NTNU
Innovation and Creativity

Financial News Mining:
Extracting useful Information from Continuous Streams of Text

Tarjei Lægreid
Paul Christian Sandal

Master of Science in Computer Science
Submission date: June 2006
Supervisor: Jon Atle Gulla, IDI
Co-supervisor: Jon Espen Ingvaldsen, IDI

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

Problem Description

News from the Norwegian paper Dagens Næringsliv and company disclosures from Oslo Stock Exchange (OSE) are examples of concise messages which describe important events in Norwegian industries. However, it is difficult to observe how different events relate. In addition it is hard to identify what kind of events that affect the different companies during a given period of time.

The objective of this thesis is to create a framework that enables extraction of entities (companies, people and other concepts) from web based business news. The framework should also offer continuous monitoring of how entity properties and relationships evolve over time.

The students should make use of statistic and linguistic techniques to analyse the contents of the documents. They must study different approaches and suggest a solution based on their experience and the findings.

Assignment given: 20. January 2006
Supervisor: Jon Atle Gulla, IDI

Abstract

Online financial news sources continuously publish information about actors involved in the Norwegian financial market. These are often short messages describing temporal relations. However, the amount of information is overwhelming and it requires a great effort to stay up to date on both the latest news and historical relations. Therefore it would have been advantageous to automatically analyse the information.

In this report we present a framework for identifying actors and relations between them. Text mining techniques are employed to extract the relations and how they evolve over time. Techniques such as part of speech tagging, named entity identification, along with traditional information retrieval and information extraction methods are employed. Features extracted from the news articles are represented as vectors in a vector space. The framework employs the feature vectors to identify and describe relations between entities in the financial market.

A qualitative evaluation of the framework shows that the approach has promising results. Our main finding is that vector representations of features have potential for detecting relations between actors, and how these relations evolve. We also found that the approach taken is dependent on an accurate identification of named entities.

Preface

This report documents the work performed in our Master's Thesis from January to June 2006 at the Information Systems Group of the Norwegian University of Science and Technology, Department of Computer and Information Science.

We would like to express our gratitude to our supervisor Jon Espen Ingvaldsen for encouraging comments and useful feedback on our work. Further, our thanks go to Professor Jon Atle Gulla for fruitful discussions and productive guidance.

Trondheim, June 15, 2006

Tarjei Lægreid Paul Christian Sandal

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem

Oslo Børs ASA[1] is the name of the Norwegian stock exchange. It is a public market place for trade in securities regulated by the fundamental principles of supply and demand. Several actors are involved in the market; listed companies, investors, brokers and monetary authorities to name a few. Each has their own roles and interests. Information related to the stock exchange is continuously published through various sources. To be able to cope with the high speed and dynamics of the market, it would be beneficial to maintain an overview of this stream of information. Due to the large volume of information this is a great challenge.

Text mining is a collective term covering various approaches to discovery of useful, preferably former unknown, information in large amounts of text. This thesis investigates the applicability of text mining operations as a means to manage such continuous streams of information.

1.2 Objective

The objective of this work is to develop a text mining framework able to continuously manage information related to the stock market from textual news streams. It should be able to capture and analyse textual data to provide simplified access to information regarding actors in the stock market and indicate how relations between different actors change over time.

1.3 Interpretations and limitations

We now quote the problem description and present important assumptions and limitations done. The original problem description is given below:

> *News from the Norwegian paper Dagens Næringsliv and company disclosures from Oslo Stock Exchange (OSE) are examples of concise*

[1] http://www.oslobors.no/

> *messages which describe important events in Norwegian industries. However, it is difficult to observe how different events relate. In addition it is hard to identify what kind of events that affect the different companies during a given period of time.*
>
> *The objective of this thesis is to create a framework that enables extraction of entities (companies, people and other concepts) from web based business news. The framework should also offer continuous monitoring of how entity properties and relationships evolve over time.The students should make use of statistic and linguistic techniques to analyse the contents of the documents. They must study different approaches and suggest a solution based on their experience and the findings.*

Stock related messages from six different sources on the Web will constitute the data foundation of the analysis:

- Dagens Næringsliv
- HegnarOnline
- Aftenposten.no
- iMarkedet
- Nettavisen
- Siste.no

The messages will be stored and indexed by the open source search engine Apache Lucene.

During analysis only persons and companies will be regarded as key actors. Lists of person- and company names from the financial domain will be used to support the identification. The mining will be focused on extraction of information related to the identified actors. Relations between different actors and how they evolve over time is of special interest.

Since the primary focus of this work is application of text mining, factors such as speed, scalability and space utilisation will not be emphasised. Further, our work on visualisation of the results will be limited to suggestions of possible approaches.

1.4 Approach

Previous research within the fields of information retrieval, information extraction and natural language processing constitute our starting point. Various methods

from these fields will be combined in an attempt to extract essential information regarding actors in the stock market.

The following approach will be carried out:

Create vector representations	The extracted information will be transformed to vector representations. This enables utilisation of the vector space model as a tool in the analysis.
Identify actors	An effort to identification of companies and persons mentioned in the extracted information should be presented. Static lists will be employed for this purpose.
Identify relations between actors	Emphasis will be put on the ability to indentify relations between various actors mentioned in the texts, and how the relations change over time.

Strategies to extraction and transformation of the texts will be tested. A qualitative evaluation of the ability to extract key information related to actors, how different actors relate to each other, and how these characteristics evolve over time, forms the basis of a discussion of the framework's performance.

1.5 Outline of the report

The report will be structured the following way:

In Chapter 2 fundamental theory will be described. Relevant techniques and the foundation the framework is built on are presented. Chapter 3 describes related work. Both general approaches dealing with temporal dynamics of text, and approaches dealing with financial news are included. A description of how the framework is implemented, and how the theory from Chapter 2 is employed in practice, is dealt with in Chapter 4. The components of the framework are evaluated in Chapter 5, followed by a discussion of our findings in Chapter 6. Finally, in Chapter 7 concluding remarks are given along with possible directions for further work.

Chapter 2

Theoretical background

Text mining can be regarded as a collective term of attempts to "discover useful and previously unknown information from unstructured texts"
[Appelt and Israel, 1999]. At first glance this may sound like an unatainable goal, but the idea is *not* to make systems that are better fit to interpret texts than human readers. It is when the amount of text grows intangible for humans, and the synthesis of information crosses the borders of multiple documents, that such systems should show to advantage.

Text mining is related to data mining, where the aim is to explore and analyse large volumes of data in order to discover meaningful patterns and rules [J. A. Berry and Linoff, 2004]. But while data mining generally is carried out in large databases of structured data, the input to a text mining process is unstructured text. The characteristics of natural language, with its liberal grammar and richness of words, impose challenges not present in traditional data mining [Holt and Chung, 1999].

Neither Information Retrieval nor Information Extraction matches the definition of text mining. Both types of systems seek to deliver output that satisfies some predefined information need. This aim differs from that of text mining, where the idea is to "...use text directly to discover heretofore unknown information" [Hearst, 1999]. Despite this fact, it may be useful to incorporate elements from these fields as components of a text mining system. Data may be made available and suitable for mining by having an IR module able to retrieve the desired text, and an IE module able to extract, transform and structure essential elements from the retrieval result. The process may be supported by techniques from natural language processing as an attempt to improve the foundation of the mining process. Finally, suitable mining techniques may be executed on the transformed text.

2.1 Information retrieval

The overall aim of an information retrieval system is to retrieve information relevant to the users information need [Appelt and Israel, 1999]. The user usually

expresses her information need through a (often small) set of keywords referred to as a query, and the system's purpose is to retrieve documents with a (perhaps partially) matching set of keywords. The underlying assumption is that if the query properly represents the user's information need, documents containing the same words probably have relevance to the user. The three tasks shown in Figure 2.1, namely indexing, searching and ranking, usually constitute the main parts of an information retrieval system [Baeza-Yates and Ribeiro Neto, 1999].

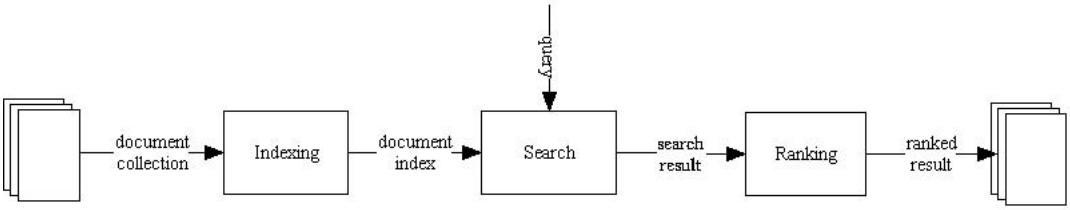

Figure 2.1: Conceptual view of typical IR architecture

2.1.1 Indexing

A naive approach when searching for words in a text document would be to perform sequential pattern matching. But as the document collection grows, the performance of this method declines, and soon it would obviously be too expensive. Thus, a more sophisticated and effective approach is needed.

Inverted file

A common way of indexing a document collection is to create an inverted file. It is a lexicographically ordered list of index terms in the document collection. An index term is "...a word whose semantics helps in remembering the documents' main themes. Thus, index terms are used to index and summarise the document contents." [Baeza-Yates and Ribeiro Neto, 1999]. Index terms may be single words, or a combination of words, i.e. phrases.

Each distinct index term, along with a reference to the document(s) where it occurs, constitute an entry in the inverted file. The data structure may be viewed as a matrix where each row represents an index term (word or phrase) and each column represents a document. Each cell of this matrix indicates the connection between the corresponding index term and document. References to the exact positions of the index terms in the documents may also be included, as suggesed by e.g. [Jackson and Moulinier, 2002]. This enable the searching and ranking algorithms to take relative term distance into consideration during their calculations. Put together, the inverted file contains the means to enable effective

search for query terms. One may simply look up each of the query terms in the index, and retrieve the documents listed there. An example illustrating the data structure is given in Figure 2.2.

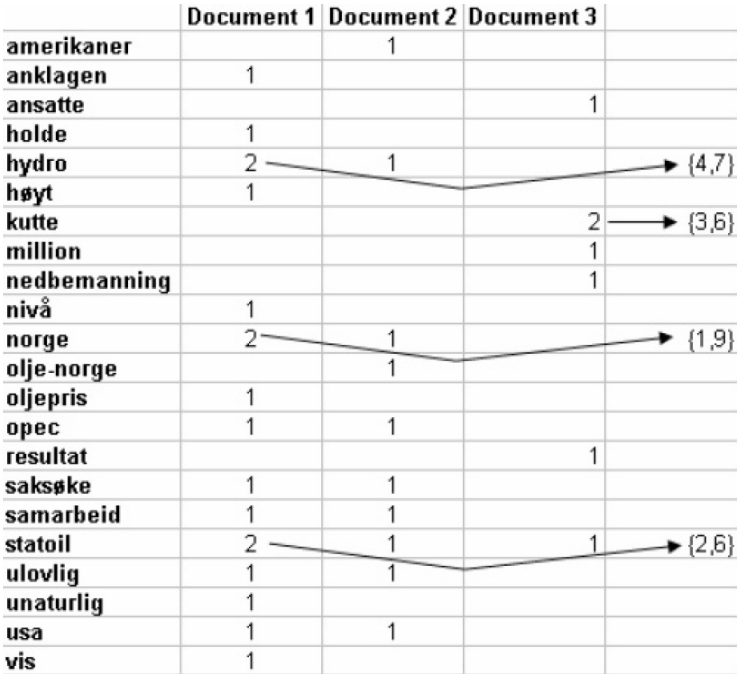

Figure 2.2: An example of an inverted file containing information from three short documents referring to the Norwegian oil company Statoil. The term-document matrix illustrated here contains frequency counts of each word. Often, as is explained later in this section, term weights that i.a. take the frequency into account are used as cell values instead of mere frequencies. In addition, each cell has a pointer to a list holding the position of each word in the documents (for clarity, only a few examples of these pointers are shown).

Selection of index terms

The selection of suitable, content bearing index terms is not straightforward. Ideally, the index terms should be able to express the semantics of the document. But automatic identification of such a set is difficult. In fact, since what is considered to be the content of a document may vary considerably from person to person, there exists no such thing as an ideal set of index terms. A solution to this problem is simply to consider all the words of a document as index terms. Yet simple, such an approach tends to impose an unacceptable level of noise [Baeza-Yates and Ribeiro Neto, 1999].

Selecting index terms is an important element of the text preprocessing. During preprocessing the text is prepared for indexing. Other important preprocessing elements are:

Lexical analysis

Lexical analysis involves converting a stream of characters into a set of words [Fox, 1992]. If possible, it may also be desirable to keep information about the logical structure of the documents (i.e. preserve information about titles, sentences, paragraphs, sections, etc.)

Stop word removal

Words with uniform frequency distributions across the documents in a collection are generally not suitable as discriminators. Hence they add little but noise to the search process and can be excluded from the index. Articles, prepositions and conjunctions are typical stop word candidates, but other words might also be removed because of their distributions in the collection. Thus, stop word lists are generally domain dependent and a word considered as a stop word in one domain may be an essential index term in another [Wilbur and Sirotkin, 1992].

Stemming and lemmatization

A word may occur in different variants in natural language texts, e.g. plurals, gerund forms and past tense suffixes. In the context of information retrieval it is often preferable that different variants of a word are indexed by a common index term. Two common approaches to word form normalisation exists, namely lemmatizing and stemming. The former is the process of transforming a word into its morphological root. One way of achieving this is to determine the words part of speech (see Subsection 2.3.1), and perform dictionary lookup. This requires knowledge of the grammar of the given language [Korenius et al., 2004].

The responsibility of a stemmer is limited to automatically determine the stem form of a given word. This is normally done by incrementally cutting off different kinds of suffixes. The resulting stem may not be the correct morphological root of the word, but the approach tend to work quite well as long as all text under consideration is stemmed by the same stemming algorithm.

Even if these steps prepare a document for indexing, the remaining text may still impose an unacceptable level of noise. To handle this, more advanced methods for selecting the index terms may be needed. Some classes of words carry more meaning then others. Nouns and groups of nouns generally represent document content better than e.g. adverbs and prepositions. A part of speech tagger (described in Subsection 2.3.1) may be used to identify word classes. This enables selection of nouns and groups of nouns that appear close to each other as index terms. In fact, sequences of words with any word class pattern may be selected through simple pattern matching with the preferred pattern of tags. More

complex information extraction methods (described in Section 2.2) may also be applied in the attempt of selecting better index terms.

Weighting of index terms

When the selection of index terms is completed the index may be built. As mentioned earlier the content of a cell in the inverted index indicates the connection between the corresponding index term and document. One question still remains: How should this connection be represented? A simple approach is to use binary values. The documents containing the index term are simply assigned the value 1, while the remaining documents are given the value 0.

A more sophisticated approach is to assign weights that reflect the index terms importance in describing the content of the document. A common scheme that tries to achieve this is called tf*idf weighting. It tries to balance two components, namely the term frequency (tf) and the inverse document frequency (idf) [Manning and Schütze, 2003].

The term frequency reflects the number of occurrences of the term in the document. The underlying assumption is that the more frequent the term appears, the more important is it in describing the document's content.

The inverse document frequency deals with the fact that terms that occur evenly throughout the document collection have limited discriminatory effect. Hence, words that occur in many documents are assigned a low idf score, while terms that occur in few documents get a higher score.

By combining these two components, the tf*idf weighting scheme favours index terms that occur frequently in the document, but infrequently across the document collection. Several weighting schemes are discussed in detail in e.g. [Lee, 1995].

2.1.2 Searching

The search for documents matching a given query may be viewed as the process of calculating the similarity between the query and documents in the index. The similarity might simply be a boolean categorisation of relevant and not relevant documents, or it might be reflected through some numerical value. The higher the score, the more relevant the document. Three classic models of information retrieval search exist: the boolean model, the probabilistic model and the vector space model.

In the boolean model queries are expressed as boolean expressions, and set opera-

tions and boolean algebra are employed to decide whether a document is relevant or not. The probabilistic model, as the name implies, is based on probabilistic methods and statistics [Baeza-Yates and Ribeiro Neto, 1999].

The vector space model

The information retrieval component employed in this project (Lucene[1]) is based on the vector space model. In this model each document is represented as a vector with the dimensionality determined by the number of index terms in the document collection. Each dimension is allocated a weight representing the corresponding index term's importance in the document. Further, queries are represented as vectors in the same vector space, and the degree of similarity between the query vector and document vectors is used as an indicator of document relevance.

The similarity between two vectors is typically quantified by evaluating the angle between the vectors. The less the angle, the higher the similarity. The *cosine similarity*, i.e. the cosine of this angle, is a common measure of vector similarity. In the vector space model, the scheme is typically used to calculate the similarity between the query vector and the document vectors of the documents indexed by the system. A desirable side effect of this approach is that it also gives a ranking of the retrieved documents; the higher the similarity, the higher the rank of the document.

The cosine similarity may of course be used to calculate the similarity between any pairs of vectors. Examples are calculation of inter-document similarity and (as will be discussed later) quantification of temporal change (vectors that change over time).

2.1.3 Ranking

The output of a search using the vector space model is a set of documents with calculated relevance scores. These scores may be used to rank the documents directly, but often it may be desirable to take other factors into account to improve the ranking. Examples of such factors are proximity (the closer the query terms occur to each other in the document, the higher the score) and document structure (e.g. boosting documents where query terms are found in titles, etc.) [Baeza-Yates and Ribeiro Neto, 1999].

[1] Apache Lucene is an open-source text search engine library, available at http://lucene.apache.org/

2.2 Information extraction

Information Extraction (IE) may be defined as "the process of deriving disambiguated quantifiable data from natural language texts in service of some prespecified precise information need." [Cunningham, 2004]. An IE-system takes unstructured text as input and the aim is to extract and present the data (snippets of information) of interest from the text. The output of the system is usually structured in a format suitable for further processing, either by a system or by the user directly.

As illustrated in Figure 2.3, Information Extraction may be split into five tasks of increasing complexity:

1.	**Named entity identification:**	Identification and classification of proper names. (Entities are usually denoted by their proper names [Appelt and Israel, 1999])
2.	**Coreference resolution:**	Identification of terms and sequences of terms that point to the same entity, e.g. 'NTNU' and 'Norwegian University of Science and Technology' refer to the same institution.
3.	**Template element construction:**	Adding attributes to entities. Combines results from 1 and 2 into a suitable format or template, e.g. adding 'NTNU' as an alias for the entity 'Norwegian University of Science and Technology'.
4.	**Template relation construction:**	Identification of relations between entities, e.g. some kind of 'located-in' relation between the city 'Trondheim' and its university 'NTNU'.
5	**Scenario template production:**	Identification of events including entities. Ties entities to event descriptions, e.g. 'NTNU is in growth'.

An IE-system generally needs to be tailored to the task at hand to some degree. If the information to be extracted is complex (e.g. scenario template production), the system is likely to require high domain adaption to give satisfactory results. If, on the contrary, the data to be extracted is simple (e.g. named entity identification), general methods may suffice.

Other factors also come to play. The type of text input to the system is essential. Extracting information from a news article is different from extraction from a

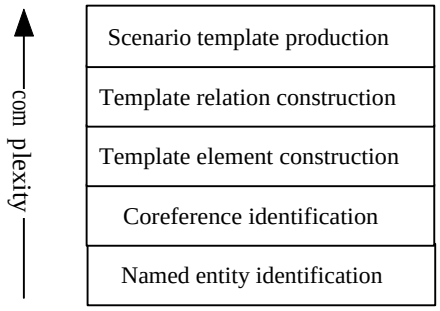

Figure 2.3: Five tasks of information extraction

weblog. The distinctive characteristics of the domain of discourse are also vital. Generally, texts that employ a consistent language and describe a narrow set of topics are better suited for information extraction than ambiguated texts. Thus, if the texts under consideration employ a fairly consistent language, have similar structure and are focused to a narrow set of topics, then the development of an IE-system able to deliver acceptable results is within range [Cunningham, 2004].

There are two main approaches to the development of an IE-system: *Knowledge Engineering approach* and *Automatic Training approach*. A system made with the first approach employs grammars and rules (i.e. a knowledge base) made by human experts, referred to as knowledge engineers. This method has the potential of delivering good results even for complex tasks, depending on the quality of the rules. But the development of a satisfactory rule base, considerable human effort of knowledge engineers with domain expertise is needed. Further, a rule base suitable for extracting complex information in one domain, is likely to be inappropriate in another.

The second approach addresses this issue of domain portability in that it relies on a training corpus to develop its domain knowledge. The system is trained (at least partially) using annotated examples from the domain of discourse, e.g. by utilising statistical methods and probabilistic models. In this way the system can be ported to different domains easily, if suitable training examples exist. The collection and/or creation of such training data, however is time-consuming and likely to become a bottleneck [Appelt and Israel, 1999].

2.2.1 Named entity identification

The task of identification and classification of named entities is of vital interest to our case. Persons and companies are essential actors in the stock market, and information related to these actors (e.g. events and relations) are of special in-

terest. Several problems are attached to the task of named entity identification, and the aim is to create rules able to cope with these problems. The rules may either be hand-crafted or based on automatic training [Appelt and Israel, 1999].

Named entity identification may be separated into (1) identification and (2) classification. In properly written mixed-case texts capital letters may aid the process of identification. But since capitalisation is also used to start sentences, in lists and opening quotation marks, to name a few, the presence of upper case letters alone gives no guarantee. The process may be supported by static lists of known names. With one list of names per category, this may solve the classification problem. While this may be adequate in restricted domains, the enormous number of alternatives makes it insufficient in general. The fact that new names emerge continuously, e.g. as new companies are established, aggravates the problem. Even with complete lists, overlaps need to be handled, since some names necessarily co-occur in various categories [Jackson and Moulinier, 2002].

The problems mentioned above demonstrate the need for rules applicable to decide whether a candidate is in fact a named entity, and to what category it belongs. Context and syntactic patterns come into consideration when developing rules. In the phrase 'George Bush says', the observation that two words with initial capitals are followed by the verb 'says', can be utilised to decide that George Bush is in fact the name of a person. The creation of such rules may be carried through in an incremental manner. First, start on with a small set of established rules. Second, run the system on some test set. Third, repeatedly create new rules to cover the cases not handled by the current set of rules until acceptable performance is met [Appelt and Israel, 1999].

2.3 Natural language processing

Information extraction is closely related to, and in some cases dependent on, elements from Natural Language Processing (NLP). NLP describes "...the function of software or hardware components in a computer system which analyse or synthesise spoken or written language" [Jackson and Moulinier, 2002]. The processing may either be based on predefined rules, or it may rely on statistical analysis of the language. The former approach, termed *knowledge based*, employs a base of rules created by domain experts. The latter, termed *empirical*, assigns probabilities to different interpretations of the text and use statistical methods to decide among them.

The processing of natural language may be separated into layers of increasing complexity, as illustrated in Figure 2.4. Lexical analysis (i.e. sentence delimiting, tokenisation and stemming) constitutes the bottom layer. This is dis-

cussed earlier (in Section 2.1) in the context of indexing. At the next level words are tagged with their part of speech. In the presentation set forth in [Jackson and Moulinier, 2002] this layer also includes identification and classification of proper names. We have already treated this subject as a task of information extraction in Section 2.2. These overlaps illustrate the weak boundaries between natural language processing, information retrieval and information extraction.

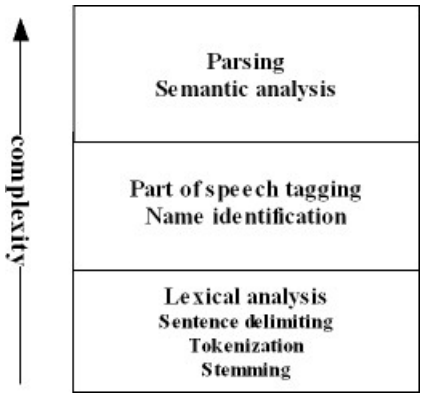

Figure 2.4: Typical layers of NLP

The top layer includes parsing and semantic analysis of the text, e.g. identification of different types of phrases and their roles in the context.

2.3.1 Part of speech tagging

Part of speech taggers label each word in a document with an appropriate word class (e.g. verb, adverb, adjective, noun, etc), based on some kind of tagging scheme.

A simple approach when determining the part of speech for a given word is to use dictionaries. The word is tagged based on the possible part(s) of speech from the dictionary. However, if this approach is employed, the tagger will lead to ambiguated tags for words which belong to multiple morphological classes.

To overcome the challenge of disambiguating tags, the tagger can be trained using a tagged corpus. Disambiguation rules can be derived from the corpus using different approaches, and thus the tagger knows which tag to assign a disambiguated word. Creating such rules requires a relatively large pre-tagged corpus.

Traditionally, two approaches have been used to train the tagger; rule-based and

stochastic models, or some combination of both. These will be explained shortly in the following paragraphs.

Rule based taggers

Rule based taggers take contextual information into consideration when assigning tags to ambiguous and unknown words. The contextual information is used as a basis for creating contextual rules. A contextual rule might be something like this: "if a determinant is followed by an unknown word, followed by a noun, then tag the word as an adjective" (determinant -> X -> noun ==> X/adjective). The rules are based on some kind of experience; they can be derived from a tagged corpus, or they can be manually hand-coded (which is a time-consuming and costly process).

As described above, typical rule based taggers resolve tagging ambiguities by using a training corpus to compute the probability of a given word in a given context. But research has also been done on automatic induction of rules. One such approach is to run an untagged text through a tagger, and then manually go through the output of this phase and correct the erroneously tagged words. Then, the properly tagged text is submitted to the tagger, which learns correction rules by comparing the two sets of data [Brill, 1992].

Some rule based taggers also use morphological information in the disambiguation process. This means that language-specific characteristics are taken into consideration when building the rules. E.g. for English text a typical rule is "if a word X, precedes a verb, and ends with -ing, then X is a verb". This rule will not necessarily be the applicable in other languages. Some taggers also include rules pertaining to factors as punctuation and capitalisation (capitalisation is very valuable in determining the tags for unknown nouns in e.g. the German language, while this is not the case in Norwegian or English).

