FROM DISORDER TO ORDER

Extracting clinical findings
from Swedish health record text

Maria Skeppstedt

Licentiate Thesis
Department of Computer and Systems Sciences
Stockholm University
April, 2012

Stockholm University
Report series
No. 12-005
ISSN 1101-8526

Typeset by the author using LaTeX
Printed in Sweden by US-AB

Abstract

Medical disorders and findings are examples of important information in health record text. Through developing methods for automatically extracting these entities from the health record text, the possibility of making use of the information by automatic computerised processes increases. That a disorder or finding is mentioned in the health record, however, does not necessarily imply that it has been observed in the patient, because disorders that are ruled out and findings that are not observed in the patient are also mentioned.

This licentiate thesis investigates the possibility of automatically extracting disorders and findings from Swedish health record text and the possibility of automatically determining whether these findings and disorders are negated or not.

A rule- and terminology-based system that uses several Swedish medical terminologies, including SNOMED CT and ICD-10 for extracting disorders, findings and body structures mentioned in Swedish clinical text was constructed and evaluated. Moreover, an English rule-based system for negation detection, NegEx, was adapted to Swedish and evaluated on clinical text written in Swedish.

The evaluation showed that disorders and findings were recognised with low recall, whereas body structures were recognised with comparatively good results. The negation detection system that was adapted to Swedish achieved the same recall as the English system, but lower precision.

The evaluated systems are accurate enough to be useful in some applications, but need to be further developed, especially when it comes to recognising disorders and findings.

Sammanfattning

Sjukdomar och andra kliniska fynd är exempel på viktig patientinformation som till största delen dokumenteras i löpande text i patientjournaler. Genom att utveckla metoder för att automatiskt identifiera kliniska fynd i den löpande texten ökar möjligheten att använda informationen i patientjournaler för automatisk informationsutvinning. Att en sjukdom eller ett kliniskt fynd nämns i patientjournaltexten innebär emellertid inte nödvändigtvis att patienten har denna sjukdom eller uppvisar det nämnda symptomet. Detta eftersom det även dokumenteras vilka sjukdomar som går att utesluta och vilka vanliga symptom som en patient inte har. Därför behövs det även ett system som automatiskt kan avgöra vilka kliniska fynd som är negerade.

Denna licentiatavhandling undersöker möjligheten att automatiskt extrahera de sjukdomar och andra kliniska fynd som finns dokumenterade i svenska patientjournaler samt möjligheten att automatiskt avgöra om dessa dokumenterade fynd är negerade eller ej.

För att undersöka automatisk extraktion av kliniska fynd konstruerades ett regel- och terminologibaserat system som extraherar kliniska fynd och kroppsdelar ur patientjournaltext genom att matcha texten mot termerna i flera olika svenska medicinska terminologier, bland annat Snomed CT och ICD-10. För att automatiskt avgöra vad som är negerat och ej, anpassades ett engelskt regelbaserat negationsdetektionssystem, NegEx, till svenska. Dessa två system utvärderades sedan på manuellt annoterad svensk patientjournaltext.

Utvärderingen visade att täckningen var låg för extraktion av kliniska fynd, medan kroppsdelar extraherades med relativt bra resultat, samt att det svenska systemet för negationsdetektion uppnådde samma täckning som det engelska systemet, medan precisionen var lägre.

Systemen är tillräckligt bra för att vara användbara för vissa applikationer, men de behöver vidareutvecklas, särskilt vad gäller extraktion av kliniska fynd.

List of Publications

This thesis is based on the studies described in the following publications:

I Hercules Dalianis and Maria Skeppstedt 2010. *Creating and Evaluating a Consensus for Negated and Speculative Words in a Swedish Clinical Corpus.* In the Proceedings of the Negation and Speculation in Natural Language Processing, NeSp-NLP 2010 Workshop, July 10, 2010, Uppsala, pp 5–13.

II Maria Skeppstedt 2011. *Negation detection in Swedish clinical text: An adaption of NegEx to Swedish.* Journal of Biomedical Semantics 2011, 2(Suppl 3):S3.

III Maria Skeppstedt, Hercules Dalianis and Gunnar H Nilsson , 2011. *Retrieving disorders and findings: Results using SNOMED CT and NegEx adapted for Swedish* In the Proceedings of the LOUHI 2011, Third International Workshop on Health Document Text Mining and Information Analysis, Co-located with AIME 2011 Bled, Slovenia, July 6, 2011, CEUR-WS, volume 744, ISSN: 1613-0073, pp 11–17.

IV Maria Skeppstedt, Maria Kvist and Hercules Dalianis, 2012. *Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text* To be published in Proceedings of LREC-2012.

Table of contents

Chapter 1

Introduction

When patients are under care, their medical status as well as their treatment is systematically documented in a health record. The information in the health record is critical for health personnel involved in the immediate care of the patient, constituting the basis for treatment decisions, but it can also be used for medical research. (Nilsson, 2007, pp. 11-12)

Health records have traditionally been kept on paper (Nilsson, 2007, p. 150). The digitalisation of the health record, however, offers the possibility of providing the health personnel with new kinds of tools that facilitate both documentation and information retrieval of patient data. When documented information from a large database of health records is aggregated, it is also possible to use it for creating new medical knowledge. (Meystre *et al.*, 2008)

Some information in health records is stored in a structured format, but much is only available in free text format, i.e. unstructured text in which the information is expressed in natural language. The unprocessed, free text cannot in most cases be used in automatic computerised processes, but must first be transformed into a more structured format. One possible approach to extracting relevant information from this free text would be to let a group of people, preferably experts in medicine, go through the free text manually and convert it into structured data. This is a feasible approach for a small amount of data, but it becomes very expensive when applied to large amounts of free text. Therefore, in order to make better use of this unstructured free text, automatic methods for extracting relevant information from it are needed. (Friedman, 2005, p. 425)

One example of important information in the health record is documented 'clinical findings', which is defined by the International Health Terminology Standards Development Organisation, IHTSDO (2008c) as observations made when examining a patient and assessments of the patient. Some of these clinical findings are stored in a structured format, for example as diagnosis codes, but the structured data does not cover the full medical status of the patient (Petersson *et al.*, 2001). Through developing methods for extracting these clinical findings from free text, the possibility of making use of the information in the free text increases.

Clinical findings that are mentioned in health record text are not always mentioned as something that the patient actually has, or might have. In many cases, these clinical entities are mentioned in the health records as a disorder that the patient does *not* have or a finding that is *not* observed in the patient. Therefore, there is also a need to determine automatically which of the mentioned disorders and findings that the patient actually suffers from and which of them that are mentioned in a negated context. (Friedman, 2005, p. 426)

From the information extraction perspective, it could be said that the information in the free text part of the health record is disordered and un-organised. To obtain usable information, pieces of ordered data need to be extracted from the text. The aim of this licentiate thesis is therefore to explore the possibilities of turning parts of the unstructured data of the health record into structured information.

The aim is thus to turn disordered clinical text into more structured information, to go from disorder to order.

Chapter 2

Background

The general research area and the definitions that are needed for formulating the aims and research questions are described in this chapter.

2.1 Information extraction from clinical text

The research area for this thesis is the sub-domain of natural language processing that is known as information extraction, more specifically information extraction from clinical texts. Information extraction is the task of automatically extracting specific, predefined types of information from unstructured data, such as free text. It differs from the closely related field of information retrieval in that information retrieval involves retrieving documents containing the required facts, whereas information extraction involves extracting the actual facts. Information extraction is a sub-task of text mining, as the task of text mining also includes mining for relations between the extracted facts. (Meystre *et al.*, 2008)

Clinical text is the text in the health record that contains patient information in free text-format. It can for instance describe the medical history of the patient, procedures that have been carried out or results of examinations. One reason for studying information extraction from clinical texts separate from information extraction in general is that the language in clinical texts is often very different from most other types of texts, and can therefore be challenging for natural language processing tools that have been developed for other kinds of texts. (Meystre *et al.*, 2008)

Clinical texts are, for instance, often less compliant with formal grammatical rules than other types of texts. They also contain many non-standard and ambiguous abbreviations and acronyms, and often many misspellings. (Meystre *et al.*, 2008)

This informal language style means that there can be considerable variation in how the same word is written, especially a word with a complex spelling. For instance, about 60 different versions of the word 'Noradrenalin' in Swedish clinical text were found in a study comparing Finnish and Swedish health records. (Allvin *et al.*, 2011)

2.2 Disorder, finding, body structure and negation

The Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT[1]) is a medical terminology. Each concept in this terminology is defined through its name and through its semantic category, for instance the categories 'disorder', 'procedure', 'body structure', 'finding' and 'substance'. (International Health Terminology Standards Development Organisation, IHTSDO, 2008a)

The focus of this thesis is entities belonging to the three semantic categories 'disorder', 'finding' and 'body structure'. The definitions used in this thesis are based on the definitions given by International Health Terminology Standards Development Organisation, IHTSDO (2008c) and therefore their exact definition is given here. The term 'clinical finding' is used in this thesis as a term that includes both the concept 'disorder' and the concept 'finding'.

> *Clinical findings may be simply defined as observations, judgments or assessments about patients. The problem with the terms "finding" and "observation" is that they seem to refer to the judgment of the observer rather than to the actual state of the body. "Organism state" has been suggested as a more neutral name, but it would*

[1] SNOMED CT is further described in section 6.2.

need to be delimited from a "course of disease." Examples of clinical findings include: difficulty swallowing, nose bleed, diabetes, headache, and so forth. More precise and reproducible definitions of clinical findings, and the precise boundaries between findings and events, between findings and observables, between findings and situations, and the distinction between "finding" and "disorder", remain ongoing challenges at the margins. The distinction between a disorder and an observation has proven to be difficult to define in a reproducible manner across the tens of thousands of concepts included under clinical findings. Nevertheless, there are several reliable characteristics of each sub-category (disorders and findings):

1.1) Disorders

1) Disorders necessarily are abnormal.

2) They have temporal persistence, with the (at least theoretical) possibility of their manifestations being treated, in remission, or quiescent even though the disorder itself still present.

3) They necessarily have an underlying pathological process.

1.2 Findings

1) Findings may be normal (but not necessarily); no disorders may.

2) Some findings may exist only at a single point in time (e.g. a serum sodium level); no disorders may.

3) Findings cannot be temporally separate from the observing of them (you can't observe them and say they are absent, nor can you have the finding present when it is not capable of being observed).

4) They cannot be defined in terms of an underlying pathological process that is present even when the observation itself is not present.

Disorders may be present as a propensity for certain abnormal states to occur, even when treatment mitigates or resolves those abnormal states. In some cases the disease process is irrefutable, e.g. meningococcal meningitis. In others an underlying disease process is assumed based on the temporal and causal association of the disorder and its manifestation, e.g. nystagmus disorder is different from the finding/observation of nystagmus, which can be a normal physiological response to rotation of the head. If you spin

> *around and around and then have nystagmus (the finding) you still do not have nystagmus disorder. And someone can have a nystagmus disorder without currently manifesting nystagmus. Similarly, deafness disorder is different from the symptom (observation) of reduced hearing, which can be due to a number of temporary causes such as excessive ear wax.* (International Health Terminology Standards Development Organisation, IHTSDO, 2008c)

The category 'body structure' is defined as a physical anatomical entity (International Health Terminology Standards Development Organisation, IHTSDO, 2008b).

Another important concept for this thesis is negation. A negated entity is defined as an entity that is mentioned in the text, but that is negated, e.g. it is stated that a patient does *not* suffer from a certain disorder. International Health Terminology Standards Development Organisation, IHTSDO (2008a) defines negation through the following examples:

> *The meaning of some concepts in SNOMED CT depends conceptually on negation (e.g. "absence of X", "lack of X", "unable to do X" etc).* (International Health Terminology Standards Development Organisation, IHTSDO, 2008c)

Lists of negation cues are often used in the detection of negations, that is, lists of expressions that are used for indicating negation in e.g. English or Swedish. Examples of such negation cues are 'not', 'rule out' and 'lack of'. (Morante, 2010)

Chapter 3

Aim and motivation

The general aim of the thesis, as well as the specific aims for each study are described in this chapter. The choice of aim is motivated by its relation to practical applications.

3.1 Aim

The general, long-term goal for this licentiate thesis is to explore methods for automatically extracting information from Swedish clinical text, thus for turning unstructured data into structured information. More specifically, the aim of the thesis is to automatically extract clinical findings that are mentioned in the health record text as well as to determine whether these findings are negated or not. This overall aim is divided into the following two sub-aims:

- Automatically extract mentioned disorders and findings from the unstructured text of electronic health records written in Swedish.
- Automatically determine whether extracted disorders and findings are negated or not.

These two aims are in turn divided into the following four research questions:

- **Negation detection**
  - To what extent can cue expressions for negation and speculation be automatically recognised? (Study I)
  - How does an English system for negation detection that is adapted to Swedish perform on Swedish clinical text? (Study II)
  - How often are disorders and findings negated in Swedish clinical text? (Study III)
- **Extracting disorders and findings**
  - To what extent is it possible to automatically retrieve findings and disorders mentioned in Swedish clinical text by employing rule-based methods and existing Swedish terminologies? (Study IV)

3.2 Relevance and applications

There are many practical applications for recognising clinical findings in health records, as well as for determining whether these clinical findings are negated or not.

3.2.1 Clinical text mining

One application area of automatic extraction of findings and disorders is clinical text mining, that is, discovering and extracting knowledge from the clinical text (Meystre *et al.*, 2008). Examples are syndromic surveillance (Chapman *et al.*, 2005) and automatic detection of adverse events (Melton and Hripcsak, 2005).

