.144		Chile
00018#10.1038#nature13123		Porcupine Abyssal Plain
00019#10.1038#nmicro1745		
00020#10.1038#ismej.2010.79		
00021#10.1038#ncomms3934		
00022#10.1038#ismej.2013.135		
00023#10.1038#ismej.2012.161		Moorea, French Polynesia
00024#10.1038#ismej.2010.28		Ace Lake, Antarctica, Ace Lake, Antarctica
00025#10.1038#nature07235		

TABLE A.30: OpenNLP Errors 2 of 4

00026#10.1038#ismej.2012.31		Antarctic
00027#10.1038#ngeo1921		
00028#10.1038#nature12857		
00029#10.1038#35038000		Antarctica
00030#10.1038#ismej.2010.185		Antarctica, Ace Lake, Antarctica, Ace Lake
00031#10.1038#srep04661		Polar Seas, Arctic Ocean
00032#10.1038#ismej.2013.121		
00033#10.1038#ismej.2012.28		Antarctic Peninsula, West Antarctica, Antarctic
00034#10.1038#ismej.2011.152	Pacific Ocean, Pacific Ocean, N	South Pacific Ocean, Western South Pacific Ocean, Western South Pacific Ocean
00035#10.1038#ismej.2008.125		
00036#10.1038#ngeo1265		Sargasso Sea
00037#10.1038#ismej.2014.39		
00038#10.1038#nature04159		
00039#10.1038#ismej.2011.20	0.34	
00040#10.1038#ismej.2009.60		Sargasso Sea
00041#10.1038#ismej.2008.15		
00042#10.1038#ismej.2009.29		
00043#10.1038#nature06267		
00044#10.1038#nature01165		
00045#10.1038#ismej.2012.130		Mauritanian
00046#10.1038#ngeo234		Earth
00047#10.1038#nature04158		
00048#10.1038#ismej.2008.117		Atlantic
00049#10.1038#ismej.2013.199		San Pedro
00050#10.1038#ismej.2009.94		

TABLE A.31: OpenNLP Errors 3 of 4

00051#10.1038#ismej.2007.70	River	Gulf of Mexico
00052#10.1038#ismej.2010.211		Pacific
00053#10.1038#ismej.2012.14		Faroe-Shetland Channel, Northeast Atlantic, Earth, Earth's
00054#10.1038#ismej.2010.53	h-1	
00055#10.1038#nmicro2439		Polar Oceans, Antarctic, Amundsen Gulf, Antarctic, Prydz Bay, Antarctic, Antarctic, Antarctic, Polar oceans
00056#10.1038#srep02339	Polar, (8.15	
00057#10.1038#ismej.2011.201		
00058#10.1038#ismej.2007.86		
00059#10.1038#ismej.2009.120	Northern Sweden	Krycklan, Sweden
00060#10.1038#ismej.2014.16	Prasinophyceae	
00061#10.1038#ismej.2010.83		Bourget, Aydat, Pavin, Bourget
00062#10.1038#nature08985		
00063#10.1038#ismej.2013.190		Antarctic, Antarctic, McMurdo Dry Valley, Fryxell, Bonney
00064#10.1038#ismej.2011.132		
00065#10.1038#ngeo2108		Earth
00066#10.1038#nature11229		
00067#10.1038#ismej.2009.8		South Pacific
00068#10.1038#nature04160		Earth
00069#10.1038#ismej.2012.44		
00070#10.1038#nclimate1507	China Sea	South China Sea
00071#10.1038#ismej.2014.11		
00072#10.1038#nature02437		Gulf of Alaska
00073#10.1038#ismej.2010.91		
00074#10.1038#nclimate1989	N-limitation.	
00075#10.1038#ismej.2011.118		

