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

Maria Skeppstedt[1], Maria Kvist[1,2], Hercules Dalianis[1]

[1]Dept. of Computer and Systems Sciences (DSV), Stockholm University, Forum 100, 164 40 Kista, Sweden
[2]Dept. of clinical immunology and transfusion medicine, Karolinska University Hospital, 171 76 Stockholm, Sweden

mariask@dsv.su.se, maria.kvist@karolinska.se, hercules@dsv.su.se

Abstract

Named entity recognition of the clinical entities disorders, findings and body structures is needed for information extraction from unstructured text in health records. Clinical notes from a Swedish emergency unit were annotated and used for evaluating a rule- and terminology-based entity recognition system. This system used different preprocessing techniques for matching terms to SNOMED CT, and, one by one, four other terminologies were added. For the class body structure, the results improved with preprocessing, whereas only small improvements were shown for the classes disorder and finding. The best average results were achieved when all terminologies were used together. The entity body structure was recognised with a precision of 0.74 and a recall of 0.80, whereas lower results were achieved for disorder (precision: 0.75, recall: 0.55) and for finding (precision: 0.57, recall: 0.30). The proportion of entities containing abbreviations were higher for false negatives than for correctly recognised entities, and no entities containing more than two tokens were recognised by the system. 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.

Keywords: Electronic patient records, Swedish, SNOMED CT, named entity recognition

1. Introduction

Health records contain valuable information that can be used for improvement of the immediate care of patients and for medical research. A considerable amount of information is stored in an unstructured format as free text, and the knowledge contained in this text is not readily available for automated analysis such as text mining and decision support systems. In order to make use of the information locked up in the free text in medical records, methods for extracting relevant entities from it are needed. The semantic classes *disorder*, *finding* and *body structure* are examples of such clinical entities that are significant for the medical history of a patient.

There are several studies in which clinical entities have been extracted from health record text through comparison with different medical terminologies. One of the resources that have been used for this purpose is the medical terminology SNOMED CT. This terminology has recently been translated into Swedish, which opens up the possibility for SNOMED CT-based extraction of clinical entities from Swedish text. However, the natural language of the free text sections of patient records does not follow the precise expressions in terminologies, and is typically written with abbreviations, misspellings and medical jargon. Therefore, it is worth investigating the extent to which SNOMED CT can be used as a resource for automatic retrieval of clinical entities from free text sections of health records, as well as the extent to which expressions that are used in daily clinical practice are included in SNOMED CT.

The present study will focus on named entity recognition of the clinical entities *disorder*, *finding* and *body structure* and has three main aims:

- To investigate to what extent it is possible to automatically recognise these three types of entities using existing medical terminologies, and whether different techniques for preprocessing the text and the terminologies affect the results.
- To examine the coverage of SNOMED CT on Swedish clinical text, that is the extent to which entities that are used in daily clinical practice correspond to how entities are expressed in SNOMED CT.
- To explore how abbreviations and number of tokens in clinical entities influence entity recognition.

2. Related research

2.1. Recognition of clinical entities

Most studies on retrieval of entities from clinical text have been performed with the intent of mapping terms in text to specific concept codes in a terminology. One example is the MetaMap program (Aronson, 2001), which discovers UMLS concepts in biomedical and clinical text through matching phrases to terms in the UMLS metathesaurus. MetaMap uses techniques such as parsing to filter out relevant phrases, and tools that generate inflections and spelling variants. Studies of matches to specific concepts within UMLS have also been carried out by for exampel Zou et al. (2003), Huang et al. (2003) and by Friedman et al. (2004). Long (2005) and Patrick et al. (2007) have focused on concepts belonging to SNOMED CT and they use different techniques for finding abbreviations, misspellings, inflections and different word orders when matching clinical text to SNOMED CT.

The focus of the present study is to carry out named entity recognition of clinical entities, thus to retrieve instances of a certain type, not to match to an exact concept. There are

several previous studies on named entity recognition in English clinical text. A rule- and lexicon-based approach for named entity recognition of disorders through a match to SNOMED CT has for example been evaluated by Savova et al. (2010). Their approach, which relies on techniques including spelling correction and generation of word permutations (Kipper-Schuler et al., 2008), resulted in a precision of 0.80 and a recall of 0.65 for an exact match (Savova et al., 2010).