Stochastic based taggers

Stochastic (or statistical) taggers rely on frequency information or probability distributions. The tagger is trained on a manually tagged corpus. The trained model is then used to assign parts of speech to the untagged text. Given an input sentence and the probabilistic model generated during training, the combination of assignments with the highest multiplied product is selected.

More sophisticated stochastic taggers calculate the probability of different sequences of tags (n-grams). The assumption behind these taggers is that the correct tag for a given word can be determined by the probability that it succeeds the n previous tags. One such approach is Hidden Markov Model (HMM)

tagging schemes. HMMs are regarded as a probabilistic variant of a finite state machine, which switch states according to transition probabilities. In the tagging domain, states correspond to tags, and the optimal tag sequence for a given sequence of words is formulated as the search which finds the state sequence which has the maximum probability given a sequence of words [Kazama et al., 2001].

The assumptions behind this model are:

- Each hidden tag state produces a word in the sentence.
- Each word is uncorrelated with neighbouring words and their tags.
- Each tag is probabilistic dependent on the N previous tags only.

This means that both tag sequence probabilities and word frequencies are considered when assigning tags [Dermatas and Kokkinakis, 1995].

[Ratnaparkhi, 1996] suggests using a Maximum Entropy Model for part of speech tagging. The underlying idea is to choose the probability distribution that has the highest entropy among the distributions that satisfy a set of constraints. The constraints are built from training data (a pre-tagged corpus), and restricts the model to behave in accordance with statistics collected from the corpus. The statistics are expressed as the expected values of functions defined on contexts and tags, and the expectations of the features for the model match the expectations of the features for the training data.

Maximum Entropy taggers are thus not restricted to n-gram sequences (such as HMM taggers), but incorporates much more complex statistics
[Toutanova and Manning, 2000].

Tag sets

Different tag sets have been used for part of speech tagging of text [Wu et al., 1992]. The size of the tag sets depends on the language, objectives and purpose of the tagging. Some tag sets use a large number of tags, while others make fewer distinctions, e.g. one tag for each word class (e.g. verbs -> V, Nouns -> N, etc). Generally, a simple morphology requires fewer tags than a complex morphology. Some of the most widespread tag sets are listed in Table 2.3 [Manning and Schütze, 2003].

The Penn Treebank tag set, which is a simplified version of the Brown tag set, is the most widely used tag set for computational work.

Tag set	Number of tags
Brown	87
Penn Treebank	45
CLAWS1	132
CLAWS2	166
CLAWS5	62
London-Lund	197
Lancaster-Oslo-Bergen	134

Table 2.3: Some common tag sets

Limitation and opportunities of Part of Speech taggers

Part of speech taggers might help the process of disambiguating the meaning of a word in its given context - at least partially. The tagger might be used to determine what role the word plays in the context [Manning and Schütze, 2003]. E.g. in the sentence *I run the store*, a tagger (ideally) tags *run* as a verb. However, the tagger is not able to recognise the meaning of *run* (i.e. it does not have the ability to differentiate between the interpretations *to move fast* and *to operate something*).

Tagging the parts of speech also opens for an opportunity to recognise phrases in the text [Cutting et al., 1992]. Typically two succeeding words with corresponding tags have a distinct meaning when they appear together. E.g. the phrase *New York* has a different meaning than *new* and *york*. If we know that the two succeeding words *new* and *york* are tagged similar, then we may conclude that they represent one entity.

Before we describe how we employ part of speech tagging in combination with other techniques described in this chapter, we will give an introduction to work related to our problem domain.

Chapter 3

Related work

In this chapter we investigate related work in the area of temporal text mining. We focus on both general approaches and applications in the specific domain of financial news.

3.1 General approaches

First, we look at several approaches to the general problem of managing the temporal dynamics of streams of text. Even though the approaches differ, the overall aim of all the systems is to be able to align different kinds of "virtual" events deduced from the streams, with real-world events on the fly. In the survey presented in [Kleinberg, 2006] such systems are categorised into four types, dependent on the approach taken. The same overall classification is employed here.

3.1.1 The Topic Detection and Tracking Initiative

Given a stream of news articles, the aim is to provide systems able to (1) identify stories that treat new events, i.e. perform *new event detection* and (2) perform *event tracking* and grouping of stories treating the same events [Kleinberg, 2006].

Evolutionary theme patterns

An example of such a system is presented in [Mei and Zhai, 2005]. There, temporal text mining is defined as the process of "discovering temporal patterns in text information collected over time." In their work, they treat the particular task of discovering how topics in a text stream evolve over time. The underlying idea is that topics and events covered in series of news articles generally have an underlying temporal and evolutionary structure characterising the beginning, progression and termination of the event.

Probabilistic methods are explotied to:

1. Discover latent patterns from text.

2. Construct an evolution graph based on the discovered patterns.

3. Analyse the life cycles of patterns and how they influence subsequent patterns.

The aim is to discover, extract and summarise these evolutionary patterns automatically. The suggested soulution is threefold:

1. Partition documents into fixed or variable (perhaps overlapping) time intervals.

2. Extract most salient themes in each interval.

3. For any themes in two different intervals, use the similarity between the two to decide whether there is a evolutionary trasition (i.e. connection)

Evaluation of their approach shows that the proposed techniques are able to generate meaningful temporal theme structures that enable summarisation and analysis of text data in a temporal perspective.

New event detection

Another algorithm for detection of new events in news streams is presented in [Allan et al., 1998]. It mainly utilise elements from information retrieval and clustering. The processing of a new story starts by extracting the most frequent words/features of the text. These then constitute a query vector. The vector is compared to its originating story, and the calculated similarity is set as a the query's threshold. Then, the story is compared to all chronologically preceding queryvectors. If the story exceeds any of the similarity-thresholds of these preceding queries, the new story is considered as a continuation of the story represented by that particular query. To cope with the fact that time matters (stories treating the same event are likely to occur close in time), the similarity is multiplied with a timefactor. The longer the timegap, the smaller the factor, and hence the computed similarity is reduced.

3.1.2 Threshold-based methods and timelines

As mentioned in [Kleinberg, 2006], threshold-based approaches aim at creating a timeline containing the most significant *episodes* in text streams in temporal order. They typically rely on extraction and evaluation of a set of features, e.g. named entities and content bearing nouns, each with an average occurence-rate over time (number of occurrences divided by e.g. number of days that the stream covers). A feature that exceeds its average by a defined threshold over a contigous interval of time is defined as an episode, and the episodes with the highest positive variance is considered most significant and added to the timeline.

TimeMines

An example of a system employing such an approach is described in [Swan and Jensen, 2000]. The output of this system is a timeline of statistically significant topics in the input text. Elements from both information extraction and natural language processing are utilised to extract the most relevant and important topics within a given timeframe. The selection of important topics is based on the assumption that the distribution of a feature follows a base rate that does not vary in time. The system then compares all features against their base rate. Features that differ significantly from their expected occurance rates are kept for further processing. This reduced set of features are then grouped together. The grouping is based on a coupling of features that co-occur in timeframes more often than by chance into topics. Finally, a threshold is used to determine the most important of these topics. The resulting set of topics are presented on the timeline for further human processing.

3.1.3 State-based methods.

Systems applying state-based methods rely on automaton theory to model the occurence rates of features. The rate of each distinct feature is modeled to be generated by its own two-state automaton. The automaton is either in its base-state (normal occurence rate) or in its burst state (occurence rates above normal). Statistical methods are used to compute the state of the automaton at any given point of time. The difference between the probabilities of the two states is used as a confidence, and the features with highest confidence are included on the timeline [Kleinberg, 2006].

3.1.4 Trend-based methods.

In trend-based systems, streams are typically grouped into discrete time-steps, and linear regression is applied on the sets of frequencies for each term. A positive slope indicate a *rising term*, while a negative slope indicates a *falling term* [Kleinberg, 2006].

PatentMiner

PatentMiner is an example of a prototype that attempts to extract information about trends. It is described in [Lent et al., 1997]. The input text collection (patent descriptions) is grouped into discrete time intervals based on the documents' date information. The main topics of each time interval are then extracted using elements from data mining. A topic is defined as a set of words. The user defines the allowed relative distance between the words describing a topic. In this way a topic may be of different granularity, e.g. containing words from single sentences, or words from different sentences, paragraphs or sections. The

topics are generated using a variant of sequential pattern mining and is based on frequency counts of co-occuring words. Only topics exceeding a user defined support threshold are kept for further processing. Next, the *history* of each of the extracted topics is generated. The history of a topic is simply the number of documents the topic occurs in during the various time intervals. Finally the user may query the collection of topic histories for topics of different kinds of trends, e.g. a topic with a rising trend would typically have increasing frequencies in a contigous interval of time.

Hierarchical Distributed Dynamic Indexing

Another system that divides the input text into discrete time intervals is presented in [Kontostathis et al., 2003]. For each interval, the system is able to extract complex noun phrases using a part of speech tagger and a regular expression indicating the sentence strucuture. The extracted phrases, named concepts, are then grouped together into semantic regions based on similarity calculations. The frequency counts of each concept along with the semantic regions of each time iterval are input to a neural network able to decide whether a concept is emerging or not. The underlying assumptions are that an emerging concept should both have a growing frequency as more text units reference it, and it should appear in an increasing number of semantic clusters, since it is likely to co-occur with an increasing number of other concepts.

3.2 Domain specific approaches

Although the systems described above give useful insight to possible techniques, none of them is applied on the particular domain of financial news. In this section we look at systems that employ text mining to analyse continuously published financial information.

NewsCats

The News Categorisation and Trading System [Mittermayer, 2004] describes the implementation of a system able to predict how press publications influence stock price trends immediately after the release. The system analyses and categorises finanicial press releases and deduce trading recommendations automatically. It consists of three main components:

Retrieval component — This component performs text preprocessing, including feature extraction (parsing, stop word removal and stemming) and feature selection based on standard tf*idf calculations. Each document is represented as a normalised feature vector.

Categorization component	This component is responsible for categorising the document vectors. Several alternatives are mentioned; decision trees, decision rules, k-nearest neighbors, Bayesian approaches, neural networks, regression-based methods and vector-based methods. The classifier actually implemented in NewsCats is based on SVM (Support Vector Machines).
Strategy component	This component gives directions such as whether to buy or sell stocks based on the output of the categorisation.

To test the system they look at historical data. With knowledge of the exact publication moment and the development of the stock price immediately after the release (subsequent 60 minutes), they are able to select a training set consisting of the three categories; *good news*, *bad news* and *no movers*. These categories are basis of three dictionaries of words. The 1000 words of each dictionary with highest tf*idf score are used to create three support vectors, which are used to train the system. The selectivity of *no movers* is reported to be fairly good, while the clustering of *good news* and *bad news* is fairly bad. The system is reported to significantly outperform a trader randomly buying and shorting stocks immediately after press releases.

The predicting power of textual information

A related approach is described in [Fung et al., 2005]. The focus of the work is to investigate textual information's impact on the financial markets. News stories are categorised as either *positive*, *negative* or *neutral*, and the effects on the stock prices is assumed to take place immediately after the publication. The categorisation of documents is based on the text classification algorihm SVM. To train the classifier, a set of useful news stories is selected based on whether they contain features that support a rise or fall of the stock prices. A χ^2-distribution is used to model the occurrences of features. Features that extend a defined χ^2-value in periods of growth is said to support rising stock prices and vice versa.

The trained system is employed to categorise new stories as they are published. The results of the classification is used to support the buying and shorting of stocks. Evaluation shows that the approach is able to give useful directions.

The domain specific approaches mentioned here focus on the influence published information has on the stock prices, and how such knowledge can be utilised for decision support. As will be described in the next chapter, we approach the problem of utilising such information on a more general basis.

Chapter 4

Implementation

In this chapter we give a brief description of our implementation. We have aimed at creating a flexible framework applicable for mining information in the domain of finacial news. The functionality is based on elements from both information retrieval, information extraction and natural language processing (Chapter 2), and incorporates ideas and aspects from related research (Chapter 3).

The components of the framework are loosely coupled in order to provide the desired flexiblity. This simplifies the process of changing the mode of operation. Since the framework should be able to support the investigation of different mining strategies, this is a desirable effect.

4.1 Architecture

Figure 4.1 shows the architecture of the implemented framework at a conceptual level. Articles are fetched by the article fetcher as they are published on the

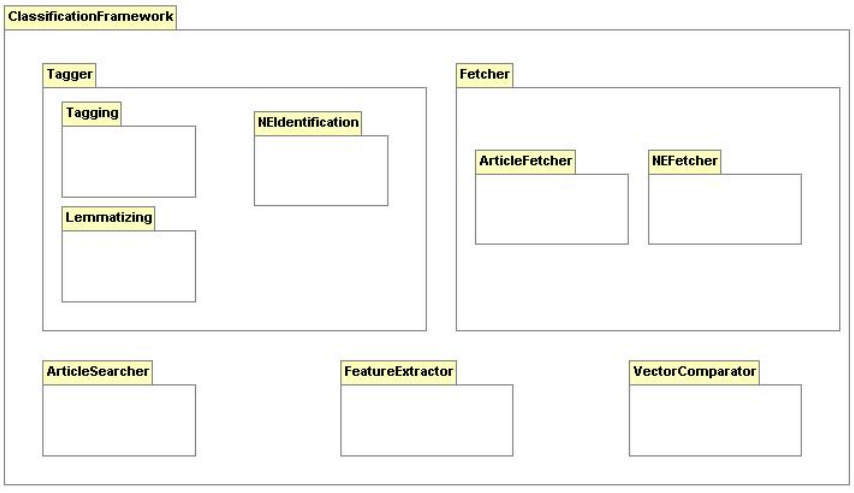

Figure 4.1: An overview of the implemented framework.

Web. Each article is time stamped and part of speech tagged. The Lucene API is used by the fetcher to index and store the articles. By keeping information

about the time of publication, the article index may be queried for information from different periods of time.

Articles judged relevant to a query are input to the feature extractor. It seeks to extract features descriptive for the result set, i.a. by utilising the part of speech tags. The tags are also used to identify persons and companies (actors) mentioned in the text. The extracted features are used to represent the result set in a vector space. Finally, the vector comparator component is able to calculate the similarity between any pair of such result set vectors.

Put together, the framework holds functionality to build and compare feature vectors of any set of query results. By allowing time-limited queries, it also enables analyses of temporal change. This makes it applicable in various scenarios.

4.2 Fetcher

The fetcher module contains functionality for fetching data from online information sources.

4.2.1 Article Fetcher

Online newspapers typically offer limited services for reading old news articles. As a consequence, we have implemented functionality for storing and indexing newspaper articles.

The article fetcher periodically (every 10 minutes) listens for newly distributed newspaper articles. This is done by crawling the newspapers' RSS-feeds. The articles are part of speech tagged and stored in the article index. As the RSS-feeds only contain article titles and ingresses, the article-urls are also visited and their content is stored in the index. The publication date and time is also stored.

4.2.2 Named Entity Fetcher

As a part of the tagging module (described in Section 4.3) static lists of organisations and persons are employed to identify companies and persons in the articles. The lists are based on information from the Brønnøysund Register Centre[1]. In practice, the lists are retrieved by crawling the web-site Proff.no[2]. The lists are

[1]The Brønnøysund Register Centre is a government body under the Norwegian Ministry of Trade and Industry, and consists of several different national computerised registers.

[2]The Proff.no-site (`http://www.proff.no`) offers information from the Brønnøysund Register Centre. For practical reasons the information has been fetched from this web site.

stored in separate indices, which are searched during the named entity identification.

4.3 Part of speech tagger

As described in Subsection 2.3.1, different approaches to part of speech tagging have been suggested. Independent of what approach is chosen, some kind of rules need to be employed. There are basically two ways of creating such rules; create them manually by hand or create them automatically by extracting statistics from a pre-tagged corpus. Creating the rules manually is obvoiously not feasible for this project. This means that we need a publicly available tagger.

Different pre-trained taggers are publicy available online. Most of these taggers are trained for English, and some of them also offer tagging for european languages. However, none of them are trained for Norwegian text, and therefore the tagger must be trained using a pre-tagged Norwegian corpus. Unfortunately, the access to such materials is limited.

[Johannessen et al., 2000] at the Text Laboratory, the Unversity of Oslo, have developed a Norwegian tagger; the Oslo-Bergen tagger. The tagger is not publicly available, but it can be tested through a web interface[3]. This means that, in the absence of a manually tagged Norwegian corpus, we can use the Oslo-Bergen tagger to create our ”own” training corpus. Using text tagged by the Oslo-Bergen tagger for training will probably lead to a tagger that is less accurate than one based on manually tagged data, but in the lack of alternatives this constitute a sufficient solution.

Building the tagged corpus

Since we are building our own training corpus, we have the advantage that we can adapt the corpus to the domain of discourse. Thus, we choose to build the corpus from Norwegian newspaper articles. The selected articles are fed into the Oslo-Bergen web-interface. The tagger returns triplets in the format <word, lemma, tag>. Only the <word, tag> tuples are included in the training corpus. In addition utilise the ’lemma’-part of the triplets to create the lemmatization functionality.

The corpus is based on financial news from the following Norwegian newspaper-websites, released in the period January 15, 2006 to May 15, 2006:

- Dagens Næringsliv (`http://www.dn.no`)

[3]The Oslo-Bergen tagger web-interface is available at `http://beta.visl.sdu.dk/visl/nor/parsing/automatic/parse.php`

- HegnarOnline (`http://www.hegnar.no`)
- Aftenposten.no (`http://www.aftenposten.no/nyheter/okonomi`)
- iMarkedet (`http://www.imarkedet.no`)
- Nettavisen (`http://pub.tv2.no/nettavisen`)
- Siste.no (`http://www.siste.no`)

The corpus consists of 2 441 087 words distributed into 162 013 sentences.

Tag set

Some common tag sets were presented in Subsection 2.3.1. The Oslo-Bergen tagger employs an extended version of the Oslo-Bergen-Lancaster tag set. This tag set is rather large, and enables a high degree of accuracy. However, for our purpose such a large tag set is unnecessary. As pointed out in [Manning and Schütze, 2003] "the larger the tag set, the more potential ambiguity, and the harder the tagging task ... For example, some tag sets make a distinction between the preposition **to** and the infinitive marker **to**, and some don't. Using the latter tag set, one can't tag **to** wrongly." This means that the tagger will probably be more accurate when using a smaller tag set. Therefore, we have decided to use the Penn Treebank tag set.

As mentioned above, the Oslo-Bergen tagger employs the Oslo-Bergen-Lancaster tag set. Therefore, the training set must be converted to the Penn Treebank tag set. For this purpose we have created at tag-to-tag conversion scheme, listed in Appendix A.

Training the tagger

Part of speech taggers differ in the model they use for tagging, but performance evaluation shows only small differences in accuracy.
[Toutanova and Manning, 2000] claim that this occurs because ".. all these methods use largely the same information sources for tagging, and often almost the same features as well, and as a consequence they also offer very similar levels of performance". With support in this statement, we will not perform any comparison of the accuracy of different taggers.

We have chosen to employ the maximum entropy based tagger from the OpenNLP[4] project. The tagger is trained with the news-corpus created by the Oslo-Bergen tagger.

[4]OpenNLP: Organizational center for open source projects related to natural language processing, `http://opennlp.sourceforge.net/`

Lemmatization

To overcome the challenge of words occuring in different inflected forms, we have implemented functionality for lemmatization.

As explained earlier in this section, lemmatization lists are created from the output-triplets of the Oslo-Bergen tagger. For each tag, a word-lemma list is constructed. These lists are used, after the tagging is done, to find the appropriate lemma for each tagged word. If any duplicate entries occur, the most frequent entry is chosen. Words that have no entry in the lemmatization lists are left as is.

As an alternative, stemming functionality could have been implemented. However, the word stems are not ”real words”, and we consider these as less useful than lemmas for our purpose. The example in Table 4.1 illustrates why.

Original sentence	Daglig leder i Det Norske Veritas, Torolf Bernhard Aadnesen hevder...
Lemmatized sentence	Daglig leder i Det Norske Veritas, Torolf Bernhard Aadnesen hevde...
Stemmed sentence	Dag led i Det Norsk Veritas, Torolf Bernhard Aadnes hevd

Table 4.1: An example illustrating the benefits of lemmatizing instead of stemming. It is very difficult to interpret what ”Dag led” really means.

Named entity identification

In addition to functionality for training and employing a part of speech tagger, the tagging module contains a feature for named entity identification. Our approach is based on using static lists of proper names. We assume that the tagger identifies proper names (NNP-tags in the Penn Treebank tag set) correctly. Conceptually, each sequence of one or more NNP-tags is then looked up in the static lists and tagged with own-defined tags. We have employed this method to persons, locations and organisations (including companies), but the approach could be extended to handle other proper names, e.g. technologies or products as well.

Given a lemmatized and part of speech tagged text, and three proper name indices (person index, location index and organisation index), then the named entity identifiaction is done as follows:

1. Extract NNP phrases	All NNP tag sequences in the text are looked up, and their corresponding words are extracted. E.g. from the text *Statoil/NNP og/CC Norsk/NNP Hydro/NNP falle/VBD mye/JJR enn/IN 3/CD prosent/NN ./.*, the phrases *Statoil* and *Norsk Hydro* will be extracted.
2. Search propername indices for phrases	The propername indices are searched for the phrases previously extracted. All relevant hits are returned (partially hits are not extracted, i.e. all the words in a phrase must occur in a hit).
3. Determine the entities	The searches in the previous step returnes tf*idf weighted hits. The entity-class of the highest-weighted hit is considered as the correct classification. If more than one hit has the highest weight, then the entity is selected in the following priority: 1. Person, 2. Company, 3. Location. The reason for this rank is that surnames often occur as company names, e.g. the company 'Vollvik gruppen'. If 'Vollvik' occurs in a text, then the referenced actor is probably the person 'Idar Harry Vollvik'.
4. Assign named entity tags	Each identified named entity is tagged by one of the following tags: *NNPP* (persons), *NNPC* (organisations/companies), *NNPL* (locations). Phrases with no hits from step 2 are left as-is (*NNP*).

This means that when the text of the example in (1.) above is run through the named entity identifiactor, the result would (hopefully) look like this: *Statoil/NNPC og/CC Norsk/NNPC Hydro/NNPC falle/VBD mye/JJR enn/IN 3/CD prosent/NN ./.*.

The functionality described so far form the preprocessing and indexing functionality of the framework. The sequence of the indexing process is illustrated in Figure 4.2.

4.4 Feature extractor

This module takes tagged text as input, and is responsible for converting the text into a vector representation. This is done by reducing the text to a set of features. Each distinct feature then constitutes a vector dimention that is assigned a numerical weight.

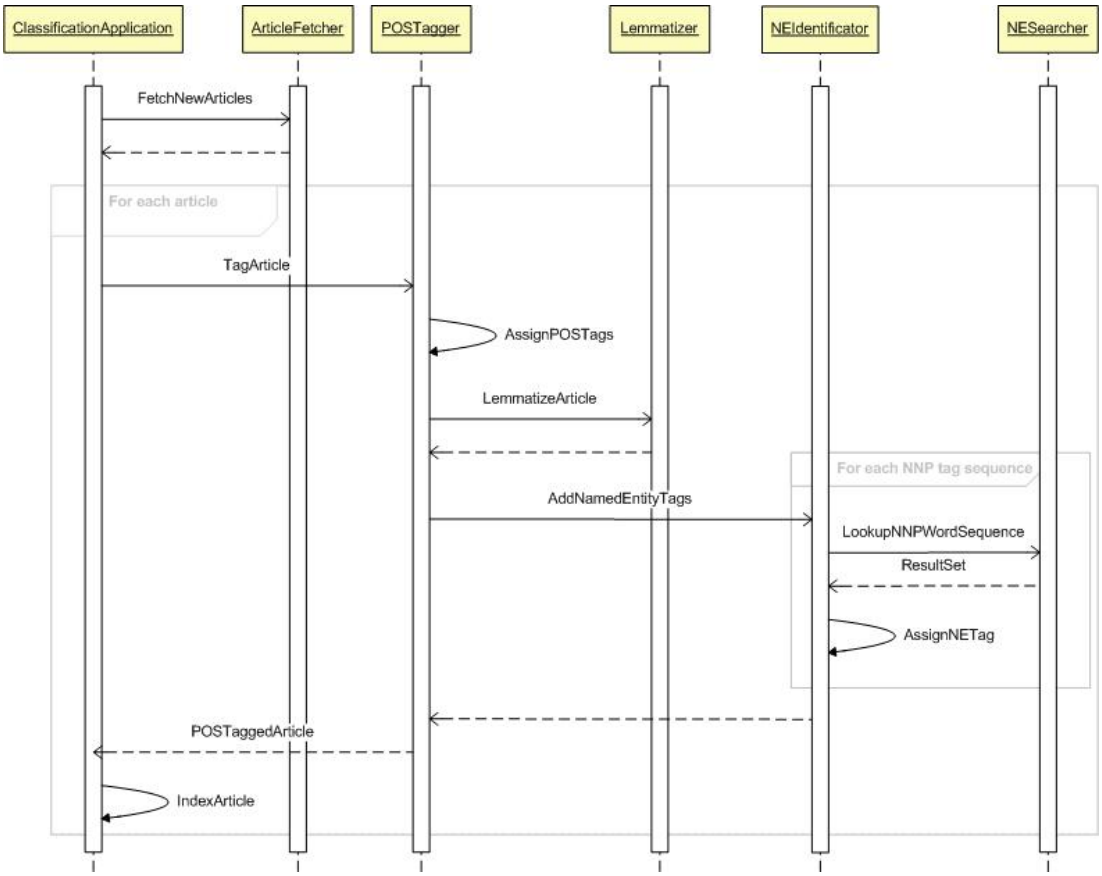

Figure 4.2: UML sequence diagram illustrating article indexing.

4.4.1 Feature extraction

The simplest approach to feature extraction is to view each distinct word of the text as a feature. Yet simple, the result is generally of high dimensionality and includes a considerable element of noise. Although these effects may be reduced by removing stop words, the result may still be to voluminous and vague. Recall from Subsection 2.1.1 that good candidates as index terms are "...words whose semantics helps in remembering the documents' main themes" [Baeza-Yates and Ribeiro Neto, 1999]. The same definition applies to the extraction of features.