TABLE A.32: OpenNLP Errors 4 of 4

00076#10.1038#ismej.2011.50		English Channel
00077#10.1038#ngeo1757		
00078#10.1038#ismej.2012.13	Pacific, Pacific	Atlantic, Pacific Oceans, Pacific Ocean
00079#10.1038#ismej.2014.1		North Pacific Subtropical Gyre
00080#10.1038#ismej.2010.36		Earth
00081#10.1038#ismej.2013.216		
00082#10.1038#ismej.2014.60		
00083#10.1038#ismej.2014.71		Peru, Peru
00084#10.1038#ismej.2010.197	Mediterranean Sea), Southern,	Central Mediterranean Sea, Tyrrhenian, Southern Tyrrhenian Sea, Tyrrhenian Sea
00085#10.1038#ismej.2009.105		Fitzroy, Great Barrier Reef, Fitzroy
00086#10.1038#ismej.2008.56		Levantine Basin, Levantine Basin
00087#10.1038#ncomms2981		Antarctic, Southern Ocean, West Antarctica, Ross Sea
00088#10.1038#ismej.2009.142		East Greenland
00089#10.1038#ncomms4271		
00090#10.1038#nmicro2831		Earth's
00091#10.1038#ismej.2011.214		North Atlantic, North Atlantic Gyre
00092#10.1038#ismej.2011.157		Canada
00093#10.1038#ismej.2011.182		South Pacific Gyre, South Pacific Gyre
00094#10.1038#ismej.2013.79		USA
00095#10.1038#ismej.2007.5		Heron Reef, Great Barrier Reef, Australia
00096#10.1038#ngeo1830		
00097#10.1038#17351	Atlantic Ocean	
00098#10.1038#ismej.2013.234		Northwest Atlantic Ocean, Bedford Basin, Bedford Basin
00099#10.1038#ismej.2011.117		Atlantic
00100#10.1038#srep01292		

TABLE A.33: IllinoisNE Tagger Errors 1 of 4

Abstract	False Positive (test:yes, gold standard:no)	False Negative (test:no, gold standard: yes)
00001#10.1038#nmicro1749		Earth's
00002#10.1038#ngeo1765		
00003#10.1038#ngeo1910		Arctic, Arctic, Arctic, Arctic
00004#10.1038#nmicro1180		
00005#10.1038#nmicro1751		
00006#10.1038#srep00490		Mediterranean, Mediterranean, Mar Menor, Albufera de Valencia, Mar Menor, Albufera
00007#10.1038#nmicro956	Earth, Earth	Earth's, Earth's
00008#10.1038#nature04161		Polar ocean
00009#10.1038#ngeo964		
00010#10.1038#ngeo1916	Carbon	
00011#10.1038#nmicro1750		
00012#10.1038#nmicro2386		
00013#10.1038#ismej.2011.12	Gamma-, Delta-, Actinobacteria	
00014#10.1038#ismej.2010.33	PA, PA, PA, AE	
00015#10.1038#35037500	polar, polar	
00016#10.1038#ismej.2013.69	CO	Antarctic, Organic Lake, Organic Lake, Organic Lake, Antarctic
00017#10.1038#ismej.2013.144		
00018#10.1038#nature13123		Porcupine Abyssal Plain
00019#10.1038#nmicro1745		
00020#10.1038#ismej.2010.79		
00021#10.1038#ncomms3934	Earth	Earth's
00022#10.1038#ismej.2013.135		
00023#10.1038#ismej.2012.161	Coral, Coral	
00024#10.1038#ismej.2010.28		Ace Lake, Ace Lake
00025#10.1038#nature07235		Arctic, Arctic

Table A.34: IllinoisNE Tagger Errors 2 of 4

00026#10.1038#ismej.2012.31	Antarctica	Antarctica Peninsula, Antarctic, Antarctic Peninsula
00027#10.1038#ngeo1921		
00028#10.1038#nature12857		
00029#10.1038#35038000		
00030#10.1038#ismej.2010.185		
00031#10.1038#srep04661		
00032#10.1038#ismej.2013.121		
00033#10.1038#ismej.2012.28		Antarctic Peninsula, Antarctic Peninsula, Antarctic
00034#10.1038#ismej.2011.152	Pacific Ocean, Western, Pacific Ocean, Western South, Pacific Ocean	South Pacific Ocean, Western South Pacific Ocean, Western South Pacific Ocean
00035#10.1038#ismej.2008.125		
00036#10.1038#ngeo1265		
00037#10.1038#ismej.2014.39	D.	
00038#10.1038#nature04159		
00039#10.1038#ismej.2011.20	Carbon, Baltic Sea	
00040#10.1038#ismej.2009.60		
00041#10.1038#ismej.2008.15		Sep Reservoir
00042#10.1038#ismej.2009.29	Wadden Sea	German Wadden Sea
00043#10.1038#nature06267	Redfield	
00044#10.1038#nature01165		
00045#10.1038#ismej.2012.130	NT, NT	Mauritanian
00046#10.1038#ngeo234		
00047#10.1038#nature04158		
00048#10.1038#ismej.2008.117		Atlantic
00049#10.1038#ismej.2013.199		
00050#10.1038#ismej.2009.94		