Another example of named entity recognition is described by Wang (2009), who compared rule-based and machine learning methods for detecting ten different classes of clinical concepts, including body structure, finding and qualifier. The clinical notes deal with a variety of areas within the Intensive Care Services. The rule-based approach, which resulted in an average precision of 0.75 and an average recall of 0.52 for exact match, uses a lexical word lookup using word lists such as UMLS, SNOMED and MOBY (standard English dictionary) as well as lists of medical abbreviations. For the machine learning approach, the output of the rule-based system was used as one feature. Wang and Patrick (2009) have later applied other machine learning methods on the same corpus.

Also Jiang et al. (2011) have used machine learning methods for named entity recognition in clinical text, focusing on medical problems, tests and treatments. As in the other described machine learning studies, they used the output from rule-based lexical lookup systems as one feature for their machine learning system.

Studies on entity recognition in clinical text have also been carried out in smaller languages such as Finnish (Suominen et al., 2006) and Swedish (Kokkinakis and Thurin, 2007). The difficulties for these smaller languages lie in limited terminologies and fewer language-specific natural language processing tools. The Swedish named entity recognition system, which used the MeSH terminology, achieved a precision of 0.98 and a recall of 0.87 for recognition of diseases in discharge summaries.

In order to develop and evaluate named entity recognition systems, annotated corpora are needed. The process of annotating entities in clinical text has been described by Wang (2009) and by Chapman et al. (2008). Both studies achieved an F-score of around 0.9 for inter-annotator agreement. Annotation of clinical entities is also described by Ogren et al. (2008).

There is a previous study on Swedish clinical text, in which clinical findings are annotated in patient records from the Stockholm EPR Corpus with the aim of assembling a list of diagnoses (Velupillai et al., 2011). These annotations have been used for an initial study of recognition of clinical findings using SNOMED CT, and since the focus lied on maximising the precision, the results were a recall of 0.13 and a precision of 0.80 (Skeppstedt et al., 2011).

2.2. Evaluation of terminologies

Automatic evaluation of coverage of the English SNOMED CT has for example been studied by Penz et al. (2004), who obtained a coverage of around 90% when automatically matching the content of problem list entries for various types of clinical domains to SNOMED CT.

For Swedish, the coverage of SNOMED CT has been automatically evaluated on free text from a scientific corpus (Kokkinakis, 2011b) and from public health portals (Kokkinakis, 2011a). In the scientific medical corpus the occurrence of SNOMED CT terms was studied and only 6.3% of all SNOMED CT terms were found in the corpus when direct match was applied. The study of text on public health portals focused on findings, signs and symptoms and the coverage of subsets describing findings in three different terminologies; MeSH, SNOMED CT and ICD-10. Depending on type of health portal, 32% to 35% of the gathered terms were found in SNOMED CT, 22% to 27% in MeSH and 3.2% to 4.1% in ICD-10.

3. Methods

Clinical text in Swedish patient records was annotated for the clinical entities *disorder*, *finding* and *body structure*. Rule-based lexical lookup, using between one and five different terminologies, was thereafter applied to recognise terms of these categories in the texts, and the resulting matches were evaluated using the annotated data as a gold standard. Different kinds of preprocessing of the text and/or the terminology were carried out to increase matching.

3.1. Terminologies

The following terminologies were used in the study:

SNOMED CT aims at providing a standardised terminology for clinical information and consists of medical concepts that are organised into hierarchies. Each concept in the terminology has a fully specified name, which also includes a semantic tag that indicates a semantic class that the concept belongs to. Examples of semantic classes are *disorder*, *finding*, *body structure*, *qualifier value* and *person*. (IHTSDO, 2008a) SNOMED CT has been translated into Swedish by the Swedish National Board of Health and Welfare (Socialstyrelsen). The translation that was used for the study described here was released in July 2011. This version contains around 280,000 clinical terms and, unlike its English counterpart, does not contain any synonyms. (Socialstyrelsen, 2011)

ICD-10 is an international standard classification of diseases, managed by WHO (WHO, 2012). The names of the Swedish translation of the ICD-10 diagnosis codes were used as one of the terminologies in this study.

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

Wikipedia: Projekt medicin contains 384 names of diseases in Swedish, often expressed in more general terms than in the other resources. (Wikipedia, 2012)

Medical abbreviations and acronyms from the second part of a book about Swedish medical abbreviations

and acronyms (Cederblom, 2005). This list contained 2,134 abbreviations.