Another example of clinical text mining is automatic hypothesis generation, such as comorbidity studies, that is, studies of co-occurrences of disorders. Such studies have been performed on structured data containing diagnosis codes, in which co-occurrences of diagnoses have been studied (Tanushi *et al.*, 2011)[1]. As some of the patient information is only available in free text format, however, not all

[1] A visualisation of a constructed comorbidity network can be found at: http://dsv.su.se/en/research/health/comorbidityview/demo/

co-occurrences of disorders will be found through analysis of the structured data alone. For instance, Roque *et al.* (2011) combined information in structured data with information extracted from clinical text when performing comorbidity studies.

A named entity recognition system that recognises disorders would facilitate studies that include the unstructured part of the corpus. A system for extracting findings as well as disorders would also make co-occurrence studies between disorders and findings possible. Such a co-occurrence study is described by Cao *et al.* (2005).

For the described examples of applications within clinical text mining, a negation detection system is necessary to classify which of the extracted disorders and findings are referred to as clinical findings in the patient and which of them are mentioned as negated clinical findings.

3.2.2 Extending and evaluating terminologies

Another application of named entity recognition of clinical entities is the use of extracted entities for evaluating and expanding medical terminologies. There are studies which evaluate the coverage of terminologies; for instance, through a manual inspection of text in order to find out to what extent the clinical entities that are found in the text are present in the terminology (Kokkinakis, 2011).

A named entity recognition system, when applied to a large corpus such as the Stockholm EPR Corpus, could be used to produce a list of common clinical terms that are not part of a certain terminology. Such a list could, in turn, be used as a basis for expanding the terminology with new concepts or synonyms to existing concepts.

3.2.3 Tools to use in daily patient care

The digitalisation of the health record offers an opportunity for new forms of presentation of the health record content. Examples of such new forms of presentation are automatic summarisations of the health record content (Hallett *et al.*, 2006; Aramaki *et al.*, 2009), automatically generated problem lists (Meystre and Haug, 2006) and visualisation of the data (Plaisant *et al.*, 1998). In order to achieve this kind of structured presentation of the information in the free text, natural language processing tools are needed (Kvist *et al.*, 2011). An important component in such systems for new forms of presentations is automatic extraction of disorders and findings mentioned in the health record, as well as automatic detection of whether the mentioned disorders and findings are negated.

Natural language processing tools can also facilitate the input of patient documentation. A draft for a discharge summary, problem list or patient certificate can be generated from the health record text, and this draft can be verified or modified before it is added to the health record.

Another application supporting input into the health record is computer-assisted coding and classification of a patient's condition, such as computer-assisted diagnosis coding based on the content of the free text in the health record (Henriksson and Hassel, 2011). Input to such a system could for instance include the disorders and findings that occur in the free text, as well as whether these are negated or not.

3.3 Motivation for selection of entities

There are other entities that would also be useful to extract from clinical text, such as procedures or occupations. This thesis, however, focuses on three categories; 'disorder', 'finding' and 'body structure'. 'Disorder' and 'finding' were chosen as they are the most important entities for describing the medical status of the patient, and 'body structure' was chosen as entities of this category are sometimes important for specifying the location of a disorder or finding.

Extraction of the three chosen entities is necessary for text-based comorbidity studies and syndromic surveillance. They are also among the most important in the construction of tools for use in daily patient care.

Chapter 4

Brief overview of included studies

A very brief overview of the included studies is given, focusing on how the studies relate to each other. The methods and results of each study are described in more detail in the following chapters.

4.1 Study I

"Creating and Evaluating a Consensus for Negated and Speculative Words in a Swedish Clinical Corpus"

In this study, a consensus corpus was constructed from three annotations of the same Swedish clinical text, carried out by three different annotators. The sentences, or clauses, in the text had been annotated as either 'certain' or 'uncertain'. Cue expressions for negation, e.g. 'not', as well as for speculation, e.g. 'likely' or 'possibly', had also been annotated.

The consensus was constructed by combining the three individual annotations. A machine learning system was thereafter applied on the constructed consensus to detect uncertainty and cues for negation and speculation. The results for the automatic detection of uncertainties and cue words for speculation were low, whereas negation cues were detected with a high precision and recall. One reason for this could be that the constructed consensus corpus only contained 13 unique ways of expressing negation, whereas it contained over 400 different cue words for expressing speculation. Also, the percentage of expressions that occurred only once in the corpus was higher for the speculation cues.

4.2 Study II

"Negation detection in Swedish clinical text: An adaption of NegEx to Swedish"

Rule-based negation detection methods for English clinical text build on the fact that there is a limited number of negation cues in English that covers most of the possible ways in which negation can be expressed. This also seems to be the case for negations in Swedish clinical text, as for instance shown in Study I, in which 13 unique ways of expressing negation were found. Therefore, it is reasonable to assume that the same methods that are used for rule-based negation detection in English clinical text would achieve similar results when applied on Swedish text.

In Study II, a rule-based English system for negation detection, NegEx, was therefore adapted to Swedish by translating English negation cues.

The input to NegEx is a sentence of clinical text and a clinical finding contained in this sentence, and for such a pair NegEx determines whether the finding is negated or not. To construct a Swedish reference standard for evaluating the adapted system, sentences containing clinical findings were automatically extracted from health records using the textual descriptions in the medical classification system ICD-10[1]. The clinical findings in these sentences were manually categorised as either negated or not negated, and the sentences were thereafter used as a reference standard for evaluating the Swedish version of NegEx.

The system adapted to Swedish achieved somewhat lower results than the English version, which could be explained by that two of the translated negation cues were

[1] See section 6.2.

not suitable as negation cues in Swedish and probably also by that the fact that the two systems were evaluated on different kinds of health record text.

4.3 Study III

”Retrieving disorders and findings: Results using SNOMED CT and NegEx adapted for Swedish”

In this study, the NegEx system that was constructed in Study II was applied on a larger Swedish clinical text set containing around 23 million tokens, in order find out how frequently clinical findings are negated in health records. Instead of using ICD-10, as in Study II, terms in the medical terminology SNOMED CT[2] were used to extract findings and disorders. These were extracted through exact string matching to terms in SNOMED CT.

It was concluded that around 9% of the clinical findings were negated, thus supporting the claim that applying negation detection in a system for extracting clinical findings is important for achieving accurate results.

The method of extracting disorders and findings through exact match to SNOMED CT terms was evaluated, and the evaluation showed that a small proportion of the mentioned disorders and findings was recognised.

4.4 Study IV

”Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text”

As it was shown in Study III that exact string matching to SNOMED CT disorders and findings resulted in only a small proportion of mentioned disorders and findings being recognised, Study IV explored the possibilities of improving rule-based methods for matching clinical text to terminologies.

[2] See section 6.2.

A system that used different preprocessing methods for matching texts to SNOMED CT, as well as to additional terminologies, was constructed for recognising clinical entities of the categories 'disorder', 'finding' and 'body structure'. For evaluation, clinical text was manually annotated for these three categories of entities, and this annotated text was used as the reference standard for evaluating the constructed system. The system was able to recognise body structures with comparatively good results, whereas the results for recognising disorders and findings remained relatively low, although they were improved by preprocessing or by the inclusion of additional terminologies.

Chapter 5

Extended background

A system for natural language processing is a system that has the ability to automatically process some form of human language (Jurafsky and Martin, 2008, p. 35). This background chapter describes general methods for constructing such natural language processing systems as well as specific methods for constructing named entity recognition systems and systems for negation detection.

5.1 Rule-based and machine-learning methods

Machine-learning is when a computer uses data, consisting of observed examples, to automatically learn to perform a certain task. Examples of tasks could be speech recognition, machine translation, or document classification. There are two main types of machine-learning, supervised and unsupervised. When performing supervised learning, the training data that is provided to the machine-learning system is labelled and when performing unsupervised learning, the system learns from unlabeled examples. (Alpaydin, 2010, pp. 2–14)

An alternative to machine-learning is to use a rule-based method. In a rule-based system, rules are manually constructed to perform the required task (Alpaydin, 2010, p. 1). An example from natural language processing could be regular expressions that are manually constructed to match certain language features, such as regular expressions to match negations (Chapman *et al.*, 2001).

A reason why machine-learning methods sometimes are preferred over rule-based methods is that there is not always an exact known method for how to perform a certain task, and thereby it is impossible to use manually constructed rules. Other reasons could be that a rule-based method would be too complex and thereby too expensive in time to construct, or that it does not adapt well to changes in the data. For instance, the content of spam e-mails might change over time and thereby it is better to retrain a machine-learning system with new kinds of spam e-mails than to construct a rule-based method for spam-filtering. (Alpaydin, 2010, p. 1)

On the other hand, machine-learning requires a large amount of training data that the system can learn from (Alpaydin, 2010, p. 1). Such data can sometimes be difficult or expensive to obtain, which makes rule-based methods preferable in some cases.

Manual annotation of text is one example of how training data for supervised machine-learning within natural language processing can be constructed. To annotate a text is to manually mark or classify characteristics of the content or structure of the text, for instance to label a token according to its part of speech or semantic category (Ogren, 2006). A related task, which also sometimes is called manual classification, is to manually classify an entire text or parts of the text, such as the sentences, or as in the evaluation of NegEx (Chapman *et al.*, 2001), to manually classify if clinical findings are negated not.

Both for rule-based systems and for machine-learning systems, labelled data is needed for evaluating the performance of the system. The set of data that is used for evaluating the performance of a system or for comparing different systems is called the reference standard. (Friedman and Hripcsak, 1998)

5.2 Named entity recognition of findings and disorders

Extracting occurrences of findings, disorders and other medical entities from free text is a kind of named entity recognition (Meystre *et al.*, 2008).

The aim of named entity recognition is to automatically find entities in text that can be referred to with a proper name, for instance a person name, a company name or a location. Named entity recognition consists of two subtasks; first spans of text that are part of a proper name are extracted and thereafter the extracted spans of text are classified according to their type, e.g. company name, person name or location. (Jurafsky and Martin, 2008, pp. 759–768)

Common methods for named entity recognition are to use gazetteers of for instance people or organisation names, or to use machine-learning methods as well as to use a combination (Mikheev *et al.*, 1999).

The concept 'named entity recognition' has later been extended to also include recognising entities that are not proper names, but that are important in a certain context (Jurafsky and Martin, 2008, pp. 759–768). Identifying medical entities is an example of such an extension of the concept of named entity recognition (Meystre *et al.*, 2008).

5.2.1 Rule-based methods for exact mapping to a terminology

Most studies of extraction of entities from clinical text have been performed with the purpose of mapping content of the text to concepts in a terminology. In these studies, the entities are thus extracted from free text through comparing the content of the text to terms in different terminologies, such as SNOMED CT[1]. The output of the described methods is then a match to a specific ID in the used terminology. The following are examples of such indexing systems for English clinical text.

The system MetaMap discovers Unified Medical Language System (UMLS) concepts in biomedical and clinical text through matching phrases to terms in the UMLS metathesaurus. The text is first parsed with a shallow parser in order to filter out noun phrases, and thereafter other inflections and spelling variants of these noun phrases are generated. Synonyms are also searched for in a synonym lexicon and abbreviations are expanded through an abbreviation lexicon, before all versions of the extracted noun phrases are searched for in the UMLS metathesaurus. (Aronson, 2001)

[1] For SNOMED CT, see section 6.2

IndexFinder is another system, for which the aim was to construct a very fast system for mapping to UMLS. Here, no parsing is performed to find noun phrases, and instead all possible permutations of short text chunks are matched to the UMLS concepts. This is combined with a user defined filter, that is, if the user for instance is only interested in diseases, only UMLS concepts with the category disease will be used for the match. (Zou *et al.*, 2003)

Other studies have investigated the possibility of improving precision of a mapping to terminologies through automatically restricting the used terminology to combinations of subsets of UMLS. A study on radiology reports showed for instance that the optimal combination of UMLS subsets varied between different sections of the report. The SAPHIRE indexing engine was used for the matching. (Huang *et al.*, 2003)

In studies described by Long (2005) and by Patrick *et al.* (2007) clinical text is matched to terms in SNOMED CT and different techniques for capturing abbreviations, misspellings, other inflections and other word orders are used.

In the system MedLEE, which originally was constructed for mapping text to UMLS terms, negation has been included in the natural language processing system. This system was adapted to not only match clinical text to codes in UMLS, but to also extract modifiers to recognised findings, for instance modifiers that contain information of negation and temporality. (Friedman *et al.*, 2004)

5.2.2 Rule-based methods for recognising clinical entities

Named entity recognition aims at solving an easier problem than mapping text to an exact concept in a terminology. The aim of named entity recognition is to extract certain types of entities, such as disorders or findings, not to map to an exact concept in a terminology.

Rule-based named entity approaches are similar to the above described methods for matching against a terminology. One rule-based approach for recognising disorders by matching to SNOMED CT has been evaluated by Savova *et al.* (2010) on a corpus annotated for disorders (Ogren *et al.*, 2008). Their approach, which relied on techniques including spelling correction and generation of word permutations (Kipper-Schuler *et al.*, 2008), resulted in an F-score[2] of 0.72 for exact match (Savova *et al.*, 2010).

There is a rule-based study for named entity recognition in Swedish clinical text that is based on extracting entities by matching the text to terms in the MeSH[3] terminology (Kokkinakis and Thurin, 2007). This Swedish named entity recognition system achieved a precision of 0.98 and a recall of 0.87 for recognition of diseases in discharge summaries.

5.2.3 Machine-learning methods for recognising clinical entities

A machine-learning technique that is often used for named entity recognition is conditional random fields. Conditional random fields is a machine-learning technique that is used for many tasks that require labelling of sequential data, which is the case for many tasks within natural language processing (Sutton and McCallum, 2010). There are a number of available implementations of conditional random fields, for instance CRF++ (Kudo, 2012) and Stanford Named Entity Recognizer (Stanford CRF NER) (Stanford, 2012).