As an attempt to enable extraction of fewer, but more descriptive features, the module is capable of utilising the knowledge of the words' part of speech. The approach is based on the work described in [Gulla et al., 2006] The implemented version is able to extract features following any desired pattern of tags. Thus, the outcome of the module may be altered simply by changing the set of tag-patterns

to search for.

Feature weighting

As described in Subsection 2.1.1 a weighting scheme that aims at calculating weights able to reflect the feature's importance in describing the content of the text is preferable. Several factors, both related to the text itself and its context come to play in the selection of an appropriate weighting scheme. What works good in one context may give poor results in another. Thus, several schemes are implemented and will be tested during evaluation.

4.5 Vector comparator

The vector comparator module is responsible for calculating the similarity between feature vectors created by the information extraction module. To achieve this, it employs cosine similarity calculation, previously described in Subsection 2.1.2. The formula is given in Equation 4.1.

$$sim(\vec{v}_a, \vec{v}_b) = \frac{\vec{v}_a \bullet \vec{v}_b}{|\vec{v}_a| \times |\vec{v}_b|} = \frac{\sum_{i=1}^{f} w_{i,a} \times w_{i,b}}{\sqrt{\sum_{i=1}^{f} w_{i,a}^2} \times \sqrt{\sum_{i=1}^{f} w_{i,b}^2}}, \text{ where} \quad (4.1)$$

$\vec{v}_a$,$\vec{v}_b$	=	the two feature vectors to be compared.
$sim(\vec{v}_a\vec{v}_b)$	=	the similarity between $\vec{v}_a$ and $\vec{v}_b$.
f	=	the total number of features.
$w_{i,a}$	=	the weight of feature i in vector a.
$w_{i,b}$	=	the weight of feature i in vector b.

Figure 4.3: Equation for calculating cosine similarity.

In the next chapter we will present an evaluation of different parts of the implemented framework.

Chapter 5

Evaluation

A formal objective evaluation of the implemented framework requires knowledge of a desired result for a given test set of input data. By comparing the outcome of the framework and the desired result, one would be able to quantify and measure the performance.

The main problem of this approach is that there exists no single correct result. Different people may have different interpretations of the same text. A result perceived as correct by one person may be wrong in the eyes of another.

To overcome this, the desired result may be determined by merging the opinion of several independent domain experts. Such an approach is used by e.g. the Text Retrieval Conference (TREC)[1]. There, the National Institute of Standards and Technology (NIST)[2] provides test sets of documents and tasks. Conference participants run their systems on the data and their results are evaluated by NIST.

Unfortunately, neither any conference document collection matching our domain and language, nor any domain experts able to define a desired result are available to us. Thus, we attempt to carry out a qualitative evaluation of the frameworks performance. Although this approach lacks the objectiveness necessary to perform a trustworthy formal evaluation, it will give hints about the frameworks capabilities and serve as a useful basis for discussion.

While we lack a predefined desired result from the framework as a whole, this is not the case for all its subcomponents. The part-of-speech tagger module is a central component in the system. As described in Section 4.3 the tagger is trained using a tagged corpus of news articles from the financial domain. Further, it utilises huge lists of names of persons, companies and geographic locations to support the identification of key actors in the analysis. By retaining parts of the tagged corpus during training, we have a subset of the corpus that may be used to measure the quality of the outcome of the tagger. This test set may be used both to evaluate the tagger's allocation of parts of speech, and its ability to

[1] `http://trec.nist.gov`
[2] `http://www.nist.gov`

classify persons and companies mentioned in the text.

The evaluation therefore consists of an isolated evaluation of the part of speech tagger and the named entity identifier, and a qualitative evaluation of the framework as a whole.

5.1 Part of speech tagger evaluation

Since the part of speech tagger forms the basis both for feature extraction and identification of names, it is of special interest to measure its performance.

5.1.1 Evaluation strategy

The evaluation of the tagger is performed on the following two different datasets:

Previously seen data. The taggers ability to tag its own training corpus gives an useful indication of the performance of the tagger's training algorithm.

Previously unseen data. The taggers ability to tag new, previously unseen text, is of course of vital interest since this indicates the quality of the model developed during training. To do this we use a subset of the training corpus not incorporated during training. In addition we test the tagger on a collection of articles from the Norwegian newspaper Adressa[3]. It is composed of news articles covering a wider set of topics than the corpus used for training. First we tag the corpus using the Oslo-Bergen tagger (as is also done with the trainig corpus). The tagged result (after translation to Penn Treebank tags) is then compared with the outcome of the trained tagger.

Characteristics regarding the size of the three corpora used in the tests are given in Table 5.2.

	# sentences	# words	Domain
Training corpus	162 013	2 441 087	Financial news
Unseen corpus	31 631	476 263	Financial news
Adressa corpus	5 478	98 264	General news (various categories)

Table 5.2: Test-corpora characteristics

[3] `http://www.adressa.no`

5.1.2 Evaluation metrics

The performance of the tagger is evaluated by the two metrics recall and precision. The calculation of these metrics is based on combinations of the properties true/false positives and true/false negatives. How these properties relate to the

Tag x	Assigned by OB	Not assigned by OB
Assigned by ME	TP_x(true positive)	FP_x(false positive)
Not assigned by ME	FN_x(false negative)	TN_x(true negative)

Table 5.3: Possible combinations of tag-assignments
OB=The Oslo-Bergen tagger
ME=The Maximum Entropy tagger

assignment of tags is shown in Table 5.3 [Jackson and Moulinier, 2002].

The calculation of recall is shown in Equation 5.1.

$$Recall_x = \frac{|TP_x|}{|TP_x| + |FN_x|}, \text{ where} \tag{5.1}$$

$|TP_x|$ = The number of times the taggers agree on the tag x.
$|FN_x|$ = The number of times a word is assigned the tag x by the Oslo-Bergen tagger, but not by the Maximum Entropy tagger.

The calculation of precision is shown in Equation 5.2.

$$Precision_x = \frac{|TP_x|}{|TP_x| + |FP_x|}, \text{ where} \tag{5.2}$$

$|TP_x|$ = The number of times the taggers agree on the tag x.
$|FP_x|$ = The number of times a word is assigned the tag x by the Maximum Entropy tagger, but not by the Oslo-Bergen tagger.

5.1.3 Evaluation results

As mentioned, the tagger evaluation is based on a benchmark test using the Oslo-Bergen tagger as the reference of measure. Ideally, the tagger should have been

compared to a manually tagged reference collection. Unfortunately, no such corpus was available to us.

The evaluation is done both on a per-tag basis and on an overall basis. Since the two taggers assign equal number of tags (one tag per word), the total precision equals the total recall. Speaking in terms of Figure 5.3, this means that $|TP_X| + |FP_X| = |TP_X| + |FN_X|$.

Training corpus

Figure 5.1 shows that the tagger has a total recall score of 96.4% when evaluated on the training corpus. The scores of the individual tags are relatively stable

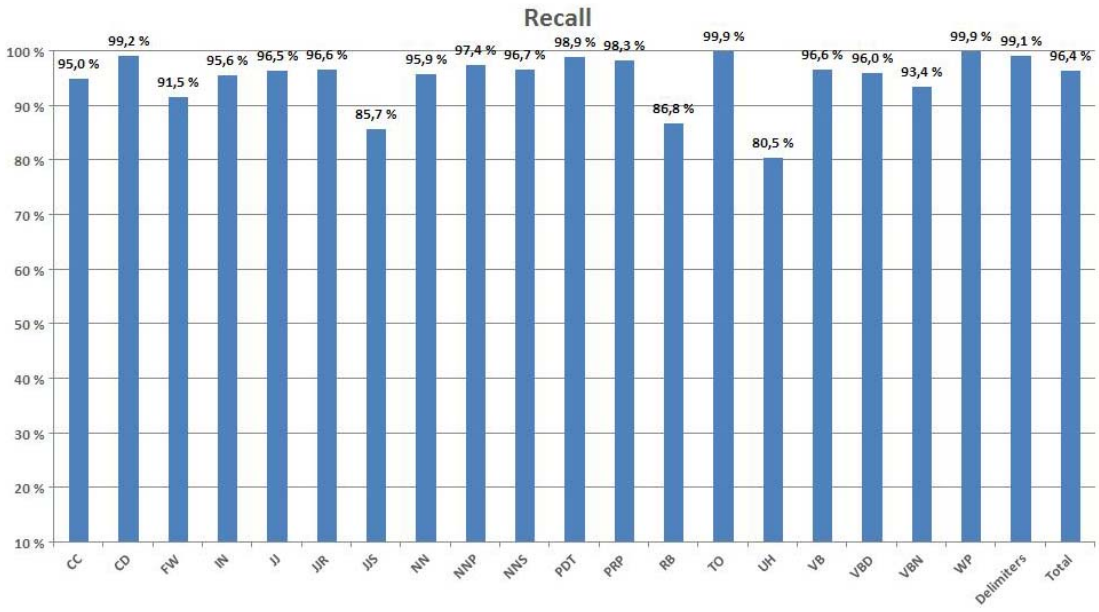

Figure 5.1: Tagger recall when evaluated on previously seen data (the training set data).

above 90%. However, it is worth noticing that the tags 'JJS' (adjective, superlative), 'RB' (adverb) and 'UH' (interjection) have substantial lower scores than the other tags.

A closer look at the data foundation (Appendix B.1) shows that the training set contains as few as 2384 ($\approx$ 0.10 %) UH-tags. Such a low number may be inadequate to train the tagger properly. Simultaneously, a closer look at the statistics shows that WP-tags are occurring even less frequently, but with a substantially higher recall than UH. This observation contradicts the above claim of an inadequate training foundation. Therefore we performed a manual inspection

of the training set. The inspection revealed that a large number of the UH-tags in the training set were results of incorrect tagging by the OB tagger. This illustrates a clear weakness of using the OB-tagged corpus as the training set.

The precision-score of the tagger is shown in Figure 5.2. Scores range from 89.3% to 99,9%.

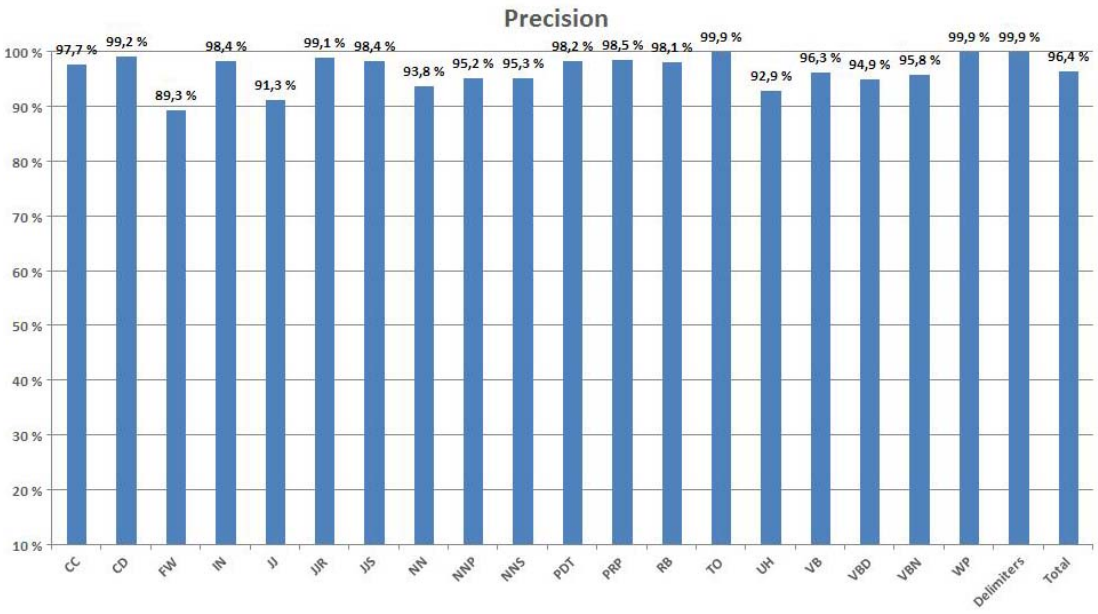

Figure 5.2: Tagger precision when evaluated on previously seen data (the training set data).

Previously unseen corpus

Although the results of the last section are promising, it is of greater interest to evaluate how well the tagger performs on previously unseen data.

With total precision/recall scores of 94.8%, the tagger performs weaker on previously unseen data then on previously seen data (as should be expected). Figure 5.3 shows that the tags 'FW', 'JJS', 'RB' and 'UH' have a substantially lower recall than was the case for the previously seen data.

The low scores of the 'UH' tags were discussed in the previous section. Here, foreign words ('FW'-tags) have low recall as well. An explanation to this may lie in the fact that some of the words in the testing-corpus don't occur in the training set. For some of these, the tagger is incapable of deciding the correct tag and thus assigns them as 'FW'. As Figure 5.4 shows, the precision of foreign words is also

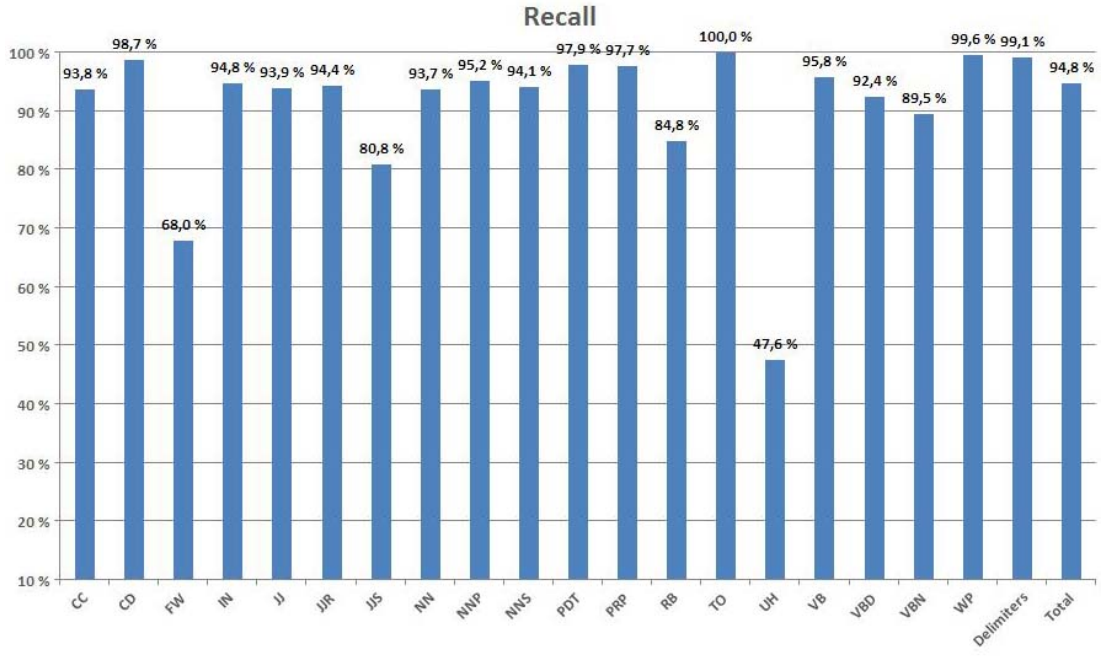

Figure 5.3: Tagger recall when evaluated on previously unseen data.

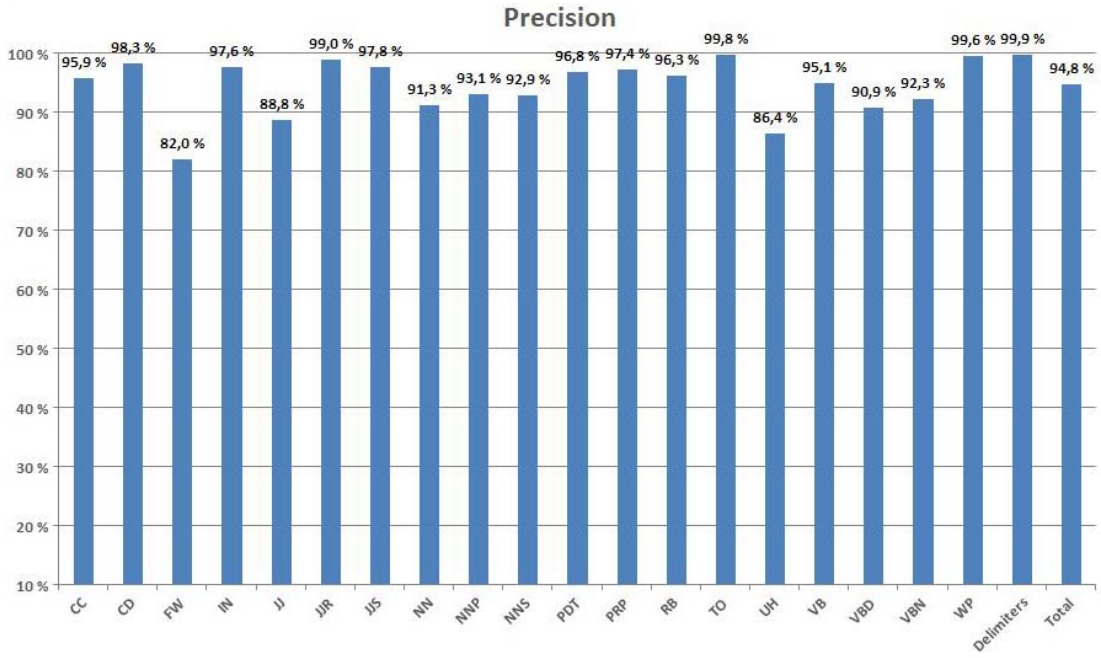

Figure 5.4: Tagger precision when evaluated on previously unseen data.

rather low. However, since the 'FW' tags are omitted by the analysis perfomed by the implemented framework, the consequences of these errors are limited in our case.

Adressa corpus

Even though the tagger is trained for the purpose of tagging financial news, it is of interest to evaluate how well it performs on a more general basis. To evaluate this, a corpus from the Norwegian newspaper Adresseavisa is employed. It contains news articles from multiple categories.

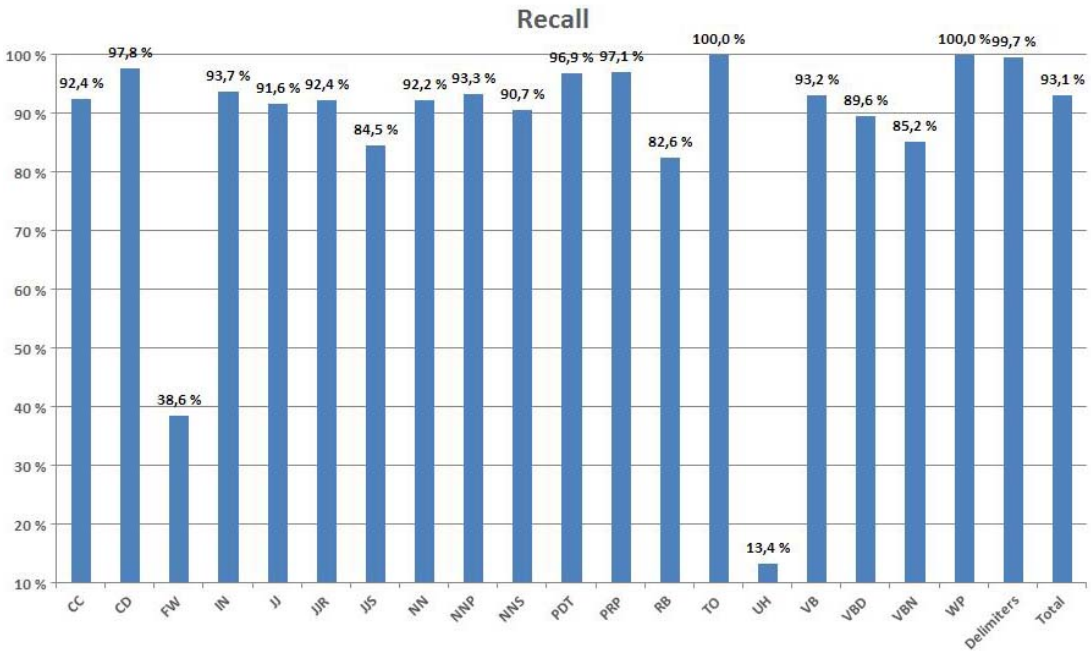

Figure 5.5: Tagger recall when evaluated on previously unseen data (the Adressa corpus).

The test results show an amplified picture of the results from the previously unseen financial news. The overall recall/precision is 93.1%. As Figure 5.5 shows, the recall scores of 'FW' (38.6%) and 'UH' (13.4%) are particularly low. This is also the case for the precision (see Figure 5.6).

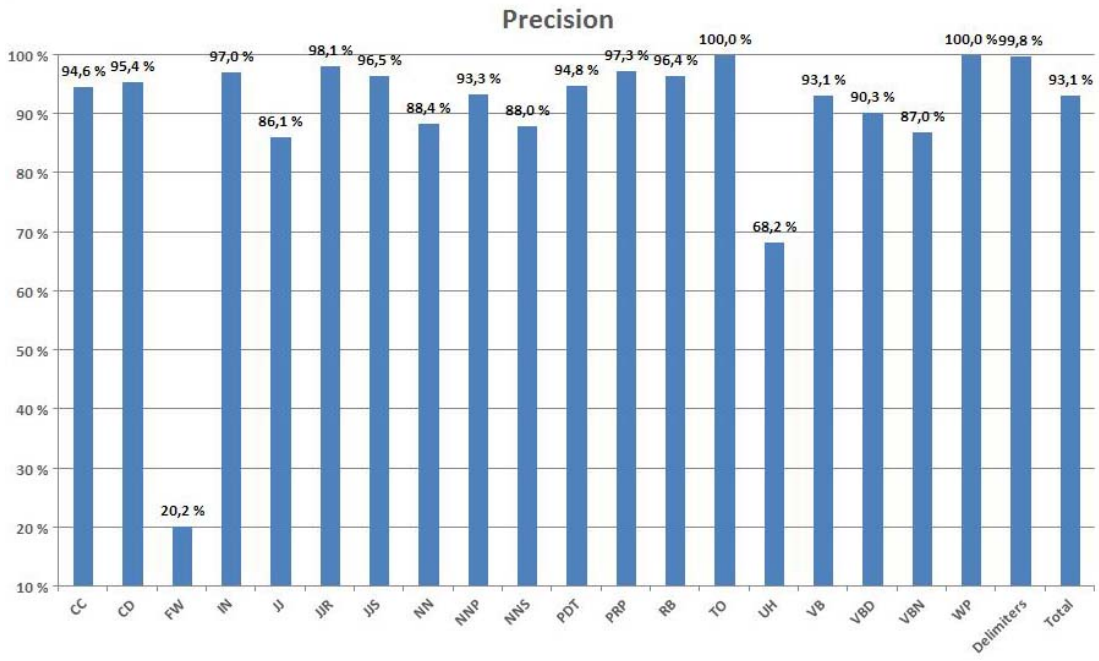

Figure 5.6: Tagger precision when evaluated on previously unseen data (the Adressa corpus).

5.2 Named Entity identification evaluation

Even though the named entity (NE) identification functionality is implemented as a part of the part-of-speech tagger, we will evaluate it separately. Recall from Section 4.3 that the NE identification relies on the part-of-speech tagger, and is thus vulnerable to errors made by the tagger. We will not take this into consideration in this evaluation. In practice this means that non-identified named entities will be evaluated as not discovered independent of whether they are tagged with ”NNP” by the tagger or not. After all, what caused the mistake is insignificant for the end result. The interesting fact is *that* the mistake actually occurs.

5.2.1 Evaluation strategy

The evaluation of the NE functionality is based on a manual walkthrough of an excerpt of the news article collection consisting of 111 articles. The articles are listed in Appendix C (original format) and Appendix D (tagged format). Even though the data set is limited, the size is adequate to give an indication of how well the NE identification performs.

5.2.2 Evaluation metrics

For each article, the number of correct identified entities (persons, companies and geographical names), the number of wrongly identified entities, and the number of non-identified entities is extracted. This is used as basis for measuring precision and recall (described in Section 5.1). The evaluation is performed on a per-entity basis. This means that if an entity is only partially recognised, then it is considered as *not* identified. For instance, 'hydro' would not be considered as a correct identification in the sentence 'rekordtall fra norsk hydro', while 'norsk hydro' would).

5.2.3 Evaluation results

The named entity identification has an average precision of 87.4%, and a recall of 86.6%. Evaluation measures for each article is shown in Figure 5.7. The pre-

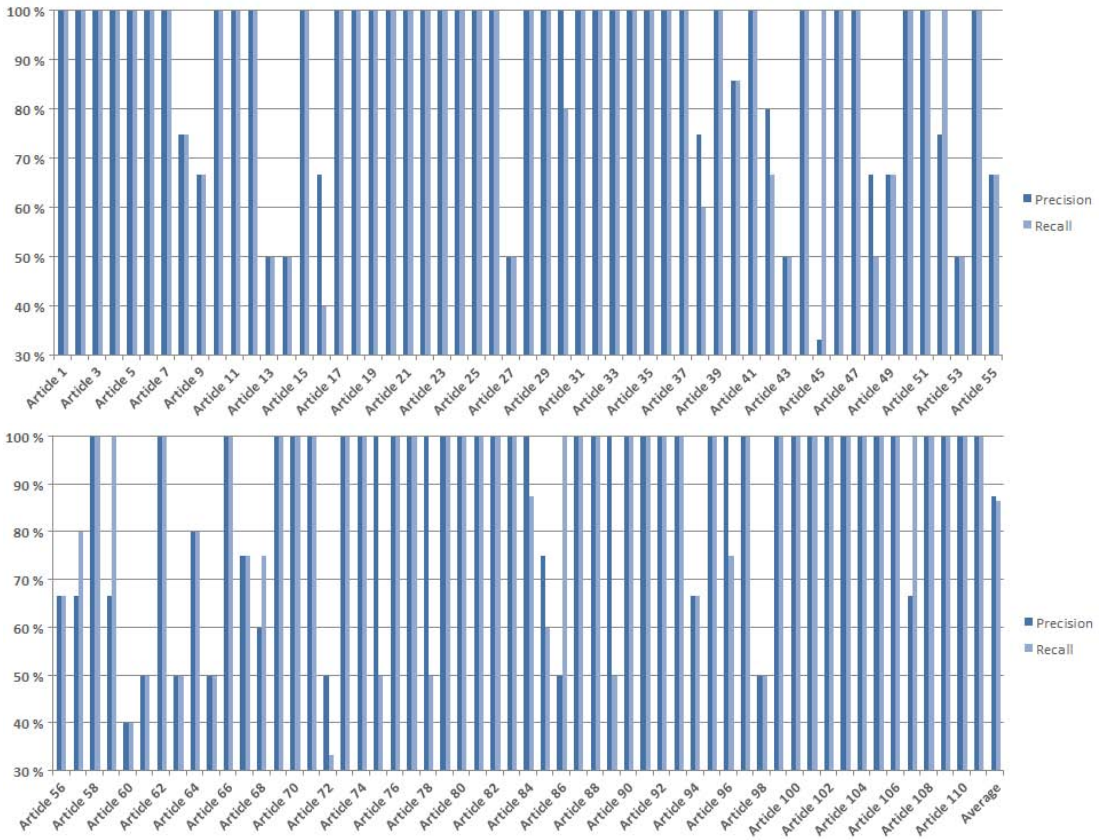

Figure 5.7: Named Entity Identification precision/recall scores.

cision and recall varies from 33% to 100%. The main reason for the large recall differences is that an entity is often referred to multiple times in an article. If

such an entity pass unidentified, the recall will suffer. A large number of the non-identified entities occurs because they do not appear in the lists that support the classification. In the approach taken occurance in the lists is a prerequisite for successfull identification.