3.2. Annotation of clinical texts

The free text sections in the assessment part of randomly chosen clinical notes from an emergency unit of internal medicine at Karolinska University Hospital were used in the study.[1] The chosen texts were all part of the Stockholm EPR Corpus (Dalianis et al., 2009), which contains patient records written in Swedish. The annotation was carried out by one senior physician with previous experience in annotation of clinical texts. A test annotation was first performed to refine the annotation guidelines, and also to give the annotator the chance to become familiar with the task. The text annotation tool Knowtator was used (Ogren, 2006).

The selected entities in this study were the three semantic classes *disorder*, *finding* and *body structure*. Definitions of the studied entities were based on the corresponding SNOMED CT semantic classes found in the SNOMED CT Style Guide (IHTSDO, 2008b), and can be summarised as follows:

Disorders: Diseases or abnormal conditions that are not momentary and that have an underlying pathological process.

Findings: Symptoms reported by the patient, observations made by the physician or results of a medical examination of the patient.

Body structures: Anatomically defined body parts, excluding body fluids and expressions indicating positions on the body.

The primary rule for the choice of annotation class was to annotate words into the class in which they are perceived in the clinical reality described in the text. The same word or expression can therefore be a finding in one case and a disorder in another and it was annotated in accordance with the context. For example, *hypertension* can be observed as a symptom of several disorders, but it is also a disorder in itself with its own underlying pathological process.

Only words in sequence were annotated for each annotation instance, and no nested annotations were allowed. Neither was it allowed to annotate only part of a word, and therefore compound words were annotated as one whole word. The shortest possible expression that still fully described the finding, disorder or body structure was annotated. Qualifiers, such as words indicating negation or level of severity, were excluded.

In the example sentence: *The patient experiences strong stabbing pain in the left knee* (*Patienten känner en kraftig stickande smärta i vänster knä*), the words *strong* and *left* are qualifiers and were therefore not annotated, *stabbing pain* is a finding, and *knee* is a body part.

[1]The study was carried out after approval from the Regional Ethical Review Board in Stockholm, permission number 2009/1742-31/5.

3.3. Rule-based lexical lookup of terms in terminologies

The aim of the rule-based lexical lookup of terms in SNOMED CT, and in some cases other terminologies, was to assign each token in the clinical text one of the three semantic classes *disorder*, *finding* and *body structure*, or the class *O*, meaning outside of an entity.

Some of the preprocessing relied on Granska, which is a part-of-speech tagger for Swedish (Carlberger and Kann, 1999).

The constructed algorithm first used the tokeniser in Granska to tokenise the clinical text, and also to divide the text into sentences through using the part-of-speech tag *major delimiter*. Thereafter, the following was performed for each sentence in the clinical text: First, the full sentence was matched to the selected terminology, and if the full sentence was found as a term in the terminology, each token in the sentence was assigned the semantic class given by the terminology. Thereafter, the sentence was divided into increasingly smaller parts, with increasingly fewer tokens and each of these smaller parts was also, in the same way, matched to the terminology. Therefore, in this first run, a single token could match both disorder, finding and body structure at the same time. This would also be the case if a single token were to match more than one concept in the terminology, for example both a disorder and a finding.

In order to resolve the semantic class for the tokens that had been assigned several semantic classes, priority rules were applied. A match for body structure always had priority over a match for finding and disorder, and a match for disorder had priority over a match for finding. These priority rules were the same priority rules that were used for the annotation task.

The BIO format was used, i.e. besides labelling the words with the semantic class, each word in the text was labelled with a B (beginning of an entity), I (inside an entity) or O (outside an entity).

Different linguistic preprocessing methods were used for the rule-based lexical lookup of terms in SNOMED CT and other terminologies. The following eleven preprocessing experiments were performed:

1: Base A baseline was established by rule-based lexical lookup in the free text of terms from the three SNOMED CT classes *body structure*, *disorder*, and *finding*. An exact match to the terms in SNOMED CT was carried out, without any preprocessing except that all letters were converted to lower case.

2: Lemm The words in the clinical text were lemmatised with Granska and both the original form and the lemmatised form of the word were compared to the SNOMED CT terminology. Matching was therefore considered positive if words were recognised either in the lemmatised or the non-lemmatised form.

