Negation scope and spelling variation for text-mining of Danish electronic patient records

Cecilia Engel Thomas[1], Peter Bjødstrup Jensen[2,3], Thomas Werge[4], and Søren Brunak[1,3]

[1]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet, Building 208, DK-2820 Lyngby, Denmark.
[2]OPEN, University of Southern Denmark, Campusvej 55, DK-5230 Odense, Denmark.
[3]NNF Center for Protein Research, Department of Disease Systems Biology, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark. [4]The Research Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Copenhagen University Hospital, DK-4000 Roskilde, Denmark.

Abstract

Electronic patient records are a potentially rich data source for knowledge extraction in biomedical research. Here we present a method based on the ICD10 system for text-mining of Danish health records. We have evaluated how adding functionalities to a baseline text-mining tool affected the overall performance.

The purpose of the tool was to create enriched phenotypic profiles for each patient in a corpus consisting of records from 5,543 patients at a Danish psychiatric hospital, by assigning each patient additional ICD10 codes based on free-text parts of these records. The tool was benchmarked by manually curating a test set consisting of all records from 50 patients. The tool evaluated was designed to handle spelling and ending variations, shuffling of tokens within a term, and introduction of gaps in terms. In particular we investigated the importance of negation identification and negation scope.

The most important functionality of the tool was handling of spelling variation, which greatly increased the number of phenotypes that could be identified in the records, without noticeably decreasing the precision. Further, our results show that different negations have different optimal scopes, some spanning only a few words, while others span up to whole sentences.

1. Introduction

Electronic patient records (EPRs) file patient treatment data over time and contain structured data, such as medication information and laboratory test results, as well as unstructured data contained in free text. Previously unstructured data has been used for a range of purposes such as diagnosis detection (e.g. Meyste, 2006; Suzuki, 2008; Liao, 2010), decision support (Tremblay, 2009), and temporal investigation of adverse drug reactions (Eriksson, to appear 2014). Structured EPR data will primarily contain diagnoses relevant to the current hospitalization, whereas free text will contain additional information about adverse drug reactions and the general health status of the patient. By utilizing unstructured EPR data, it is possible to obtain a much richer phenotypic profile of each patient, which can be applied to the investigation of disease-disease correlations, patient stratification, and underlying molecular level disease etiology (Jensen, 2012).

Several tools for text mining of free text in English medical records have been developed previously. We present a non-English contribution to the field. We have developed a simple parser based on the ICD10 classification system for a Scandinavian language; Danish, which performs well and is relatively fast to implement. The parser handles a number of variations such as spelling and ending when matching between the corpus and the dictionary. We have evaluated the importance of taking these variations into account in a Danish context.

An additional focus of this work was to evaluate how negations should be handled in a Danish context. It has previously been shown that it is important to consider negations when medical text mining and several methods such as NegScope (Agarwal, 2010), NegFinder (Mutalik, 2001) and NegEx (Chapman, 2001) have been developed. These methods have shown good performance, but they have all been specifically developed for application to English text, and can thus not be directly transferred to our purpose. Instead we have here implemented a simple method for handling negations, and subsequently evaluated the scope of negations.

Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014, pages 64–68,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

2. Materials and methods

The text-mining tool presented here uses a dictionary based on the Danish version of the ICD10 system to search for mentioning of disease terminology terms in the corpus consisting of EPRs. Five add-on functionalities for the text-mining tool were evaluated. These were; handling of A) spelling, B) ending variations, C) allowing a gap in terms when matching, D) allowing shuffling of tokens in term when matching, and E) handling of negations.

The EPRs used here were 5,543 records from the Sct. Hans Psychiatric Hospital (Roque, 2011). The free text in these records consists of many different note types, written by a range of different types of medical and non-medical personnel including doctors, psychiatrists, nurses and social workers.

A test set of all records from a randomly selected set of 50 patients (roughly 1% of cohort) was manually curated. 5,765 disease related terms (hits) were found in the test set. On average each patient was associated with a total of 115.3 hits, which covered an average of 16.96 different ICD10 codes. Each hit was traced back to its origin in the corpus, and based on the context (sentence or entire note) it was evaluated whether the hit was correctly associated with the patient in the text.

2.1 Generation of spelling and ending variants

The ICD10 terms in the dictionary are supplemented with synonyms comprised of spelling and ending variants to allow a degree of fuzzy mapping between the corpus and the dictionary. Spelling (A) and ending (B) variants are generated by comparing all unique tokens of the corpus that exceed three letters with all unique tokens of the dictionary. Spelling variants (A) are generated by allowing a Damarau Levehnstein[1] edit distance of one between corpus and dictionary tokens. Ending variations (B) are generated by testing if a token becomes identical to a dictionary term if they are both stemmed for typical Danish endings.