When using a machine-learning system it must be decided what features in the data that the system is going to use for making predictions. Regardless of what machine-learning algorithm is used for named entity recognition, common features to use are the word for which the semantic class is going to be predicted, its neighboring words, prefixes and suffixes of the word, or if the word is present in word lists, such as a gazetteer of cities. (Jurafsky and Martin, 2008, pp. 759–768)

There are studies in which machine-learning methods for named entity recognition in clinical text have been explored. In a study by Wang (2009), clinical text from an intensive care unit was annotated for ten different kinds of SNOMED CT semantic categories, including the categories 'body part', 'finding' and 'qualifier'. The annotated corpus was used to train CRF++ to automatically recognise the semantic types. The performance of the CRF++ system was compared to a rule-based lexical look-up system, and the two systems achieved average F-scores of 0.81 and 0.64, respectively.

[2] For F-score, see section 6.3.1.

[3] For MeSH, see section 6.2.

The same categories were used in a study by Wang and Patrick (2009). In this study, the entities were instead recognised using a majority voting among a conditional random fields system and two other kinds of machine-learning systems. The combination of the systems had a precision of 0.84 and recall of 0.82 for the category ’finding’ and a precision of 0.76 and recall of 0.66 for the category ’body part’. This was better than the baseline, consisting only of a CRF++ system, which had a precision and recall of 0.80 for ’finding’, while ’body part’ was recognised with a precision of 0.75 and a recall of 0.60. Examples of features that were used are the current word, the preceding and following words and part-of-speech information. Whether or not a word matched a SNOMED CT term and which SNOMED CT semantic category this term belonged to were also used as features. A list of abbreviations and acronyms was also used for the SNOMED CT matching.

A conditional random fields system for named entity recognition of medical problems, tests and treatments was also explored by Jiang *et al.* (2011). Their machine-learning system used the output from other natural language processing systems, including MedLee and KnowledgeMap, as one of the features. Further on, the machine-learning system was supplemented with rule-based post-processing and with a combination of different classifiers that were trained using different features. This hybrid system resulted in an F-score of 0.84 for recognising medical entities.

5.3 Methods for negation detection

Some of the above mentioned systems for extracting findings and disorders perform additional classifications of the extracted entities. As stated above, the system by Friedman *et al.* (2004) includes negation and temporality modifiers and the previously mentioned study by Jiang *et al.* (2011) includes assertion classification of the recognised entities.

There are also many studies that focus solely on negation detection. Both rule-based and machine-learning methods have been explored for negation detection in clinical text.

5.3.1 Rule-based methods

Rule-based methods for negation detection in clinical texts build on lists of cue words for negation, that is words that indicate that there is a negation in the text. Examples of such cue words are 'not', 'without' and 'no evidence of'.

The widely used NegEx algorithm employes three different lists of cue phrases for negation, and in the first version of NegEx, a medical problem is classified as negated if it is in close proximity to a negation cue. (Chapman *et al.*, 2001)

Subsequent rule-based negation detection systems have focused on improving the method for determining the scope of the negation trigger. The next version of NegEx used a list of conjunctions to limit the scope of what medical problems a negation cue negates. A similar approach has been employed in another negation detection system, in which a list of words are limiting the extension of the negation cues. (Elkin *et al.*, 2005)

Another example of a rule-based negation detection approach, is a manual construction of a grammar for possible ways of expressing negations, in which negation cue is one type of phrase constituent (Huang and Lowe, 2007). The program NegFinder also uses structural rules for finding negated concepts. It uses a parser designed for parsing programming languages and manually constructed rules that use negation cues, negation terminators (mostly prepositions and conjunctions) as well as sentence terminators and words matching a medical terminology. (Mutalik *et al.*, 2001)

5.3.2 Machine-learning methods

Machine-learning methods have also been applied for negation detection. A machine-learning based extension of NegEx has been constructed, following the observation that the negation cue 'not' had a lower precision than the other negation cues in the NegEx system. This extension was built on the two machine-learning algorithms naive bayes classifier and decision trees, and the classification was solely focused on determining in which cases the cue 'not' indicated a negation. (Goldin and Chapman, 2003)

Decision trees have also been employed in another approach to automatically detect negations. In this approach, patterns for how negations are expressed were automatically derived from clinical text that was annotated for different diagnoses as well as for the affirmed or negated polarity of these diagnoses. The patterns consisted of automatically extracted negation cues and the number of allowed words between these cues and the diagnosis, as well as patterns for when a diagnosis occurred in a positive context. These patterns were then used as features for three different decision trees that were used in a cascade. In the first tree, only patterns where the cue was very close to the diagnosis were used, in the second tree only patterns where a diagnosis was negated were used and in the third tree also positive patterns were used. (Rokach *et al.*, 2008)

Of the here described systems for negation detection, only one of them attempts to detect negation in general, as opposed to detecting whether instances of a specific semantic class, such as diagnoses, are negated. This system was trained on the BioScope Corpus, which contains annotations for negation and speculation cues as well as the scope of these cues. The system detects negations in two steps, first the negation cues are detected using one machine-learning algorithm that classifies each token as a negation cue or not. Thereafter, the scope of the cues are learnt through another machine-learning approach, in which the results of three different machine-learning algorithms are combined into classifying whether a token is within the scope of the negation cue or not. (Morante and Daelemans, 2009)

5.3.3 Other modifiers than negations

Apart from negation, there are other reasons why a medical problem that is mentioned in a health record is not always experienced by the patient. It could be the case that a disorder is expressed as an uncertainty rather than as a negation, for instance that it is possible that a patient experiences a certain medical problem. Other reasons could be that it is a relative who suffers from the mentioned medical problem or that the medical problem was experienced in the past.

To cover some of these cases, NegEx has been extended through another rule-based system called Context, which apart from detecting negations also detects historical and hypothetical clinical conditions, as well as whether a condition is experienced by someone other than the patient. (Chapman *et al.*, 2007)

Work has been conducted on detection of clinical findings that are expressed as uncertainties, and there are also studies on clinical text written in Swedish. Velupillai (2011) has used CRF++ to automatically classify clinical findings in Swedish health record text into six different factuality levels; ’certainly positive’, ’probably positive’, ’possibly positive’, ’possibly negative’, ’probably negative’ and ’certainly negative’. The performance of the system was evaluated both for each one of these six certainty levels and for these six levels merged into four. For detecting the merged class ’probably and possibly negative’ the precision was 0.58 and the recall was 0.55. For detecting the findings belonging to the class ’certainly negative’, the precision was 0.79 and the recall was 0.60. An analysis of the annotated data showed that some clinical findings often were negated, whereas others were rarely negated. For instance, 93% of mentioned instances of ’atrial fibrillation’ were certainly positive as were 89% of mentioned instances of ’hypertension’, whereas ’ischemia’ was assigned probably negative in 28% of the cases and certainly negative in 58% of the cases.

Chapter 6

Materials and general methods

This chapter describes materials, in form of terminologies and corpora, that were used for the four studies. It also describes the evaluation methods that were employed for the studies as well as existing natural language processing tools that were used.

The chapter ends with a table showing an overview of which methods and materials that were used in which studies.

6.1 Corpora

An overview of used and created corpora is given in Table 6.1, rows 3 and 4.

6.1.1 Stockholm EPR Corpus

In all experiments different extracts from the Stockholm Electronic Patient Record Corpus (Stockholm EPR Corpus) were used. Some of these extracts were also annotated or classified as part of the experiment.

Stockholm EPR Corpus is a large corpus of patient records that has been made available for research for the research group 'Health Care Analytics and Modeling' at Stockholm University. The corpus contains health records from more than 600,000 patients patients from over 900 different health units in the Stockholm area. The records were written in the years 2006, 2007 and the first half of 2008. (Dalianis *et al.*, 2009)

The health records consist of both structured data, for instance the age and gender of the patient, and of unstructured data in form of free text (Dalianis *et al.*, 2009). The structured data also includes ICD-10 codes[1], which classify the diagnoses and symptoms of a patient. The free text could be described as semi-structured, since the information is structured under fixed headings, such as 'assessment' or 'current status'.

In Study III, a large subset consisting of 23,171,559 tokens was extracted from the Stockholm EPR Corpus in order to generate statistics of negated clinical findings. The subset was obtained by randomly extracting 500,000 fields with a headline ending with the word 'assessment'.

From the Stockholm EPR Corpus, several smaller subsets have been extracted for the purpose of creating annotated corpora. The following two previously created subsets of the Stockholm EPR Corpus were used in the studies:

Stockholm EPR Uncertainty Corpus consists of sentences that were randomly extracted from assessment entries in the Stockholm EPR Corpus. The entire corpus, consisting of 6,740 sentences, was annotated by three annotators, none of them with a medical background. Each sentence was judged as 'certain', 'uncertain', and in a few cases as 'undefined'. A sentence could also be divided into clauses if, for instance, the main clause was certain and a subordinate clause was uncertain. In addition to this, cue words for 'speculation' and for 'negation' were annotated. (Dalianis and Velupillai, 2010)

Stockholm EPR Diagnosis Symptom Corpus is annotated for terms belonging to any of the three semantic classes 'diagnosis', 'symptom', and 'diagnosis

[1] See section 6.2.

and symptom'[2]. The corpus consists of 23,100 tokens of text randomly extracted from assessment[3] fields from an emergency ward.

The following three corpora were created in the studies carried out for this licentiate thesis:

Stockholm EPR Uncertainty Corpus Consensus is a not a new corpus, but a compilation of the three individual annotations that were performed for the Stockholm EPR Uncertainty Corpus.

Stockholm EPR Negated Findings Corpus consists of 900 sentences extracted from the annotated sentences in Stockholm EPR Uncertainty Corpus as well as from surrounding sentences. The extracted sentences all contained a clinical finding that matched a textual description in the medical classification system ICD-10[4]. The extracted sentences were divided into two groups, sentences containing negation cues and sentences that did not contain negation cues. The clinical findings in each one of the sentences were thereafter manually classified into negated, not negated or uncertain. Each sentence with a negation cue was manually classified by two persons, one of them a senior physician. A subset of the sentences without negation cues were also manually classified by two persons, and the remaining sentences were classified by one person.

Stockholm EPR Clinical Entity Corpus, version 1 is a corpus annotated for findings, disorders, body structures and pharmaceutical drugs by a senior physician. The texts are randomly extracted from assessment[5] fields from an emergency ward.

6.1.2 BioScope Corpus

One corpus that is not part of the Stockholm EPR Corpus was used for this licentiate thesis, the BioScope Corpus. The BioScope Corpus contains clinical radiology

[2] The annotations were performed by two senior physicians for collecting a list of clinical entities for a study of factuality levels of diagnoses (Velupillai *et al.*, 2011).

[3] Bedömning in Swedish.

[4] See section 6.2

[5] Bedömning in Swedish.

reports and other English biomedical texts. The corpus was annotated by two students and their work was led by a chief annotator, who resolved the annotation cases in which the two students did not agree. The clinical text was annotated for negation and speculation cues as well as the scope of the cues. (Vincze *et al.*, 2008)

6.2 Terminologies

An overview of used terminologies is given in Table 6.1, row 5.

The following terminologies were used:

SNOMED CT stands for Systematized Nomenclature of Medicine – Clinical Terms and is a terminology of clinical terms. It was compiled through merging an English medical terminology constructed by the College of American Pathologists and another English terminology constructed by the National Health Service of United Kingdom (Stearns *et al.*, 2001). The purpose of SNOMED is to provide a standardised terminology for clinical information (International Health Terminology Standards Development Organisation, IHTSDO, 2008a).

In March 2011, a translation of SNOMED CT into Swedish was released by the Swedish National Board of Health and Welfare (Socialstyrelsen). This translation contains around 280,000 clinical terms and, unlike the English version, does not yet contain any synonyms (Socialstyrelsen, 2011).

The basic building blocks of SNOMED CT are the concepts. A concept in SNOMED CT is the abstract idea that refers to a single meaning in the real word. The concepts are organized into hierarchies, in which the 'SNOMED CT Concept' is the root concept, and this root concept has 19 child concepts, called 'top-level hierarchies'. Each concept is defined by a unique numeric identifier (ConceptID), which is the same regardless of what language is used. The same concept can have many synonyms, it can thus map to several terms. Each concept also has a preferred term as well as a fully specified name, which includes a semantic tag that indicates which semantic category the word belongs to, for instance 'disorder', 'finding',

’body structure’ or ’qualifier’. (International Health Terminology Standards Development Organisation, IHTSDO, 2008a)

ICD-10 codes are used to classify the diagnoses and symptoms of a patient. ICD stands for ’International Classification of Diseases’ and is an international standard diagnostic classification for general epidemiological and clinical use (WHO, 2012). The classification is managed by WHO and the most recent version, version 10, was endorsed by WHO in 1990. There is a Swedish translation of ICD-10, named ICD-10-SE (Socialstyrelsen, 2012).

MeSH is a controlled vocabulary which was created for the purpose of indexing medical literature. There is a Swedish version of MeSH that has been translated from English by Karolinska Institutet University Library. (Karolinska Institutet, 2012)

Wikipedia: Projekt medicin is a Wikipedia list of names of diseases in Swedish (Wikipedia, 2012).

Abbreviations and acronyms extracted from a book named ’Medicinska förkortningar och akronymer’, which lists Swedish medical abbreviations and acronyms (Cederblom, 2005).

6.3 Employed methods for evaluation

Normally, when running a system that automatically labels or classifies data, not everything will be correctly labelled or classified. Therefore, in order to evaluate the system, the output of the system is compared to the content of a reference standard, which for instance could consist of data that has been manually labelled or classified by experts. (Friedman and Hripcsak, 1998)

Methods and measures that are used for comparing the performance of a system to a reference standard are described below, and an overview of methods and measures used in this licentiate thesis is given in Table 6.1, row 6.