3: Stop The same settings as in the previous experiment, except that terms in SNOMED CT belonging to the semantic class *body structure* were stop-word filtered. A list of unique tokens for all terms in the terminology

belonging to that class was created. All tokens that occurred more often than a threshold value were added to the stop-word list. The total number of unique tokens divided by 10 was found to be an optimal threshold value.

4: Qual A token that matched a SNOMED CT term from the semantic classes *qualifier* or *person* in addition to matching a *finding* or *body structure* was assigned the class O (outside an entity). The reason for also including the class *qualifier* was that qualifiers were not included in the annotated entities, whereas they are sometimes included in a SNOMED CT *finding* or *body structure*. The *person* terms were added since some of the terms categorised as SNOMED CT body structures are terms that also could be categorised as belonging to the person class. Apart from this addition, the same settings as in the previous experiment were used.

5: Leve The same settings as in the previous experiment, but in addition generated versions of chunks of the text with a Levenshtein distance of one from what was originally written were compared to the SNOMED CT terms. This was carried out in order to find misspelled SNOMED CT terms in the clinical text.

6: Perm The same preprocessing as in experiment *4: Qual*, but permutations of tokens in the clinical text were also generated and compared to the content of SNOMED CT. Permutations were constructed for text chunks containing a minimum of two tokens and a maximum of five.

7: Comp The same preprocessing as in experiment *4: Qual*, but a compound splitter for Swedish was used (Sjöbergh and Kann, 2004) for all words that contained at least ten letters. If a word was split into smaller parts by the compound splitter, each of these smaller parts was matched to the terminology and if any of them matched a term in the terminology, it was assigned the semantic class of that term. Versions of the included parts of the word with a Levensthein distance of one were also matched to the terminology.

8: ICD10 The same settings as in experiment *4: Qual*, but ICD-10 codes (WHO, 2012) were also added as one terminology to which the clinical text was matched. For assigning the semantic class *disorder*, ICD-10 codes in chapter 1–17 and 19, except codes T357–T629, were used. For assigning the semantic class *finding*, codes in chapter 18 were used. No preprocessing of the ICD-10 code texts was carried out.

9: MeSH The same settings as in the previous experiment, but terms from MeSH were also added. For assigning the semantic class *disorder*, terms in category F03 and category C were used, except terms in category C23, which were used for assigning the semantic class *finding*. For assigning the semantic class *body structure*, terms in the MeSH categories A01–A10 were used. The terms for *body structure* and *disorder* are often expressed in plural in MeSH and were therefore lemmatised with Granska.

10: Wiki The same settings as in the previous experiment, with the addition that terms in the Wikipedia list of diseases were used for assigning the semantic class *disorder*.

11: Abbr Three lists of abbreviations were generated; abbreviations that matched the words in the other terminologies for disorders, abbreviations that matched findings and abbreviations that matched body structures. The same settings as in the previous experiment were used with the addition that these lists of abbreviations were used for capturing abbreviated entities and assigning the corresponding class.

For evaluation, the script from the CoNLL 2000 shared task[2] was applied. This script calculates precision and recall as well as F-score for exact match, and to this a calculation of a 95% confidence interval for precision and recall was added.
An error analysis, mainly focusing on occurrences of abbreviations and number of tokens in annotated instances, was carried out, since it was hypothesised that those features would affect the result.

4. Results

4.1. Properties of the annotated entities

In the evaluation data, which contained 26,011 tokens, a total of 2,342 annotations had been made. The number of annotation instances for each class is shown in Table 1.

Semantic class	Annotated instances
Disorder	759
Finding	1,319
Body structure	264

Table 1: Total number of instances that were used for the evaluation.

4.2. Automatic recognition of the annotated entities

The results for each one of the eleven preprocessing experiments are shown in Tables 2, 3 and 4. Even though the results for each entity type are presented in a separate table, the system always attempted to recognise all three types of entities, which means that the results for one entity might affect the results for another. This is a more realistic usage of the system than to try to recognise each semantic class separately.
The results above the horizontal line in Tables 2–4 show experiments using only the SNOMED CT terminology, whereas the experiments presented under the horizontal line show results with additional terminologies.
The baseline results for *body structure* show exceptionally low recall, since terms for body structures are not independent SNOMED CT terms, but are included in descriptive expressions (e.g. *arm as structure* or *arm as a whole*). Stop word filtering was therefore added to SNOMED CT terms describing *body structure*, which improved the recall.