2.2 Text-mining

A potential hit is a token or a set of tokens in a sentence, which match a full term in the dictionary. When matching one gap, comprised of an interposed word, is allowed (C) in the token string that is not found in the dictionary term. When matching a string of tokens to a dictionary term, shuffling of words is allowed (D), such that the order of the words is not important.

If a potential hit is found, the preceding part of the sentence is checked for negations (E). If a negation is found the potential hit is discarded. The end result is a list of hits with their matching ICD10 codes.

The negations evaluated here are both true negations like “ingen” and “ikke” (“none” and “no”), and alternative subjects such as family members. These alternative subjects are included as a form of negations, as a clinical term mentioned in the same sentence as an alternative subject, will often refer to that subject rather than the patient covered by the record.

2.3 Evaluation of features

All different combinations of functionalities A-D were tested and compared to the baseline text-mining tool with no add-on functionalities. The total number of hits and unique hits that a run of the tool results in were evaluated. Total hits include all hits, whereas unique hits consider simply how many unique 3-digit ICD10 terms are represented.

As described above each hit generated from the test set was evaluated to determine if it was correctly associated with the patient or not. Two different types of precisions were calculated: I) incidence precision, which is the number of correct hits divided by the total number of hits; II) association precision, where a hit is counted as correct as long as the corresponding ICD10 code is correctly associated with the patient at least once. Here it is assumed that as long as an ICD10 code is correctly associated with the patient once, it does not matter if the same ICD10 code is also incorrectly associated with the patient elsewhere.

2.4 Evaluation of negations and their scope

A random sample of 500 potential hits that were disqualified by the negation step was manually curated, and it was evaluated whether it was correct to negate the potential hits or not. The total number of negated hits, incidence precision and the distance, in terms of number of tokens, between the negation and the term it negates, i.e. the scope of the negations were calculated for all the negations. The same measures were calculated for each individual negation word occurring in the test set (data not shown).

[1] The Daramau Levehnstein edit distance is the number of edits needed to turn one token into another token. An edit can be a substitution, deletion or insertion of a letter, or the reversal of a pair of letters.

In order to investigate the influence of the distance between the hit and the negation further, the incidence precision for each distance was also calculated.

3. Results

The incidence precision of the tool with all features enabled was 0.867 and the association precision was slightly higher at 0.888. Enabling or disabling of fuzzy mapping features does not seem to affect the precision of the method. In contrast to this, both the total number of hits and the number of unique hits increase as more features are enabled. This is especially true for enabling feature A (spelling) and B (ending). Results for all runs can be seen in Figure 1 and Table 1

Figure 2 shows the results from evaluation of the negations for all 500 negated sentences. The correlation between the precision of a negation and its distance from the hits can be seen in Figure 2. As can be seen not all distances are represented in the test set. It seems that incidence precision is at least partly inversely related to the distance between the candidate hit and the negation.

Two negations are by far the most used in the records. These are ‘ikke’ and ‘ingen’, which are both true negations. Whereas ‘ingen’ has a very high incidence precision at 0.946 ‘ikke’ has a precision of only 0.573. These two negations also have very different negation scopes as can be seen on the plot in Figure 2 illustrating that different negation words can have very different scopes.

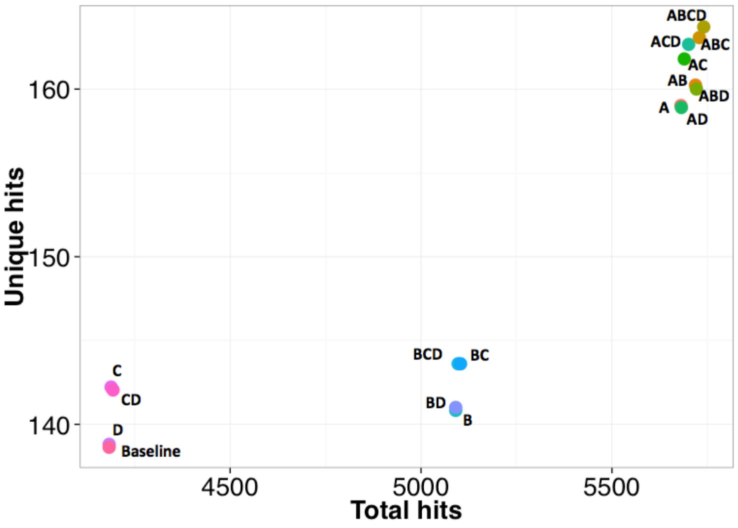

Figure 1: Number of hits generated for each run. A: spelling, B: ending, C: gap, D: shuffling.