6.3.1 Measuring the performance of a system

The most common measures in e.g. information retrieval are **precision** and **recall**. (Alpaydin, 2010, pp. 489–493)

Recall is also sometimes called **sensitivity** (Alpaydin, 2010, pp. 489–493), for instance when used for diagnostic tests in medicine (Campbell *et al.*, 2007, p. 50).

Precision and recall are defined as follows (Alpaydin, 2010, pp. 489–493):

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

tp = true positives, the number of classification instances for which the predicted class matches the actual class.

fp = false positives, the number of classification instances for which the classifier incorrectly predicts the class.

tn = true negatives, the number classification instances for which the classifier correctly does not predict the class.

fn = false negatives, the number of classification instances for which the classifier fails to predict the actual class.

Precision for named entity recognition is thus the proportion of the total number of labelled chunks that are correctly labelled, whereas recall is the proportion of chunks actually present in the reference standard that are correctly identified by the system. (Jurafsky and Martin, 2008, p. 489)

Specificity is a measure that is rarely used within natural language processing. Specificity measures how well a system detects negatives occurring in the reference standard (Alpaydin, 2010, p. 493).

This measure is calculated as (Campbell *et al.*, 2007, p. 50):

$$\text{Specificity} = \frac{tn}{tn + fp}$$

Another infrequently used measure within natural language processing is the **negative predictive value**. It is, as sensitivity and specificity, for instance used for diagnostic tests in medicine. For diagnostic testing, it is defined as the proportion of patients with negative tests who are correctly diagnosed. (Altman and Bland, 1994)

Expressed in terms of true negatives and false negatives, it is calculated as:

$$\text{Negative predictive value} = \frac{tn}{tn + fn}$$

A very common measure, on the other hand, is F-score or F-measure, which is a combination of precision and recall. This is defined as follows when precision and recall are given equal weight (Jurafsky and Martin, 2008, p. 489):

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

6.3.2 Computing confidence interval for a measured value

One criterion for a reliable evaluation of a natural language processing system is to include confidence intervals for all measures (Friedman and Hripcsak, 1998). That is, the fact that a system tested on a certain test data shows a particular value for precision or for recall does not mean that it would show the same values when presented with another set of data, not even if the data is drawn from the same population. Therefore, a confidence interval for each estimated value has to be computed, that is an estimation of in what range it is likely that the true values lie (Blom, 1989, p. 222).

If the result of a random trial either can take the value ’success’ or the value ’failure’ and nothing else, and p is the probability for ’success’, then if n independent executions of the trial are performed, the random variable for the total number of successes is binomially distributed, written as $Bin(n, p)$. (Newbold *et al.*, 2003, p. 147)

The distribution of true positives given the number of returned objects $M(= tp + fp)$ is binomial with parameters $n = M$ and p = precision. Likewise, the distribution of true positives given the number of relevant objects $N(= tp + fn)$ is binomial with parameters n = N and p = recall. (Goutte and Gaussier, 2005)

Precision and recall are thus binomial proportions, which means that statistical methods for binomial proportions can be applied for estimating confidence interval, as well as for significance testing, which will be discussed in the next section. A binomial distribution can be approximated with the normal distribution if the number of observed instances are large enough, which can be justified through the central limit theorem (Andersson, 1968, chapter 2.7).

To compute a confidence interval for binomial proportions, the easiest method is to approximate the binomial distribution with the normal distribution (Campbell *et al.*, 2007, p. 92).

The formula for the confidence interval for the proportions is then:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- The variable $\hat{p}$ is the estimated proportion, for instance the precision or the recall.
- The variabel n is the total number of instances that were used in the test, for instance the total number of relevant entities in a text or the total number of extracted entities.
- The constant $z_{1-\alpha/2}$ depends on how small or large confidence interval that is desired. A common size for the confidence interval is a 95% confidence interval, and in that case the constant $z_{1-\alpha/2}$ is 1.96. (Campbell *et al.*, 2007, pp. 91–92, 94–96)

As a rule of thumb, both $\hat{p} * n$ and $(1-\hat{p}) * n$ should be larger than 10 in order to approximate the binomial distribution with the normal distribution (Andersson, 1968, chapter 2.7).

6.3.3 Comparing the performance of different systems

If two different systems are compared and one of them shows a higher recall and a higher precision for a certain set of test data, that does not automatically imply that this system will perform better on all kinds of data. In order to find out if it is likely that one of the systems actually performs better than the other when run on a certain population, significance testing between the results of the two systems has to be performed.

If the binomial proportions can be approximated with the normal distribution, there are two possible significance tests. Either a Z-test or a χ^2-test can be performed (Campbell *et al.*, 2007, pp. 126, 132–136). A significance test results in a p-value that indicates the strength of evidence, and a p-value less than 0.01 indicates strong evidence of a difference between the two systems (Campbell *et al.*, 2007, p. 107).

6.3.4 Evaluating annotation

Annotation is often a complex task, and in order to measure reliability of the labelled data, several annotators normally annotate the same text. Thereafter, the agreement between the different annotators, the inter-annotator agreement, is measured. (Artstein and Poesio, 2008)

There are several methods for measuring inter-annotator agreement. One commonly used measure that takes the probability of random agreement into account is the Cohen's kappa coefficient. (Artstein and Poesio, 2008)

6.3.5 Validating and testing a machine learning system

In order to optimise a machine learning system, for instance with respect to parameter values and used features, a validation data set is needed. The validation data set must be a data set that is separate from the test data, since in order to give a final estimation of the performance of a system, it must be tested on a data set that has not been seen by the system during training and validation. (Alpaydin, 2010, p. 477)

A common method for performing validation is to use K-fold cross-validation. When performing K-fold cross-validation, the entire data set, except the test data that is left for the final test, is divided randomly into K equally large parts. Thereafter, K pairs of training and validation data are created. In each pair, one of the K parts is used as validation data and the other K-1 parts are used as training data. K is typically set to 10 or 30, and the method is then called ten-fold cross-validation or thirty-fold cross-validation. (Alpaydin, 2010, p. 487)

6.4 Used natural language processing tools

An overview of used natural language processing tools is given in Table 6.1, row 2.

The following natural language processing tools were used:

Stanford Named Entity Recognizer (Stanford CRF NER) is an implementation of the machine learning algorithm conditional random fields. The implementation also includes automatic feature extractors that are optimised for named entity recognition in English and German. (Stanford, 2012)

NegEx is a negation detection system constructed for English clinical text. Given a sentence, and a clinical finding that is mentioned in this sentence, as input, NegEx determines if that finding is negated or not. The NegEx algorithm employes three different lists of cue phrases for negation; cue phrases preceding the clinical finding e.g. 'no evidence of', cue phrases following it e.g. 'unlikely' and pseudo-negation phrases, that is phrases that could be mistaken for negation cues even though they are not, e.g. 'not only'. The algorithm classifies the mentioned finding to be negated if it is in the range of one to six words from a post- or pre-negation trigger. (Chapman *et al.*, 2001)

When evaluated on sentences containing negation cues, NegEx achieved a precision of 0.845 and a recall of 0.82, and for sentences without negation cues, the system achieved a negative predictive value of 0.97. Of the correctly classified negations, 82% were triggered by three cues for negation; 'no', 'without' and 'no evidence of'. (Chapman *et al.*, 2001)

Granska Tagger is a part-of-speech tagger for Swedish that is built on Hidden Markov Models (Carlberger and Kann, 1999). Apart from performing part-of-speech tagging, it can also lemmatise the tagged words (KTH, 2012a). Lemmatisation is to transform different inflections of a word into the same base form, its lemma form (Jurafsky and Martin, 2008, p. 645).

Granska Inflector is a word inflector for Swedish that generates inflections for Swedish words (KTH, 2012b).

Compound Splitter is a tool for automatic compound splitting of Swedish words (Sjöbergh and Kann, 2004).

Knowtator is a text annotation tool that is a plug-in to the program Protégé (Ogren, 2006).

6.5 Overview of employed materials and methods

	Study I	Study II	Study III	Study IV
Overall method	- Machine learning	- Rule-based	- Rule-based	- Rule-based
NLP tools	- Stanford CRF NER	- NegEx - Inflector	- NegEx	- Knowtator - Granska - Compound splitter
Used corpora	- Stockholm EPR Uncertainty Corpus - BioScope	- Stockholm EPR Uncertainty Corpus	- Stockholm EPR Corpus (subset)	
Created corpora	- Stockholm EPR Uncertainty Corpus Consensus	- Stockholm EPR Negated Findings Corpus		- Stockholm EPR Clinical Entity Corpus
Used terminologies		- ICD-10 - MeSH	- SNOMED CT	- SNOMED CT - MeSH - ICD-10 - Wikipedia: Projekt medicin - Abbrv. & acron.
Evaluation measures and methods	- Precision/Recall - F-score - Ten-fold cross-validation	- Precision/Recall - Specificity - Negative predictive value - Confidence interval - χ^2-test - Cohen's kappa	- Precision/Recall - Confidence interval	- Precision/Recall - F-score - Confidence interval

Table 6.1: Used methods, NLP (natural language processing) tools, corpora, terminologies and evaluation measures.

Chapter 7

Specific methods

This chapter includes a description of the detailed methods that were used for each one of the four studies.

7.1 Experiment structure

Each study follows the structure of the experimental Cranfield evaluation paradigm (Voorhees, 2002) that has been dominant since the 60s for information retrieval. A similar approach is used in information extraction and natural language processing. A system performing an natural language processing task is constructed and the performance of this system is evaluated, normally based on comparing the system to a manually created reference standard (Friedman and Hripcsak, 1998).

The method for each of the studies thus consists of two parts, the first one is the innovative part, in which something new is created, and the second one is the evaluation part, in which standard quantitative methods for comparing the output of the constructed system to the reference standard are applied.

An exception to this is Study III, in which nothing new is created, but instead a system constructed in a previous study is used to investigate the language in clinical text.

7.2 Study I

The study ”Creating and Evaluating a Consensus for Negated and Speculative Words in a Swedish Clinical Corpus” investigated to what extent cue expressions for negation and speculation in clinical text can be automatically recognised.

7.2.1 Corpus construction and training of a machine learning model

A consensus corpus was created from three individual annotations that had previously been carried out for constructing the Stockholm EPR Uncertainty Corpus. A few basic rules for combining the three annotations were devised, mostly based on a majority vote among the three annotators. These rules were thereafter automatically executed to create the consensus corpus, the Stockholm EPR Uncertainty Corpus Consensus.

Thereafter, the constructed corpus was used by a machine learning system to learn to automatically detect cues for negation and speculation as well as to detect expressions of uncertainty in the text. The used machine learning system was Stanford CRF NER. As a comparison, and to verify the suitability of the chosen machine learning method, Stanford CRF NER was used for learning to automatically detect negation and speculation cues, as well as their scope, in the annotated BioScope Corpus.

7.2.2 Evaluation method

The consensus was investigated through exploring differences between the consensus and the individual annotations. The constructed Stockholm EPR Uncertainty Corpus Consensus was thereafter used as a reference standard for evaluating the performance of the constructed model for named entity recognition of cues and expressions of uncertainty. The training and testing was carried out using ten-fold cross validation. The same evaluation method was used for evaluating the model created for recognising cues and scope in the BioScope Corpus.

7.3 Study II

In the study ”Negation detection in Swedish clinical text: An adaption of NegEx to Swedish”, an adaption of an English system for negation detection was evaluated on Swedish clinical text

7.3.1 System adaption method

In this study, the negation detection system NegEx, which has been developed for English, was adapted to Swedish. Since NegEx has shown relatively good results, even though it uses a very straightforward method based on three lists of different negation cues, it was decided to be a good starting point for negation detection in Swedish.

Swedish negation cues were obtained through a translation of English negation cues, through the use of Google translate and a dictionary, and different inflections for the translations were also generated using the Granska inflector. A total of 148 cue expression from NegEx version 2 (NegEx, 2009) were translated. To make the system more comparable to the evaluated English version, only the 42 translated cue expressions that were most frequently occurring in a subset of Stockholm EPR Corpus were used for the evaluation.[1]

It was hypothesised that an adaption of NegEx to Swedish would show the same results as for English, since English and Swedish are grammatically close and since negation seems to be expressed also in Swedish clincial text through a limited number of cue phrases.

7.3.2 Evaluation method

The evaluation was performed with Stockholm EPR Negated Findings Corpus as the reference standard. This corpus was obtained by a manual classification of 900 sentences. These sentences were extracted by searching for mentions of the textual descriptions of ICD-10 codes in the Stockholm EPR Uncertainty Corpus.

[1] The used negation cues can be found at http://people.dsv.su.se/~mariask/resources.html

The significance of the difference between the precision of the Swedish and the English version was measured using the χ^2-test.

7.4 Study III

In the study, "Retrieving disorders and findings: Results using SNOMED CT and NegEx adapted for Swedish", the Swedish adaption of NegEx was used for estimating how often disorders and findings are negated in Swedish clinical text.

7.4.1 Experimental set-up

In this study, the NegEx system constructed in Study II was applied on a larger subset of the Stockholm EPR Corpus in order to explore the prevalence of negated findings and disorders in the corpus. This larger subset contained a total of 23,171,559 tokens and was obtained by randomly extracting 500,000 fields from the Stockholm EPR Corpus that had a headline ending with the word 'assessment'.

The explored entities were retrieved by an exact string matching to terms in SNOMED CT that fulfilled the following two criteria:

- The term must have one of the semantic categories 'disorder' or 'finding'.
- If the term is a 'finding', it must not be a common Swedish word. (A common Swedish word was defined as a word that occurred more than five times in the Swedish PAROLE corpus, which is a non-medical corpus.) The reason for removing common Swedish words was that when including them, SNOMED CT terms for findings often matched words in the text that were not clinical findings, e.g. 'walk'[2], which resulted in a low precision.