[2] http://www.cnts.ua.ac.be/conll2000/chunking/

Only results from stop word filtering of the SNOMED CT terms for *body structure* are presented here, since the same technique applied to *disorder* and *finding* resulted in a decreased precision without an improvement of recall.
A number of minimum lengths of words for which to generate Levenshtein distance versions were tested. Only the best results are presented here, which were obtained for a minimum length of 8 letters.

Disorder			
Nr.	Prec. (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 2: Results for the semantic class disorder. Preprocessing had no or little effect, but the inclusion of additional terminologies (*8:ICD10 – 11:Abbr*) improved recall.

Finding			
Nr.	Prec. (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 3: Results for the semantic class 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.

4.3. Error analysis

An error analysis of the false positives and false negatives was carried out for the experiment *4: Qual*. False positives are tokens that the system incorrectly assigned one of the three classes *disorder*, *finding* and *body structure*. False negatives are tokens that were assigned one of the three semantic classes by the annotator but that were not assigned a semantic class by the system, thus the instances that the system failed to match.
The average number of tokens that were annotated for expressions of clinical entities varied between the different

Body structure			
Nr.	Prec. (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 4: Results for the semantic class body structure. Stop word filtering (*3:Stop*) improved recall considerably, whereas an inclusion of a match to SNOMED CT terms for qualifiers and persons (*4:Qual*) improved precision. The best F-score was obtained for *11:Abbr*.

semantic classes, as shown in Table 5. Body structures were almost exclusively annotated as one-token expressions, whereas disorders were annotated as two-token expressions in many cases. For findings, two-token as well as three-token expressions were common, and expressions containing up to 13 tokens existed. The distributions of the number of annotated tokens for disorders and findings, divided into false negatives and true positives, are shown in Figures 1 and 2. No entities containing more than two tokens were recognised by the constructed system.

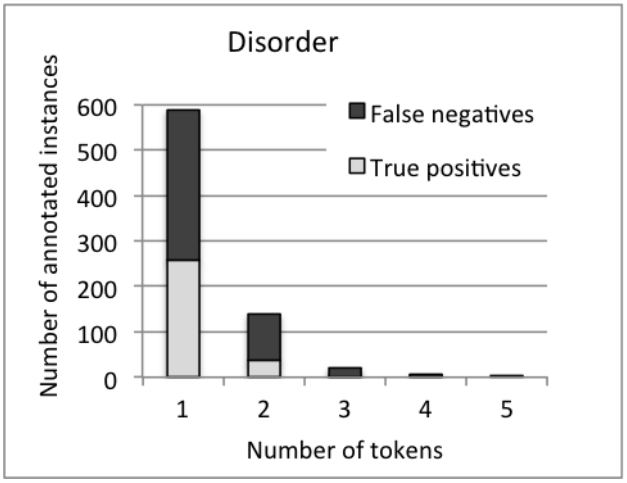

Figure 1: Distribution of the number of tokens for annotated disorders, divided into true positives and false negatives (for *4:Qual*). No disorders longer than two tokens were recognised by the constructed rule-based system.

It could also be concluded from the error analysis that some findings were expressed in a combinations of two or more separate findings (e.g. *ECG and urine sample both OK*). Since the annotation scheme did not allow nested annotations, combined findings were annotated as one entity, which resulted in annotated entities that were not likely to be present in the terminology.
The number of false negatives that contained abbreviations and the number of correctly matched instances that con-

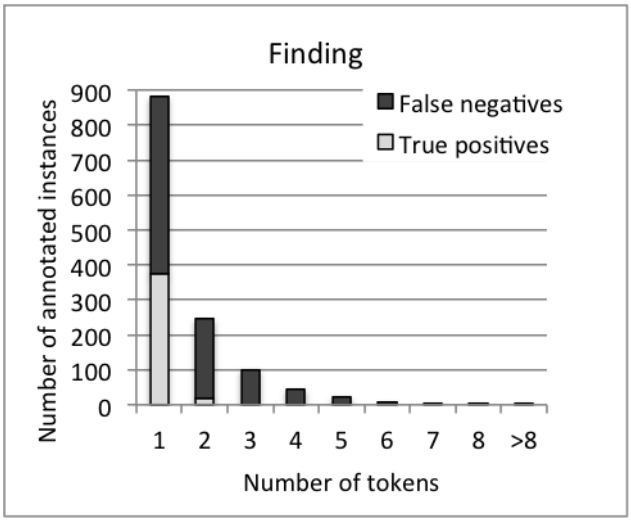

Figure 2: Distribution of the number of tokens for annotated findings, divided into true positives and false negatives (for *4:Qual*). No findings longer than two tokens were recognised by the constructed rule-based system.