Features	Incidence precision	Association precision
Baseline	0.872	0.889
D	0.872	0.889
C	0.872	0.89
CD	0.872	0.89
B	0.874	0.891
BD	0.874	0.891
BC	0.874	0.892
BCD	0.874	0.893
A	0.867	0.889
AD	0.867	0.886
AC	0.868	0.891
ACD	0.867	0.888
AB	0.867	0.889
ABD	0.867	0.887
ABC	0.868	0.891
ABCD	0.867	0.888

Table 1: Precision for all runs.

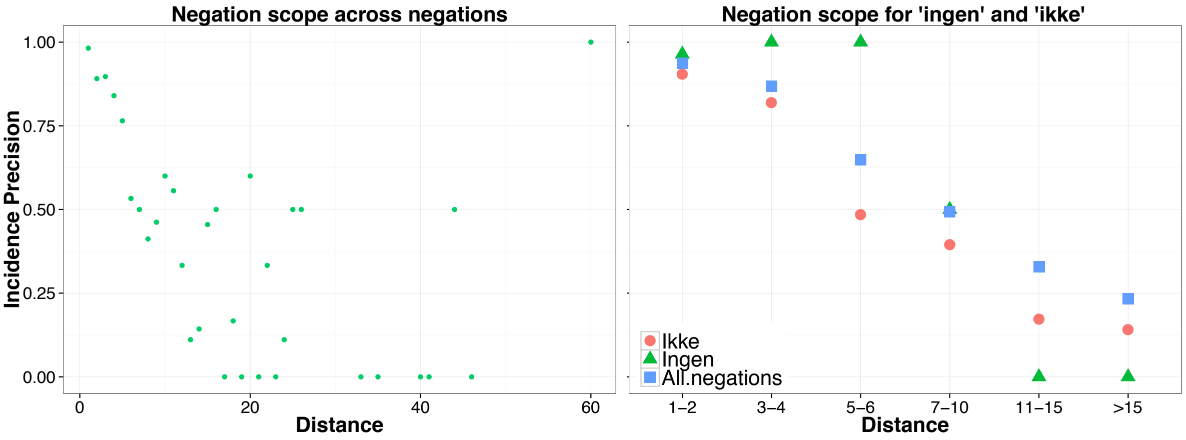

Figure 2: Evaluation of negation scopes for all negations (left) and ‘ingen’ and ‘ikke’ (right).

Negation	Total occurrences	Incidence precision	Average distance
All negations	500	0.722	4.9
True neg.	449	0.724	4.7
Subject neg.	51	0.706	6.4

Table 2: Evaluation parameters for negations.

Negation cutoff	Total hits	Unique hits	Incidence precision
None	5741	164	0.867
4	5964	171	0.854
10	5836	166	0.864

Table 3: Performance with hard negation cutoffs.

4. Discussion

4.1 Fuzzy mapping features

Quantitatively the precision of the tool presented here is on par with other similar tools such as MedLEE; 0.89 (Friedman 2004) and the tool presented in Meystre 2006; 0.76, despite that a relatively simple approach presented here.

Allowing ending variants (B) gives a significant increase in total hits, but only a minor increase in unique hits. This was investigated further, and it was revealed that the term ‘ryger’ (“smoking” or “smoker”) was responsible for this peculiarity, as the term ‘ryge’ matches ‘ryger’ when spelling variation is allowed. More than 4/5 of the total hits generated when enabling ending variation were due to this one synonym generated. The same problem is apparent when allowing spelling variants (A) as this also allows ‘ryger’ as a synonym.

It is debatable whether it is even worth including gap variations (D), since only very few hits are generated. However, there seems to be a synergistic effect between allowing gaps and shuffling and one must keep in mind that gap and shuffling variations only come into effect when a hit has more than one token, and only around 12% of all hits identified, have more than one token. Therefore gap and shuffling variations would make a bigger difference in a corpus where hits with more words are more frequent.

4.2 Negation evaluation

The data indicates that higher distance leads to lower precision. In order to improve the use of negations we tested two hard precision cutoffs (4 and 10) to limit the scope of negations. Using these hard cutoffs increased the precision of the negations from 0.722 to 0.921 and 0.820, respectively. This is comparable to the precisions reported for other tools such as NegEx; 0.845 (Chapman 2001) and NegFinder; 0.977/0.918 (Mutalik 2001), though one must keep in mind that these are tested on different corpora. Setting negation cutoffs also resulted in an increase in number of hits identified, but did lead to slightly lower precisions for the hits generated compared to no cutoff (see Table 3).