The precision suffers from i.a. ambiguous proper names. An example is 'møre', which is an abbreviation of 'møre og romsdal' (a district in Norway). However, there are some companies which has 'møre' as part of their names, e.g. 'Digitaltrykk Møre' and 'Sparebanken Møre'. Since these two have a higher degree of similarity with 'møre' than 'møre og romsdal', 'møre' is faulty classified as an organisation. This example illustrates one of the main problems of the list approach.

5.3 Framework evaluation

Regarding the framework as a whole, two aspects are of special interest. First, the framework's ability to deduce relationships between actors should be evaluated. This includes both identification of connections between actors, and calculation of connection strength. Second, the framework's ability to indicate how these characteristics change over time should be considered.

5.3.1 Evaluation strategy

To do this we perform a qualitative evaluation. It is based on the creation of a scenario where a chosen actor is considered as the one of special interest. We sample articles mentioning the selected actor within three separate time frames of one week. Each of these "snapshots" is inspected manually, with focus on how the actor relates to other actors mentioned in the text. Our interpretation of the situation is then compared with the results of the framework.

We only sample titles and ingresses of the articles. The reason for this limitation is that financial news articles often end with a summary of recent developments at the stock market. Such summaries tend to list companies not necessarily related to the company described in the article. Thus, including the whole articles would probably reduce the quality of analysis. Since the main elements of the texts should be covered by the ingress, the danger of excluding vital information is limited.

Figure 5.8 illustrates how the framework is employed. It can be split into five subsequent activities:

1. A query representing the chosen actor is formed. The query is run within each of the three snapshots.

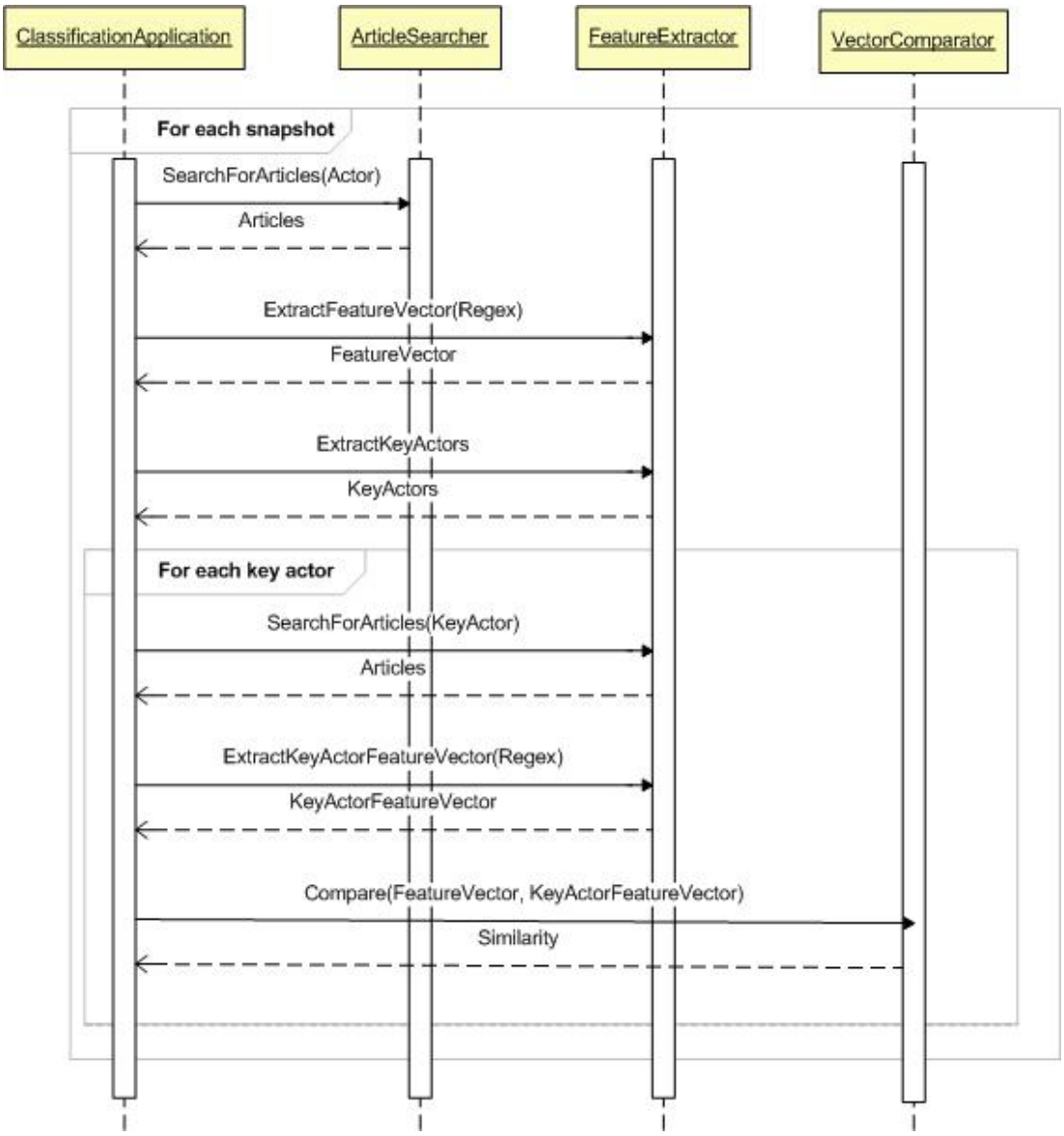

Figure 5.8: UML sequence diagram illustrating how the framework is employed in the evaluation scenario.

2. The result set (articles) of each of the queries is analysed by the framework, and one feature vector per query result set is built. The analysis also includes identification of other companies and persons, i.e. key actors, mentioned in the text.

3. For each of the other identified actors, queries are formed and run within the three snapshots, and feature vectors are built from the result sets.

4. The vectors belonging to the same time frame are then compared. This

is used to indicate how strong the different actors are related within the snapshot.

5. Finally the results of the different snapshots are compared in order to indicate how the situation evolves.

Feature extraction

Since the feature vectors extracted by the framework play a vital role in the analysis, it is of interest to evaluate the extraction of features' impact on the results. As described in Section 4.4 the framework is able to extract features following any pattern of part of speech tags.

In this evaluation the framework extracts features following the pattern (NNx|JJx)*NNx, i.e. any number of subsequent nouns and/or adjectives followed by one or several nouns. This pattern is also applied to extract features in the TimeMines system [Kontostathis et al., 2003], earlier described in Subsection 3.1.2. The results of using this pattern is compared with a baseline run using feature vectors composed merely of the tokenized text (i.e. the words in the text).

Feature weighting

Weighting of the extracted features is another important factor. In the evaluation two different approaches to weighting are tested. Weighting is previously described in Subsection 2.1.1.

tf*idf is a common weighting scheme in information retrieval [Baeza-Yates and Ribeiro Neto, 1999]. It aims at balancing the two factors of locality (features that occur frequently in a document are likely to be important in describing the content of the document) and globality (words with uniform distribution throughout the collection of documents are bad discriminators).

tf*df on the other hand gives a boost to the local frequency. The boost is proportional to the number of documents that the feature occurs in. Such an approach makes sence in our case since we extract features from result sets containing multiple documents. Thus we seek to give the features that best describe this *collection of documents* a higher weight, as opposed to the features that best discriminate between the documents in the collection.

News articles differ in their degree of formality. Some are to a large extent based on rumours, while others are restrained to objective facts. Therefore, we boost the news sources which we consider as most reliable (Dagens Næringsliv and Aftenposten). In addition we boost phrase length (the more words a phrase consists of, the higher the weight).

5.3.2 Evaluation results

To pick a scenario for evaluation, we limited the set of candidates to a set of large companies. For each of these, the documents mentioning its name were retrieved, and the hits counted on a per week basis. Under the assumption that a peak in number of hits indicates that the company is involved in some important event in that week (as described in e.g. [Kleinberg, 2006]), we selected Statoil ASA as our scenario. The hits for Statoil per week is shown in Figure 5.9. The three weeks with the largest number of documents are used as snapshots in the evaluation scenario.

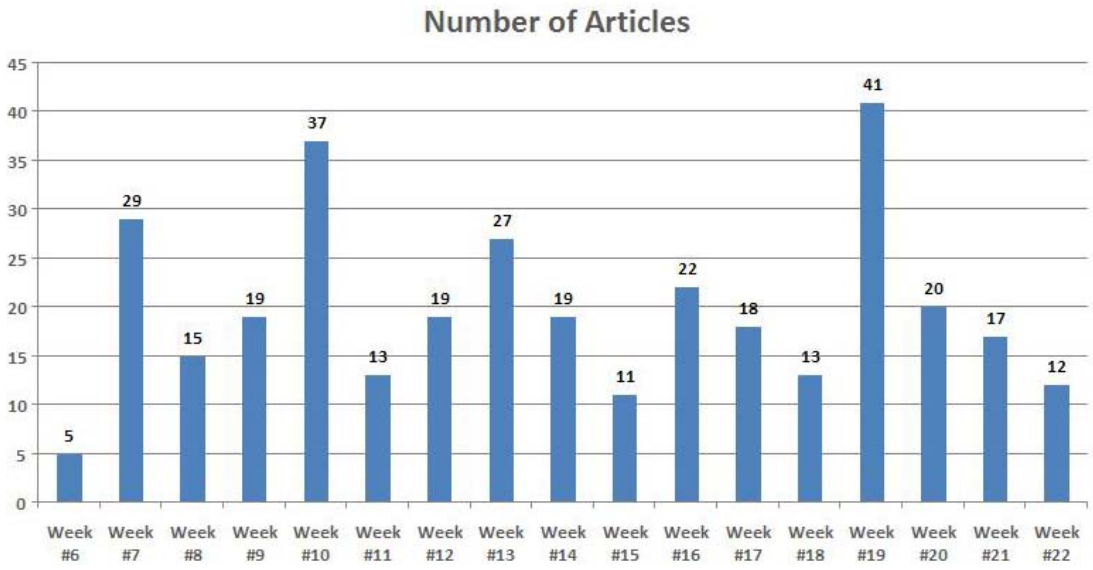

Figure 5.9: Number of articles per week containing information about Statoil. The weeks 7, 10 and 19 constitute the evaluation snapshots.

Snapshot 1

This period ranges from 13.02.06 to 20.02.06. The two most mentioned events in this timeframe are summarised below:

1. Statoil and Norsk Hydro are sued because of alleged illegal cooperation with Opec.

2. Statoil's results of the last quarter of 2005 along with the overall results of 2005 are published.

Table 5.4 shows the evaluation results. Column 1 lists all actors actually mentioned together with Statoil in the period. In some cases the framework identifies variations that differs from the real name. In these cases the names identified

		Man	Baseline	tf*idf	tf*df
Companies					
Norsk Hydro	Norsk Hydro Hydro	S	0,212(3) 0,362(1)	0,152(3) 0,271(1)	0,115(1) 0,085(4)
Opec		S	0,172(7)	0,081(10)	0,082(5)
Aker Kværner		W	0,156(9)	0,111(5)	0,055(8)
Christiania		W	0,175(6)	0,101(7)	0,089(2)
DnB NOR Markets	DnB Markets	W	0,118(11) 0,119(10)	0,179(2) 0,029(12)	0,051(9) 0,025(11)
DNO		W	0,226(2)	0,100(8)	0,058(7)
Fjord 1	Fjord	W	0,166(8)	0,043(11)	0,015(12)
Meglerhuset Carnegie	Meglerhuset	W	0,100(12)	0,098(9)	0,088(3)
Oslo Børs		W	0,186(4)	0,101(6)	0,034(10)
Telenor		W	0,176(5)	0,137(4)	0,067(6)
TFDS	N/A	W	N/A	N/A	N/A
Persons					
Helge Lund		S	0,231(1)	0,423(1)	0,216(1)
Kristin Halvorsen		W	0,101(2)	0,423(2)	0,216(2)

Table 5.4: Relations in Snapshot 1

by the framework are given in column 2. The strength of the relation, as judged by the manual inspection is given in column 3. The strengths are classified into strong (S), medium (M) and weak (W). Column 4 gives the results of the baseline run (using tf*idf weighting), while columns 5 and 6 give the results of feature extraction with tf*idf- and tf*df-weighting respectively. The weights are used for internal ranking and should not be compared across columns since the schemes employed are different. For convenience the internal ranking is given in brackets.

The fact that both 'Norsk Hydro' and 'DnB NOR Markets' are represented by two different variations, and that e.g. 'Meglerhuset Carnegie' is represented by 'Meglerhuset' alone, illustrates challenges concerning the identification and classification of proper names. It should however be noted that the framework is able to identify (some variation of) all actors except for one in this snapshot without additional noise (i.e. faulty identified actors).

The framework's ranking of actors with relations judged as strong or medium is of special interest. As can be seen from the Table 5.4, both 'Norsk Hydro' (in some form) and 'Helge Lund' are ranked highest by each and all of the approaches. 'Opec' on the other hand is ranked highest (fifth) when feature extraction with tf*df weighting is employed. The same approach with tf*idf weighting ranks it as low as number ten.

Snapshot 2

This period ranges from 06.03.06 to 13.03.06. The most prominent events are summarised below:

1. Statoil and Shell present revolutionary plans for the handling of CO_2 in gas power plants.

2. Statoil buys properties from BP in Norskehavet.

		Man	Baseline	tf*idf	tf*df
Companies					
BP		S	0,141(4)	0,216(3)	0,112(2)
Shell		S	0,569(1)	0,681(1)	0,250(1)
Norsk Hydro	Norsk Hydro Hydro	M	0,113(8) 0,154(3)	0,130(6) 0,191(4)	0,047(6) 0,063(4)
DnB NOR Markets	DnB NOR	W	0,104(10)	0,023(10)	0,018(10)
Gazprom		W	0,123(6)	0,220(2)	0,108(3)
Greenpeace	N/A	W	N/A	N/A	N/A
LO		W	0,116(7)	0,021(12)	0,041(8)
Opec		W	0,158(2)	0,134(5)	0,057(5)
Oslo Børs		W	0,088(12)	0,029(9)	0,010(11)
Pan Fish		W	0,107(9)	0,022(11)	0,007(12)
PGS		W	0,126(5)	0,072(8)	0,040(9)
Transparency International	International	W	0,096(11)	0,083(7)	0,045(7)
Persons					
Jens Stoltenberg		M	0,181(1)	0,219(2)	0,077(3)
Odd Roger Enoksen		M	0,102(2)	0,274(1)	0,156(1)
Helen Bjørnøy		W	0(3)	0(3)	0,154(2)
Alexander Mededev	N/A	W	N/A	N/A	N/A
Mads Greaker	N/A	W	N/A	N/A	N/A
Tor Kartevold	N/A	W	N/A	N/A	N/A

Table 5.5: Relations in Snapshot 2

The evaluation results are given in Table 5.5. All but one company were successfully identified by the framework, but in this case with some additional noise (not shown in the table). The nouns *fisk*(fish), *olje*(oil), and CO_2 along with the geographical names *møre* and *draugen* were all identified as companies. Further, the framework was unable to identify three of the persons mentioned in the articles. It should also be noted that the similarity score of the actor *Helen Bjørnøy* is zero for both approaches with tf*idf. The explanation to this is that the actor is mentioned in only one article in this period, resulting in a idf score of zero. One can claim that this is a desirable effect, since an actor only mentioned in a single article is likely to have a weak relation in the period.

Regarding the strong- and medium relations there are limited variations between the three approaches. The three companies in this category are ranked among the top four by all approaches. The fact that only three of the persons are identified by the framework, and two of them have medium relations, makes it difficult to judge the quality of the ranking of persons in this snapshot.

Snapshot 3

This period ranges from 08.05.06 to 15.05.06. The most mentioned events are summarised below:

1. Venezuela threatens with increased taxation of Statoil.
2. Statoil presents record results for the first quarter of 2006.
3. Statoil signs important contracts with Aker Kværner.
4. Variations in the price of oil have impact on the stock market

		Man	**Baseline**	**tf*idf**	**tf*df**
Companies					
Aker Kværner		M	0,127(7)	0,135(2)	0,071(5)
Norsk Hydro	Norsk Hydro Hydro	M	0,131(5) 0,258(1)	0,080(5) 0,183(1)	0,085(4) 0,107(1)
PGS		M	0,108(8)	0,100(3)	0,043(6)
Capital Markets	Capital	W	0(10)	0(10)	0,088(3)
DNO		W	0,129(6)	0,047(8)	0,031(10)
Lehman Brothers	N/A	W	N/A	N/A	N/A
Nordsjørigg	N/A	W	N/A	N/A	N/A
Microsoft		W	0,095(9)	0,087(4)	0,035(9)
Oslo Børs		W	0,180(3)	0,057(7)	0,022(11)
REC		W	0,142(4)	0,065(6)	0,042(7)
Roxar		W	0(11)	0(11)	0,088(2)
Telenor		W	0,192(2)	0,041(9)	0,041(8)
UBS	N/A	W	N/A	N/A	N/A
Persons					
Helge Lund		S	0,265(1)	0,335(1)	0,180(1)
Hugo Chavez		S	0,143(4)	0,157(3)	0,050(5)
Jannik Lindbæk		S	0,228(2)	0,277(2)	0,150(2)
Arne Indreeide		W	0(6)	0(6)	0,125(3)
Idar Vollvik		W	0,085(5)	0,060(4)	0,023(6)
John Fredriksen		W	0,155(3)	0,053(5)	0,0133(7)
Stein Bredal		W	0(7)	0(7)	0,110(4)

Table 5.6: Relations in Snapshot 3

The evaluation results are given in Table 5.6. All persons are successfully identified, while three companies are ignored. In addition four entities are faulty

classified as companies, namely the geographical names 'norne' and 'wall street', the award 'universum' and the noun 'kjøp'(purchase). As with 'Helen Bjørnøy' in the previous snapshot, actors that are only mentioned in one article gets a calculated score of zero for both approaches with tf*idf weighting (e.g. 'Roxar').

Feature weighting with tf*idf ranks both the three companies and the three persons with important relations as top three, and can be said to be the most promising approach in this snapshot. It is followed by feature weighting with tf*df in the ranking of companies, while the baseline outperforms this approach in the ranking of persons.

5.3.3 Summary

As is pointed out in the introduction of this chapter, the evaluation performed on the framework has clear limitations. The results of the framework are compared to the subjective opinion of two persons. Such a foundation is evidently not adequate to draw conclusions. Still the comparison has turned out useful since it has illuminated several important aspects. The main findings are summarised below:

- The identification and classification of names needs to be improved. As is, the portion of actors that are ignored or falsely/partly identified is too high. This adds considerable noise to the analysis.
- The approaches that employ feature extraction seems to outperform the baseline. It should be stressed, however, that the differences, with some exceptions, are small.
- The evaluation gives no clear indication to what should be the preferred weighting scheme. While tf*df is slightly better in snapshots one and two, tf*idf gives the best ranking in the third.
- The vector space approach to deduce and quantifiy relations generally seems to give promising results. The analysis performed by the framework coincides to a uplifting degree with our interpretations of the situations in the snapshots.

The fact that the quality of the frameworks interpretation of the situations in the different snapshots seems to be satisfactory in this scenario, shows that the approach taken has potential. A natural next step is to consider how the vector representations may be utilised to present temporal change. This is further discussed in the next chapter.

Chapter 6

Discussion

The evaluation indicates that the approach taken has potential. Even so, there are obvious possibilities for improvements. Several factors have impact on the results. The part of speech tagging lays the foundation of both the name classification and the feature extraction. The quality of the tagging is therefore vital. The pattern of tags sought for during feature extraction, the weighting of features, and the information sources used as data foundation are all examples of other factors that come to play.

In this chapter we discuss possible improvements, and how important parts of the framework may be modified and optimised. During the evaluation the framework was set up to follow one selected strategy. Other ways to employ the framework are also possible. Alternative strategies, along with possible extensions and other areas of application will be discussed. Finally, approaches to visualisation will be proposed and sketched.

6.1 Part of speech tagger

In Section 5.1 we evaluated the part of speech tagger on three different corpora:

Previously seen data (accuracy: 96.4%)	Applied on the training corpus (financial news articles).
Previously unseen data (accuracy: 94.8%)	Applied on a test set from the training corpus not incorporated in the training.
Adresseavisa corpus (accuracy: 93.1%)	Applied on a corpus of news articles from several news domains (not only financial news).

The evaluation is based on a comparison of the Oslo-Bergen (OB) tagger and the Maximum Entropy (ME) tagger employed in this project. Since it is unreasonable to assume that the OB-tagger is 100% accurate, the evaluation scores are likely to differ from the real accuracy of the ME-tagger. However, the comparison does give a hint of the tagger performance.

According to the Text Laboratory at the University of Oslo, the precision of the Oslo-Bergen tagger is 95.4%[1]. Under the assumption that this accuracy is correct for the tagging of the training corpus used in our framework, the real accuracy of the ME-tagger will be as shown in Table 6.2.

Corpus	**Accuracy, compared to OB-tagger**	**Accuracy assuming OB-tagger accuracy of 95.4%**
Previously seen data	96.4%	92.0%
Previously unseen data	94.8%	90.5%
Adresseavisa corpus	93.1%	88.8%

Table 6.2: Accuracy of the maximum entropy tagger

This means that the tagger can be expected to have an actual accuracy of about 90.5% on previously unseen financial news articles. One can, of course, claim that the fact that the OB-tagger is not perfect enable tags that are actually assigned correctly by the ME-tagger to be faulty judged as errors, and that the accuracy of the ME-tagger thus may be somewhat higher. However, for such an assertion to be trustworthy, a manual inspection of the data foundation is needed. In [Manning and Schütze, 2003] an accuracy above 90% is said to be tolerable, suggesting that the results of the ME-tagger are acceptable.

The texts of the Adresseavisa corpus and the training corpus differs the most. Therefore, as [Elworthy, 1994] and [Samuelsson and Voutilainen, 1997] describe, there is no surprise that the reported accuracy of the Adresseavisa corpus is lower. Likewise, if the accuracy of a corpus from a different domain than news was measured, one would have to expect further reduction of the accuracy.

6.1.1 Potential improvements

In the strategy applied during evaluation, we employed the tag pattern (NNx|JJx)*NNx to extract features. In addition the proper name tags (NNP) formed the basis for the classification of actors. Thus, the Penn Treebank tag set used in the Maximum Entropy tagger contains a richer tag set than actually needed. For example, we are neither interested in separating plural nouns (NNS) from singular nouns (NN), nor separating comparative and superlative adjectives. This means that a smaller tag set could have been employed in an attempt to increase the accuracy of the tagger, as described in [Kupiec, 1992].

[1] `http://www.hf.uio.no/tekstlab/tagger.html`

[Brants, 2000] has evaluated the accuracy of previously seen versus unseen words, and their results show that the expected accuracy is 8.7% higher when the word is seen before; *"Especially foreign words have a low accuracy"*. This observation coincide with the evaluation of our tagger. The recall of foreign words drops 23.5% from previously seen to previously unseen data. Although a larger training set could lead to increased recall, this problem is likely to remain for statistical taggers (unless all foreign words are included in the training set, which is obviously unrealistic).

6.1.2 Named entity identification

The quality of the part of speech tagger has direct impact on the identification of named entities. As described in Section 4.3, sequences of words identified as proper names by the tagger are looked up in lists of persons, companies and geographic locations. This approach has several shortcomings.

The evaluation of the part of speech tagger (Section 5.1) shows that for the tagging of proper names (NNP) in previously unseen text, both precision and recall lie between 93% and 95%, compared to the Oslo-Bergen tagger. If we take into consideration that the Oslo-Bergen tagger has a reported accuracy of about 95% the range drops to between 88% and 90%. Thus, one can expect about 10% of the proper names that occur in the text to be tagged wrongly (affecting the recall), while at the same time about 10% will be faulty tagged as proper names (affecting the precision).

One could claim that the faulty tagged proper names (the low precision) will cause no problem in our case since there should be a corresponding entry in the lists of names for the classification to success. But by demanding exact matches in the lists, one doesn't only reject uninteresting and/or false candidates. The use of partial names is widespread in the news articles. E.g. the company with the complete name 'Norsk Hydro ASA' in the list are typically denoted 'Norsk Hydro' or merely 'Hydro', while the use of the complete name is absent. Thus, if only exact matches were to be considered, the actor 'Norsk Hydro ASA' would pass unidentified, even if it was referred to several times in the articles. Such a policy is obviously too restrictive.

