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

Maria Skeppstedt[1], Hercules Dalianis[1], and Gunnar H Nilsson[2]

[1] Department of Computer and Systems Sciences (DSV)/Stockholm University, Forum 100, 164 40 Kista, Sweden
[2] Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden

Abstract. Access to reliable data from electronic health records is of high importance in several key areas in patient care, biomedical research, and education. However, many of the clinical entities are negated in the patient record text. Detecting what is a negation and what is not is therefore a key to high quality text mining. In this study we used the NegEx system adapted for Swedish to investigate negated clinical entities. We applied the system to a subset of free-text entries under a heading containing the word 'assessment' from the Stockholm EPR corpus, containing in total 23,171,559 tokens. Specifically, the explored entities were the SNOMED CT terms having the semantic categories 'finding' or 'disorder'. The study showed that the proportion of negated clinical entities was around 9%. The results thus support that negations are abundant in clinical text and hence negation detection is vital for high quality text mining in the medical domain.

Keywords: Negation detection, Clinical text, Electronic patient records, SNOMED CT, Swedish

1 Introduction

A huge amount of health documentation in digital form is produced today in the form of electronic patients records (EPR). This abundant database of medical competence and problem solving activities, expressed both in free text but also in structured form, can today be used for improving quality in patient care, for biomedical research, and also for educational purposes. In addition, the access to all this knowledge in patient records gives the possibility of finding information that can be useful for an individual patient case or in a specific clinical situation. Collections of EPR contain valuable information about the symptoms and diseases of patients and also traces of the reasoning process carried out by the physicians to find the disease and select the best treatment.

This reasoning includes negating possible diseases and symptoms, and if clinical entities are very frequently negated, then high quality information retrieval from patient records is heavily dependent on negation detection. A study by Wu et al. [14] on a search system for clinical findings in radiology reports showed for example that the precision improved from 27% to 81% when negation and uncertainty detection was applied. A similar search system is for instance needed when searching for patients with a specific symptom or disease, for example, during epidemics, detecting a specific disease and its management [6]. Identifying patients eligible for research studies or for enrollment in clinical trials is another important example of where a search system is needed [1]. Here, a patient with a complex profile is often searched for, of which most information is stored as narratives in the electronic patient record system. Usually the profile includes both inclusion as well as exclusion criteria, of which some may be found as negations. Another example of where negation detection is important is automated coding and classification, for instance automated ICD-10 coding [4].

The frequency of negated clinical entities has, for example, been studied by Chapman et al. In their study of negation in ten different types of clinical texts, including discharge summaries, radiology reports, and history and physical exams, the percentage of negated UMLS phrases was measured using automatic negation detection. The frequency of negated Unified Medical Language System (UMLS) phrases depended on the type of clinical text and varied from just above 80% to just below 40% [2].

In the BioScope corpus, which was annotated for uncertainty and negation, there was a negation in 14% of the sentences in the clinical text. However, the explored negated entities were not restricted to clinical entities. [13]

The aim of this study was to explore to what extent clinical entities in Swedish EPR are negated. The hypothesis was that negated clinical entities are as frequent in Swedish clinical text as in English clinical text, and since all the areas mentioned above depend on detecting negated entities, a high frequency of negated clinical entities implies the need for a robust system for automated negation detection.

2 Material and Methods

This study was conducted on a subset of the Stockholm EPR corpus. This subset was constructed through randomly extracting 500,000 fields with a headline ending with the word 'assessment', which resulted in a corpus containing 23,171,559 tokens. The full Stockholm EPR corpus contains patient records from over 900 clinics, and the records were written in Swedish from 2006 to 2008 [4].[3]

The clinical entities chosen to be explored in this study were the terms in the Swedish translation of SNOMED CT (version 20100820) [11], belonging to the semantic categories 'finding' or 'disorder' [7]. The clinical corpus was matched to the list of chosen SNOMED clinical terms. This was performed through an exact string matching, with the exception that everything was converted to lowercase and when the clinical term contained a comma, a match was also performed

[3] The research was carried out after approval from the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2009/1742-31/5.

without the comma. All sentences that contained a least one clinical entity found among the selected SNOMED terms were extracted to form the test set.