4.3 Limitations

The tool presented here was developed for EPRs from a psychiatric hospital, which does not guarantee its direct applicability to EPRs from other indication areas, as these psychiatric EPRs contain a high proportion of notes entered by nurses and other personnel that are not medical doctors. One possible issue related to this is that the EPRs used here do not show widespread use of abbreviations and acronyms for disease terms, thus a method for handling abbreviations was not implemented. However, this might be necessary for EPRs from other clinical domains.

Additionally the tool is limited to handle the 10 real and 24 subject negations present in the manually constructed negation list and negations are only allowed to negate terms in the succeeding part of the sentence, which will not be true for all negation usages.

In the approach described here it is assumed that a disease term found in a patients journal, is related to the given patient unless negated. This assumption is accepted here to preserve the simplicity of the approach, but is actually handled to so some extent by including subject negations.

5. Conclusion

We have shown here that it is possible to make a text-mining tool for a non-English language that has good performance in a quick and simple way. The full tool described here has rather good precision and many patient-disease relations were identified that could be used to enrich the phenotypes of the patients. Large variations in the precision of the different negations were found, but restricting the scopes of negations, contributes to increasing the precision of the negations. Furthermore, this also resulted in an increase in the number of hits generated without severely affecting the precision of the hits.

Acknowledgements

We thank Henriette Schmock, Sct. Hans Hospital, for help interpreting details in the corpus and the Novo Nordisk Foundation for support.

References

Agarwal, Shashank and Yu, Hong (2010) Biomedical negation scope detection with conditional random fields, J Am Med Inform Assoc., 17:696-701.

Chapman, Wendy W., Bridewell, Will, Hanbury, Paul, Cooper, Gregory F., and Buchanan, Bruce G. (2001) A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries, Journal of Biomedical Informatics 34, 301–310.

Eriksson, Robert, Werge, Thomas, Jensen, Lars J., and Brunak, Søren (to appear 2014) Dose-specific adverse drug reaction identification in electronic patient records: temporal data mining in an inpatient psychiatric population.

Friedman, Carol, Shagina, Lyudmila, Lussier, Yves, and Hripcsak, George (2004) Automated encoding of clinical documents based on natural language processing, J Am Med Inform Assoc., 11(5):392-402.

Jensen, Peter B., Jensen, Lars J., and Brunak, Søren (2012) Mining electronic health records: towards better research applications and clinical care, Nature Rev. Genetics 13(6):395-405.

Liao, Katherine P., Cai, Tianxi, Gainer, Vivian, Goryachev, Sergey, Zeng-Treitler, Qing, Raychaudhuri, Soumya, Szolovits, Peter, Churchill, Susanne, Murphy, Shawn, Kohane, Isaac, Karlson, Elizabeth W., and Plenge, Robert M. (2010) Electronic medical records for discovery research in rheumatoid arthritis, Arthritis Care Res (Hoboken) 62;1120-1127.

Meystre, Stephane and Haug, Peter J. (2006) Natural language processing to extract medical problems from electronic clinical documents: performance evaluation, J Biomed Inform 39;589-599.

Mutalik, Pradeep G., Deshpande, Aniruddha, and Nadkarni, Prakash M. (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. 8:598–609.

Park, Juyong, Lee, Deok-Sun, Christakis, Nicholas A., and Barabási, Albert-László (2009) The impact of cellular networks on disease comorbidity, Molecular Systems Biology 5:262.

Prather, J.C., Lobach, D.F., Goodwin, L.K., Hales, J.W., Hage, M.L., and Hammond, W.E. (1997) Medical data mining: knowledge discovery in a clinical warehouse, Proc AMIA Annu Fall Symp; 101-105.

Roque, Francisco S. Jensen, Peter B., Schmock, Henriette, Dalgaard, Marlene, Andreatta, Massimo, Hansen, Thomas, Søeby, Karen, Bredkjær, Søren, Juul, Anders, Werge, Thomas, Jensen, Lars J., and Brunak, Søren (2011) Using electronic patient records to discover disease correlations and stratify patient cohorts, PLOS computational biology.

Suzuki, T., Yakoi, H., Fujita, S., and Takabayashi, K. (2008) Automatic DPC code selection from electronic medical records: text mining trial of discharge summary, Methods Inf Med 47;541-548.

Tremblay, Monica C., Berndt, Donald J., Luther, Stephen L., Foulis, Philip R., and French, Dustin D. (2009) Identifying fall-related injuries: Text mining the electronic medical record, Inf Technol Manag 10;253-26.