tained abbreviations are shown in Table 5. As can be seen in the table, almost no correctly matched instances contained abbreviations, and abbreviations were more common among false negatives than in general. In experiment *11: Abbr*, the abbreviation list matched two types of disorders and one body structure correctly.

The manual estimation of the false positives showed that 32% of the false positives for *disorder*, 12% of those for *finding* and 24% of those for *body structure* were terms that could be classified as actually belonging to the semantic class that they had been assigned to by the system. However, among the 32% for disorders, there were also instances that were annotated as findings but that were assigned the class *disorder* by the system, showing that whether a term is a finding or a disorder can be context-dependent.

Class		Contains abbreviation	Avg. number of tokens
Disorder	Tot.	14%	1.28
	False neg.	22%	1.34
	True pos.	0%	1.16
Finding	Tot.	12%	1.61
	False neg.	17%	1.84
	True pos.	0.8%	1.05
Body structure	Tot.	1.9%	1.03
	False neg.	8.3%	1.05
	True pos.	0%	1.01

Table 5: Percentage of annotated instances that contain abbreviations and average number of tokens in the annotated entities. 11% of the total number of annotated entities contained abbreviations.

5. Discussion

The best results were obtained for recognition of the entity *body structure* which received a maximum F-score of 0.77. This was also the class that was most influenced by preprocessing. *Disorder* showed the second best results, with an F-score of 0.63. The recognition of entities of this class was only marginally improved by preprocessing techniques, but were instead improved when more terminologies were added. The lowest results were obtained for recognition of entities of the class *finding*, with a maximum F-score of 0.41. Recognition of this class was somewhat improved by preprocessing techniques such as lemmatisation and compound splitting, but no improvement was shown through inclusion of additional terminologies.

The best average F-score, 0.60, was obtained for the last experiment, *11: Abbr*, which showed an average precision of 0.69 and an average recall of 0.55.

The results obtained by Wang (2009) for the rule- and terminology-based named entity recognition show higher recall and lower precision than the results presented here, whereas Savova et al. (2010) achieved both higher precision and higher recall. The study by Savova et al. (2010) is more comparable, since results for recognition of the entity *disorder* are presented, whereas Wang (2009) presents average results for a number of different entities. Studies on English text are not directly comparable, since medical terminologies for Swedish are not as extensive as for English. However, the most comparable study, carried out by Kokkinakis and Thurin (2007) on Swedish discharge summaries, showed much higher precision and recall than the results 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.

The evaluation demonstrates limitations in the coverage of SNOMED CT on expressions used in Swedish clinical text. Even though there are probably many false negatives that could be matched to SNOMED CT terms through a further improvement of the preprocessing, the increase in recall for the class *disorder* with the inclusion of additional terminologies shows that there are still expressions for disorders that occur in clinical language that are not included in SNOMED CT.

The study by Penz et al. (2004) on coverage of the English SNOMED CT shows better results than the results presented here, but since that study was carried out on problem lists and not on free text, the results are only partly comparable. The results presented by Kokkinakis (2011a) are more comparable to the present study and they also show a low coverage in the terminologies for expressions of findings. The results of that study also correlate with the results presented here in that the recognition of findings does not improve when terms from ICD-10 and MeSH are included.

The error analysis showed that many of the false negatives contained abbreviations. It also showed that longer expressions were not recognised at all by the system. From this result it could be argued that the guidelines ought not to have allowed annotations of these complex expressions. However, it was deliberately chosen not to include such restrictions in the annotation guidelines, but instead to annotate all findings and disorders mentioned in the text, regardless

of how they were expressed. Thereby a corpus was created that enabled an evaluation of how much of the actual content of the text that automatic methods recognise.
From the error analysis of the false positives, it can be concluded that there is a need to further develop the annotated corpus in order to capture entities missed by the annotator. It can also be concluded that a system that aims at distinguishing findings from disorders needs to incorporate the context of words in order to classify them correctly.