In order to make sure that the method of exact string matching against SNOMED terms detected entities with a high precision, it was evaluated using another subset of the Stockholm EPR Corpus that had the size of 23,100 tokens and contained notes from one ICU clinic. In this subset, all terms that belonged to any of the three semantic classes 'diagnosis', 'symptom', and 'finding' had been manually annotated in a previous study [12]. Since there is not always a clear mapping between the annotation classes 'diagnosis', 'symptom', and 'finding' and the SNOMED classes 'disorder' and 'finding', a gold standard for the evaluation was constructed through grouping all three annotation classes into one class, the class 'clinical entity'.

An exact matching against SNOMED terms with the semantic category 'disorder' showed a precision of 0.93 (± 0.03, 95% CI), when evaluated using the annotations of clinical entities in the gold standard. Also, a manual review of the false positives for the exact string matching against disorders showed that most could be defined as either some kind of clinical finding or as modifiers to clinical findings. Therefore, an exact string matching against the complete list of SNOMED disorders were considered to give high enough precision for an evaluation of the occurrence of negated disorders. This list contained 75,361 terms.

An exact string matching against SNOMED terms with either the semantic category 'disorder' or the semantic category 'finding' did, however, result in a precision of only 0.58 (± 0.04, 95% CI) when evaluated against the same clinical entities. A review of the false positives showed that they in many cases were due to ambiguity in the meaning of some SNOMED findings, since many words describing findings also have a different non-clinical meaning. Examples are the Swedish translations of *'walks'* (*'går'*) and *'moves'* (*'rör sig'*), both of which are used in many common Swedish set phrases such as *'It is not possible to...'* (*'Det går inte att ...'*) or *'It is the case of ...'* (*'Det rör sig om ...'*). To investigate to what extent such terms affect the result, the same string matching was performed again, except that all SNOMED disorders and findings occurring as unigrams or bigrams more than five times in a non-clinical corpus were excluded from the list of terms. The used non-clinical corpus was the PAROLE corpus, consisting of 600,000 tokens [5]. Excluding common non-clinical terms when performing exact string matching against the gold standard resulted in a precision of 0.80 (± 0.04, 95% CI). Moreover, a manual review of the false positives showed that most of them could be classified as some kind of clinical finding or modifier of a clinical finding. Therefore, when investigating negated findings, the modified list of SNOMED findings was used, in which common non-clinical terms were excluded. This list contained 37,616 terms.

Exact string matching resulted in a very low recall when compared against the gold standard annotation. Using the complete list of SNOMED disorders and findings resulted in a recall of 0.23, and when common non-clinical terms were excluded the recall decreased to 0.13. However, in order to investigate the frequency of negation, the main priority was considered to be ensuring that the extracted disorders and findings actually were disorders and findings, as opposed to maximizing the number of extracted clinical findings.

The NegEx system, which is a rule-based system developed for detecting negations in English clinical text [3], which also has been adapted to Swedish [10], was used to explore how many of the retrieved entities were negated. There exist many systems for automatic negation detection, as for example described by Rokach et al. [9]. However, since NegEx has been adapted and evaluated for Swedish clinical text, it was chosen for this study. The NegEx system, which is based on cue phrases for negation, works at the sentence level and determines whether a given clinical entity in a sentence is negated or not. The adaption of NegEx, in which the cue phrases are translated to Swedish, has previously been evaluated on a subset of the Stockholm EPR corpus [10]. The settings that were used for the present study achieved a precision of 0.78 (± 0.05, 95% CI) and a recall of 0.82 (± 0.05, 95% CI) on sentences containing negation cues, and a precision of 0.95 (± 0.02, 95% CI) when classifying sentences without cue phrases as not negated.