The Swedish version of NegEx was thereafter executed on sentences containing an identified disorder or finding.

[2] Translated as 'går' in Swedish, which is common in many set expressions.

7.4.2 Precision and recall of the method

The precision and recall of the method of retrieving clinical entities through exact string matching was evaluated against the Stockholm EPR Diagnosis Symptom Corpus. This corpus contained notes from one emergency care unit and consisted of 23,100 tokens. A reference standard for the evaluation was constructed through grouping all three annotation classes in this corpus into one class, the class ’clinical entity’. The exact match against SNOMED CT was thereafter evaluated against this reference standard.

7.5 Study IV

In the study ”Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text”, a rule- and terminology-based system for automatically recognising clinical entities was evaluated against a manually annotated reference standard.

7.5.1 System construction method

A rule- and lexicon-based named entity recognition system for detecting entities of the semantic types disorders, findings and body structures was constructed. The definitions of these entities all corresponded to the SNOMED CT definitions of these three semantic categories.[3] The system is based on string matching to terms in terminologies and was constructed to be customisable, in the sense that different terminologies as well as priority rules between matched terms can be added. The system also makes it possible to choose between four different pre-processing techniques; lemmatisation, a generation of all possible permutations of the tokens in a sentence, a match to terminology terms with a Levenshtein distance[4] of one from the original word and finally compound splitting.

The system uses Granska for tokenization as well as for lemmatisation, and the Swedish compound splitter is used for compound splitting. The terminologies that were evaluated were subsets of SNOMED CT, ICD-10 and MeSH as well as lists

[3] See section 2.2.

[4] A measure for distance between strings (Jurafsky and Martin, 2008, p. 697).

derived from Wikipedia and a book on medical abbreviations. The used subsets of the terminolgies all belonged to one of the semantic categories disorder, finding, body structure, qualifier or person.

7.5.2 Evaluation method

For the evaluation, a subset of the Stockholm EPR Clinical Entity Corpus, version 1 was used as the reference standard. The corpus consists of clinical notes from a Swedish emergency unit and was manually annotated for the entities 'disorder', 'finding' and 'body structure' by a senior physician using the annotation tool Knowtator. The used subset, which consisted of about a third of the annotated data, contained 26,011 tokens and a total of 2,342 annotations for the three types of entities.

Different settings in terms of used terminologies and pre-processing methods were used for the evaluation.

The precision, recall and F-score for the constructed system were calculated for each setting, and an evaluation was also carried out on the annotated instances that the system failed to recognise, especially with respect to the number of tokens and occurrences of abbreviations in these instances.

7.6 Choice of system construction method

For English, NegEx has shown relatively good results compared to more complex systems. As Swedish and English are grammatically close and as Study I showed that negation is expressed in Swedish clinical text, as in English, with a limited set of negation cues, it was hypothesised that similar results would be achieved for Swedish as for English. Therefore, considering the low complexity of the NegEx system, combined with its relatively good results, it was considered as the best method for constructing a negation detection system for Swedish.

For named entity recognition of clinical entities on the other hand, English studies have shown that machine-learning systems perform much better than rule-based systems (Wang, 2009). As a consequence, it could be claimed that a machine-learning system would be a more appropriate choice for a Swedish named entity recognition system for clinical findings. All described English machine-learning systems for recognising clinical findings, however, use a rule- and terminology-based system for feature extraction. These studies therefore indicate that a rule- and terminology-based system is necessary in order to achieve good results for a machine-learning system.

7.7 Other possible evaluation methods

Friedman and Hripcsak (1998) claim that since natural language processing systems within the clinical domain have a more specific application than general natural language processing systems, they could be evaluated in a more realistic setting. Evaluation studies within the clinical domain could thus be performed through implementing a specific clinical application and evaluating it in a realistic clinical setting, instead of performing evaluations through a comparison to a reference standard. Thereby, the effectiveness of the system for this specific application would be more accurately evaluated.

The field of natural language processing on clinical text written in Swedish is, however, new in comparison to natural language processing on English clinical text. Therefore, it is not yet evaluated how well standard natural language processing techniques perform on Swedish clinical text. Before it has been established that existing tools have an acceptable performance, or before these tools have been developed, it is a more important contribution to construct or evaluate appropriate tools for Swedish than to evaluate a specific application.

To carry out evaluation through an implementation of applications for natural language processing tools in the clinical domain is more appropriate when the research field for Swedish clinical text is more established. Such an approach would influence the entire research paradigm. Instead of carrying out experimental research, in which different experimental approaches are evaluated against a reference standard, the research would be more focused on constructing an artefact that solves a specific practical problem within a clinical setting or within medical research. With this focus, it would be suitable to position the research within the design science paradigm, since the core of design science is to create artefacts with the purpose of finding a solution to a real world problem and to evaluate this artefact (Peffers *et al.*, 2008).

7.8 Ethics

The data that is contained in health records is often very sensitive, which makes it extremely important that the content is read by as few as possible and that the identity of patients is not unnecessarily revealed to researchers. Therefore, the health record data in Stockholm EPR Corpus has been provided for research in a format in which structured data that can identify patients, such as patient names and social security numbers, has been removed.

When extracting data for annotation, small extracts from health records from many patients have been randomly chosen as opposed to using the entire health record of only one patient, thereby reducing the risk of involuntarily identifying patients from context.

The studies described in this licentiate were conducted after approval from the Regional Ethical Review Board in Stockholm, permission number 2009/1742-31/5.

Chapter 8

Results and conclusions

A summary of the results and conclusions from each of the four studies is given in this chapter.

8.1 Study I

In the study "Creating and Evaluating a Consensus for Negated and Speculative Words in a Swedish Clinical Corpus" a consensus was created between three individual annotations for negation and speculation cues as well as for uncertain or certain sentences or clauses. Moreover, it was investigated to what extent uncertainty and cue expressions for negation and speculation can be automatically recognised with a machine learning system.

When constructing the Stockholm EPR Uncertainty Corpus Consensus through combining the three different annotations for negation and uncertainty in Stockholm EPR Uncertainty Corpus, it was found that 92% of the sentences in the corpus had been identically annotated by at least two of the annotators. For these sentences, a consensus was created through a majority vote and for the remaining 8%, other rules for creating a consensus were applied. Table 8.1 shows statistics of the difference between the individual annotations and the constructed consensus.

Type of annotated class	**Mean for individual annotations**	**Consensus**
Cues for negation	853	910
Cues for uncertainty	1,174	1,077
Uncertain expression	697	582
Certain expression	4,787	4,938
Undefined expression	257	146

Table 8.1: Comparison of the number of occurrences of each annotation class between the mean of the number of occurrences in the three individual annotations in Stockholm EPR Uncertainty Corpus and the number of occurrences in Stockholm EPR Uncertainty Corpus Consensus. The figures for the individual annotations are the mean of the three annotators, normalised on the number of sentences in the consensus.

A comparison between cue expressions for speculation and negation, both for the Swedish Stockholm EPR Uncertainty Corpus Consensus and for the English BioScope Corpus was carried out and is shown in Table 8.2. Stockholm EPR Uncertainty Corpus Consensus contained only 13 different unique ways of expressing negation, whereas it contained over 400 different cue expressions for uncertainty. Also, the percentage of expressions that occurred only once among the annotated instances were larger among the speculation cues, with 38% among negation cues and 72% among the speculation cues. The number of unique cue words for negation in the BioScope Corpus was close to the number of unique negation cues in Stockholm EPR Uncertainty Corpus Consensus, but for speculation cues Stockholm EPR Uncertainty Corpus Consensus contained many more unique cues than the BioScope Corpus. Moreover, in the BioScope Corpus, only 24% of the speculation cues were cues that only occurred once.

The ability of the machine learning system Stanford CRF NER to automatically recognise the annotated entities in Stockholm EPR Uncertainty Corpus Consensus was compared to its ability to recognise annotations in the BioScope Corpus. The results show a much higher F-score for recognition of uncertainty cues in the BioScope Corpus than in Stockholm EPR Uncertainty Corpus Consensus (Table 8.3). The difference in the distribution of uncertainty cues could be one explanation for this difference. Also for recognition of negation cues, the results were better for the BioScope Corpus.

Type of word	Consensus (sv)	BioScope (en)
Unique words (Types) annotated as ’negation cues’	13	19
’Negation cues’ that occurred only once	5	10
Unique words (Types) annotated as ’Speculation cues’	408	79
’Speculation cues’ that occurred only once	294	19

Table 8.2: Number of unique words both in the Swedish Stockholm EPR Uncertainty Corpus Consensus and in the English BioScope Corpus that were annotated as ’negation cues’ and as ’speculation cues’, and how many of these that occurred only once.

The Stanford CRF NER was also applied on the annotations on sentence and clause level, which categorise an entire sentence or clause as uncertain or certain (and in a few cases as undefined). Of the 5,641 sentences in Stockholm EPR Uncertainty Corpus Consensus, 147 were split up into clauses, and for the rest, the entire sentence was categorised as belonging to one of the three classes. The results for detecting sentences categorised as uncertain were similar to the results for detecting cues for uncertainty, with an F-score of 0.424.

Class Neg-Spec	Consensus (sv)	Individual (sv)	BioScope (en)
F-score negation cues	0.897	0.838	0.971
F-score uncertainty cues	0.464	0.455	0.908

Table 8.3: The results for ’negation’ and ’uncertainty’ when executing Stanford CRF NER using ten-fold cross validation. ’Individual’ stands for the average of the three individual annotations.

The main conclusion was that it proved to be difficult to detect cue words for uncertainties in the Swedish text by the applied automatic methods, whereas the system achieved good results for English cue words for uncertainty. For negation cues, on the other hand, the automatic systems achieved high results, perhaps because of the limited amount of unique negation cues.

The annotations for uncertainty had been performed to target expressions of uncertainty in general. For future work, it was therefore suggested to focus the annotation of uncertainty on a specific type of concept that is to be extracted, in order to improve both the inter-annotator agreement and the possibility of automatically learning to detect the concept of uncertainty through natural language processing methods.

8.2 Study II

In the study "Negation detection in Swedish clinical text: An adaption of NegEx to Swedish", an adaption of an English system for negation detection was evaluated on Swedish clinical text

As a reference standard for evaluating the adapted system, the Stockholm EPR Negated Findings Corpus was constructed. A total of 900 sentences, divided into sentences with negation cues (group 1.) and sentences without negation cues (group 2.), were manually categorised and used as evaluation data. All sentences in group 1. and 95 of the sentences in group 2. were manually categorised by two annotators, one of them a physician. The clinical findings were categorised into 'negated', 'not negated' or 'uncertain', and the two categories 'uncertain' and 'not negated' were combined into one category for the evaluation, the category 'affirmed'. The proportion of sentences with negated clinical findings and the proportion of sentences with affirmed clinical findings is shown in Table 8.4.

	Negated	**Affirmed**	**Total**
Group 1: Sentences with negation cues	269	289	558
Group 2: Sentences without negation cues	12	330	342

Table 8.4: Number of sentences manually classified as 'negated' and 'affirmed (uncertain and not negated)'.

In group 1., the inter-annotator agreement was 87.4% and for the 95 sentences that were categorised by two annotators in group 2., there was an inter-annotator agreement of 100%. Cohen's Kappa for group 1., with respect to the two categories 'negated' and 'not negated' was 0.745. For the evaluation of the adaption of NegEx, the categorisations made by a physician were used for the reference standard.

Sentences with negation cues	**English**	**Swedish (95% CI)**
Recall	0.824	0.819 (0.773 - 0.864)
Specificity	0.825	0.747 (0.696 - 0.797)
Precision	0.845	0.752 (0.702 - 0.801)
Negative predictive value	0.802	0.814 (0.767 - 0.861)

Sentences without negation cues	**English**	**Swedish (95% CI)**
Negative predictive value	0.970	0.965 (0.945 - 0.986)

Table 8.5: Results for the Swedish adaption of NegEx. Figures for English from Chapman *et al.* (2001).

The Swedish adaption of NegEx showed a precision of 0.752 and a recall of 0.819, for sentences containing a negation cue and a negative predictive value of 0.965 for sentences not containing a negation cue. The results are compared to the corresponding English results in Table 8.5. The lower precision for the Swedish adaption can partly be attributed to that two of the translated negation cues, 'icke' and 'utan' were not entirely suitable as negation cues for Swedish. When removing the cue 'icke' and including rules for disambiguating the cue 'utan', the precision was increased to 0.779. The precision was, however, still lower for the Swedish version, which might be attributed to different types of clinical texts in the English and Swedish studies. The English version was evaluated on discharge summaries, which often are written in a more formal language than the assessment fields that were used for the Swedish evaluation.

The frequency and precision for negation cues was calculated and is shown in Table 8.6. The completeness of the used negation cues was investigated through a comparison with manually annotated negation cues in Stockholm EPR Uncertainty Corpus and through a manual search for additional negation cues in group 2. in the evaluation data. Three infrequent negation cues were found that were not among the translated cues, and among the 42 most common translated negation cues there were two negation cues that had not been annotated in the Stockholm EPR Uncertainty Corpus.

It could thus be concluded that the recall for the Swedish adaption was similar to the original NegEx system for English, whereas the precision was lower for the Swedish system than for the English. However, since many negated findings were identified through a limited set of cue expressions, it was concluded that the

Phrase		**Number of occurrences**	**Precision**
förnekar	(denies)	3	100,0%
aldrig	(never)	3	100,0%
avsaknad av	(absence of)	2	100,0%
inga tecken	(no signs of)	25	96,0%
ingen	(no, common gender)	84	90,5%
inga	(no, plural)	48	87,5%
inget	(no, neuter gender)	6	83,3%
inte har	(not have)	6	66,7%
utan tecken	(without signs of)	3	66,7%
utan	(without)	20	65,0%
inte	(not)	45	57,8%
ej	(not)	31	54,8%
inte visar	(does not show)	6	50,0%
har inte	(have not)	4	50,0%
icke	(non-, not)	7	0,0%

Table 8.6: All negation cues that occur more than once, their precision and the number of times they occur in the evaluation data.

same general approach with cue expressions is possible to use for Swedish clinical text, but that it needs to be further tailored to Swedish. It was also concluded that the method of translating English negation cues and using the most frequently occurring cues was sufficient for covering common negation cues, but that there were a few unusual Swedish negation cues that were not included in the list.