TABLE A.35: IllinoisNE Tagger Errors 3 of 4

00051#10.1038#ismej.2007.70	River, Gulf, Mexico	Gulf of Mexico
00052#10.1038#ismej.2010.211		
00053#10.1038#ismej.2012.14	Earth	Northeast Atlantic, Earth's
00054#10.1038#ismej.2010.53		
00055#10.1038#nmicro2439		
00056#10.1038#srep02339		Polar Oceans, Arctic, Antarctic, Arctic, Antarctic, Arctic, Antarctic, Arctic, Antarctic, Antarctic, Arctic, Polar oceans
00057#10.1038#ismej.2011.201	Cyanobacteria, Bacteroidetes, Flavobacteria, Mimivirus	Southern California, California
00058#10.1038#ismej.2007.86		
00059#10.1038#ismej.2009.120	AA, Northern, Sweden, AA	Kryckian, Sweden
00060#10.1038#ismej.2014.16		Arctic, Arctic
00061#10.1038#ismej.2010.83	Haptophyta, Cercozoa, Haptophyta	
00062#10.1038#nature08985		
00063#10.1038#ismej.2013.190		Antarctic, Antarctic, McMurdo Dry Valley, Fryxell
00064#10.1038#ismej.2011.132		
00065#10.1038#ngeo2108		Earth
00066#10.1038#nature11229		
00067#10.1038#ismej.2009.8		South Pacific
00068#10.1038#nature04160		
00069#10.1038#ismej.2012.44		
00070#10.1038#nclimate1507	South China	South China Sea
00071#10.1038#ismej.2014.11		
00072#10.1038#nature02437	Gulf, Alaska	Gulf of Alaska
00073#10.1038#ismej.2010.91		
00074#10.1038#nclimate1989		
00075#10.1038#ismej.2011.118	DMSP-sulphur, DMSP-sulphur	Mediterranean

TABLE A.36: IllinoisNE Tagger Errors 4 of 4

00076#10.1038#ismej.2011.50			
00077#10.1038#ngeo1757			
00078#10.1038#ismej.2012.13			
00079#10.1038#ismej.2014.1	Alphaproteobacteria SAR11	North Pacific Subtropical Gyre	
00080#10.1038#ismej.2010.36			
00081#10.1038#ismej.2013.216			
00082#10.1038#ismej.2014.60	Ne, Ne		
00083#10.1038#ismej.2014.71			
00084#10.1038#ismej.2010.197			
00085#10.1038#ismej.2009.105	central Queensland	Queensland, Great Barrier Reef	
00086#10.1038#ismej.2008.56	OH	Levantine Basin, Levantine Basin	
00087#10.1038#ncomms2981	Iron		
00088#10.1038#ismej.2009.142	East, Greenland, BPLR, BPLR, BPLR	North Atlantic Ocean, North Atlantic Ocean, East Greenland, North Atlantic, Arctic	
00089#10.1038#ncomms4271			
00090#10.1038#nmicro2831	Earth	Earth's	
00091#10.1038#ismej.2011.214		North Atlantic, North Atlantic, North Atlantic Gyre	
00092#10.1038#ismej.2011.157			
00093#10.1038#ismej.2011.182	Gyre	South Pacific Gyre, South Pacific Gyre	
00094#10.1038#ismej.2013.79			
00095#10.1038#ismej.2007.5		Great Barrier Reef	
00096#10.1038#ngeo1830			
00097#10.1038#17351			
00098#10.1038#ismej.2013.234	OM43		
00099#10.1038#ismej.2011.117	Time-Series Study	Atlantic	
00100#10.1038#srep01292			

Bibliography

M. Alan Ritter, Sam Clark and O. Etzioni. Named entity recognition in tweets: an experimental study. volume EMNLP '11, pages 1524–1534, 2011. URL http://dl.acm.org/citation.cfm?id=2145432.2145595.

S. A. P. B. G. B. T. H. S. L. A. N. e. a. Alphonse, Erick. Event-based information extraction for the biomedical domain: the caderige project. pages 43–49. Association for Computational Linguistics, 2004. URL http://dl.acm.org/citation.cfm?id=1567602.

J. Baldridge. The opennlp project. http://opennlp.apache.org/index.html, 2005.

J. Berant, V. Srikumar, P.-C. Chen, A. Vander Linden, B. Harding, B. Huang, P. Clark, and C. D. Manning. Modeling biological processes for reading comprehension. pages 1499–1510. Association for Computational Linguistics, 2014. URL http://www.aclweb.org/anthology/D14-1159.