Since SNOMED CT also contains negated concepts, all retrieved SNOMED terms that contained any of the negation cues used by the Swedish NegEx system were excluded from the set of entities to study.

3 Results

In the total number of 23,171,559 tokens in the corpus, 228,531 clinical entities belonging to the SNOMED CT semantic category 'disorder' were found (8,688 unique terms), of which 20,814 were negated according to the NegEx system. The proportion of negated disorders was thus 0.091 (± 0.001, 95% CI). For the semantic category 'finding', 66,751 clinical entities belonging to that category were found (3,273 unique terms), and of these 6,180 were negated, resulting in a proportion of 0.093 (± 0.002, 95% CI) negated findings. The most frequent disorders and findings are listed in Tables 1 and 2.

4 Discussion

The result gives an indication of the frequency of negated disorders and findings in Swedish clinical text and underlines that negation detection is critical for high quality text mining in the medical domain. Compared to the study by Chapman et al. [2], the frequency of negated findings was lower in our corpus. Since their study showed that the frequency of negated findings varied substantially between different types of clinical texts, it is not unlikely that the lower frequency of negated findings in our corpus is due to the type of the text.

The method of using SNOMED terms for extracting findings and disorders might over- or under-estimate the proportion of negated entities. This would be the case if there were a systematic variation in which terms are used depending on whether they occur in a negated or affirmed context. The low recall of the

Table 1. The most common **disorders** in the SNOMED CT list of disorders. Affirmed means in this case 'Not negated'. The columns n give the number of affirmed or negated occurrences, and the columns % give in what proportion these occurrences were negated.

Affirmed				*Negated*			
In Swedish	*In English*	n	%	*In Swedish*	*In English*	n	%
hypertoni	hypertensive disorder	7508	3	sjukdom	disease	1227	17
sjukdom	disease	5886	17	ischemi	ischemia	575	44
astma	asthma	5401	8	astma	asthma	501	8
förmaksflimmer	atrial fibrillation	5205	3	allergi	allergic state	453	28
hjärtsvikt	heart failure	4274	8	hörselnedsättning	hearing loss	432	16
pneumoni	pneumonia	3457	7	främmande kropp	foreign body	409	43
otit	otitis	3447	9	sår	wound	390	12
sår	wound	2859	12	lungemboli	pulmonary embolism	383	18
anemi	anemia	2797	8	angina	angina	358	12
njursvikt	renal failure	2733	3	hjärtsvikt	heart failure	355	8
All retrieved:		207,717				20,814	

Table 2. The most common **findings** in the modified SNOMED CT list of findings in which common non-clinical terms are excluded. Column discriptions identical to the descriptions for Table 1.

Affirmed				*Negated*			
In Swedish	*In English*	n	%	*In Swedish*	*In English*	n	%
sinusrytm	sinus rhythm	2269	1	bröstsmärta	chest pain	454	19
bröstsmärta	chest pain	1896	19	blåsljud	bruit	227	16
buksmärta	abdominal pain	1750	7	återbesök planerat	recall arranged	221	79
tinnitus	tinnitus	1562	5	ödem	edema	215	18
nästa besök	next appointment	1270	2	hematuri	hematuria syndrome	213	16
dyspné	dyspnea	1239	11	proteinuri	proteinuria	190	16
reflux	reflux	1226	11	dyspné	dyspnea	160	11
blåsljud	bruit	1204	16	reflux	reflux	155	11
hematuri	hematuria syndrome	1131	16	uppföljning planerad	follow-up arranged	149	76
ödem	edema	1013	18	buksmärta	abdominal pain	134	7
All retrieved:		60,571				6,180	

exact string matching would then affect the results. There is some evidence that there is a systematic variation for some terms. It has, for example, been shown that the term 'hypertension' in the 'assessment' part of notes from an ICU clinic is almost never negated, as the absence of hypertension instead is described with the expression 'normal blood pressure' [8].

That neither the precision, nor the recall for the used negation detection system was perfect could also affect the results. Therefore, the real proportion of negated clinical entities is probably slightly different. However, the results still show that findings and disorders often are negated in clinical text.