8.3 Study III

In the study, "Retrieving disorders and findings: Results using SNOMED CT and NegEx adapted for Swedish", the Swedish adaption of NegEx was used for estimating how often disorders and findings are negated in Swedish clinical text.

In order to study the frequency of negated disorders and findings, a randomly extracted subset of the Stockholm EPR Corpus was used. Disorders were extracted through an exact string matching to SNOMED CT terms belonging to the semantic

Affirmed	# occurrences	Negated	# occurrences
hypertension	7,508	disease	1,227
disease	5,886	ischemia	575
asthma	5,401	asthma	501
atrial fibrillation	5,205	allergic state	453
heart failure	4,274	hearing loss	432
pneumonia	3,457	foreign body	409
otitis	3,447	wound	390
wound	2,859	pulmonary embolism	383
anemia	2,797	angina	358
renal failure	2,733	heart failure	355
All retrieved:	207,717		20,814

Table 8.7: The most common affirmed and negated **disorders** in the SNOMED CT list of disorders. Affirmed in this case is 'Not negated'. The columns '# occurrences' give the number of affirmed or negated occurrences. The English translations of the used Swedish SNOMED CT terms are given here, their Swedish translations can be found in Table 8.8

category disorder, whereas findings were extracted through an exact string matching to a list of SNOMED CT findings, from which common non-clinical words and expressions were removed.

The Swedish adaption of NegEx was thereafter applied on the extracted disorders and findings. The results showed that the proportion of negated disorders was 9.1% (± 0.1% , 95% CI) and the proportion of negated findings was 9.3% (± 0.2%, 95% CI).

The most common affirmed, i.e. not negated, disorders as well as the most common negated disorders are shown in Table 8.7. The most common affirmed and negated findings are shown in Table 8.9. For the disorders and findings listed in Table 8.7 and Table 8.9, proportion of negated occurrences is shown in Tables 8.8 and 8.10.

It can be observed that some of the findings and disorders show a much higher proportion of negated occurrences than the general proportion of negated findings and disorders, for instance 'ischemia', which is negated in 44% of the cases. There are

Disorder	(in Swedish)	% negated
ischemia	(ischemi)	44
foreign body	(främmande kropp)	43
allergy	(allergi)	28
pulmonary embolism	(lungemboli)	18
disease	(sjukdom)	17
hearing loss	(hörselnedsättning)	16
angina	(angina)	12
wound	(sår)	12
otitis	(otit)	9
anemia	(anemi)	8
asthma	(astma)	8
heart failure	(hjärtsvikt)	8
pneumonia	(pneumoni)	7
atrial fibrillation	(förmaksflimmer)	3
hypertension	(hypertoni)	3
renal failure	(njursvikt)	3

Table 8.8: The proportion of negated occurrences for the **disorders** in Table 8.7.

also findings and disorders that are more seldom negated than average, for instance 'atrial fibrillation' and 'hypertension'. This information of frequency of negation could for instance be used as one parameter in a negation detection system.

The evaluation of the method of extracting findings and disorders through exact string matching against SNOMED CT disorders and findings showed that the method had a precision of 0.80. A manual review of the false positives showed that most of them could be classified as a clinical finding or modifier to a clinical finding. The recall, on the other hand, when using modified SNOMED CT lists, was only 0.13 for extracting the clinical entities and 0.23 when using the complete lists of SNOMED CT disorders and findings.

The main conclusion of the study was that around 9% of the disorders and findings are negated in the Stockholm EPR Corpus and that methods for properly handling negations in information extraction applications therefore are important. Some expressions for disorders and findings were more often negated than aver-

Affirmed	**# occurrences**	**Negated**	**# occurrences**
sinus rhythm	2,269	chest pain	454
chest pain	1,896	bruit	227
abdominal pain	1,750	recall arranged	221
tinnitus	1,562	edema	215
next appointment	1,270	hematuria syndrome	213
dyspnea	1,239	proteinuria	190
reflux	1,226	dyspnea	160
bruit	1,204	reflux	155
hematuria syndrome	1,131	follow-up arranged	149
edema	1,013	abdominal pain	134
All retrieved:	60,571		6,180

Table 8.9: The most common **findings** in the modified SNOMED CT list of findings in which common non-clinical terms are excluded. Affirmed in this case is 'Not negated'. The columns '# occurrences' give the number of affirmed or negated occurrences. The English translations of the used Swedish SNOMED CT terms are given here, their Swedish translations can be found in Table 8.10

age, whereas some were more seldom negated. From the evaluation of the method, it could be concluded that an extraction of disorders and findings by an exact string matching to SNOMED CT results in a low recall.

Finding	(in Swedish)	% negated
recall arranged	(återbesök planerat)	79
follow-up arranged	(uppföljning planerad)	76
chest pain	(bröstsmärta)	19
edema	(ödem)	18
bruit	(blåsljud)	16
hematuria syndrome	(hematuri)	16
proteinuria	(proteinuri)	16
dyspnea	(dyspné)	11
reflux	(reflux)	11
abdominal pain	(buksmärta)	7
tinnitus	(tinnitus)	5
next appointment	(nästa besök)	2
sinus rhythm	(sinusrytm)	1

Table 8.10: The proportion of negated occurrences for **findings** in Table 8.9.

8.4 Study IV

In the study ”Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text”, a rule- and terminology-based system for automatically recognising clinical entities was evaluated against a manually annotated reference standard.

For evaluating the rule- and terminology-based system constructed for performing named entity recognition of disorders, findings and body structures, eleven different settings for the system were tested. These eleven different settings were:

1: Base A base-line only performing exact string matching to SNOMED CT.

2: Lemm The clinical text was lemmatised.

3: Stop The SNOMED CT terms for body structure were stop word filtered.

4: Qual SNOMED CT terms signifying qualifiers and person were used to decrease the number of matches for findings and body structures.

5: Leve Text chunks with a Levenshtein distance of one from the original text chunk were also matched to the SNOMED CT terms.

6: Perm Permutations of tokens in the clinical text were also matched to SNOMED CT terms.

7: Comp A compound splitter was used to also match parts of words in the text to the terminology.

8: ICD10 The describing text of a subset of ICD-10 was used as an additional terminology.

9: MeSH A subset of the MeSH terms was added as an additional terminology.

10: Wiki A Wikipedia list of diseases was added.

11: Abbr Finally, a list of medical abbreviations was also added as a terminology.

The results for the 11 experiments are presented in one table for each semantic category, Tables 8.11, 8.12 and 8.13

Nr.	Precision (95% CI)	Recall (95% CI)	F-Score
1: Base	0.78 (± 0.04)	0.38 (± 0.03)	0.51
2: Lemm	0.78 (± 0.04)	0.39 (± 0.03)	0.52
3: Stop	0.78 (± 0.04)	0.39 (± 0.03)	0.52
4: Qual	0.78 (± 0.04)	0.39 (± 0.03)	0.52
5: Leve	0.77 (± 0.04)	0.41 (± 0.04)	0.54
6: Perm	0.78 (± 0.04)	0.39 (± 0.03)	0.52
7: Comp	0.74 (± 0.04)	0.41 (± 0.03)	0.52
8: ICD10	0.79 (± 0.04)	**0.41** (± 0.04)	0.54
9: MeSH	0.73 (± 0.04)	**0.46** (± 0.04)	0.56
10: Wiki	0.74 (± 0.04)	**0.49** (± 0.04)	0.59
11: Abbr	0.75 (± 0.04)	**0.55** (± 0.04)	**0.63**

Table 8.11: Results for the semantic category **disorder**. Preprocessing had little or no effect, but the inclusion of additional terminologies (8:ICD10 – 11:Abbr) improved recall.

Nr.	Precision (95% CI)	Recall (95% CI)	F-Score
1: Base	0.51 (± 0.04)	0.23 (± 0.02)	0.31
2: Lemm	0.52 (± 0.04)	**0.29** (± 0.02)	0.37
3: Stop	0.53 (± 0.04)	0.29 (± 0.02)	0.37
4: Qual	**0.57** (± 0.04)	0.30 (± 0.02)	0.39
5: Leve	0.57 (± 0.04)	0.30 (± 0.02)	0.39
6: Perm	0.57 (± 0.04)	0.30 (± 0.02)	0.39
7: Comp	0.55 (± 0.03)	**0.33** (± 0.03)	**0.41**
8: ICD10	0.57 (± 0.04)	0.30 (± 0.02)	0.39
9: MeSH	0.57 (± 0.04)	0.30 (± 0.02)	0.39
10: Wiki	0.57 (± 0.04)	0.30 (± 0.02)	0.39
11: Abbr	0.57 (± 0.04)	0.30 (± 0.02)	0.39

Table 8.12: Results for the semantic category **finding**. Lemmatisation (2:Lemm) and compound splitting (3:Comp) improved recall, whereas an inclusion of a match to SNOMED CT terms for qualifiers and persons (4:Qual) slightly improved precision.

Nr.	Precision (95% CI)	Recall (95% CI)	F-Score
1: Base	0.11 (± 0.14)	0.01 (± 0.01)	0.01
2: Lemm	0.09 (± 0.12)	0.01 (± 0.01)	0.01
3: Stop	0.41 (± 0.04)	**0.79** (± 0.05)	0.54
4: Qual	**0.73** (± 0.05)	0.77 (± 0.05)	0.75
5: Leve	0.72 (± 0.05)	0.78 (± 0.05)	0.75
6: Perm	0.73 (± 0.05)	0.77 (± 0.05)	0.75
7: Comp	0.6 (± 0.05)	0.78 (± 0.05)	0.68
9: MeSH	0.74 (± 0.05)	0.80 (± 0.05)	0.76
11: Abbr	0.74 (± 0.05)	0.80 (± 0.05)	**0.77**

Table 8.13: Results for the semantic category **body structure**. Stop word filtering (3:Stop) improved recall considerably, whereas a match to SNOMED CT terms for qualifiers and persons (4:Qual) improved precision. The best F-score was obtained for ’11:Abbr’.

For disorder, preprocessing had no or little effect, but the inclusion of additional terminologies improved recall. That the inclusion of additional terminologies increased recall, shows that there are terms for disorders in Swedish clinical text that are present in other terminologies, but that are not included in SNOMED CT.

For clinical findings, lemmatisation and compound splitting improved recall, whereas an inclusion of a match to SNOMED CT terms for qualifiers and persons slightly improved precision. Stop word filtering improved recall considerably for the recognition of body structures and a match to SNOMED CT terms for qualifiers and persons further improved precision.

The average number of tokens that were annotated for expressions of clinical entities varied between different semantic categories. Body structures were almost exclusively annotated as one-token expressions, whereas disorders sometimes were two-token expressions and findings sometimes even longer, as shown in Figures 8.1 and 8.2. No entities containing more than two tokens were recognised by the constructed system. It was also shown that the proportion of entities containing abbreviations were higher among false negatives than for correctly recognised entities.

From the study, it could be concluded that preprocessing, together with the inclusion of additional terminologies, resulted in improved results compared to the baseline. The best average results were achieved when all terminologies were used together. The entity body structure, which was most affected by preprocessing, was then recognised with a precision of 0.74 and a recall of 0.80. Lower results, a precision of 0.75 and a recall of 0.55, were achieved for recognition of disorders, and even lower results were achieved for finding, a precision of 0.57 and a recall of 0.30. Low recall for disorders and findings shows both that additional methods are needed for entity recognition and that there are many expressions in clinical text that are not included in SNOMED CT.

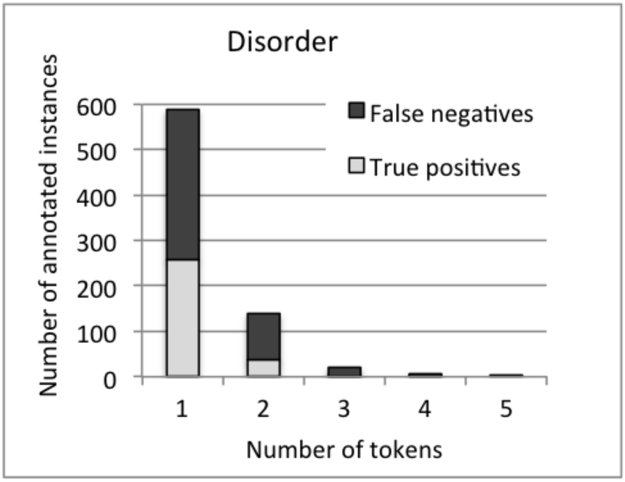

Figure 8.1: Distribution of the number of tokens for annotated **disorders**, divided into true positives and false negatives (for '4:Qual'). No disorders longer than two tokens were recognised by the constructed rule-based system. The bars show the number of annotated entities and the light grey part of the bars shows the number of entities that were recognised by the system.