M. Berardi and D. Malerba. Learning recursive patterns for biomedical information extraction. *Lecture Notes in Computer Science*, 4455:79, 2007. URL http://link.springer.com/book/10.1007/978-3-540-73847-3#page=90.

A. Borthwick. *A maximum entropy approach to named entity recognition.* PhD thesis, New York University, 1999.

O. B. H. Y. T. D. S. V. K. T. A. P. S. Cameron, Delroy and T. C. Rindflesch. A graph-based recovery and decomposition of swanson's hypothesis using semantic predications. *Journal of biomedical informatics*, 46(2):238–251, 2013. URL http://www.sciencedirect.com/science/article/pii/S1532046412001517.

I. P. O. C. Change. Climate change 2007: The physical science basis. *Agenda*, 6 (7):333, 2007. URL http://www.slvwd.com/agendas/Full/2007/06-07-07/Item%2010b.pdf.

M. Ciaramita and Y. Altun. Named-entity recognition in novel domains with external lexical knowledge. 2005. URL http://nagoya.uchicago.edu/~altun/pubs/CiaAlt-NIPSWork05.pdf.

D. L. R. CL. The biological pump. volume 6, pages 83–111. Pergamon Press, 2006.

S. Coates-Stephens. The analysis and acquisition of proper names for the understanding of free text. *Computers and the Humanities*, 26(5-6):441–456, 1992. URL http://link.springer.com/article/10.1007/BF00136985.

P. Corbett and P. Murray-Rust. High-throughput identification of chemistry in life science texts. *Computational Life Sciences II*, pages 107–118, 2006. URL http://link.springer.com/chapter/10.1007/11875741_11.

R. L. M. C. E. C. dos Santos, Cícero N. and E. R. Fernandes. Etl ensembles for chunking, ner and srl. pages 100–112. Springer Berlin Heidelberg, 2010. URL http://link.springer.com/chapter/10.1007/978-3-642-12116-6_9#page-1.

D. K. S. Ducklow, Hugh W. and K. O. Buesseler. Upper ocean carbon export and the biological pump. *OCEANOGRAPHY-WASHINGTON DC-OCEANOGRAPHY SOCIETY*, 14(4):50–58, 2001. URL http://www.tos.org/oceanography/archive/14-4_ducklow.html.

A. Ekbal and S. Bandyopadhyay. A web-based bengali news corpus for named entity recognition. *Language Resources and Evaluation*, 42(2):173–182, 2008.

D. P. Elaine Marsh. Muc-7 evaluation of ie technology: Overview of results. 1998. URL http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/marsh_slides.pdf.

G. N. Gerner, Martin and C. M. Bergman. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85, 2010. URL http://www.biomedcentral.com/1471-2105/11/85.

R. Grishman and B. Sundheim. Message understanding conference-6: A brief history. volume 96, pages 466–471, 1996. URL http://www.alta.asn.au/events/altss_w2003_proc/altss/courses/molla/C96-1079.pdf.

K. F. H.-T. M. R. Z. Hanisch, Daniel and J. Fluck. Prominer: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6(1):14, 2005. URL http://www.biomedcentral.com/1471-2105/6/S1/S14.

M. D. Holzinger, Andreas and I. Jurisica. Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics*, 15(6):l1, 2014. URL http://www.biomedcentral.com/1471-2105/15/S6/I1.

J. T. Houghton and B. A. Callander. *Climate change 1992: the supplementary report to the IPCC scientific assessment.* Cambridge University Press, Cambridge, United Kingdom, 1992.

L. A. i Alemany and R. Carrascosa. A system for adaptive information extraction from highly informal text. pages 145–152. Springer Berlin Heidelberg, 2011. URL http://link.springer.com/chapter/10.1007/978-3-642-22327-3_14.

S. E. A. E. L. W. L. H. Jessop, David M. and P. Murray-Rust. Oscar4: a flexible architecture for chemical text-mining. *J. Cheminformatics*, 3(1):41, 2011. URL http://www.biomedcentral.com/content/pdf/1758-2946-3-41.pdf.

S. Z. F. W. Liu, Xiaohua and M. Zhou. Recognizing named entities in tweets. volume 1, pages 359–367. Association for Computational Linguistics, 2011. URL http://dl.acm.org/citation.cfm?id=2002519.