The components used for this study could be developed into a system for retrieving the clinical status of a patient. However, many relevant entities are not found through the simple method of exact string matching used in this study. Therefore, the next step for building such a system is to explore possibilities for increasing the recall of clinical entities, for example through using stemming, spell checking, or additional resources such as lists of abbreviations.

Also, in order to achieve a more reliable retrieval of patient status, the system for detecting negations needs to be improved and also supplemented with the detection of if someone other than the patient is experiencing the medical problem as well as the detection of the temporality and level of certainty of the problem.

5 Conclusion

Clinical entities in Swedish EPR were frequently negated. About 9% of the terms matching the SNOMED terms with the semantic categories 'finding' and 'disorder' were negated in fields with a headline containing the word 'assessment' in the Stockholm EPR Corpus. These figures are somewhat lower than published results for the proportion of negated entities in English clinical texts.

The methods for recognizing clinical entities as well as for detecting negations have to be further developed before they can be used to draw any substantial conclusions about the occurrence of negations. However, the methods can so far give a reasonable indication of the frequencies of negated clinical entities. The results show that negated disorders and findings are common in clinical text, and hence support the claim that negation detection is critical for high quality text mining applications in the medical domain.

Acknowledgements

We would like to thank Maria Kvist and Sumithra Velupillai for development and access to annotated clinical data.

References

1. Aronis, J.M., Cooper, G.F., Kayaalp, M., Buchanan, B.G.: Identifying patient subgroups with simple Bayes'. In: Proc AMIA Symp. pp. 658–662 (1999)

2. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: Evaluation of negation phrases in narrative clinical reports. Proc AMIA Symp pp. 105–109 (2001)
3. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 34(5), 301–310 (2001)
4. Dalianis, H., Hassel, M., Velupillai, S.: The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In: Proceedings of ISHIMR. pp. 243–249 (2009)
5. Gellerstam, M., Cederholm, Y., Rasmark, T.: The bank of Swedish. In: Proceedings of LREC. pp. 329–333 (2000)
6. Hripcsak, G., Knirsch, C.A., Jain, N.L., Stazesky, R.C.J., Pablos-Mendez, A., Fulmer, T.: A health information network for managing innercity tuberculosis: bridging clinical care, public health, and home care. Comput Biomed Res 32(1), 67–76 (1999)
7. IHTSDO: SNOMED Clinical Terms User Guide, July 2008 International Release. http://www.ihtsdo.org/ (Accessed 2011-01-24)
8. Kvist, M., Skeppstedt, M., Velupillai, S., Dalianis, H.: Modeling human comprehension of Swedish medical records for intelligent access and summarization systems — Future vision, a physician's perspective. In: Proceedings of SHI 2011, Scandinavian Health Informatics meeting (To be published) (2011)
9. Rokach, L., Romano, R., Maimon, O.: Negation recognition in medical narrative reports. Information Retrieval 11(6), 499–538 (2008)
10. Skeppstedt, M.: Negation detection in Swedish clinical text: An adaption of Negex to Swedish. Journal of Biomedical Semantics 2 (Supp 3):S3 (To be published) (2011)
11. Socialstyrelsen: SNOMED CT-licens. http://www.socialstyrelsen.se/halsoinformatik /nationelltfacksprak/snomedct-licens (Accessed 2011-01-24)
12. Velupillai, S., Dalianis, H., Kvist, M.: Factuality levels of diagnoses in Swedish clinical text. In: Proceedings of XXIII International Conference of the European Federation for Medical Informatics, MIE (To be published) (2011)
13. Vincze, V., Szarvas, G., Farkas, R., Móra, G., Csirik, J.: The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 9(Suppl 11):S9 (2008)
14. Wu, A.S., Do, B.H., Kim, J., Rubin, D.L.: Evaluation of negation and uncertainty detection and its impact on precision and recall in search. J. Digital Imaging 24(2), 234–242 (2011)