The above problem could be resolved by adding all possible sub-forms of a name to the list. However, a less demanding solution is to allow partial matches. This is the approach employed during the evaluation. While a demand for exact matches appears to be too strict, our evaluation results indicates that allowing partial matches is too liberal. An example is the noun 'fisk' (eng:fish), which is identified as a company in snapshot two of the framework evaluation. This is a

result of incorrect assigned part of speech, followed by a partial match in the list of companies. That is, there are company names that include the word 'fisk' e.g. 'Fisk og sånt AS', resulting in a incorrect classification of 'fisk' as a company.

The observations above suggest that utilisation of lists alone is inadequate to achieve a desirable level of accuracy. To keep such lists up to date is also a challenge. For instance, new companies are continuously established, while already existing ones may close out. Even if one is interested only in a limited, manageable subset of the occurrences, the problems mentioned above still need to be addressed. Adding heuristics to the lookup, such as merging candidates that are subsumed by already classified ones, may increase the performance in some cases, but at the risk of decreasing it in others.

In a more comprehensive approach the lists could be supported by rules that take the context of the candidates into account. By first applying a set of strict rules, a subset of the actors actually mentioned may be classified with high certainty. Then, in the next step, parital orders of the already classified actors are sought for, along with a more relaxed set of rules. The performance of a similar method [Mikheev et al., 1998] is reported to be good in [Jackson and Moulinier, 2002]. As an attempt to lift the acurracy of the classification of names in our framework, incorporating such a rule base may be a wise extension.

6.2 Extracting the relations

In the evaluation scenario the similarity between the feature vectors of different actors are used as a measure of relation strength. In addition, since the actors are classified as persons and companies, the type of the relations are also available. As the evaluation shows, the framework is capable of deducing these pieces of information with a reasonable degree of success. However, additional insight could be achieved if more information regarding the *background* for the relation was also made available. While it is useful to detect relations between actors, the ability to describe them as well, would add value to the analysis.

Since the feature vectors forms the basis of the extracted relations, they constitute a natural place to look for such background information. To investigate the potential, we have attempted to extract key phrases describing the relations following two different strategies.

In the first approach we extract the key phrases in the following way:

1. Extract the feature vectors of each pair of related actors.

2. Compare the vectors and use the highest weighted features occuring in both

vectors to describe the relation.

As an alternative, we also attempt to extract the descriptive features by reducing the set of underlying text to articles where the related actors are both mentioned. That is, if we consider the relation between 'Statoil' and 'Norsk Hydro', the key phrases are extracted from the set of articles where **both** companies are mentioned. Thus, the second approach boils down to:

1. Fetch all documents containing the related actors (i.e. perform a boolean and-search on the related actors).
2. Extract the feature vector of the result set. Let the highest weighted features describe the relation.

Intuitively we expected the two approaches to return different results, especially for actors with weak links (relations with low similarity score). Further, we expected that the first approach would perform worse due to an higher element of noise (non-descriptive features) in the feature vectors. The actual difference, however, seems to be limited. Table 6.3 shows a comparison of the two approaches.

The two approaches to a large extent return the same phrases. This differs from our espectations. We will not perform a thorough evaluation of what should be the preferred approach, but rather focus on the applicability of the results.

If we consider the features in Table 6.3, it would be interesting to see whether they reflect actual relations in the given period of time (Snapshot 3, May 8th - 15th, 2006). Recall the events described in Subsection 5.3.2:

- Venezuela threatens with increased taxation of Statoil.
- Statoil presents record results for the first quarter of 2006.
- Statoil signs important contracts with Aker Kværner.
- Variations in the price of oil have inpact on the stock market

As an example we look at the the first event. It should be interesting in connection with relation between Statoil and Hugo Chavez (the president of Venezuela). We see that both 'statoil-skatt' (eng: statoil tax) and 'stor Statoil-skatt' (eng: great Statoil-tax) appear among the highest ranked results. Further, we find 'Venezuela president' and 'oljeselskap' (eng: oil company) among the results. These features describe the two actors (Statoil is an oil company and Hugo Chavez is the president of Venezuela), and can indeed be claimed to help describe the background of the deduced relation.

A single example is obviously inadequate to decide whether the highest weighted

Entities	Rank	Approach 1	Approach 2
Statoil	1	Oljeserviceselskapet	Miljøkontrakt
/	2	Verden stor system	Oljegasser
Aker	3	Anlegg	Mongstad
Kværner	4	Miljøkontrakt	Behandling
	5	Oljegasser	Oppfangelse
Statoil	1	Oslo Børs	Oslo Børs
/	2	Børs tungvekter	Børs tungvekter
Norsk	3	Oljepris	Oljepris
Hydro	4	Mexicogulfen	Oljeaksje
	5	USA	Tre
Statoil	1	Venezuela	Venezuela president
/	2	Venezuela president	Skrekk
Hugo	3	Skrekk	Stor Statoil-skatt
Chavez	4	Stor Statoil-skatt	Skatteskjerpelse
	5	Oljepris	Oljeselskap
Statoil	1	Statoil-konsernsjef Helge Lund	Prosent vekst
/	2	Konsernsjef Helge Lund rapport	Spørsmål
Helge	3	Sterk resultat	Selskap
Lund	4	Konsernsjef Helge Lund	Rekordsterkt
	5	Tid resultat	Rekordtall

Table 6.3: Comparison on the two approaches to the extraction of relation-descriptions. The relations are fetched from snapshot 3 (May 8 - May 15, 2006)

elements of the feature vectors may be used as-is to describe the relations between entities. As can be seen, Table 6.3 also contains elements of noise. However, the example above give reason to think that the approach has potential.

6.3 Detecting relations over time

So far, we have discussed the extraction of relations and presented a potential way to describe them. Another main concern is the frameworks ability to deduce temporal change, i.e. how the relations evolve over time. In this section we will focus on how the framework may be utilised to achieve this.

One alternative is to create series of discrete snapshots, as we do in the evaluation scenario. As time evolves, new snapshots are created and added to the series. The timeline of snapshots in this way constitute a representation of the history of the actor covered by the snapshots. By browsing the history one can get an impression of how the situation has evolved to present day. Further, to

get an indication of the degree of change, the approach may be supported by calculations of the similarity between the actor's feature vectors on a snapshot-to-snapshot basis. Such a quantification of the change could be useful e.g. to direct the attention of the user to important events, or to look for patterns in the evolution.

Another possibility is to maintain a single snapshot per interesting actor. The snapshot covers articles from a desired period of time up to present day, e.g. the last week. As new articles are published, they are added to the set of articles underlying the extracted snapshot. Likewise, as the age of an article exceed the desired period of time, it is excluded from the foundation. In this way the snapshot gives an continuously updated representation of the actor's situation.

The approaches described above are closely related to the strategy employed in our evaluation. A more comprehensive alternative is to investigate how the semantics of a relation change. As described in the previous section, we consider the feature vectors as relevant descriptors of the relations between actors. Taken further, the idea is to employ the same vectors as sources for mining of temporal changes. The aim is to be able to detect how a relation evolves, not only by means of relations appearing and dissapearing, but also by tracking how the *characteristics* of the relations evolve.

As an example, we look at a couple of situations that is assumed to be related in unequal snapshots from the evaluation scenario. As presented in Subsection 5.3.2, Statoil published good quarter-year-results in both snapshot 1 and snapshot 3. One could therefore expect the two snapshots to share some characteristics. Table 6.4 shows the similarity scores between Statoil and Oslo Børs (Oslo Stock Exchange) for the three snapshots. The corresponding highest-weighted features are also presented. From the table we see that the calculated difference between

Statoil/Oslo Børs

		Snapshot 1	**Snapshot 2**	**Snapshot 3**
Weight		0.034	0.010	0.022
Features (Rank)	1	Investor	Pan Fish	Børs
	2	Vinnere	Oljefall	Tid
	3	Skyld	Hydro	Nedtur
	4	Verdi	Børs	Hydro
	5	Kjøp	Topp	Rekordtall
	6	Eier	Fisk	Rekordnivå

Table 6.4: Comparison of the relation between Statoil and Oslo Børs for the three snapshots.

snapshot 1 and snapshot 3 is smaller than is the case for snapshot 2. This ob-

servation supports the assumption that the situations described by snapshots 1 and 3 have a higher degree of overlap than snapshot 2. A comparison of the top five extracted features of the two snapshots, however, cannot be said to give an impression of *the content* of this overlap.

An example where the features of relations from different snapshots *do* give an indication of overlap is given in Table 6.5. A comparison of the features leaves the impression that the semantics of the relation between the companies Statoil and PGS is relatively stable. Inspection of the underlying articles reveal that both are companies in the oil-sector, and that both are traded on the Wall Street market. Further, we see that 'oljeaksje' (oil stock) exists in both feature vectors. This

Statoil/PGS

	Snapshot 1	**Snapshot 2**	**Snapshot 3**
Weight	0.000	0.040	0.043
Features (Rank)	1	New York	Wall Street fredag
	2	Oljepriseffekt	Oljeaksje
	3	Oljeaksje	Aksje
	4	Retning	Statoil-løft
	5	Utslag	USA
	6	Hydro	Olje

Table 6.5: Comparison of the relation between Statoil and PGS for the three snapshots.

indicates that 'oljeaksje' is a stable description of the relation between Statoil and PGS, which is indeed the fact.

The discussed approaches both represent possibilities and challenges connected with the use of feature vectors to describe the temporal change of actor relationships. Although we give examples extracted by the framework that is able to illustrate the possibilities, a more extensive evaluation is needed to decide the actual applicability of the proposed strategies.

6.4 Domain portability

Flexibility was emphasised during the development of the framework. The components are loosely coupled so that individual parts may be altered or replaced, and new components added without considerable external change. Both significant parameters and the mode of operation may be modified in simple ways. This makes the framework adaptable to various contexts. Obviously, even though the evaluated procedure gave satisfactory results in our scenario, it may be unsuitable in others. Instead, e.g. a clustering component working on the extracted

feature vectors might be of interest. The system architecture render such changes possible without extensive effort.

The use of general methods and statistical based training enables this flexibility and simplifies portability to other domains and even other languages. However, it must be stressed that the quality of the part of speech tagging, and with that the entire framework, is highly dependent on the training corpus. Thus, the availability of adequate training corpora is vital. In some languages, including Norwegian, the number of such materials is very limited. The creation of a training corpus with sufficient size is both labor intensive and time consuming. Further, the classification of proper names is based on domain dependent lists of names. The avilability of such lists is not given. These observations demonstrate clear limitations of the portability.

Yet flexible, obviously all domains will not fit the qualities of the framework. Key characteristics such as large quantities of text and some element of dynamical change should be in place, in order to justify an expectation of the framework's capabilities.

6.5 Visualisation

The ability to extract key actors and important relations from comprehensive amounts of text is apparently useful. But the value of knowing that e.g. 'Statoil' has a relation strength of 0.226 with 'Gazprom' in week 7, or that 'Hydro' has a evolution of 0.045 from week 8 to week 9, is clearly limited. To enable simplified overview and interpretation of the analysis, the results needs to be synthesised in a perspicuous way.

Although the outcome of the framework constitutes a suitable basis for visualisation, no such component is implemented. Figure 6.1 shows a possible realisation. It illustrates the relationship between Statoil and other companies between snapshot 2 and snapshot 3. The length between the entities reflects their similarity (a high degree of similarity leads to a short distance), and the relations are described by the highest-weighted terms of the feature vectors. The relation-description-terms are terms extracted from the intersection of the feature vectors of both snapshots (i.e. the terms occur in both feature vectors).

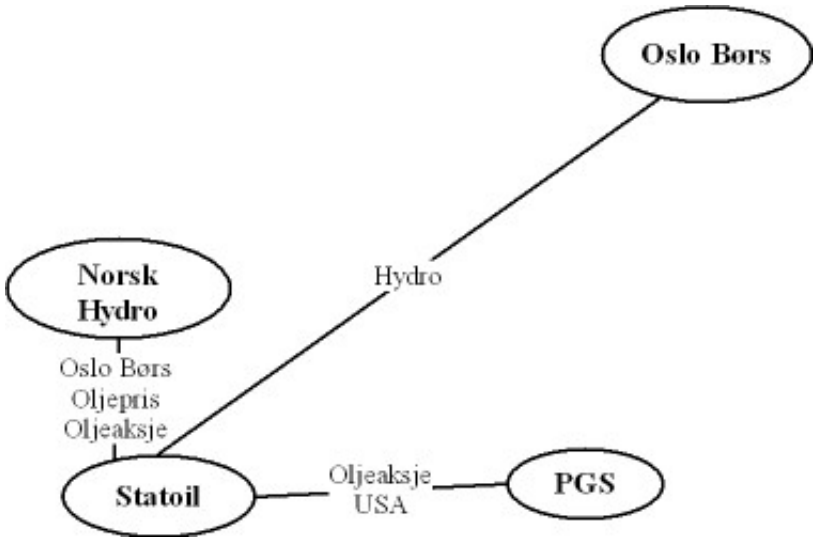

Figure 6.1: A suggested visualisation of relationships between entities over time. The figure illustrates the relationship between Statoil and related companies between snapshot 2 and snapshot 3.

Another way of visualising relations and how they evolve is to create a timeline-based representation. By dragging a timeline, it could be possible to see how the relations change, i.e. what describes the relations over time, how do the relationships evolve, and how does the similarity between entities evolve. By using distance as similarity-visualisation, one could possibly create a map of relations between entities, and thus also illustrate transitive relationships.

Chapter 7

Conclusion

This thesis addresses the problem of extracting, analysing and synthesising valuable information from continuous text streams covering financial information. A text mining framework combining elements from information retrieval, information extraction and natural language processing has been implemented. The framework is utilised to extract information regarding key actors in the domain, how they relate to each other, and how these characteristics evolve over time.

Main findings

The qualitative evaluation described in Chapter 5 indicates that the framework is capable of extracting useful information in a selcted evaluation scenario. However, the results also reveal elements of noise that affect the quality of the analysis. Although a qualitative evaluation based on a single scenario is inadequate to draw conclusions, it gives a useful impression of the performance. The main observations are summarised below:

Extraction of feature vectors Extraction of feature vectors based on knowledge of the words' part of speech adds value to the analysis. It enables selection of a smaller set of descriptive features that seems to improve the identification of relations. Further, the feature vectors may be utilised to give an impression of the relations' semantics.

Identificaton of actors This task is central to the analysis. An approach based on part of speech tagging and static lists of known names in the domain cannot be said to produce the desired quality. Erroneously tagging and ambiguous entries in the lists impose noise that effectively degrade the quality.

Identification of relations Use of the vector space model to calculate relation strength between the feature vectors of different actors is promising. The approach agrees to an uplifing degree with our interpretations of important relations in the scenario.

Further directions

As discussed in Chapter 6 there are several aspects, both regarding possibilities and challenges, that could be adressed in a continuation of this work. Possible further directions are:

Part of speech, identification of actors	Attempts to improve accuracy of the part of speech tagging and the identification of actors are desireable. This includes creation of a proper training corpus and an improved strategy for identification and classification of actors.
Evaluation	A more comprehensive evaluation of the framework is needed to properly state the potential. It should include a thorough investigation of possibilities related to extraction and description of temporal change.
Extensions	Various extensions to the framework could be desireable in other contexts of use. Examples are algorithms for clustering and mining of temporal patterns.
Visualisation	To increase the availability and usefulness of the extracted information the results of the analysis should be visualised.

With a proper visualisation component on top of the framework, and a more comprehensive evaluation of the results, one would be able to decide whether the approach is applicable to provide increased availablity and simplified interpretaion of essential information. If so, it would indeed be a step towards the goal of knowledge discovery.

Bibliography

[DBL, 2001] (2001). *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, November 27-30, 2001, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan*.

[Allan et al., 1998] Allan, J., Papka, R., and Lavrenko, V. (1998). On-line new event detection and tracking. In *Research and Development in Information Retrieval*, pages 37–45. Available from: `citeseer.ist.psu.edu/allan98line.html`.

[Appelt and Israel, 1999] Appelt, D. E. and Israel, D. J. (1999). Introduction to information extraction. Available from: `http://www.dfki.de/%7Eneumann/esslli04/reader/overview/IJCAI99.pdf`.

[Baeza-Yates, 2005] Baeza-Yates, R. (2005). Searching the future. *ACM SIGIR Workshop MF/IR 2005*. Available from: `http://cir.dcs.vein.hu/cikkek/searching_the_future.pdf`.

[Baeza-Yates and Ribeiro Neto, 1999] Baeza-Yates, R. and Ribeiro Neto, B. (1999). *Modern information retrieval*. ACM Press Books.

[Blake and Pratt, 2001] Blake, C. and Pratt, W. (2001). Better rules, few features: A semantic approach to selecting features from text. In *ICDM*, pages 59–66. Available from: `citeseer.ist.psu.edu/blake01better.html`.

[Brants, 2000] Brants, T. (2000). Tnt – a statistical part-of-speech tagger. Available from: `citeseer.ist.psu.edu/brants00tnt.html`.

[Brill, 1992] Brill, E. (1992). A simple rule-based part of speech tagger. Available from: `http://citeseer.ist.psu.edu/cache/papers/cs/24451/http:zSzzSzwww.rohan.sdsu.eduzSz~corporazSzcorpus_coursezSzBRILL92.PDF/brill92simple.pdf`.

[Cunningham, 2004] Cunningham, H. (2004). Information extraction, automatic. Available from: `http://gate.ac.uk/sale/ell2/ie/main.pdf`.

[Cutting et al., 1992] Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). A practical part-of-speech tagger. Available from: `citeseer.ist.psu.edu/cutting92practical.html`.

[Dermatas and Kokkinakis, 1995] Dermatas, E. and Kokkinakis, G. (1995). Automatic stochastic tagging of natural language texts. *Comput. Linguist.*, 21(2):137–163.

[Elworthy, 1994] Elworthy, D. (1994). Does baum-welch re-estimation help taggers? In *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing (13–15 October 1994, Stuttgart)*. Available from: `citeseer.ist.psu.edu/elworthy94does.html`.

[Feldman and Dagan, 1995] Feldman, R. and Dagan, I. (1995). Knowledge discovery in textual databases (KDT). In *Knowledge Discovery and Data Mining*, pages 112–117. Available from: `citeseer.ist.psu.edu/feldman95knowledge.html`.

[Fox, 1992] Fox, C. (1992). Lexical analysis and stoplists. pages 102–130.

[Fung et al., 2005] Fung, G. P. C., Yu, J. X., and Lu, H. (2005). The predicting power of textual information on financial markets. Available from: `citeseer.ist.psu.edu/726091.html`.

[Gulla et al., 2006] Gulla, J., Borch, H., and Ingvaldsen, J. (2006). Unsupervised keyphrase extraction for search ontologies.

[Hearst, 1999] Hearst, M. (1999). Untangling text data mining. Available from: `citeseer.ist.psu.edu/hearst99untangling.html`.

[Holt and Chung, 1999] Holt, J. D. and Chung, S. M. (1999). Efficient mining of association rules in text databases. *Proceedings of the eighth international conference on Information and knowledge management*, pages 234–242. Available from: `http://portal.acm.org/citation.cfm?id=319981`.

[J. A. Berry and Linoff, 2004] J. A. Berry, M. and Linoff, G. S. (2004). *Data mining techniques: For marketing, sales, and customer relationship management*. Wiley Publishing, Inc., Indianapolis, Indiana.

[Jackson and Moulinier, 2002] Jackson, P. and Moulinier, I. (2002). *Natural Language Processing for Online Applications. Text retrieval, extraction and categorization*. John Benjamins Publishing Co, Philadelphia USA. Available from: `http://site.ebrary.com/lib/ntnu/Doc?id=10022351`.

[Johannessen et al., 2000] Johannessen, J. B., Hagen, K., and Nøklestad, A. (2000). A constraint-based tagger for norwegian. Available from: `http://folk.uio.no/jannebj/Scand.conf.-98.ps`.

[Kazama et al., 2001] Kazama, J., Miyao, Y., and ichi Tsujii, J. (2001). A maximum entropy tagger with unsupervised hidden markov models. In [DBL, 2001], pages 333–340.

[Kleinberg, 2006] Kleinberg, J. (2006). *Temporal Dynamics of On-Line Information Streams*. Springer. Available from: `http://www.cs.cornell.edu/home/kleinber/stream-survey04.pdf`.

[Kontostathis et al., 2003] Kontostathis, A., Galitsky, L., Pottenger, W. M., Roy, S., and Phelps, D. J. (2003). *A Survey of Emerging Trend Detection in Textual Data Mining*. Springer-Verlag. Available from: `citeseer.ist.psu.edu/kontostathis03survey.html`.

[Korenius et al., 2004] Korenius, T., Laurikkala, J., Järvelin, K., and Juhola, M. (2004). Stemming and lemmatization in the clustering of finnish text documents. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 625–633, New York, NY, USA. ACM Press.

[Kupiec, 1992] Kupiec, J. (1992). Robust Part–of–speech Tagging Using a Hidden Markov Model. *Computer Speech and Language*, 6.

[Lee, 1995] Lee, J. H. (1995). Combining multiple evidence from different properties of weighting schemes.

[Lent et al., 1997] Lent, B., Agrawal, R., and Srikant, R. (1997). Discovering trends in text databases. In Heckerman, D., Mannila, H., Pregibon, D., and Uthurusamy, R., editors, *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining, KDD*, pages 227–230. AAAI Press. Available from: `citeseer.ist.psu.edu/lent97discovering.html`.

[Lin et al., 2002] Lin, W., Orgun, M. A., and Williams, G. J. (2002). An overview of temporal data mining. Available from: `http://www.act.cmis.csiro.au/edm/papers/adm02.pdf`.

[Manning and Schütze, 2003] Manning, C. D. and Schütze, H. (2003). *Foundations of statistical natural language processing*. MIT Press, Cambridge, Massachusetts.

[Mei and Zhai, 2005] Mei, Q. and Zhai, C. (2005). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 198–207, New York, NY, USA. ACM Press.

[Mikheev et al., 1998] Mikheev, A., Grover, C., and Moens, M. (1998). Description of the ltg system used for muc. In *In Seventh Message Understanding Conference (MUC 7): Proceedings of a Conference held in Fairfax, Virginia, 29 April - 1 May, 1998.* Available from: `http://www.muc.saic.com/proceedings/muc_7_toc.html`.

[Mittermayer, 2004] Mittermayer, M. (2004). Forecasting intraday stock price trends with text mining techniques. *hicss*, 03:30064b.

[Ratnaparkhi, 1996] Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In Brill, E. and Church, K., editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics, Somerset, New Jersey. Available from: `citeseer.ist.psu.edu/ratnaparkhi96maximum.html`.

[Samuelsson and Voutilainen, 1997] Samuelsson, C. and Voutilainen, A. (1997). Comparing a linguistic and a stochastic tagger. In Cohen, P. R. and Wahlster, W., editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 246–253, Somerset, New Jersey. Association for Computational Linguistics. Available from: `citeseer.ist.psu.edu/article/samuelsson97comparing.html`.

[Sennelart and Blondel, 2004] Sennelart, P. P. and Blondel, V. D. (2004). Automatic discovery of similar words. *Survey of Text Mining: Clustering, Classification and Retrieval*, pages 25–43.

[Swan and Jensen, 2000] Swan, R. and Jensen, D. (2000). Timemines: Constructing timelines with statistical models of word usage. Available from: `citeseer.ist.psu.edu/514824.html`.

[Toutanova and Manning, 2000] Toutanova, K. and Manning, C. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Available from: `citeseer.ist.psu.edu/article/toutanova00enriching.html`.

[Wilbur and Sirotkin, 1992] Wilbur, W. J. and Sirotkin, K. (1992). The automatic identification of stop words. volume 18, pages 45–55, Thousand Oaks, CA, USA. Sage Publications, Inc.

[Wong et al., 2000] Wong, P. C., Cowley, W., Foote, H., Jurrus, E., and Thomas, J. (2000). Visualizing sequential patterns for text mining. In *INFOVIS '00: Proceedings of the IEEE Symposium on Information Vizualization 2000*, page 105, Washington, DC, USA. IEEE Computer Society.

[Wu et al., 1992] Wu, Z. B., Hsu, L. S., , and Tan, C. L. (1992). A survey of statistical approaches to natural language processing. (TRA4/92). Available from: `citeseer.ist.psu.edu/wu92survey.html`.

Appendix A

Part Of Speech tagger conversion table

The following table is used for converting from Oslo-Bergen-Lancaster tags to Penn Treebank tags.