R. Mack and M. Hehenberger. Text-based knowledge discovery: search and mining of life-sciences documents. *Drug discovery today*, 7(11):89–98, 2002. URL http://www.sciencedirect.com/science/article/pii/S1359644602022869.

M. S. J. B. J. F. S. J. B. Manning, Christopher D. and D. McClosky. The stanford corenlp natural language processing toolkit. pages 55–60, 2014. URL http://www.aclweb.org/website/old_anthology/P/P14/P14-5.pdf#page=67.

P. O. E. A. G. S. Marsi, Erwin and M. V. Ardelan. Towards text mining in climate science: Extraction of quantitative variables and their relations. 2014. URL http://www.nactem.ac.uk/biotxtm2014/papers/Marsietal.pdf.

V. T. C. U. H. C. Maynard, Diana and Y. Wilks. Named entity recognition from diverse text types. pages 257–274, 2001. URL https://gate.ac.uk/sale/ranlp2001/maynard-etal.pdf.

R. S. R. W. Miller, David and R. Stone. Named entity extraction from broadcast news. In *Proceedings of the DARPA Broadcast News Workshop*, pages 37–40. Association for Computational Linguistics, 1999.

D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007. URL http://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad.

T. K. C. J. B. Naderi, Nona and R. Witte. Organismtagger: detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics*, 27(19):2721–2729, 2011. URL http://bioinformatics.oxfordjournals.org/content/27/19/2721.short.

M. M. Y. T. Nguyen, N. and S. Tojo. Open information extraction from biomedical literature using predicate-argument structure patterns. volume 19, pages 51–58, 2013. URL http://bioinformatics.oxfordjournals.org/content/19/suppl_1/i340.short.

V. J. F. O. A. L. B. S. C. D. R. A. F. A. G. e. a. Orr, James C. Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. *Nature*, 437(7059):681–686, 2005. URL http://www.nature.com/nature/journal/v437/n7059/abs/nature04095.html.

S. P. F. L. F. S. F. C. P. A. V. C. A. Pafilis, Evangelos and L. J. Jensen. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6):e65390, 2013. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0065390.

M. L. B. R. B. C. R. D. J. G. G. P. P. A. Pouchard, Line C. and N. F. Noy. A linked science investigation: enhancing climate change data discovery with semantic technologies. *Earth science informatics*, 6(3):175–185, 2013.

URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3883668&tool=pmcentrez&rendertype=abstract.

D. M. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011. URL http://dspace2.flinders.edu.au/xmlui/handle/2328/27165.

L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. pages 147–155. Association for Computational Linguistics, 2009. URL http://dl.acm.org/citation.cfm?id=1596399.

M. W. Rocktäschel, Tim and U. Leser. Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640, 2012. URL http://bioinformatics.oxfordjournals.org/content/28/12/1633.short.

J. B. M. W. C. D. M. J. L. B. H. Scaria, Aju Thalappillil and P. Clark. Learning biological processes with global constraints.

B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. Association for Computational Linguistics, 2004. URL http://dl.acm.org/citation.cfm?id=1567618.

D. R. Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18, 1986. URL http://muse.jhu.edu/journals/perspectives_in_biology_and_medicine/summary/v030/30.1.swanson.html.

A.-H. Tan. Text mining: The state of the art and the challenges. volume 8, page 65, 1999. URL http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf.

C. Thielen. An approach to proper name tagging for german. *arXiv preprint cmp-lg/9506024*, 1995. URL http://arxiv.org/abs/cmp-lg/9506024.

B. Upbin. Ibm's watson gets its first piece of business in healthcare. http://www.forbes.com/sites/bruceupbin/2013/02/08/ibms-watson-gets-its-first-piece-of-business-in-healthcare/, 2013.

R. A. F. S. M. V. Usié, Anabel and A. Valencia. Chener: chemical named entity recognizer. *Bioinformatics*, 30(7):1039–1040, 2014. URL http://bioinformatics.oxfordjournals.org/content/30/7/1039.short.

Wikipedia. Precision and recall. https://en.wikipedia.org/wiki/Precision_and_recall, 2013.

H. Yu and E. Agichtein. Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, 19(1):i340–i349, 2003. URL http://bioinformatics.oxfordjournals.org/content/19/suppl_1/i340.short.

G. Zhou and J. Su. Named entity recognition using an hmm-based chunk tagger. pages 473–480. Association for Computational Linguistics, 2002. URL http://dl.acm.org/citation.cfm?id=1073163.