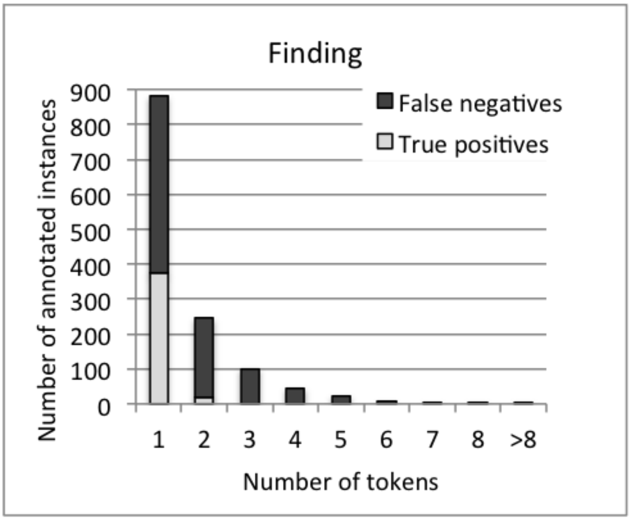

Figure 8.2: Distribution of the number of tokens for annotated **findings**, divided into true positives and false negatives (for '4:Qual'). No findings longer than two tokens were recognised by the constructed rule-based system. The bars show the number of annotated entities and the light grey part of the bars shows the number of entities that were recognised by the system.

Chapter 9

Discussion

In this licentiate thesis, rule- and terminology-based methods for retrieving disorders, findings and body structures in Swedish clinical text have been explored, as well as rule-based methods for negation detection. A comparison of the results to previous studies is carried out in this chapter, and it is also discussed to what extent constructed systems can be used in suggested applications. The chapter also includes future directions, main contributions and general conclusions.

9.1 Comparison with other studies

There are other studies of named entity recognition of clinical entities and of negation detection, as described in the background, some of which it is possible to compare with the present study.

9.1.1 Named entity recognition of clinical entities

The best results from the present study were obtained for recognition of the entity 'body structure' which received a maximum F-score of 0.77. The entity 'disorder' showed the second best results, with an F-score of 0.63, and the lowest results were obtained for recognition of entities of the category ’finding’, with a maximum F-score of 0.41. The best average F-score, 0.60, was obtained for the last experiment, in which all terminologies and some of the preprocessing techniques were used.

There are two comparable studies of rule-based named entity recognition of clinical entities in English text, a study by Savova *et al.* (2010) and a study by Wang (2009). Savova *et al.* (2010) achieved an F-score of 0.72 for rule-based named entity recognition of ’disorder’ in English clinical text, a somewhat better result than those presented here. For the baseline rule- and terminology-based named entity recognition, constructed by Wang (2009), only the average results of several entities are presented, and they show an average F-score of 0.64, slightly higher than the average results obtained here. The average results obtained by Wang (2009) are, however, not entirely comparable, as a wider range of entities were studied than the three entities ’disorder’, ’finding’ and ’body structure’. For two studies of machine learning-based systems for recognising clinical entities in the same data, performed by Wang (2009) and by Wang and Patrick (2009), the results are presented separately for each of the ten entity categories. The category ’body structure’ had a maximum F-score of 0.71 and the category ’clinical finding’, which corresponds to both ’disorder’ and ’finding’ in the present study, had a maximum F-score of 0.83. The rule-based system constructed here was thus more successful in recognising entities of the category ’body structure’, whereas the machine-learning system achieved much better results for ’clinical findings’. One reason for the difference for ’body structure’ might be that the study by Wang and Patrick (2009) allowed nested annotations and that ’body structure’ therefore sometimes appeared inside a nested entity of another category, making it more difficult to recognise.

A study of a rule- and terminology-based system based on MeSH for recognising diseases in Swedish clinical text is the most comparable study (Kokkinakis and Thurin, 2007). This system achieved an F-score of 0.92 for recognising diseases in Swedish discharge summaries, thus a much better result than those presented here. One reason for the large difference could be that more formal language is used in discharge summaries than in the type of clinical text used in the present study.

9.1.2 Negation detection

The Swedish version of NegEx showed a precision of 0.75 and a recall of 0.82 for sentences containing a negation cue and a negative predictive value of 0.97 for sentences without negation cues as shown in Table 8.5. The English NegEx had similar results for recall and for negative predictive value, whereas the English NegEx obtained a higher precision, a value of 0.845. The lower precision for the Swedish version can partly be explained by the fact that not all of the translated negation cues were suitable as negation cues for Swedish. When one of these cues was removed and disambiguation rules were added for the other, the precision increased to 0.78. That the precision was still lower for the Swedish version might, like the differences in the results for named entity recognition obtained by Kokkinakis and Thurin (2007), be attributed to the fact that the English version was evaluated on discharge summaries whereas the Swedish version was evaluated on assessment fields.

The machine-learning-based study of factuality levels in Swedish clinical text conducted by Velupillai (2011) was, however, carried out with clinical text from assessment fields as the reference standard. The factuality level 'certainly negative' of that study is very close to the definition of negated findings that was used for evaluating the Swedish NegEx. The Swedish adaption of NegEx was, like the original English system, evaluated by dividing the reference standard into two groups of sentences, sentences with negation cues and sentences without. The precision for the group containing sentences with negation cues in the NegEx study can still be compared, however, with the precision obtained in the study by Velupillai, as neither true positives nor false positives can be included in the group without negation cues. The precision is slightly higher for the system constructed by Velupillai, a precision of 0.79 for certainly negative compared with 0.75 for the Swedish NegEx.

For a recall, on the other hand, the results of the two studies cannot be compared, as the number of false negatives in a random sample is unknown for the Swedish NegEx. Such a comparison would be made possible by running NegEx on the reference standard that was used for evaluating the machine-learning system constructed by Velupillai.

9.2 Mapping the results to suggested applications

Three different application areas of the systems constructed for this licentiate thesis were suggested in the introduction; clinical text mining, extending and evaluating medical terminologies, and tools for health personnel to use in daily patient care. The implications of the results for these three application areas are discussed in the following paragraphs.

9.2.1 Clinical text mining

For some types of text mining, e.g. some types of automatic generation of hypotheses for clinical research, the constructed systems could probably be used. Low recall of the named entity recognition system and low precision of the negation detection system result in fewer extracted clinical entities, and in turn means that fewer potential hypotheses are generated. A named entity recognition system with low precision and a negation detection system with low recall, on the other hand, result in the generation of incorrect hypotheses. Also a system that does not generate all possible hypotheses and that occasionally generates false hypotheses can be useful, especially when applied to the large amount of clinical text that is contained in the Stockholm EPR Corpus it is likely for instance that limitations in the negation detection system only have small effects on the general performance.

Study III, in which the NegEx system that was constructed in Study II, was applied to a large subset of the Stockholm EPR corpus, is an example of how constructed tools, when applied to a large text set, can be useful for information extraction even if they do not achieve perfect precision and recall. In the study, it could be concluded that 'ischemia' is much more frequently negated than the average disorder, whereas 'atrial fibrillation' and 'hypertension' are less frequently negated. These results, which were obtained by automatic methods, are consistent with frequencies of negated disorders in manually annotated clinical data (Velupillai, 2011)[1].

[1] This annotated corpus, which was created for studying automatic classification of factuality levels, contained annotations for six different levels of factuality.

9.2.2 Extending and evaluating terminologies

In Study IV, the constructed named entity recognition system was used for evaluation of SNOMED CT. There is of course room for improvements in the constructed rules for string matching, and more focus could be put on evaluating what kind of annotated expressions that are not present in the terminologies and to what extent these expressions ought to be included. It was concluded, however, that as there was an increase in recall with the inclusion of other terminologies than SNOMED CT, there are still expressions for the studied entities that occur in clinical language that are not included in SNOMED CT.

A deeper analysis of terms that are not included in the terminologies could also form the basis for expanding these terminologies. The suggested additions to the terminologies would be limited to manually annotated expressions, however, as no methods for recognising clinical entities apart from methods originating in existing terminologies were explored in the constructed systems. Therefore, more advanced methods are needed for using named entity recognition for expanding terminologies. Perfect precision, however, is not required for these methods, as incorrect suggestions for new terms to be added to the terminology can be manually filtered out. Also, a list of suggestions of terms to add to the terminology is useful even if it is not complete, and therefore perfect recall is not required either.

9.2.3 Tools to use in daily patient care

There are very high demands on accuracy of a system that automatically extracts structured data from the unstructured free text and presents it to health personnel; for instance, a system that summarises all disorders and findings in a patient's history. That is, this system might require a named entity recognition system with almost perfect recall and acceptable precision, as well as a negation detection component with almost perfect precision and acceptable recall. It could be argued that such a system ought to be an improvement on a system that does not provide any kind of patient overview, even a system with the low recall presented in Study IV, as such a system would probably provide more information of the patient history than can be gathered by manually skimming through the health record text. The implications of the low recall, however, would probably be difficult to communicate to users and they might therefore incorrectly perceive the presented information as always covering the whole patient history and therefore rely too much on it.

The demands on applications facilitating the input of patient documentation are probably lower, as there is a possibility of a manual review of the completeness and accuracy of the information before it is added to the health record. Therefore, the performance of the constructed negation detection system is likely to be high enough for an application that automatically generates a problem list for example. For extracting findings and disorders, however, the recall presented in Study IV is likely to be too low for such a system to be useful.

When it comes to computer-assisted ICD-10 coding, on the other hand, the negation detection system constructed in Study II, together with the detection of SNOMED CT disorders and findings that was used in Study III, was shown to improve the automatic suggestion of ICD-10 codes when applied to a large data set. Whether the mentioned SNOMED terms were negated or not was taken into account when the model on which the similarity calculations were based was constructed. As the method used for detecting findings and disorders had very low recall, as was shown in Study III, the authors hypothesised that the effect of negation detection might be larger if a better method for detecting clinical entities was used. (Henriksson and Hassel, 2011)

Such a method is evaluated in Study IV, although recall is still low for this method, it would probably be useful for improving a system for computer-assisted ICD-10 coding.

9.3 Future directions

As stated above, no comparison of the performance on the same corpus by machine-learning methods (Velupillai, 2011) and the adaption of NegEx has been has been carried out. Such a comparison, as well as an extension of the rule-based system, for instance through including additional cues or incorporating information from a parser, could be future research topics.

As regards machine learning-based named entity recognition of clinical entities, no work has yet been conducted for Swedish clinical text. Future work will therefore focus on investigating to what extent disorders, findings and body structures can be automatically detected by means of machine-learning methods. The Stockholm EPR Clinical Entity Corpus needs to be further developed in order to achieve this. No studies of inter-annotator agreement to evaluate the quality of the annotations of the corpus have been performed, and such evaluation is therefore a topic for future studies.

There are studies outside the domain of clinical text that exploit large corpora for weak supervised learning of named entity recognition (Niu *et al.*, 2003; Nadeau, 2007). As the Stockholm EPR Corpus is a large corpus, such methods might also be used.

Another possible future direction is to implement and evaluate some of the above-mentioned applications, such as using the constructed systems for generating hypotheses.

9.4 Main contributions

The main contribution of this licentiate thesis is that two useful systems for working with clinical text written in Swedish have been constructed and evaluated:

- A system for negation detection, that was adapted from English (Study II).
- A system for extracting disorders, findings and body structures mentioned in Swedish clinical text (Study IV).

The constructed systems can for instance be used for hypothesis generation, terminology evaluation and improvement of computer-assisted ICD-10 coding. The systems can also be useful for generating features that can be used as input to machine learning systems.

The two systems were constructed and evaluated separately, but have now been combined into one system, performing named entity recognition of disorders, findings and body structures as well as negation detection.[2]

[2] The program, which is still under development, is available here: http://people.dsv.su.se/~mariask/resources.html.

The Swedish negation detection system is also a demonstration of how a system developed for English can be adapted for another, grammatically similar language.

Apart from the two systems, the annotated Stockholm EPR Clinical Entity Corpus is a resource that can be used for future research.

Another contribution is that an evaluation of the coverage of the Swedish translation of SNOMED CT in clinical text from an emergency care unit has been carried out, showing that there are entities that occur in clinical language but that are not yet included in SNOMED CT.

9.5 General conclusions

In this licentiate thesis, it has been investigated to what extent rule-based methods and the currently available medical terminologies for Swedish can be used for automatically extracting mentioned findings and disorders in Swedish clinical text, as well as for determining whether these are negated or not.

The results from the evaluation of the named entity recognition system showed that disorders and findings were recognised with low recall, whereas body structures were recognised with comparatively good results. The constructed named entity recognition has been used for evaluating the coverage of SNOMED CT and could probably also be used for e.g. hypothesis generation.

The negation detection system that was adapted to Swedish showed somewhat lower results than the English version, but the system is still accurate enough to be useful in some applications. It has been used for improving computer-assisted ICD-10 coding (Henriksson and Hassel, 2011) as well as for estimating the frequency of negated findings in a large corpus of clinical text.

The constructed systems need to be further developed, especially when it comes to automatically recognising disorders and findings. As similar systems for recognising clinical findings in English text have been based on machine learning (Wang and Patrick, 2009), the plan for future studies is to apply the same approach using Swedish annotated clinical text. The results from the rule-based systems that have been constructed for this licentiate thesis will then be used for generating features which the machine-learning methods can utilise when learning to recognise disorders and findings automatically.

References

Helen Allvin, Elin Carlsson, Hercules Dalianis, Riitta Danielsson-Ojala, Vidas Daudaravicius, Martin Hassel, Dimitrios Kokkinakis, Helja Lundgren-Laine, Gunnar H Nilsson, Oystein Nytro, Sanna Salantera, Maria Skeppstedt, Hanna Suominen, and Sumithra Velupillai. 2011. Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. *J Biomed Semantics*, 2 Suppl 3: S1. ISSN 2041-1480 (Electronic).

Ethem Alpaydin. 2010. *Introduction to Machine Learning*, 2. ed. edition. MIT Press, Cambridge, MA. ISBN 978-0-262-01243-0 (hardcover : alk. paper).

Douglas G Altman and J Martin Bland. 1994. Statistics notes: Diagnostic tests 2: predictive values. *BMJ*, 308(896).

Göran Andersson. 1968. *Repetitorium i statistik. D.1, Teori och formler (In Swedish).*

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, and Kazuhiko Ohe. 2009. Text2table: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the BioNLP 2009 Workshop*, pages 185–192, Boulder, Colorado, June 2009. Association for Computational Linguistics.