Oslo-Bergen-Lancaster Tag	Example	Penn Treebank tag
<ANF>	”	”
<KOMMA>	,	,
<PARANTES-BEG>	(	(
<PARANTES-SLUTT>	)	)
CLB <ELLIPSE>	...	...
CLB <KOLON>	:	:
CLB <KOMMA>	, (som setningsgrense)	,
CLB <OVERSKRIFT>	[slutt på overskrift]	.
CLB <PUNKT>	.	.
CLB <SEMI>	;	;
CLB <SP M>	?	?
CLB <STREK>	-	-
CLB <UTROP>	!	!
CLB konj	og	CC
adj <ordenstall> pos mask fem nøyt be ent	(det) første	JJ
adj <ordenstall> pos mask fem nøyt be ent gen	(det) førstes	JJ
adj <ordenstall> pos mask fem ub ent	(en) første	JJ
adj <ordenstall> pos nøyt ub ent	(et) første	JJ
adj <ordenstall> pos ub be fl	(de) første	JJ
adj <ordenstall> pos ub be fl gen	(de) førstes	JJ
adj <perf-part> mask fem nøyt be ent	(den) fargelagte	JJ
adj <perf-part> mask fem nøyt be ent gen	(den) fargelagtes	JJ
adj <perf-part> mask fem ub ent	(en) fargelagt	JJ
adj <perf-part> nøyt ub ent	(et) fargelagt	JJ
adj <perf-part> ub be fl	(de) fargelagte	JJ
adj <perf-part> ub be fl gen	(de) fargelagtes	JJ
adj <pres-part> mask fem nøyt ub be ent fl	administrerende	JJ
adj komp	eldre	JJR
adj komp gen	eldres	JJR
adj pos fem ub ent	lita	JJ
adj pos mask fem nøyt be ent	lille	JJ
adj pos mask fem nøyt be ent gen	lilles	JJ
adj pos mask fem ub ent	stor	JJ

adj pos mask ub ent	liten	JJ
adj pos nøyt ub ent	lite	JJ
adj pos ub be fl	små	JJ
adj pos ub be fl gen	(de) gamles	JJ
adj sup be	minste	JJS
adj sup ub	minst	JJS
adv	ikke	RB
det be	selve	PRP
det dem <adj> fem ub ent	anna	PRP
det dem <adj> mask fem nøyt be ent	(den/det) andre	PRP
det dem <adj> mask fem nøyt be ent gen	(den/det) andres	PRP
det dem <adj> mask ub ent	annen	PRP
det dem <adj> nøyt ub ent	annet	PRP
det dem <adj> ub be fl	(de) andre	PRP
det dem <adj> ub be fl gen	(de) andres	PRP
det dem be <adj>	neste	JJ
det dem be <adj> gen	(den) nestes	JJ
det dem fem ent	den	PRP
et dem fem ent gen	egen	JJ
det forst <adj> nøyt ub ent	eget	JJ
det forst <adj> ub be fl	(deres) egne	JJ
det kvant	16.00	CD
det kvant be <adj>	eneste	JJ
det kvant be ent	(den) ene	JJ
det kvant be ent gen	(den) enes	JJ
det kvant ent	1	CD
det kvant fem ent	hver (bok)	PRP
det kvant fl	fem	CD
det kvant fl gen	(på) alles (lepper)	PRP
det kvant mask ent	en (aksjon)	PDT
det kvant mask ent gen	ens	PDT
det kvant nøyt ent	et (ansvar)	PDT
det poss fem ent	boka (si)	PRP
det poss fl	sine (lesere)	PRP
det poss høflig fem ent	Deres (form)	PRP
det poss høflig fl	Deres (bøker)	PRP
det poss høflig mask ent	Deres (fantasi)	PRP
det poss høflig nøyt ent	Deres (navn)	PRP
det poss mask ent	hans (opplevelse)	PRP
det poss nøyt ent	(navnet) sitt	PRP
det sp fem ent	hvilken (bok)	PRP
det sp fl	hvilke (kvalifikasjoner)	PRP
det sp mask ent	hvilken (feil)	PRP
det sp nøyt ent	hvilket (fly)	PRP
fork adv @ADV	etc.	RB
fork adv @ADV>	ca.	RB
fork konj+adv+adj adv @ADV	osv.	CC + RB + JJ
fork prep	m.	IN
fork prep @ADV	pr.	IN

fork prep+adj adv @ADV	mfl.	IN + JJ
fork prep+adj prep @ADV	bl.a.	IN + JJ
fork prep+det+subst adv @ADV	m.a.o.	IN + PRP + NNS
fork prep+prop adv @ADV	f.Kr.	IN + NNP
fork prep+subst adv @ADV	f.eks.	IN + NN
fork pron+verb+verb adv @ADV	dvs.	PRP + VB + VB
fork subst	art.	NN
fork subst @<SUBST	jr.	NN
fork subst @TITTEL	dr.	NN
fork subst appell	adr.	NN
fork subst mask appell	ill.	NN
fork subst nøyt appell ent fl ub be	bnr.	NN
fork subst nøyt appell ent fl ub be @<SUBST	AL	NN
fork subst prop	AP	NNP
inf-merke	å	TO
interj	ja	UH
konj	og	CC
konj @KON		CC
prep	på	IN
prep @ADV	bortefra	IN
pron ent	ingenting	PRP $
pron pers 1 ent hum akk	meg	PRP
pron pers 1 ent hum nom	jeg	PRP
pron pers 1 fl hum akk	oss	PRP
pron pers 1 fl hum nom	vi	PRP
pron pers 2 ent hum akk	deg	PRP
pron pers 2 ent hum nom	du	PRP
pron pers 2 fl hum akk	dere	PRP
pron pers 2 fl hum nom	dere	PRP
pron pers 3 ent fem hum akk	henne	PRP
pron pers 3 ent fem hum nom	hun	PRP
pron pers 3 ent mask fem	denne	PRP
pron pers 3 ent mask hum akk	ham	PRP
pron pers 3 ent mask hum nom	han	PRP
pron pers 3 ent nøyt	dette	PRP
pron pers 3 fl	disse	PRP
pron pers 3 fl akk	dem	PRP
pron pers 3 fl høflig akk	Dem	PRP
pron pers 3 fl høflig nom	De	PRP
pron pers 3 fl nom	de	PRP
pron pers ent hum	man	PRP
pron refl ent/fl akk	seg	PRP
pron res fl hum	hverandre	PRP
pron sp hum	hvem	WP
pron sp	hva	WP
pron sp poss hum	hvis	WP$
sbu	at	CC
subst	%	NN
subst <dato>	7.8.97	DAT

subst <klokke>	15.10	KLK
subst @TITTEL	kong	NN
subst appell fl ub	(på) vegne	NNS
subst appell ubøy	behold	NN
subst appell ubøy gen	beholds	NN
subst fem appell ent be	tida	NN
subst fem appell ent be gen	tidas	NN
subst fem appell ent ub	tid	NN
subst fem appell ent ub gen	tids	NN
subst fem appell fl be	tidene	NNS
subst fem appell fl be gen	tidenes	NNS
subst fem appell fl ub	tider	NNS
subst fem appell fl ub gen	tiders	NNS
subst fem appell ubøy	(dårlig) råd	NN
subst fem prop	Aud	NNP
subst fem prop gen	Auds	NNPS
subst fl ub	(alle) mann	NNS
subst gen prop	Kristi	NNP
subst mask appell ent be	dagen	NN
subst mask appell ent be gen	dagens	NN
subst mask appell ent ub	dag	NN
subst mask appell ent ub gen	dags	NN
subst mask appell fl be	dagene	NNS
subst mask appell fl be gen	dagenes	NNS
subst mask appell fl ub	dager	NNS
subst mask appell fl ub gen	dagers	NNS
subst mask appell ubøy	april	NN
subst mask appell ubøy gen	aprils	NN
subst mask prop	Arne	NNP
subst mask prop gen	Arnes	NNP
subst nøyt appell ent be	landet	NN
subst nøyt appell ent be gen	landets	NN
subst nøyt appell ent ub	land	NN
subst nøyt appell ent ub gen	lands	NN
subst nøyt appell fl be	landa	NNS
subst nøyt appell fl be gen	landas	NNS
subst nøyt appell fl ub	land	NNS
subst nøyt appell fl ub gen	lands	NNS
subst nøyt appell ubøy	(få) lov	NN
subst nøyt appell ubøy gen	lovs	NN
subst nøyt prop	Dagbladet	NNP
subst nøyt prop gen	Dagbladets	NNP
subst prop	Hansen	NNP
subst prop gen	Hansens	NNP
ymb	Au	UH
symb subst	ha	UH
ukjent ord	perfect	FW
verb imp	reguler	VB
verb imp <s-verb>	synes	VB

verb imp gen	regulers	VB
verb inf	regulere	VB
verb inf gen	reguleres	VB
verb inf <s-verb>	synes	VB
verb inf pres pass	reguleres	VB
verb perf-part	regulert	VBN
verb perf-part <s-verb>	synes	VBN
verb perf-part gen	regulerts	VBN
verb pres	regulerer	VB
verb pres <s-verb>	synes	VB
verb pres gen	regulerers	VB
verb pret	regulerte	VBD
verb pret <s-verb>	syntes	VBD
verb pret gen	regulertes	VBD
verb ubøy	nåde	VB

Table A.1: Oslo-Bergen-Lancaster to Penn Treebank tag conversion scheme.

Appendix B

Data foundation from the POS tagger evaluation

This appendix contains the data foundation of the tagger evaluation. The evaluation is based on a comparison between the Maximum Entropy tagger and the Oslo-Bergen (OB) tagger, using OB as the reference tagger.

B.1 Previously seen data

Table B.1 contains the data used as the foundation of the tagger-evaluation for previously seen data (i.e. the training corpus).

Tag	**#Correct**	**#Wrong**	**#Tags** (ME)	**#Tags**(OB)	**Precision**	**Recall**
CC	116301	2755	119056	122405	0,9769	0,9501
CD	82200	653	82853	82883	0,9921	0,9918
FW	7112	852	7964	7769	0,8930	0,9154
IN	345013	5639	350652	360954	0,9839	0,9558
JJ	157992	15081	173073	163755	0,9129	0,9648
JJR	20835	199	21034	21569	0,9905	0,9660
JJS	7598	125	7723	8863	0,9838	0,8573
NN	307787	20477	328264	321029	0,9376	0,9588
NNP	224233	11314	235547	230245	0,9520	0,9739
NNS	194520	9596	204116	201129	0,9530	0,9671
PDT	50302	904	51206	50882	0,9823	0,9886
PRP	135572	2006	137578	137898	0,9854	0,9831
RB	58265	1119	59384	67094	0,9812	0,8684
TO	25201	17	25218	25214	0,9993	0,9995
UH	1918	146	2064	2384	0,9293	0,8045
VB	259708	10082	269790	268935	0,9626	0,9657
VBD	51388	2752	54140	53513	0,9492	0,9603
VBN	47736	2092	49828	51095	0,9580	0,9343
WP	1399	1	1400	1400	0,9993	0,9993
Delimiters	231787	125	231912	233786	0,9995	0,9914
Total	2326867	85935	2412802	2412802	**0,9644**	**0,9644**

Table B.1: Tagger performance on previously seen data. (OB = Oslo-Bergen tagger, ME = Maximum Entropy tagger)

B.2 Previously unseen data

Table B.2 contains the data used as the foundation of the tagger-evaluation for previously unseen data (the same news sources as the training corpus, not including the articles used in the training process).

Tag	**#Correct**	**#Wrong**	**#Tags** (ME)	**#Tags**(OB)	**Precision**	**Recall**
CC	22526	958	23484	24021	0,9592	0,9378
CD	15779	279	16058	15986	0,9826	0,9871
FW	1019	223	1242	1499	0,8205	0,6798
IN	66530	1613	68143	70201	0,9763	0,9477
JJ	29996	3792	33788	31961	0,8878	0,9385
JJR	3921	39	3960	4153	0,9902	0,9441
JJS	1521	35	1556	1882	0,9775	0,8082
NN	59227	5641	64868	63242	0,9130	0,9365
NNP	43647	3216	46863	45862	0,9314	0,9517
NNS	35663	2730	38393	37882	0,9289	0,9414
PDT	9609	320	9929	9817	0,9678	0,9788
PRP	27009	732	27741	27638	0,9736	0,9772
RB	11096	431	11527	13085	0,9626	0,8480
TO	4502	7	4509	4504	0,9984	0,9996
UH	241	38	279	506	0,8638	0,4763
VB	49963	2599	52562	52169	0,9506	0,9577
VBD	9455	948	10403	10235	0,9089	0,9238
VBN	9100	756	9856	10169	0,9233	0,8949
WP	261	1	262	262	0,9962	0,9962
Delimiters	44906	52	44958	45307	0,9988	0,9911
Total	445971	24410	470381	470381	**0,9481**	**0,9481**

Table B.2: Tagger performance on previously unseen data. (OB = Oslo-Bergen tagger, ME = Maximum Entropy tagger)

B.3 Previously unseen data (adressa corpus)

Table B.3 contains the data used as the foundation of the tagger-evaluation for previously unseen data (i.e. the adressa corpus).

Tag	**#Correct**	**#Wrong**	**#Tags** (ME)	**#Tags**(OB)	**Precision**	**Recall**
CC	5036	285	5321	5450	0,9464	0,9240
CD	1629	79	1708	1666	0,9537	0,9778
FW	34	134	168	88	0,2024	0,3864
IN	11678	362	12040	12464	0,9699	0,9369
JJ	6066	978	7044	6622	0,8612	0,9160
JJR	726	14	740	786	0,9811	0,9237
JJS	246	9	255	291	0,9647	0,8454
NN	10772	1413	12185	11684	0,8840	0,9219
NNP	4459	322	4781	4777	0,9327	0,9334
NNS	5148	700	5848	5675	0,8803	0,9071
PDT	1939	107	2046	2001	0,9477	0,9690
PRP	6092	166	6258	6274	0,9735	0,9710
RB	2717	102	2819	3291	0,9638	0,8256
TO	1213	0	1213	1213	1,0000	1,0000
UH	15	7	22	112	0,6818	0,1339
VB	9002	663	9665	9658	0,9314	0,9321
VBD	1868	201	2069	2085	0,9029	0,8959
VBN	1403	210	1613	1646	0,8698	0,8524
WP	64	0	64	64	1,0000	1,0000
Delimiters	7836	13	7849	7861	0,9983	0,9968
Total	77943	5765	83708	83708	**0,9311**	**0,9311**

Table B.3: Tagger performance on previously unseen data (The Adressa corpus). (OB = Oslo-Bergen tagger, ME = Maximum Entropy tagger)

Appendix C

Evaluation scenario articles

This appendix contains the articles extracted from the three evaluation snapshots. The articles were extracting by searching for 'Statoil' in the time period of each snapshot.

C.1 Snapshot 1, February 12 - February 20 2006

The following articles were extracted in shapshot 1:

Norge, Statoil og Hydro saksøkes
Statoil, Hydro og Norge saksøkes i USA. Anklagen er at de på ulovlig vis har samarbeidet med Opec om å holde oljeprisen på et unaturlig høyt nivå.

Olje-Norge saksøkes i USA
Amerikanere saksøker Hydro, Statoil og Norge for ulovlig samarbeid med Opec.

Amerikanere saksøker olje-Norge
Statoil, Hydro og Norge saksøkes i USA for ulovlig samarbeid med Opec.

Saksøker Olje-Norge
Tre privatpersoner går til søksmål mot Statoil, Hydro og Norge. Bakgrunnen er det de mener er et ulovlig samarbeid med Opec.

Statoil kan kutte ansatte
Statoil vil kutte kostnadene med 500 millioner kroner, noe som kan resultere nedbemanning.

Her er Statoil-tallene Statoil melder om solid omsetningsvekst i fjerde kvartal 2005. Her er tallene.

Statoil tynger børsen
Statoil må ta mesteparten av skylden for at Oslo Børs faller i formiddag.

Statoil blør - anbefaler salg
Statoil falt kraftig på Oslo Børs etter at produksjonsestimatene ble lavere enn ventet. Meglerhuset Carnegie snur fra kjøp til salg.

Statoil svakere enn ventet
Oljekjempen Statoil og Helge Lund var ute med svakere resultater enn ventet i fjerde kvartal. Les mer her:

Statoil-kutt tross rekord
Statoil leverte et resultat på 30,7 milliarder kroner etter skatt for 2005, men har likevel tatt mål av seg til å kutte 500 millioner kroner i administrasjonen.

Børsen sank med Statoil
Når Statoil faller fire prosent, kan det nesten bare gå én vei. Og det kom flere skuffelser for investorene på Oslo Børs.

Statoil-overskudd på 30 milliarder kroner
Statoil leverer et rekordresultat med 30,7 milliarder kroner i overskudd for fjoråret.

Meglerhus kjøper Statoil-aksjer
DnB NOR kjøper Statoil-aksjer til bruk i Statoils aksjespareprogram.

Kutter Statoil-aksjen
DnB NOR Markets mener Statoil fortjener et annet kursmål etter gårsdagens tall.

Statoil styrker Barentssamarbeidet
Statoil har inngått flere avtaler med de lokale myndighetene i Murmansk om utviklingssamarbeid.

Inngår leieavtale med Statoil
TFDS og Fjord 1 har inngått en avtale med Statoil om leie av "MS Jupiter" som innkvarteringsskip for Snøhvitprosjektet.

- Statoil-analytikere har ikke gjort leksen sin
En Statoil-analytiker mener kollegaer som er overrasket over Statoils lave produksjons-guiding ikke har gjort hjemmeleksen sin godt nok.

Statoil med 50 milliarder i utbytte
5 år med Statoil på Oslo Børs har gitt eierne 48,2 milliarder kroner i utbytte og en formidabel verdistigning. Men i går falt verdien med 11 milliarder kroner.

Private har tjent 70 milliarder på Statoil
Etter Statoil-privatiseringen i 2001, har private eiere stukket av med 10 milliarder kroner i utbytte og en verdistigning på eventyrlige 60 milliarder kroner. Samtidig varsler Statoil kutt i staben.

Statoil-fall på kjempetall
Statoilsjef Helge Lund leverte tidenes beste resultat i 2005. Men lavere oljeproduksjon enn forventet skuffet aksjemarkedet, som sendte kursen ned med fire prosent.

Statoil-verdien ned 50 mrd.
Det kraftige kursfallet har på få dager resultert i et solid innhugg i Statoils markedsverdi.

Han tjente 4675 kr for deg
Statoil tjente 30 milliarder i fjor. Og du tjente 4.675 kroner og 75 øre.

Rødt på Oslo Børs
Statoil-aksjen tynger hovedindeksen mandag formiddag. Men det finnes andre tapere - og også noen vinnere.

Svak ukestart etter Asia-fall
Statoil faller etter resultatpresentasjon, mens Aker Kværner stiger. I Tokoyo gikk markedet på en solid smell i morges.

Børslokomotivene senket børsen
Statoil, Norsk Hydro og Telenor falt alle med to til tre prosent og gjorde Oslo Børs til taperen blant de europeiske børsene.

Forsiktig mottagelse
Akkurat som Statoil, fikk også Hydro rekordresulat i fjor. Aksjemarkedet sender likevel Hydro forsiktig ned.

- God kjøpsmulighet
Statoil er nå nær kraftige støtteområder. - En god kjøpsmulighet, mener Christiania og hever anbefalingen.

Finansministeren vil stoppe ledernes lønnsfest
Topplederne i Statoil, Hydro og Telenor er blant dem som får en siste advarsel fra finansminister Kristin Halvorsen (SV) om at lønnsfesten må stoppe.

Bakveien til Irak
Erbil/oslo: Mens Statoil og Hydro sitter på gjerdet, har DNO gått bakveien inn i Irak. Det kan bli et eventyr eller et mareritt. De neste par månedene kan avgjøre.

C.2 Snapshot 2, March 6 - March 13 2006

The following articles were extracted in shapshot 2:

Fraråder støtte til Statoil og Shell
Energiforsker Mads Greaker sier nei til å bruke statens penger på Statoil og Shells CO2-frie gasskraftverk. I stedet vil han utvikle ny teknologi.

Tre tilsyn kritiserer Statoil
Tre statlige tilsynsorgan ønsker å diskutere helse og sikkerhet med Statoil, men oljegiganten har avlyst møtene flere ganger.

Mener Statoil opptrer rotete
Tre statlige tilsynsorgan ønsker å diskutere helse og sikkerhet med Statoil, men oljegiganten har avlyst møtene flere ganger.

Statoil og Shell samarbeider
Oljegiganten Statoil skal samarbeide med oljekollega Shell om et industriprosjekt.

Statoil kjøper i Norskehavet
Oljegiganten Statoil kjøper BPs andel i flere leteblokker, og dette inkluderer gassfunnet Luva.

Statoil kjøper andeler av BP
Statoil kjøper flere leteblokker av BP i Norskehavet.

Bjørnøy roser Statoil og Shell
Miljøvernminister Helen Bjørnøy mener at forslaget fra Statoil og Shell viser at industrien tar regjeringens signaler på alvor.

Kinderegg fra Statoil og Shell
Shell og Statoil går sammen om et storstilt prosjekt, som skal redusere miljøgassutslipp, utvinne mer olje, og samtidig løse energikrisen i Midt-Norge.

Shell og Statoil sammen om C02
Shell og Statoil inngår et historisk samarbeid om CO2-rensing, som skal gi økt oljeutvinning.

Statoil og Shell samarbeider på miljø
Statoil og Shell lanserer onsdag morgen en felles plan som skal bli et stort industriprosjekt knyttet opp til energi og miljø.

Statoil og Shell med gasskraftverk med CO2-rensing
Statoil og Shell lanserer en plan for hvordan Statoils gasskraftverk på Tjeldbergodden i Midt-Norge kan bygges og stå klart i 2009.

Statoil-ansatte deler 270 millioner
Statoils norske ansatte kan glede seeg til en solid ekstrautbetaling i mars.

Statoil tror Opec vil produsere for fullt
Statoils oljeprisekspert Tor Kartevold tror Opec vil beslutte å fortsette å produsere for fullt når kartellet møtes i Wien onsdag.

-Prosjekt av internasjonal interesse
Statoil og Shell lanserer onsdag gasskraftverk med CO2-rensing.

Oljebosser på tiggerferd
Statoil og Shell ber staten stille opp med penger for å skaffe mer kraft og olje.

Mange oljefelt trenger CO2
Injisering av CO2 som Statoil og Shell vil gjøre på Draugen og Heidrun er høyaktuelt på flere andre norske oljefelt.

Eksperter kritiserer CO2-planen
Den store CO2-planen til Statoil og Shell er i strid med anbefalingene fra offentlig utvalg og andre eksperter.

Fisk- og oljefall på børsen
Pan Fish, Statoil og Hydro tynger i toppen på Oslo Børs.

Greenpeace refser gasskraft-planer
Greenpeace refser miljøbevegelsen for å juble for Statoil og Shells planer om gasskraftverk med CO2-håndtering.

Unison glede over gassplaner
Statoil og Shell får unison støtte til sine gasskraftverkplaner på Tjeldbergodden.

Kjøper mer i Norskehavet
Statoil har kjøpt BPs andel i lisens 218 og får med gassfunnet i Luva. Les me her:

Handler for 63 mrd.
Statoil kjøpte varer og tjenester for godt over 60 milliarder i fjor. Årets budsjett blir ikke mindre.

Når det regner på presten....
....så drypper det også på klokkeren. I 2005 kjøpte Statoil varer og tjenester for 63 milliarder kroner.

Krever svar om CO2-bidrag
Venstre vil snarest ha løfter fra statsminister Jens Stoltenberg om statlig bidrag til CO2-rensing. Uten slik drahjelp kan Statoil og Shells planer strande.

- Må oppgi annonsebudsjett
Statoil må fortelle hvor mye de har brukt på annonser for å åpne nordområdene for oljeutvinning, mener Transparency International Norge.

Øker i Norskehavet
Statoil styrke sin posisjon på dypt vann i Norskehavet ved å kjøpe BPs andel av et gassfunn.

Lanserer gasskraft med rensing
Oljeselskapene Statoil og Shell vil i dag lansere planene for et gasskraftverk på Tjeldbergodden med CO2-rensing .

Løser CO2-krisen
Statoil og Shell legger i dag frem en løsning for CO2-rensing av et gasskraftverk på Tjeldbergodden. CO2-gassen skal deponeres i det Shell-opererte feltet !Draugen.

Enoksen: -Staten vil bidra
Olje- og energiminister Odd Roger Enoksen sier at staten vil delta i det spleiselaget som Statoil og Shell legger opp til på Tjeldbergodden.

Tror ikke på rensing fra dag én
Fremskrittspartiet synes gasskraftprosjektet til Statoil og Shell er svært gledelig, men tror ikke at det blir CO2-rensing fra dag én. Partiet ber regjeringen lempe på rensekravet.

De nye gassplanene
Shell og Statoil går sammen om et prosjekt for å bruke CO2 fra gasskraftverk til å øke oljeutvinningen. Forutsetningen er at myndighetene er med på lagspillet.

LO utfordrer myndighetene til aktiv CO2-deltakelse
- Det er svært gledelig at Statoil og Shell så konkret og målrettet igangsetter planer for bygging av gasskraftverk på Tjeldbergodden, mener LO.

Jubler for ren gasskraft

Miljøbevegelsen jubler over planene for ren gasskraft på Tjeldbergodden i Møre og Romsdal. Shell og Statoil går sammen om å bruke CO2-gassen til økt oljeutvinning.

- Designfeil bak Visund-lekkasjen
Statoil har avsluttet den interne granskingen etter gasslekkasjen på Visund-plattformen i Nordsjøen i januar. Den bakenforliggende årsaken er ifølge Statoil feil design av en væskeutskiller.

Shtokman beslutning i april
Visekonsernsjef Alxander Mededev i det russiske gasskonsernet Gazprom lover nå en Shtokman-beslutning innen 15. april. Statoil og Norsk Hydro venter i spenning.

Norsk i USA - Blandet oljepriseffekt
Oljeprisens rekyl torsdag ga ikke de store utslagene i positiv retning for de norske oljeaksjene i New York. Statoil og Hydro fortsatte ned, mens PGS ble dratt opp av sterkje seismikktall.

Satser på stigende oljepris
DnB Nor Markets tar denne uken inn Statoil i sin favorittporteføljе.

Design-feil bak ulykke på Visund
Statoil har avsluttet granskingen av den alvorlige gassulykken på Visund-plattformen.

Håper nye gasskraftverk kan finansiere seg selv
Statsminister Jens Stoltenberg (Ap) avviser ikke at staten vil bidra i Statoil og Shells "spleiselag" på Tjeldbergodden, men håper samtidig at det nye gasskraftverket vil klare seg uten statsstøtte.

- Staten vil bidra
Olje- og energiminister Odd Roger Enoksen (Sp) bekrefter at staten vil delta i det spleiselaget som Statoil og Shell legger opp til på Tjeldbergodden. Men han vil ikke antyde noen sum.

C.3 Snapshot 3, May 8 - May 15 2006

The following articles were extracted in shapshot 3:

Anmelder Statoil
Forurensningstilsynet anmelder Statoil etter et av Norges største oljeutslipp.

Ryddesjau i styrerom
Jannik Lindbæk, styreleder i Statoil, er blant de som må gå når regjeringen rydder opp i de statlige styrene.

Presser Statoil
Venezuela varsler nasjonalisering. Det kan gå hardt utover Statoil.

- Kjøp Statoil
Handelsbanken Capital Markets anbefaler sine kunder å laste opp med Statoil-aksjer.

Statoil under forventningene
Statoil leverer rekordhøye kvartallstall, men skuffer likevel markedet.

Kjøpe aksjer i Statoil?
Meglerhus gir sin vurdering av Statoil-aksjen etter dagens rekordtall.