6. Future work

The planned future work includes:

- Allowing multiple annotators to annotate the corpus in order to measure inter-annotator agreement and establish the degree of reliability of the annotation.
- Evaluating the constructed rule-based system on clinical text from another domain.
- Applying machine learning methods in order to recognise the annotated entities. The rule-based entity recognition system that was developed for this study will then be used as a baseline against which the resulting machine learning model can be compared. The output of the rule-based system will also be used as one feature for the machine learning system. This is a similar strategy to those applied by the previous machine learning studies described above.
- Developing methods for further expansion of abbreviations.

More long-term future work includes applying this entity recognition system to a larger clinical corpus in order to study the prevalence of different clinical findings as well as connections between them. The constructed system can easily be expanded into also retrieving the SNOMED CT concept ID of a term in the text in addition to retrieving its semantic class. This would open up the possibility of using the information from the hierarchical structure of SNOMED CT.

7. Conclusions and main contributions

The most important contribution of the study is that a rule-based system for recognising the clinical entities *disorder*, *finding* and *body structure* in Swedish clinical text has been constructed and evaluated. Even though the system shows relatively low precision and low recall for the entities *finding* and *disorder*, the constructed system is still very useful for two purposes; firstly to function as a baseline when evaluating a machine learning system that recognises these entities and secondly to be utilised for generating features to be used by this machine learning system.
Another contribution of this work is that the way in which the three kinds of clinical entities are expressed in everyday clinical text, and how this manner of expression corresponds to the use of language in SNOMED CT, has been explored.
There are two main conclusions from this work. First, it can be concluded that rule-based methods using existing Swedish medical terminologies are not sufficient for recognising clinical entities with high precision and recall, especially not when it comes to the entities *disorder* and *finding*. Further work is needed in order to achieve a system that performs better. Secondly, it has been shown that the studied clinical entities often are expressed in a way that varies from how they are expressed in the SNOMED CT terminology.

8. Acknowledgements

We are very grateful to Gunnar Nilsson, Sumithra Velupillai and Aron Henriksson for many useful comments regarding the annotations. We are also grateful to Staffan Cederblom and Studentlitteratur for giving us access to their database of medical abbreviations.
This work was partly supported by the project High-Performance Data Mining for Drug Effect Detection at Stockholm University, funded by Swedish Foundation for Strategic Research under grant IIS11-0053.

9. References

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, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. alan@nlm.nih.gov.

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.

Wendy W Chapman, John N Dowling, and George Hripcsak. 2008. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. *Int J Med Inform, Epub 2007 Feb 20*, 77(2):107–113, February.

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.

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), Sep/Oct.

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.

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

IHTSDO. 2008b. SNOMED CT style guide: Clinical findings. http://www.ihtsdo.org. 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*.

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.

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.

Dimitrios Kokkinakis. 2011a. 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. 2011b. What is the coverage of SNOMED CT on scientific medical corpora? In A. Moen, S. K. Andersen, J. Aarts, and P. Hurlen, editors, *Procedings of XXIII International Conference of the European Federation for Medical Informatics*. IOS Press.

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

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 28–30. European Language Resources Association (ELRA).

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.

Janet F E Penz, Steven H Brown, John S Carter, Peter L Elkin, Viet N Nguyen, Shannon A Sims, and Michael J Lincoln. 2004. Evaluation of snomed coverage of veterans health administration terms. *Studies In Health Technology And Informatics*, 107(Pt 1):540–544.

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, Sep-Oct.

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.

Maria Skeppstedt, Hercules Dalianis, and Gunnar Nilsson. 2011. Retrieving disorders and findings: Results using SNOMED CT and NegEx adapted for Swedish. In *Proceedings of the LOUHI 2011, Third International Workshop on Health Document Text Mining and Information Analysis*. CEUR-WS.

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/.

Hanna Suominen, Tapio Pahikkala, Marketta Hiissa, Tuija Lehtikunnas, Barbro Back, Helena Karsten, Sanna Salanterä, and Tapio Salakoski. 2006. Relevance ranking of intensive care nursing narratives. In *Proceedings of the 10th international conference on Knowledge-Based Intelligent Information and Engineering Systems - Volume Part I*, KES'06, pages 720–727, Berlin, Heidelberg. Springer-Verlag.

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. IOS Press.

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.

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

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.