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of AMIA Annual Symposium*, pages 17–21.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596.

Gunnar Blom. 1989. *Sannolikhetsteori och statistikteori med tillämpningar*, 4. uppl. edition. Studentlitteratur, Lund. ISBN 91-44-03594-2 ;.

Michael J. Campbell, David Machin, and Stephen J. Walters. 2007. *Medical statistics : A textbook for the health sciences*, 4. ed. edition. Wiley, Chichester. ISBN 978-0-470-02519-2.

Hui Cao, Marianthi Markatou, Genevieve B Melton, Michael F Chiang, and George Hripcsak. 2005. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. *AMIA Annu Symp Proc*, pages 106–110. ISSN 1942-597X (Electronic); 1559-4076 (Linking).

Johan Carlberger and Viggo Kann. 1999. Implementing an efficient part-of-speech tagger. *Software–Practice and Experience*, 29:815–832.

Staffan Cederblom. 2005. *Medicinska förkortningar och akronymer (In Swedish)*. Studentlitteratur, Lund. ISBN 91-44-03393-1.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*, 34(5):301–310. ISSN 1532-0464 (Print).

Wendy W Chapman, Lee M Christensen, Michael M Wagner, Peter J Haug, Oleg Ivanov, John N Dowling, and Robert T Olszewski. 2005. Classifying free-text triage chief complaints into syndromic categories with natural language processing. *Artif Intell Med*, 33(1):31–40. ISSN 0933-3657 (Print); 0933-3657 (Linking).

Wendy W Chapman, John Dowling, and David Chu. 2007. Context: An algorithm for identifying contextual features from clinical text. In *Biological, translational, and clinical language processing*, pages 81–88, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

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

Hercules Dalianis and Sumithra Velupillai. 2010. How certain are clinical assessments? Annotating Swedish clinical text for (un)certainties, speculations and negations. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10) 19-21 May 2010; Valletta, Malta*. European Language Resources Association (ELRA).

Peter L Elkin, Steven H Brown, Brent A Bauer, Casey S Husser, William Carruth, Larry R Bergstrom, and Dietlind L Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC Med Inform Decis Mak*, 5:13+. ISSN 1472-6947 (Electronic).

Carol Friedman. 2005. Semantic text parsing for patient records. In Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh, editors, *Medical informatics : knowledge management and data mining in biomedicine*. Springer Science+Business Media, Inc., Boston, MA. ISBN 978-0-387-25739-6.

Carol Friedman and George Hripcsak. 1998. Evaluating natural language in processors in the clinical domain. *Methods of Information in Medicine*, 37:334–344.

Carol Friedman, Lyuda Shagina, Yves Lussier, and George Hripcsak. 2004. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5).

Ilya M. Goldin and Wendy W. Chapman. 2003. Learning to detect negation with 'not' in medical texts. In *Workshop at the 26th ACM SIGIR Conference*.

Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In David Losada and Juan Fernández-Luna, editors, *Advances in Information Retrieval*, volume 3408 of *Lecture Notes in Computer Science*, pages 345–359. Springer Berlin / Heidelberg. ISBN 978-3-540-25295-5.

Catalina Hallett, Richard Power, and Donia Scott. 2006. Summarisation and visualisation of e-health data repositories. In *UK E-Science All-Hands Meeting*, Nottingham, UK.

Aron Henriksson and Martin Hassel. 2011. Exploiting Structured Data, Negation Detection and SNOMED CT Terms in a Random Indexing Approach to Clinical Coding. In *Proceedings of Workshop on Biomedical Natural Language Processing*, Hissar, Bulgaria, September 15 2011.

Yang Huang and Henry J Lowe. 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. *J Am Med Inform Assoc*, 14(3): 304–311. ISSN 1067-5027 (Print).

Yang Huang, Henry J. Lowe, and William R. Hersh. 2003. A pilot study of contextual umls indexing to improve the precision of concept-based representation in xml-structured clinical radiology reports. *Journal of the American Medical Informatics Association*, 10(6):580 – 587. ISSN 1067-5027.

International Health Terminology Standards Development Organisation, IHTSDO. 2008a. SNOMED Clinical Terms User Guide, July 2008 International Release. http://www.ihtsdo.org. Accessed 2011-01-24.

International Health Terminology Standards Development Organisation, IHTSDO. 2008b. SNOMED CT style guide: Body structures – anatomy. http://www.ihtsdo.org. Accessed 2012-04-13.

International Health Terminology Standards Development Organisation, IHTSDO. 2008c. SNOMED CT Style Guide: Clinical Findings. Accessed 2011-01-24.

Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. 2011. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*.

Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, second edition. Prentice Hall. ISBN 013122798X.

Karolinska Institutet. 2012. Hur man använder den svenska MeSHen (In Swedish, translated as: How to use the Swedish MeSH). http://mesh.kib.ki.se/swemesh/manual_se.html. Accessed 2012-03-10.

Karin Kipper-Schuler, Vinod Kaggal, James J. Masanz, Philip V. Ogren, and Guergana K. Savova. 2008. System evaluation on a named entity corpus from clinical notes. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008.

Dimitrios Kokkinakis. 2011. Evaluating the coverage of three controlled health vocabularies with focus on findings, signs and symptoms. In NEALT Proceedings Series, editor, *NODALIDA*, volume 12, pages 27–31.

Dimitrios Kokkinakis and Anders Thurin. 2007. Identification of entity references in hospital discharge letters. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA)*, Estonia.

Human Language Technology Group KTH. 2012a. Granska Tagger: A part-of-speech tagger for Swedish. http://www.csc.kth.se/tcs/humanlang/tools.html. Accessed 2012-04-01.

Human Language Technology Group KTH. 2012b. Inflector: A simple word inflector for Swedish. http://www.csc.kth.se/tcs/humanlang/tools.html. Accessed 2012-04-01.

Taku Kudo. 2012. CRF++: Yet Another CRF toolkit. http://crfpp.sourceforge.net/. Accessed 2012-03-31.

Maria Kvist, Maria Skeppstedt, Sumithra Velupillai, and Hercules Dalianis. 2011. Modeling human comprehension of Swedish medical records for intelligent access and summarization systems - future vision, a physician's perspective. In *Proceedings of SHI 2011, Scandinavian Health Informatics meeting*.

William Long. 2005. Extracting diagnoses from discharge summaries. In *AMIA Annual Symp Proc*, pages 470–474.

Genevieve B Melton and George Hripcsak. 2005. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc*, 12(4):448–457. ISSN 1067-5027 (Print); 1067-5027 (Linking).

Stephane Meystre and Peter J Haug. 2006. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *J Biomed Inform*, 39(6):589–599. ISSN 1532-0480 (Electronic); 1532-0464 (Linking).

Stephane M. Meystre, Guergana K. Savova, Karin C. Kipper-Schuler, and John F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, pages 128–144. ISSN 0943-4747 (Print); 0943-4747 (Linking).

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roser Morante. 2010. Descriptive analysis of negation cues in biomedical texts. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Roser Morante and Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 21–29, Morristown, NJ, USA. Association for Computational Linguistics.

Pradeep G Mutalik, Aniruddha Deshpande, and Prakash M Nadkarni. 2001. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the umls. *J Am Med Inform Assoc*, 8(6): 598–609. ISSN 1067-5027 (Print); 1067-5027 (Linking).

David Nadeau. 2007. *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. PhD thesis, University of Ottawa.

NegEx. 2009. Negex version 2. http://www.dbmi.pitt.edu/chapman/negex.html. Accessed 2009-12-01.

Paul Newbold, William L. Carlson, and Betty Thorne. 2003. *Statistics for business and economics*, 5. ed. edition. Prentice-Hall, Upper Saddle River, N. J. ISBN 0-13-029320-2.

Inga Nilsson. 2007. *Medicinsk dokumentation genom tiderna : En studie av den svenska patientjournalens utveckling under 1700-talet, 1800-talet och 1900-talet*. Enheten för medicinens historia, Medicinska fakulteten, Lunds universitet, Lund. ISBN 978-91-633-1987-7.

Cheng Niu, Wei Li, Jihong Ding, and Rohini K. Srihari. 2003. A bootstrapping approach to named entity classification using successive learners. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 335–342.

Philip Ogren, Guergana Savova, and Christopher Chute. 2008. Constructing evaluation corpora for automated clinical named entity recognition. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0.

Philip V. Ogren. 2006. Knowtator: A protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, Morristown, NJ, USA. Association for Computational Linguistics.

Jon Patrick, Yefeng Wang, and Peter Budd. 2007. An automated system for conversion of clinical notes into SNOMED clinical terminology. In *Proceedings of the fifth Australasian symposium on ACSW frontiers - Volume 68*, ACSW '07, pages 219–226, Darlinghurst, Australia, Australia. Australian Computer Society, Inc. ISBN 1-920-68285-X.

Ken Peffers, Tuure Tuunanen, Marcus Rothenberger, and Samir Chatterjee. 2008. Journal of management information systems. *A Design Science Research Methodology for Information Systems Research.*

Håkan Petersson, Gunnar Nilsson, Lars-Erik Strender, and Hans Åhlfeldt. 2001. The connection between terms used in medical records and coding system: A study on Swedish primary health care data. *Med Inform Internet Med*, 26(2): 87–99. ISSN 1463-9238 (Print).

Catherine Plaisant, Richard Mushlin, Aaron Snyder, Jia Li, Dan Heller, Ben Shneiderman, and Kaiser Permanente Colorado. 1998. Lifelines: Using visualization to enhance navigation and analysis of patient records. In *In Proceedings of the 1998 American Medical Informatic Association Annual Fall Symposium*, pages 76–80.

Lior Rokach, Roni Romano, and Oded Maimon. 2008. Negation recognition in medical narrative reports. *Information Retrieval*, 11(6):499–538.

Francisco S. Roque, Peter B. Jensen, Henriette Schmock, Marlene Dalgaard, Massimo Andreatta, Thomas Hansen, Karen Søeby, Søren Bredkjær, Anders Juul, Thomas Werge, Lars J. Jensen, and Søren Brunak. 2011. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol*, 7(8):e1002141.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5):507–513. ISSN 1527-974X (Electronic); 1067-5027 (Linking).

Jonas Sjöbergh and Viggo Kann. 2004. Finding the correct interpretation of Swedish compounds a statistical approach. In *Proceedings of LREC-2004*, pages 899–902, Lisbon, Portugal.

Socialstyrelsen. 2011. Språkliga riktlinjer för översättningen av Snomed CT finns nu att beställa, (In Swedish, translated as: Linguistic guidelines for the translation of SNOMED CT can now be ordered). http://www.socialstyrelsen.se/nyheter/2011mars/.

Socialstyrelsen. 2012. Diagnoskoder (ICD-10). http://www.socialstyrelsen.se/klassificeringochkoder/diagnoskoder. Accessed 2012-03-31.

NLP Group Stanford. 2012. Stanford Named Entity Recognizer (NER). http://www-nlp.stanford.edu/software/CRF-NER.shtml. Accessed 2012-03-29.

Michael Q. Stearns, Colin Price, Kent A. Spackman, and Amy Y Wang. 2001. Snomed clinical terms: Overview of the development process and project status. In *Proceedings of the AMIA Symposium*, pages 662–666. American Medical Informatics Association.

Charles Sutton and Andrew McCallum. 2010. An introduction to conditional random fields. cite arxiv:1011.4088.

Hideyuki Tanushi, Hercules Dalianis, and Gunnar Nilsson. 2011. Calculating prevalence of comorbidity and comorbidity combinations with diabetes in hospital care in Sweden using a health care record database. In Hans Moen Øystein Nytrø, Laura Slaughter, editor, *LOUHI, Third International Workshop on Health Document Text Mining and Information Analysis*.

Sumithra Velupillai. 2011. Automatic Classification of Factuality Levels – A Case Study on Swedish Diagnoses and the Impact of Local Context. In *Proc. The Fourth International Symposium on Languages in Biology and Medicine – LBM 2011*, Singapore, December 2011.

Sumithra Velupillai, Hercules Dalianis, and Maria Kvist. 2011. Factuality Levels of Diagnoses in Swedish Clinical Text. In A. Moen, S. K. Andersen, J. Aarts, and P. Hurlen, editors, *Proc. XXIII International Conference of the European Federation for Medical Informatics (User Centred Networked Health Care)*, pages 559 – 563, Oslo, August 2011. IOS Press.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra1, and János Csirik. 2008. The BioScope Corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Ellen M. Voorhees. 2002. The philosophy of information retrieval evaluation. In *Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, CLEF '01, pages 355–370, London, UK. Springer-Verlag. ISBN 3-540-44042-9.

Yefeng Wang. 2009. Annotating and recognising named entities in clinical notes. In *Proceedings of the ACL-IJCNLP Student Research Workshop*, pages 18–26, Singapore.

Yefeng Wang and Jon Patrick. 2009. Cascading classifiers for named entity recognition in clinical notes. In *Proceedings of the Workshop on Biomedical Information Extraction*, pages 42–49.

WHO. 2012. WHO international classification of diseases (ICD). http://www.who.int/classifications/icd/en/. Accessed 2012-02-04.

Wikipedia. 2012. Projekt medicin/lista över sjukdomar (In Swedish, translated as Project drugs/ listing of deseases in Swedish. http://sv.wikipedia.org/w/index.php?title= Wikipedia:Projekt_medicin/Lista_över_sjukdomar&oldid =15872672. Accessed 2012-02-17, 13:14.

Qinghua Zou, Wesley W. Chu, Craig Morioka, Gregory H. Leazer, and Hooshang Kangarloo. 2003. Indexfinder: A method of extracting key concepts from clinical texts for indexing. In *Proceedings of AMIA Annual Symposium*, pages 763–767.