To Statoil-tips
Lehman Brothers og UBS endrer kursmålene på Statoil-aksjen.

Statoil senker børsen
Statoil er på nedtur tross tidenes resultat. Fallende oljepris trekker også Hydro og Oslo Børs ned.

Rekordtall fra Statoil
Statoil melder om over 50 prosent vekst i bunnlinjen etter første kvartal. - Rekordsterkt, melder konsernsjef Helge Lund.

Vil ha større Statoil-skatt
Oljeselskapens skrekk, Venezuelas Hugo Chavez, varsler nye skatteskjerpelser mot Statoil.

Statoil sletter 23 millioner aksjer
I dag vinker Statoil farvel til 23 millioner aksjer fra gammelt bonusprogram.

Gassknipe for Statoil
Ikke nok med at Snøhvit-forsinkelsen koster Statoil store penger. Nå får de en ny ekstraregning dumpende i fanget.

Statoil faller tross rekordtall
Statoil går i rødt på en rekordsterk Oslo Børs ved åpning mandag, til tross for tidenes beste resultat.

Statoil senker børsen
Statoil er på nedtur etter tidenes resultat. Den fallende oljeprisen trekker Hydro og Oslo Børs med ned.

Sletter 23 millioner Statoil-aksjer
I morgen slettes 23 millioner Statoil-aksjer. Årsaken er en riv ruskende gal avgjørelse da aksjen ble børsnotert i 2001.

Roxar leverer til Statoil
Roxar leverer i dag sin første Oil-in-water monitor til et pilotprosjekt med Statoil på Sleipner A.

Vil beskatte Statoil hardere
Venezuelas president og internasjonale oljeselskapers store skrekk Hugo Chavez varsler nå nye skatteskjerpelser mot Statoil og de andre selskapene som foredler landets tungolje. Les mer her:

Statoil tapte arbeidsrettssak
Arne Indreeide (68) fra Hafrsfjord i Rogaland fikk fredag rettens medhold i at han skal få fortsette i Statoil til han fyller 70 år.

Anmelder Statoil for slapp beredskap
Statoil får hard kritikk av Statens Forurensingstilsyn (SFT) for treg respons etter en oljelekkasje på Norne-feltet.

Nytt Statoil-løft i USA
Oljeaksjene løftet seg igjen på Wall Street fredag etter en pustepause. Statoil og PGS var blant aksjene som ble med opp.

Statoil knaller inn tidenes resultat
Statoil-konsernsjef Helge Lund legger frem det sterkeste resultatet noe norsk selskap noengang har gjort. Spørsmålet er om det er godt nok.

Statoil traff gass i Mexicogolfen
Statoil og de andre eierne i blokk 140 i Atwater Valley-området i Mexicogolfen har funnet tørrgass i den første av letebrønnene som er boret på blokken.

Vil ha Hydro og Statoil ut av Iran
Fremskrittspartiet vil ha slutt på virksomheten til Hydro og Statoil i Iran.

Statoil traff gass i Mexico-gulfen
Statoil og de andre eierne i blokk 140 i Atwater Valley-området i Mexico-gulfen har funnet tørrgass i den første av letebrønnen som er boret på blokken.

Vi ha Hydro og Statoil ut av Iran
Fremskrittspartiet vil ha slutt på virksomheten til Hydro og Statoil i Iran, og mener det statlige eierskapet bør brukes til å trekke selskapene ut av landet.

- Største siden Statoil
Når solenergiselskapet REC kommer på børs i morgen, er det det største som er skjedd siden Statoils børsinntreden.

Meglerhus med Statoil-advarsel
Meglerhus kommer med en advarsel etter kursfest og rekordtall fra oljeselskapet.

Statoil omtrent som ventet
Statoils og konsernsjef Helge Lunds rapport for første kvartal viser at selskapets hovedtall kom inn rett i nærheten av hva markedet hadde ventet på forhånd.

Rekordresultat for Statoil
Statoils resultat før finans, skatt og minoritetsinteresser var 31,0 milliarder kroner i første kvartal 2006, sammenlignet med 21,5 milliarder kroner i første kvartal i fjor.

Venezuela øker trykket på Statoil
Etter at Venezuelas president Hugo Chaves i helgen varslet en drastisk skatteskjerpelse, følger kongressen nå opp med forslag om ytterligere nasjonalisering.

Oljeprisen sendte Oslo Børs ned
Den svake oljeprisen rammet den oljetunge Oslo Børs. Statoil og Norsk Hydro falt mer enn 3 prosent.

Aker Kværner får miljøkontrakt

Oljeserviceselskapet Aker Kværner skal designe og levere verdens største system for behandling av oljegasser til Statoil.

Rekordsterkt i Oslo
Oslo Børs fortsetter oppturen og stiger videre til nye høyder. Den høye oljeprisen hjelper når Statoil skuffer.

Investerer i falleferdige stater
Norske selskaper som Statoil, Telenor og Hydro risikerer ryktet i verdens mest ustabile land, mener NUPI-forsker.

Ny rekord på Oslo Børs
Flere av John Fredriksens selskaper utmerker seg positivt på børsen i åpningshandelen, men Statoil faller på tall.

Oslo Børs flater ut
Børsen snur fra rekordnivå og er nå tilbake ved utgangspunktet. Oljekameratene Hydro og Statoil legger en demper på stemningen.

Storkontrakt til Nordsjørigg
Statoil har tildelt boreriggen Scarabeo 5 en lang kontrakt for boring på norsk sokkel.

Blodige oljeaksjer
Tre av børsens tungvektere, Norsk Hydro, Statoil og DNO, faller som blylodd i tung mandagshandel.

Storkontrakt til Nordsjørigg
Statoil har tildelt boreriggen Scarabeo 5 en lang kontrakt for boring på norsk sokkel.

Norsk i USA - Nytt løft i olje
Oljeaksjene løftet seg igjen på Wallstreet fredag etter en pustepause. Statoil og PGS var blant aksjene som ble med opp. Les mer her:

Store, stygge Staten?
Kommentar: Staten som storeier ble brukt som investorskremsel da Telenor og Statoil gikk på børs. Staten har slått hardt tilbake.

Ferskingenes drømmebedrifter
Statoil troner på toppen som norske studenters drømmearbeidsplass. Her er listen over selskapene flest studenter vil arbeide i.

Vollvik: Det er ikke sexy å tjene noen titall millioner
? Det er liksom ikke noe sexy å tjene noen titall millioner på Hydro og Statoil, sier Idar Vollvik.

Norges beste arbeidsgivere
Landets mest attraktive bedrifter ble hedret under utdelingen av Universum Awards 2006 onsdag. Og prisen for beste arbeidsgivere gikk til Statoil og Microsoft.

Tilbakeholden Lindbæk
Han blir hudflettet av styrekollega Stein Bredal, og rødgrønne krefter vil ha ham fjernet. Likevel vil ikke Statoil-styreleder Jannik Lindbæk si om han har tenkt å fortsette i vervet.

Appendix D

Named Entity Identification evaluation data

This appendix contains the named entity identification evaluation data. The evaluation is performed on the articles in the three evaluation scenarios presented in Appendix C.

D.1 Evaluation results

This section contains the data foundation used for calculating precision and recall of the named entiy identification.

Article	Correct assigned NE tags	Wrongly assigned NE tags	Correct tags	Precision	Recall
1	8	0	8	1,00	1,00
2	5	0	5	1,00	1,00
3	5	0	5	1,00	1,00
4	4	0	4	1,00	1,00
5	2	0	2	1,00	1,00
6	1	0	1	1,00	1,00
7	4	0	4	1,00	1,00
8	3	1	4	0,75	0,75
9	2	1	3	0,67	0,67
10	1	0	1	1,00	1,00
11	3	0	3	1,00	1,00
12	1	0	1	1,00	1,00
13	1	1	2	0,50	0,50
14	1	1	2	0,50	0,50
15	3	0	3	1,00	1,00
16	2	1	5	0,67	0,40
17	1	0	1	1,00	1,00
18	3	0	3	1,00	1,00
19	2	0	2	1,00	1,00
20	1	0	1	1,00	1,00
21	1	0	1	1,00	1,00
22	1	0	1	1,00	1,00
23	1	0	1	1,00	1,00
24	2	0	2	1,00	1,00
25	4	0	4	1,00	1,00

26	3	0	3	1,00	1,00
27	1	1	2	0,50	0,50
28	4	0	4	1,00	1,00
29	5	0	5	1,00	1,00
30	4	0	5	1,00	0,80
31	2	0	2	1,00	1,00
32	2	0	2	1,00	1,00
33	4	0	4	1,00	1,00
34	3	0	3	1,00	1,00
35	4	0	4	1,00	1,00
36	6	0	6	1,00	1,00
37	5	0	5	1,00	1,00
38	3	1	5	0,75	0,60
39	4	0	4	1,00	1,00
40	6	1	7	0,86	0,86
41	1	0	1	1,00	1,00
42	4	1	6	0,80	0,67
43	1	1	2	0,50	0,50
44	2	0	2	1,00	1,00
45	2	4	2	0,33	1,00
46	2	0	2	1,00	1,00
47	4	0	4	1,00	1,00
48	2	1	4	0,67	0,50
49	2	1	3	0,67	0,67
50	2	0	2	1,00	1,00
51	1	0	1	1,00	1,00
52	3	1	3	0,75	1,00
53	1	1	2	0,50	0,50
54	1	0	1	1,00	1,00
55	2	1	3	0,67	0,67
56	2	1	3	0,67	0,67
57	4	2	5	0,67	0,80
58	2	0	2	1,00	1,00
59	2	1	2	0,67	1,00
60	2	3	5	0,40	0,40
61	2	2	4	0,50	0,50
62	2	0	2	1,00	1,00
63	2	2	4	0,50	0,50
64	4	1	5	0,80	0,80
65	1	1	2	0,50	0,50
66	1	0	1	1,00	1,00
67	3	1	4	0,75	0,75
68	3	2	4	0,60	0,75
69	3	0	3	1,00	1,00
70	2	0	2	1,00	1,00
71	3	0	3	1,00	1,00
72	1	1	3	0,50	0,33
73	2	0	2	1,00	1,00
74	1	0	1	1,00	1,00

75	1	0	2	1,00	0,50
76	4	0	4	1,00	1,00
77	3	0	3	1,00	1,00
78	1	0	2	1,00	0,50
79	2	0	2	1,00	1,00
80	2	0	2	1,00	1,00
81	3	0	3	1,00	1,00
82	3	0	3	1,00	1,00
83	1	0	1	1,00	1,00
84	7	0	8	1,00	0,88
85	3	1	5	0,75	0,60
86	2	2	2	0,50	1,00
87	4	0	4	1,00	1,00
88	2	0	2	1,00	1,00
89	2	0	4	1,00	0,50
90	6	0	6	1,00	1,00
91	2	0	2	1,00	1,00
92	5	0	5	1,00	1,00
93	3	0	3	1,00	1,00
94	2	1	3	0,67	0,67
95	2	0	2	1,00	1,00
96	3	0	4	1,00	0,75
97	4	0	4	1,00	1,00
98	1	1	2	0,50	0,50
99	3	0	3	1,00	1,00
100	3	0	3	1,00	1,00
101	3	0	3	1,00	1,00
102	3	0	3	1,00	1,00
103	1	0	1	1,00	1,00
104	3	0	3	1,00	1,00
105	1	0	1	1,00	1,00
106	3	0	3	1,00	1,00
107	2	1	2	0,67	1,00
108	1	0	1	1,00	1,00
109	4	0	4	1,00	1,00
110	4	0	4	1,00	1,00
111	3	0	3	1,00	1,00
SUM	292	42	337	0,874	0,866

Table D.1: Precision/Recall scores of the Named Entity Identification feature.

D.2 Evaluation data foundation

This section contains the articles used for the named entity identification evaluation, in lemmatized and part-of-speech-tagged format.

norge/NNPL ,/, statoil/NNPC og/CC hydro/NNPC saksøke/VB

statoil/NNPC ,/, hydro/NNPC og/CC norge/NNPL saksøke/VB i/IN usa/NNPL ./. anklage/NN være/VB at/CC de/PRP på/RB ulovlig/JJ vis/NN ha/VB samarbeide/VBN med/IN opec/NNPC om/IN å/TO holde/VB oljepris/NN på/IN en/PDT unaturlig/JJ høy/JJ nivå/NNS ./.

olje-norge/NNP saksøke/VB i/IN usa/NNPL
amerikaner/NNS saksøke/VB hydro/NNPC ,/, statoil/NNPC og/CC norge/NNPL for/IN ulovlig/JJ samarbeid/NN med/IN opec/NNPC ./.

amerikaner/NNS saksøke/VB olje-norge/NN
statoil/NNPC ,/, hydro/NNPC og/CC norge/NNPL saksøke/VB i/IN usa/NNPL for/IN ulovlig/JJ samarbeid/NN med/IN opec/NNPC ./.

saksøker/NN olje-norge/NNP
tre/CD privatperson/NNS gå/VB til/IN søksmål/NN mot/IN statoil/NNPC ,/, hydro/NNPC og/CC norge/NNPL ./. bakgrunn/NN være/VB det/PRP de/PRP mene/VB være/VB en/PDT ulovlig/JJ samarbeid/NN med/IN opec/NNPC ./.

statoil/NNPC kunne/VB kutte/VB ansatt/JJ
statoil/NNPC ville/VB kutte/VB kostnad/NNS med/IN 500/CD million/NNS kroner/NNS ,/, noe/PRP som/CC kunne/VB resultere/VB nedbemanning/FW ./.

her/IN være/VB statoil-tallene/NNS
statoil/NNPC melde/VB om/IN solid/JJ omsetningsvekst/NN i/IN fjerde/JJ kvartal/NNS 2005/JJ ./. her/IN være/VB tallene/FW ./.

statoil/NNPC tynge/VB børse/NN
statoil/NNPC måtte/VB ta/VB mestepart/NN av/IN skyld/NN for/IN at/CC oslo/NNPC børs/NNPC falle/VB i/IN formiddag/FW ./.

statoil/NNPC blør/VBD -/- anbefale/VB salg/NNS
statoil/NNPC falle/VBD kraftig/JJ på/IN oslo/NNPC børs/NNPC etter/IN at/CC produksjonsestimat/NNS bli/VBD lav/JJR enn/IN ventet/FW ./. meglerhuset/NNPC carnegie/NNP snu/VB fra/IN kjøp/NN til/IN salg/FW ./.

statoil/NNPC svak/JJR enn/IN vente/VBD
oljekjempen/NNP statoil/NNP og/CC helge/NNPP lund/NNPP være/VBD ute/IN med/IN svak/JJR resultat/NNS enn/IN vente/VBD i/IN fjerde/JJ kvartal/FW ./. le/VB mye/JJR her/IN :/:

statoil-kutt/NNP tross/IN rekord/NN
statoil/NNPC levere/VBD en/PDT resultat/NN på/IN 307/CD 30,7/30,7 milliard/NNS krone/NNS etter/IN skatt/NN for/IN 2005/CD ,/, men/CC ha/VB likevel/RB ta/VBN mål/NNS av/IN seg/PRP til/IN å/TO kutte/VB 500/CD million/NNS krone/NNS i/IN administrasjonen/FW ./.

børse/NN synke/VBD med/IN statoil/NNPC
når/NNP statoil/NNP falle/VB fire/CD prosent/NNS ,/, kunne/VB det/PRP nesten/RB bare/RB gå/VB én/NNP vei/. ./. og/CC det/PRP komme/VBD mange/JJR skuffelse/NNS for/IN investor/NNS på/IN oslo/NNPC børs/NNPC ./.

statoil-overskudd/NN på/IN 30/CD milliard/NNS krone/NNS
statoil/NNPC levere/VB en/PDT rekordresultat/NN med/IN 307/CD 30,7/30,7 milliard/NNS krone/NNS i/IN overskudd/NNS for/IN fjoråret/FW ./.

meglerhus/NNP kjøper/NN statoil-aksjer/NNS
dnb/NNPC nor/NNS kjøpe/VB statoil-aksjer/NNS til/IN bruk/NN i/IN statoil/NNPC aksjespareprogram/NNS ./.

kutte/VB statoil-aksjen/NN
dnb/NNPC nor/NNS markets/NNPC mene/VB statoil/NNPC fortjene/VB en/PDT annen/PRP kursmål/NN etter/IN gårsdag/NN tall/FW ./.

statoil/NNPC styrke/VB barentssamarbeidet/NNP
statoil/NNPC ha/VB inngå/VBN mange/JJR avtale/NNS med/IN de/PRP lokal/JJ myndighet/NNS i/IN murmansk/NNPL om/IN utviklingssamarbeid/NNS ./.

inngå/VB leieavtale/NN med/IN statoil/NNPC
tfds/NNP og/CC fjord/NNPC 1/CD ha/VB inngå/VBN en/PDT avtale/NN med/IN statoil/NNPC om/IN leie/NN av/IN ms/NNPC "/" jupiter/NNPC "/" som/CC innkvarteringsskip/NN for/IN snøhvitprosjektet/NNP ./.

-/- statoil-analytikere/NNP ha/VB ikke/RB gjøre/VBN lekse/NN sin/PRP
en/PDT statoil-analytiker/NN mene/VB kollega/NNS som/CC være/VB overraske/VBN over/IN statoil/NNPC lav/JJ produksjons-guiding/NN ikke/RB ha/VB gjøre/VBN hjemmelekse/NN sin/PRP god/JJ nok/RB ./.

statoil/NNPC med/IN 50/CD milliard/NNS i/IN utbytte/NN
5/CD år/NNS med/IN statoil/NNPC på/IN oslo/NNPC børs/NNPC ha/VB gi/VBN eier/NNS 482/CD 48,2/48,2 milliard/NNS krone/VB i/IN utbytte/NN og/CC en/PDT formidabel/JJ verdistigning/FW ./. men/CC i/IN går/NN falle/VBD verdi/NN med/IN 11/CD milliard/NNS kroner/FW ./.

privat/JJ ha/VB tjene/VBN 70/CD milliard/NNS på/IN statoil/NNPC
etter/IN statoil-privatisering/NN i/IN 2001/CD ,/, ha/VB privat/JJ eier/NNS stikke/VBN av/IN med/IN 10/CD milliard/NNS krone/NNS i/IN utbytte/NN og/CC en/PDT verdistigning/NN på/IN eventyrlig/JJ 60/CD milliard/NNS kroner/FW ./. samtidig/JJ varsle/VB statoil/NNPC kutt/NNS i/IN staben/FW ./.

statoil-fall/NNP på/VB kjempetall/NN
statoilsjef/NN helge/NNPP lund/NNPP levere/VBD tid/NNS god/JJS resultat/NNS i/IN 2005/JJ ./. men/CC lav/JJR oljeproduksjon/NN enn/IN forvente/VBD skuffe/VBN aksjemarkedet/NN ,/, som/CC sende/VBD kurs/NN ned/IN med/IN fire/CD prosent/FW ./.

statoil-verdien/NN ned/IN 50/CD mrd/RB ./.
det/PRP kraftig/JJ kursfall/NN ha/VB på/IN få/JJ dag/NNS resultere/VBN i/IN en/PDT solid/JJ innhugg/NN i/IN statoil/NNPC markedsverdi/NNS ./.

han/PRP tjene/VBD 4675/CD kr/RB for/IN du/PRP
statoil/NNPC tjene/VBD 30/CD milliard/NNS i/IN fjor/FW ./. og/CC du/PRP tjene/VBD 4.675/CD krone/NNS og/CC 75/CD øre/FW ./.

rød/JJ på/IN oslo/NNPC børs/NNPC
statoil-aksjen/NN tynge/VB hovedindeks/NN mandag/NN formiddag/FW ./. men/CC det/PRP finnes/VB annen/PRP taper/NNS -/- og/CC også/RB noen/CD vinnere/NNS ./.

svak/JJ ukestart/NN etter/IN asia-fall/NN
statoil/NNPC falle/VB etter/IN resultatpresentasjon/NNS ,/, mens/CC aker/NNPC kværner/NNPC stiger/FW ./. i/IN tokoyo/NNP gå/VBD marked/NN på/IN en/PDT solid/JJ smell/NN i/IN morges/FW ./.

børslokomotiv/NNS senke/VBD børse/NN
statoil/NNPC ,/, norsk/NNPC hydro/NNPC og/CC telenor/NNPC falle/VBD all/CD med/IN to/CD til/IN tre/CD prosent/NNS og/CC gjøre/VBD oslo/NNPC børs/NNPC til/IN taper/NN blant/IN de/PRP europeisk/JJ børsene/FW ./.

forsiktig/JJ mottagelse/NN
akkurat/RB som/CC statoil/NNPC ,/, få/VBD også/IN hydro/NNPC rekordresulat/JJ i/IN fjor/FW ./. aksjemarked/NN sende/VB likevel/RB hydro/NNPC forsiktig/JJ ned/FW ./.

-/- god/JJ kjøpsmulighet/NN
statoil/NNPC være/VB nå/RB nær/RB kraftig/JJ støtteområder/FW ./. -/- en/PDT god/JJ kjøpsmulighet/NN ,/, mene/VB christiania/NNPL og/CC heve/VB anbefalingen/FW ./.

finansminister/NN ville/VB stoppe/VB leder/NNS lønnsfest/NN
toppleder/NNS i/IN statoil/NNPC ,/, hydro/NNPC og/CC telenor/NNPC være/VB blant/IN de/PRP som/CC få/VB en/PDT sist/JJ advarsel/NN fra/IN finansminister/NN kristin/NNPP halvorsen/NNPP sv/NNP ()/() om/IN at/CC lønnsfest/NN måtte/VB stoppe/FW ./.

bakvei/NN til/IN irak/NNPL
erbil/oslo: mens/CC statoil/NNPC og/CC hydro/NNPC sitte/VB på/IN gjerdet/NN ,/, ha/VB dno/NNPC gå/VBN bakvei/NN inn/IN i/IN irak/NNPL ./. det/PRP kunne/VB bli/VB en/PDT eventyr/NN eller/CC en/PDT mareritt/FW ./. de/PRP neste/JJ par/NNS måned/NNS kunne/VB avgjøre/FW ./.

fraråde/VB støtte/VB til/IN statoil/NNPC og/CC shell/NNPC
energiforsker/NNP mads/NNPP greaker/NNP si/VB nei/NNS til/IN å/TO bruke/VB stat/NN penge/NNS på/IN statoil/NNPC og/CC shell/NNPC co2-frie/NN gasskraftverk/FW ./. i/IN sted/NN ville/VB han/PRP utvikle/VB ny/JJ teknologi/FW ./.

tre/CD tilsyn/NNS kritisere/VB statoil/NNPC
tre/CD statlig/JJ tilsynsorgan/NNS ønske/VB å/TO diskutere/VB helse/NN og/CC sikkerhet/NN med/IN statoil/NNPC ,/, men/CC oljegigant/NN ha/VB avlyse/VBN møte/NNS mange/JJR ganger/FW ./.

mene/VB statoil/NNPC opptre/VB rotete/JJ
tre/CD statlig/JJ tilsynsorgan/NNS ønske/VB å/TO diskutere/VB helse/NN og/CC sikkerhet/NN med/IN statoil/NNPC ,/, men/CC oljegigant/NN ha/VB avlyse/VBN møte/NNS mange/JJR ganger/FW ./.

statoil/NNPC og/CC shell/NNPC samarbeide/VB
oljegigant/NN statoil/NNPC skulle/VB samarbeide/VB med/IN oljekollega/NN shell/NNPC om/IN en/PDT industriprosjekt/FW ./.

statoil/NNPC kjøpe/VB i/IN norskehavet/NNP
oljegigant/NN statoil/NNPC kjøpe/VB bp/NNPC andel/NN i/IN mange/JJR leteblokker/NNS ,/, og/CC dette/PRP inkludere/VB gassfunn/NN luva/NNP ./.

statoil/NNPC kjøpe/VB andel/NNS av/IN bp/NNPC
statoil/NNPC kjøpe/VB mange/JJR leteblokk/NNS av/IN bp/NNPC i/IN norskehavet/NNP ./.

bjørnøy/NNPP rose/VB statoil/NNPC og/CC shell/NNPC
miljøvernminister/NN helen/NNPP bjørnøy/NNPP mene/VB at/CC forslag/NN fra/IN statoil/NNPC og/CC shell/NNPC vise/VB at/CC industri/NN ta/VB regjering/NN signal/NNS på/IN alvor/NNS ./.

kinderegg/NNP fra/IN statoil/NNPC og/CC shell/NNPC shell/NNPC og/CC statoil/NNPC gå/VB sammen/RB om/IN en/PDT storstilt/JJ prosjekt/NN ,/, som/CC skulle/VB redusere/VB miljøgassutslipp/NN ,/, utvinne/VB mye/JJR olje/NN ,/, og/CC samtidig/JJ løse/NN energikrise/NN i/IN midt-norge/NNPL ./.

shell/NNPC og/CC statoil/NNP sammen/RB om/IN c02/NNPC
shell/NNPC og/CC statoil/NNPC inngå/VB en/PDT historisk/JJ samarbeid/NN om/IN co2/NN som/CC skulle/VB gi/VB økt/NNP oljeutvinning/FW ./.

statoil/NNPC og/CC shell/NNPC samarbeide/VB på/IN miljø/NN
statoil/NNPC og/CC shell/NNPC lansere/VB onsdag/NN morgen/NN en/PDT felles/JJ plan/NN som/CC skulle/VB bli/VB en/PDT stor/JJ industriprosjekt/NN knytte/VBN opp/IN til/IN energi/NN og/CC miljø/NNS ./.

statoil/NNPC og/CC shell/NNPC med/IN gasskraftverk/NNS med/IN co2-rensing/NN
statoil/NNPC og/CC shell/NNPC lansere/VB en/PDT plan/NN for/IN hvordan/RB statoil/NNPC gasskraftverk/NNS på/IN tjeldbergodden/NNPC i/IN midt-norge/NNPL kunne/VB bygge/VB og/CC stå/VB klar/JJ i/IN 2009/ ./.

statoil-ansatte/NNP dele/VB 270/CD million/NNS
statoil/NNPC norsk/JJ ansatt/JJ kunne/VB glede/VB seeg/FW til/IN en/PDT solid/JJ ekstrautbetaling/NN i/IN mars/ ./.

statoil/NNPC tro/VB opec/NNPC ville/VB produsere/VB for/IN full/JJ
statoil/NNPC oljeprisekspert/NN tor/NNPP kartevold/NNP tro/VB opec/NNPC ville/VB beslutte/VB å/TO fortsette/VB å/TO produsere/VB for/IN full/JJ når/NN kartell/NN møtes/VB i/IN wien/NNPP onsdag/FW ./.

-prosjekt/NN av/IN internasjonal/JJ interesse/NN
statoil/NNPC og/CC shell/NNPC lansere/VB onsdag/NN gasskraftverk/NNS med/IN co2/NNPC

oljebosser/NNS på/IN tiggerferd/NN
statoil/NNPC og/CC shell/NNPC be/VB stat/NN stille/VB opp/IN med/IN penge/NNS for/IN å/TO skaffe/VB mye/JJR kraft/NN og/CC olje/NNS ./.

mange/JJ oljefelt/NNS trenge/VB co2/NNPC

injisering/NN av/IN co2/NNPC som/CC statoil/NNPC og/CC shell/NNPC ville/VB gjøre/VB på/VB draugen/NNPC og/CC heidrun/NNPP være/VB høyaktuell/JJ på/IN mange/JJR annen/PRP norsk/JJ oljefelt/FW ./.

ekspert/NNS kritisere/VB co2-plan/NN
den/PRP stor/JJ co2-plan/NN til/IN statoil/NNPC og/CC shell/NNPC være/VB i/IN strid/NN med/IN anbefaling/NNS fra/IN offentlig/JJ utvalg/NN og/CC annen/PRP eksperter/FW ./.

fisk/NNPC og/CC oljefall/NN på/IN børse/NN
pan/NNPC fish/NNPC ,/, statoil/NNPC og/CC hydro/NNPC tynge/VB i/IN topp/NN på/IN oslo/NNPC børs/NNPC ./.

greenpeace/NNP refse/VB gasskraft-plan/NNS
greenpeace/NNP refse/VB miljøbevegelse/NN for/IN å/TO juble/VB for/IN statoil/NNPC og/CC shell/NNPC plan/NNS om/IN gasskraftverk/NNS med/IN co2/NNPC

unison/JJ glede/NN over/IN gassplan/NNS
statoil/NNPC og/CC shell/NNPC få/VB unison/JJ støtte/NN til/IN sin/PRP gasskraftverkplan/NNS på/IN tjeldbergodden/NNPC ./.

kjøpe/VB mye/JJR i/IN norskehavet/NNP
statoil/NNPC ha/VB kjøpe/VBN bp/NNPC andel/NN i/IN lisens/NN 218/CD og/CC få/VB med/IN gassfunn/NN i/IN luva/NNP ./. le/VB me/NN her/IN :/:

handel/NNS for/IN 63/CD mrd/RB ./.
statoil/NNPC kjøpe/VBD vare/NNS og/CC tjeneste/NNS for/IN god/JJ over/IN 60/CD milliard/NNS i/IN fjor/FW ./. år/NN budsjett/NNS bli/VB ikke/RB mindre/FW ./.

når/NNP det/PRP regne/VB på/JJ presten..../NNS
....så/RB dryppe/VB det/PRP også/RB på/VB klokkeren/FW ./. i/IN 2005/CD kjøpe/VBD statoil/NNPC vare/NNS og/CC tjeneste/NNS for/IN 63/CD milliard/NNS kroner/FW ./.

kreve/VB svar/NNS om/IN co2-bidrag/NN
venstre/NNP ville/VB snar/JJS ha/VB løfte/NNS fra/IN statsminister/NN jens/NNPP stoltenberg/NNPP om/IN statlig/JJ bidrag/NNS til/IN co2/NNPC uten/IN slik/RB drahjelp/NN kunne/VB statoil/NNPC og/CC shell/NNPC plan/NNS strande/FW ./.

-/- må/NNP oppgi/VB annonsebudsjett/NN
statoil/NNPC måtte/VB fortelle/VB hvor/RB mye/JJ de/PRP ha/VB bruke/VBN på/IN annonse/NNS for/IN å/TO åpne/NNPC nordområde/NNS for/IN oljeutvinning/NN ,/, mene/VB transparency/NNP international/NNPC norge/NNP ./.

øke/VB i/IN norskehavet/NNP
statoil/NNPC styrke/VB sin/PRP posisjon/NN på/IN dyp/JJ vann/NN i/IN norskehavet/NNP ved/IN å/TO kjøpe/VB bp/NNPC andel/NN av/IN en/PDT gassfunn/FW ./.

lansere/VB gasskraft/NN med/IN rensing/NN
oljeselskap/NNS statoil/NNPC og/CC shell/NNPC ville/VB i/IN dag/NN lansere/VB plan/NNS for/IN en/PDT gasskraftverk/NN på/IN tjeldbergodden/NNPC med/IN co2-rensing/NN ./.

løser/NNS co2-krise/NN
statoil/NNPC og/CC shell/NNPC legge/VB i/IN dag/NN frem/IN en/PDT løsning/NN for/IN co2-rensing/NN av/IN en/PDT gasskraftverk/NN på/IN tjeldbergodden/NNPC ./. co2-gasse/NN skulle/VB deponere/VB i/IN det/PRP shell-operere/JJ felt/NN draugen/FW !./!.

enoksen/NNPP :/: -staten/NN ville/VB bidra/VB
olje/NNPC og/CC energiminister/NN odd/NNPP roger/NNPP enoksen/NNPP si/VB at/CC stat/NN ville/VB delta/VB i/IN det/PRP spleiselag/NN som/CC statoil/NNPC og/CC shell/NNPC legge/VB opp/IN til/IN på/NN tjeldbergodden/NNPC ./.

tro/VB ikke/RB på/VB rensing/NN fra/IN dag/NN én/NNP
fremskrittsparti/NN synes/VB gasskraftprosjekt/NN til/IN statoil/NNPC og/CC shell/NNPC være/VB svær/JJ gledelig/NN ,/, men/CC tro/VB ikke/RB at/CC det/PRP bli/VB co2-rensing/NN fra/IN dag/NN én/NNP ./. parti/NN be/VB regjering/NN lempe/VB på/IN rensekravet/FW ./.

de/PRP ny/JJ gassplan/NNS
shell/NNPC og/CC statoil/NNPC gå/VB sammen/RB om/IN en/PDT prosjekt/NN for/IN å/TO bruke/VB co2/NNPC fra/IN gasskraftverk/NNS til/IN å/TO øke/NNP oljeutvinningen/FW ./. forutsetning/NN være/VB at/CC myndighet/NNS være/VB med/IN på/NN lagspillet/FW ./.

lo/NNPP utfordre/VB myndighet/NNS til/IN aktiv/JJ co2-deltakelse/NN
-/- det/PRP være/VB svær/JJ gledelig/JJ at/CC statoil/NNPC og/CC shell/NNPC så/VB konkret/JJ og/CC målrettet/VBD igangsette/VB plan/NNS for/IN bygging/NN av/IN gasskraftverk/NNS på/IN tjeldbergodden/NNPC ,/, mene/VB lo/NNPP ./.

jubel/NNS for/IN ren/JJ gasskraft/NN
miljøbevegelsen/NNP juble/VB over/IN plan/NNS for/IN ren/JJ gasskraft/NN på/IN tjeldbergodden/NNPC i/IN møre/NNPC og/CC romsdal/NNPP ./. shell/NNPC og/CC statoil/NNPC gå/VB sammen/RB om/IN å/TO bruke/VB co2-gasse/NN til/IN økt/NN oljeutvinning/FW ./.

-/- designfeil/NNP bak/IN visund-lekkasje/NN
statoil/NNPC ha/VB avslutte/VBN den/PRP intern/JJ gransking/NN etter/IN gasslekkasje/NN på/IN visund-plattform/NN i/IN nordsjøen/NNP i/IN januar/FW ./. den/PRP bakenforliggende/JJ årsak/NN være/VB ifølge/IN statoil/NNPC feil/NNS design/NNS av/IN en/PDT væskeutskiller/FW ./.

shtokman/NNP beslutning/NN i/IN april/NN
visekonsernsjef/NN alxander/NNP mededev/NNP i/IN det/PRP russisk/JJ gasskonsern/NN gazprom/NNP love/VB nå/VB en/PDT shtokman-beslutning/NN innen/IN 15/JJ ./. april/FW ./. statoil/NNPC og/CC norsk/NNPC hydro/NNPC vente/VB i/IN spenning/FW ./.

norsk/JJ i/IN usa/NNPL -/- blande/VBN oljepriseffekt/NN
oljeprisens/NNP rekyl/NN torsdag/NN gi/VBD ikke/RB de/PRP stor/JJ utslag/NNS i/IN positiv/JJ retning/NN for/IN de/PRP norsk/JJ oljeaksje/NNS i/IN new/NNPC york/NNPC ./. statoil/NNPC og/CC hydro/NNPC fortsette/VBD ned/NN ,/, mens/CC pgs/NNPC bli/VBD dra/VBN opp/IN av/IN sterkje/JJ seismikktall/FW ./.

sats/NNS på/IN stige/JJ oljepris/NN
dnb/NNPC nor/NNPC markets/NNP ta/VB denne/PRP uke/NN inn/IN statoil/NNPC i/IN sin/PRP favorittporteføljе/FW ./.

design-feil/NNP bak/IN ulykke/NN på/IN visund/NNP
statoil/NNPC ha/VB avslutte/VBN gransking/NN av/IN den/PRP alvorlig/JJ gassulykke/NN på/IN visund-plattformen/FW ./.

håpe/VB ny/JJ gasskraftverk/NNS kunne/VB finansiere/VB seg/PRP selv/JJ
statsminister/NN jens/NNPP stoltenberg/NNPP ap/NNP ()/() avvise/VB ikke/RB at/CC stat/NN ville/VB bidra/VB i/IN statoil/NNPC og/CC shell/NNPC spleiselag/NNS på/VB tjeldbergodden/NNPC ,/, men/CC håpe/VB samtidig/JJ at/CC det/PRP ny/JJ gasskraftverk/NN ville/VB klare/VB seg/PRP uten/IN statsstøtte/FW ./.

-/- stat/NN ville/VB bidra/VB
olje/NNPC og/CC energiminister/NN odd/NNPP roger/NNPP enoksen/NNPP sp/NNP ()/() bekrefte/VB at/CC stat/NN ville/VB delta/VB i/IN det/PRP spleiselag/NN som/CC statoil/NNPC og/CC shell/NNPC legge/VB opp/IN til/IN på/NN tjeldbergodden/NNPC ./. men/CC han/PRP ville/VB ikke/RB antyde/VB noen/PDT sum/NNS ./.

anmelder/NN statoil/NNPC
forurensningstilsynet/NN anmelde/VB statoil/NNPC etter/IN en/PDT av/IN norge/NNPL stor/JJS oljeutslipp/FW ./.

ryddesjau/NNS i/IN styrerom/NN
jannik/NNPP lindbæk/NNPP ,/, styreleder/NN i/IN statoil/NNPC ,/, være/VB blant/IN de/PRP som/CC måtte/VB gå/VB når/NNS regjering/NN rydde/VB opp/IN i/IN de/PRP statlig/JJ styrene/FW ./.

press/NNS statoil/NNPC
venezuela/NNPL varsle/VB nasjonalisering/FW ./. det/PRP kunne/VB gå/VB hard/JJ utover/IN statoil/NNPC ./.

-/- kjøp/NNPC statoil/NNP
handelsbank/NN capital/NNPC markets/NNP anbefale/VB sin/PRP kunde/NNS å/TO laste/VB opp/IN med/IN statoil/NNPC

statoil/NNPC under/IN forventning/NNS
statoil/NNPC levere/VB rekordhøy/JJ kvartallstall/NN ,/, men/CC skuffe/VB likevel/RB markedet/FW ./.

kjøpe/NNP aksje/NNS i/IN statoil/NNPC ?/?
meglerhus/NNP gi/VB sin/PRP vurdering/NN av/IN statoil-aksjen/NN etter/IN dag/NN rekordtall/FW ./.

to/CD statoil-tips/NNP
lehman/NNP brothers/NNP og/CC ubs/NNP endre/VB kursmål/NNS på/IN statoil/NNPC

statoil/NNPC senke/VB børse/NN

statoil/NNPC være/VB på/JJ nedtur/NN tross/IN tid/NNS resultat/FW ./. falle/JJ oljepris/NN trekke/VB også/IN hydro/NNPC og/CC oslo/NNPC børs/NNPC ned/FW ./.

rekordtall/NNP fra/IN statoil/NNPC
statoil/NNPC melde/VB om/IN over/IN 50/CD prosent/NNS vekst/NN i/IN bunnlinje/NN etter/IN første/JJ kvartal/FW ./. -/- rekordsterkt/NNP ,/, melde/VB konsernsjef/NN helge/NNPP lund/NNPP ./.

ville/VB ha/VB stor/JJR statoil-skatt/NN
oljeselskapens/NN skrekk/VBD ,/, venezuela/NNP hugo/NNP chavez/NNP ,/, varsel/NNS ny/JJ skatteskjerpelse/NNS mot/IN statoil/NNPC ./.

statoil/NNPC slette/VB 23/CD million/NNS aksje/NNS
i/IN dag/NN vinker/VB statoil/NNPC farvel/NNS til/IN 23/CD million/NNS aksje/NNS fra/IN gammel/JJ bonusprogram/FW ./.

gassknipe/NNP for/IN statoil/NNPC
ikke/RB nok/RB med/IN at/CC snøhvit-forsinkelse/NN koste/VB statoil/NNPC stor/JJ penger/FW ./. nå/RB få/VB de/PRP en/PDT ny/JJ ekstraregning/NN dumpende/JJ i/IN fanget/FW ./.

statoil/NNPC falle/VB tross/IN rekordtall/NN
statoil/NNPC gå/VB i/IN rød/JJ på/IN en/PDT rekordsterk/JJ oslo/NNPC børs/NNPC ved/IN åpning/NN mandag/NN ,/, til/IN tross/IN for/IN tid/NNS god/JJS resultat/FW ./.

statoil/NNPC senke/VB børse/NN
statoil/NNPC være/VB på/JJ nedtur/NN etter/IN tid/NNS resultat/FW ./. den/PRP falle/JJ oljepris/NN trekke/VB hydro/NNPC og/CC oslo/NNPC børs/NNPC med/IN ned/FW ./.

sletter/NNS 23/CD million/NNS statoil-aksjer/NNS
i/IN morgen/NN slette/VB 23/CD million/NNS statoil/NNPC årsak/NN være/VB en/PDT riv/RB ruskende/JJ gal/JJ avgjørelse/NN da/CC aksje/NN bli/VBD børsnotere/VBN i/IN 2001/ ./.

roxar/NNPC levere/VB til/IN statoil/NNPC
roxar/NNPC levere/VB i/IN dag/NN sin/PRP første/JJ oil-water/NNP monitor/NN til/IN en/PDT pilotprosjekt/NN med/IN statoil/NNPC på/NN sleipner/NNP a./RB

ville/VB beskatte/VB statoil/NNPC hard/JJR
venezuela/NNPL president/NN og/CC internasjonal/JJ oljeselskap/NNS stor/JJ skrekk/NN hugo/NNPP chavez/NNP varsle/VB nå/RB ny/JJ skatteskjerpelse/NNS mot/IN statoil/NNPC og/CC de/PRP annen/PRP selskap/NNS som/CC foredler/VB land/NN tungolje/FW ./. le/VB mye/JJR her/IN :/:

statoil/NNPC tape/VBD arbeidsrettssak/NN
arne/NNP indreeide/NNP 68/NNP ()/() fra/IN hafrsfjord/NNPL i/IN rogaland/NNPC få/VBD fredag/NN rett/NN medhold/NNS i/IN at/CC han/PRP skulle/VB få/VB fortsette/VB i/IN statoil/NNPC til/CC han/PRP fylle/VB 70/CD år/NN ./.

anmelder/NN statoil/NNPC for/IN slapp/JJ beredskap/NNS

statoil/NNPC få/VB hard/JJ kritikk/NN av/IN staten/NNP forurensingstilsyn/NNP sft/NNP ()/() for/IN treg/JJ respons/NN etter/IN en/PDT oljelekkasje/NN på/IN norne/NNPC

ny/JJ statoil-løft/NN i/IN usa/NNPL
oljeaksje/NNS løfte/VBD seg/PRP igjen/RB på/IN wall/NNPC street/NNPC fredag/NN etter/IN en/PDT pustepause/FW ./. statoil/NNPC og/CC pgs/NNPC være/VBD blant/IN aksje/NNS som/CC bli/VBD med/IN opp/FW ./.

statoil/NNPC knalle/VB inn/IN tid/NNS resultat/NNS
statoil-konsernsjef/NN helge/NNPP lund/NNPP legge/VB frem/IN det/PRP sterk/JJS resultat/NN noen/PDT norsk/JJ selskap/NN noengang/CC ha/VB gjort/FW ./. spørsmål/NN være/VB om/IN det/PRP være/VB god/JJ nok/RB ./.

statoil/NNPC treffe/VBD gass/NN i/IN mexicogolfen/NNP
statoil/NNPC og/CC de/PRP annen/PRP eier/NNS i/IN blokk/NN 140/CD i/IN atwater/NNP valley-området/NN i/IN mexicogolfen/NNP ha/VB finne/VBN tørrgass/NN i/IN den/PRP første/JJ av/IN letebrønnene/NNS som/CC være/VB bore/VBN på/IN blokken/FW ./.

ville/VB ha/VB hydro/NNPC og/CC statoil/NNPC ut/IN av/IN iran/NNPL
fremskrittsparti/NN ville/VB ha/VB slutt/NN på/IN virksomhet/NN til/IN hydro/NNPC og/CC statoil/NNPC i/IN iran/NNPL ./.

statoil/NNPC treffe/VBD gass/NN i/IN mexico-gulfen/NN
statoil/NNPC og/CC de/PRP annen/PRP eier/NNS i/IN blokk/NN 140/CD i/IN atwater/NNP valley-området/NNP i/IN mexico-gulfen/NN ha/VB finne/VBN tørrgass/NN i/IN den/PRP første/JJ av/IN letebrønn/NN som/CC være/VB bore/VBN på/IN blokken/FW ./.

vi/PRP ha/VB hydro/NNPC og/CC statoil/NNPC ut/IN av/IN iran/NNPL
fremskrittsparti/NN ville/VB ha/VB slutt/NN på/IN virksomhet/NN til/IN hydro/NNPC og/CC statoil/NNPC i/IN iran/NNPL ,/, og/CC mene/VB det/PRP statlig/JJ eierskap/NN bør/NN bruke/VB til/IN å/TO trekke/VB selskap/NNS ut/IN av/IN landet/FW ./.

-/- største/JJ siden/RB statoil/NNPC
når/NN solenergiselskap/NN rec/NNPC komme/VB på/IN bør/NN i/IN morgen/NN ,/, være/VB det/PRP det/PRP største/JJ som/CC være/VB skje/VBN siden/RB statoil/NNPC børsinntreden/FW ./.

meglerhus/NNP med/IN statoil-advarsel/NN
meglerhus/NNP komme/VB med/IN en/PDT advarsel/NN etter/IN kursfest/NN og/CC rekordtall/NN fra/IN oljeselskapet/FW ./.

statoil/NNP omtrent/RB som/CC vente/VBD
statoil/NNPC og/CC konsernsjef/NN helge/NNPP lund/NNPP rapport/NN for/IN første/JJ kvartal/NNS vise/VB at/CC selskap/NN hovedtall/NNS komme/VBD inn/IN rett/JJ i/IN nærhet/NN av/IN hva/WP marked/NN ha/VBD vente/VBN på/IN forhånd/FW ./.

rekordresultat/NNP for/IN statoil/NNPC

statoil/NNPC resultat/NNS før/NNS finans/IN ,/, skatt/NN og/CC minoritetsinteresse/NNS være/VBD 310/CD 31,0/31,0 milliard/NNS krone/NNS i/IN første/JJ kvartal/NNS 2006/CD ,/, sammenligne/VBD med/IN 215/CD 21,5/21,5 milliard/NNS krone/NNS i/IN første/JJ kvartal/NNS i/IN fjor/FW ./.

venezuela/NNPL øke/VB trykke/VBN på/IN statoil/NNPC
etter/IN at/CC venezuela/NNPL president/NN hugo/NNPP chaves/NNP i/IN helg/NN varsle/VBD en/PDT drastisk/JJ skatteskjerpelse/NN ,/, følge/VB kongress/NN nå/VB opp/IN med/IN forslag/NNS om/IN ytterligere/JJ nasjonalisering/FW ./.

oljepris/NN sende/VBD oslo/NNPC børs/NNPC ned/IN
den/PRP svak/JJ oljepris/NN ramme/VBD den/JJ oljetung/JJ oslo/NNPC børs/NNPC ./. statoil/NNPC og/CC norsk/NNPC hydro/NNPC falle/VBD mye/JJR enn/IN 3/CD prosent/FW ./.

aker/NNPC kværner/NNPC få/VB miljøkontrakt/NN
oljeserviceselskapet/NNP aker/NNPP kværner/NNP skulle/VB designe/VB og/CC levere/VB verden/NN stor/JJS system/NNS for/IN behandling/NN av/IN oljegasser/NNS til/IN statoil/NNPC ./.

rekordsterk/JJ i/IN oslo/NNPL
oslo/NNPC børs/NNPC fortsette/VB opptur/NN og/CC stige/VB vid/JJR til/IN ny/JJ høyder/FW ./. den/PRP høy/JJ oljepris/NN hjelpe/VB nå/VB statoil/NNPC skuffer/FW ./.

investerer/NN i/IN falleferdige/JJ stat/NNS
norsk/JJ selskap/NNS som/CC statoil/NNPC ,/, telenor/NNPC og/CC hydro/NNPC risikere/VB rykte/NN i/IN verden/NN mye/JJS ustabil/JJ land/NNS ,/, mene/VB nupi/NNP

ny/JJ rekord/NN på/IN oslo/NNPC børs/NNPC
mange/JJR av/IN john/NNPP fredriksen/NNPP selskap/NNS utmerke/VB seg/PRP positiv/JJ på/NN børse/NN i/IN åpningshandelen/ ,/, men/CC statoil/NNPC falle/VB på/FW tall/FW ./.

oslo/NNPC børs/NNPC flate/VB ut/IN
børse/NN snu/VB fra/IN rekordnivå/NN og/CC være/VB nå/JJ tilbake/IN ved/IN utgangspunktet/FW ./. oljekamerat/NNS hydro/NNPC og/CC statoil/NNPC legge/VB en/PDT demper/NN på/JJ stemningen/FW ./.

storkontrakt/NN til/IN nordsjørigg/NN
statoil/NNPC ha/VB tildele/VBN borerigg/NN scarabeo/NNP 5/CD en/PDT lang/JJ kontrakt/NN for/IN boring/NN på/IN norsk/JJ sokkel/FW ./.

blodig/JJ oljeaksje/NNS
tre/NNS av/IN børse/NN tungvektere/NN ,/, norsk/NNPC hydro/NNPC ,/, statoil/NNPC og/CC dno/NNPC ,/, falle/VB som/CC blylodd/JJ i/IN tung/JJ mandags-handel/FW ./.

storkontrakt/NN til/IN nordsjørigg/NN
statoil/NNPC ha/VB tildele/VBN borerigg/NN scarabeo/NNP 5/CD en/PDT lang/JJ kontrakt/NN for/IN boring/NN på/IN norsk/JJ sokkel/FW ./.

norsk/JJ i/IN usa/NNPL -/- ny/JJ løft/NN i/IN olje/NN

oljeaksje/NNS løfte/VBD seg/PRP igjen/RB på/IN wallstreet/NNP fredag/NN etter/IN en/PDT pustepause/FW ./. statoil/NNPC og/CC pgs/NNPC være/VBD blant/IN aksje/NNS som/CC bli/VBD med/IN opp/FW ./. le/VB mye/JJR her/IN :/:

store/NNPP ,/, stygg/JJ staten/NNP ?/?
kommentar/JJ :/: stat/NN som/CC storeier/UH bli/VBD bruke/VBN som/IN investorskremsel/NN da/CC telenor/NNPC og/CC statoil/NNPC gå/VBD på/IN børs/FW ./. stat/NN ha/VB slå/VBN hard/JJ tilbake/FW ./.

ferskingenes/NNS drømmebedrifter/NNS
statoil/NNPC trone/VB på/IN topp/NN som/CC norsk/JJ student/NNS drømmearbeidsplass/JJ ./. her/IN være/VB liste/NN over/IN selskap/NNS mange/JJS student/NNS ville/VB arbeide/VB i./FW

vollvik/NNPP :/: det/PRP være/VB ikke/RB sexy/JJ å/TO tjene/VB noen/CD titall/NNS million/NNS
?/? det/PRP være/VB liksom/RB ikke/RB noen/PDT sexy/JJ å/TO tjene/VB noen/CD titall/NNS million/NNS på/VB hydro/NNPC og/CC statoil/NNPC ,/, si/VB idar/NNPP vollvik/NNPP ./.

norge/NNPL god/JJS arbeidsgiver/NNS
land/NN mye/JJS attraktiv/JJ bedrift/NNS bli/VBD hedre/VBN under/IN utdeling/NN av/IN universum/NNPC award/NNP 2006/CD onsdag/FW ./. og/CC pris/NN for/IN god/JJS arbeidsgiver/NNS gå/VBD til/IN statoil/NNPC og/CC microsoft/NNPC ./.

tilbakeholden/NN lindbæk/NNPP
han/PRP bli/VB hudflette/VBN av/IN styrekollega/NN stein/NNPP bredal/NNPP ,/, og/CC rødgrønn/JJ kraft/NNS ville/VB ha/VB han/PRP fjernet/JJ ./. likevel/RB ville/VB ikke/RB statoil-styreleder/NN jannik/NNPP lindbæk/NNPP si/VB om/CC han/PRP ha/VB tenke/VBN å/TO fortsette/VB i/IN vervet/FW ./.