Entity Recognition of Pharmaceutical Drugs in Swedish Clinical Text

Sidrat ul Muntaha[1], 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
simu9349@dsv.su.se, mariask@dsv.su.se, maria.kvist@karolinska.se, hercules@dsv.su.se

Abstract

An entity recognition system for expressions of pharmaceutical drugs, based on vocabulary lists from FASS, the Medical Subject Headings and SNOMED CT, achieved a precision of 94% and a recall of 74% when evaluated on assessment texts from Swedish emergency unit health records.

1. Introduction

A patient's health and treatment progress is documented in the clinical record in the form of structured data as well as in the form of narrative text. The data documented in narrative form is difficult to use for e.g. structured summarization, advanced search, statistical analysis and data mining. To be able to use narrative information for these purposes, automatic information extraction tools are called for to retrieve relevant information from free text. (Meystre et al., 2008)

An important part of the health record is documentation of a patient's medication. Automatic text summarization of clinical notes, including parts reasoning about medication, would enable clinicians to form a quick overview, also of records with long and detailed patient histories. Documentation of medication in health records could also be used for mining for new knowledge on pharmaceutical drugs used in health care, e.g. knowledge of adverse drug reactions caused by medication.

The first step for extracting information on medication, both for the purpose of summarization and for text mining, is to automatically recognize drugs that are mentioned in the clinical text. The aim of the work presented here is to study automatic recognition of pharmaceutical drugs mentioned in Swedish clinical text.

2. Method

The general approach of this study was to recognize mentions of drugs using a rule-based matching of clinical text to vocabulary lists, and evaluate this matching on annotated text data.[1]

2.1 Annotation

The data used for evaluation was clinical text annotated for mentions of pharmaceutical drugs. Generic substances, e.g. 'Paracetamol', and pharmaceutical drug names, e.g. 'Alvedon', as well as more general terms denoting medication, e.g. 'smärtstillande' ('pain killer') were annotated.

Free text in the assessment part of clinical notes from an emergency unit of internal medicine at Karolinska University Hospital was used. The texts are part of the Stockholm EPR Corpus (Dalianis et al., 2009) which contains electronic patient records written in Swedish. The same texts were previously used in a study focusing on clinical findings and body structures (Skeppstedt et al., 2012). The annotation had been carried out by a senior physician, using the annotation tool Knowtator (Ogren, 2006).

2.2 Vocabulary lists

The vocabulary for pharmaceutical drugs (25,161 unique expressions) was retrieved from three main sources: The Swedish version of MeSH, Medical Subject Headings (Karolinska Institutet, 2012), the Swedish translation of SNOMED CT (IHTSDO, 2008) and FASS, Farmaceutiska specialiteter i Sverige (FASS, 2012), which provides detailed about approved pharmaceutical drugs in Sweden.

From MeSH, terms in the category *pharmacologic-substance* (2,554 terms) as well as in the category *antibiotic* (239 terms) were used. From SNOMED CT, terms under the main category node *pharmacuetical* (16,977 terms) were used. From FASS, a list of Swedish product names for drugs (7,056 terms) as well as a list of classifications (5,062 terms) were used (NPL, 2011). The FASS terms for classifications of drugs, also contains a few very general terms, and to avoid false positives, terms in this list that were also included in the Swedish non-medical corpus Parole (Gellerstam et al., 2000) were therefore removed.

2.3 Matching to lists

Information in health records is often expressed using abbreviations, medical jargon or misspellings. This writing style has the advantage of quick recording, but makes it more difficult to process by a natural language processing system. As a consequence, an exact match to vocabulary lists might not be sufficient. Therefore, apart from exact string matching, the Levenshtein distance algorithm was used for comparing the clinical text to the terms in the vocabulary list.

The Levenshtein distance is a measure of similarity/distance between two strings, defined as the number of deletions, insertions and substitutions that are needed to transform one string to the other. Experiments were carried out in which expressions that had a Levenshtein distance of one or a Levenshtein distance of two from a term in the vocabulary lists were considered as a matching expression.

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

The automatic matching was evaluated against the annotated text, using the conll 2000 script (CONLL, 2000).

3. Results

The matching methods were evaluated on the annotated data, and precision, recall and F-measure were calculated. The results are shown in Table 1. A total of 580 mentions of drugs were present in the evaluation data, consisting of 26,011 tokens.

Method	Precision (CI)	Recall (CI)	F-score
Exact match	0.51 (± 0.03)	0.72 (± 0.04)	0.60
Excl. parole	0.94 (± 0.02)	0.74 (± 0.04)	0.83
Lev dist. 1	0.91 (± 0.03)	0.74 (± 0.04)	0.82
Lev dist. 2	0.89 (± 0.03)	0.75 (± 0.04)	0.81

Table 1: Precision, recall and F-score of the matching methods: 'Exact string match', 'Exact match, but words occurring both in classification list and Parole removed', 'Levenshtein distance of 1' and 'Levenshtein distance of 2'. For precision and recall a 95% confidence interval is provided.

4. Discussion

Just above 70% of the words and expressions for drugs were found using exact string matching. The Levenshtein distance matching method did not result in an improvement of recall, but only in decreased precision, which indicates that misspellings are not a common source of error when performing string matching of drugs.

4.1 Error analysis

That misspellings were rare, was also confirmed by the error analysis of the unmatched words. Also abbreviations were few among the false negatives.

Instead, compound words accounted for a large number of unmatched drug expressions, e.g. 'furixbehandling' ('furix treatment'), as well as expressions denoting drugs that were expressed with the effect of the drug or the disease for which it is given e.g. 'blodförtunnande' ('blood thinners') and 'hjärtsviktsmedicinering' ('heart failure medication'). Swedish is a language full of compound words, which provides special difficulties in building/porting tools.

Among the false positives were the term 'läkemedel' ('pharmaceutical') and expressions denoting narcotics.

4.2 Related work

When evaluating vocabulary-based entity recognition of drugs on text in discharge letters, a precision of 95% and a recall of 93% was achieved by Kokkinakis and Thurin (2007). That better results were achieved by Kokkinakis and Thurin (2007) might be due to that different rule-based approaches were used, but it may also be due to different types of evaluation data (discharge letters often have a more formal writing style than assessment notes). A part of the difference could perhaps also be explained by that a more wide definition of what expressions denote a pharmaceutical drug was used in the present study, compared to the study by Kokkinakis and Thurin.

5. Conclusion and future work

The vocabulary-based recognition of pharmaceutical drugs evaluated in this study identified more than 70% of the expressions for drugs in the free text of health records. Since compound words were frequent among the false negatives, compound splitting could be applied to improve results. Also additional methods ought to be applied, such as machine-learning-based recognition of drugs, which has been used by e.g. Wang and Patrick (2009). The vocabulary-based method developed for this study could, however, serve as a baseline method, and more importantly, the method evaluated here could also serve as one of the key features for such a machine learning system.

6. References

CONLL. 2000. CoNLL-2000. http://www.cnts.ua.ac.be/conll2000/chunking/, Accessed 2011-10-09.

Hercules Dalianis, Martin Hassel, and Sumithra Velupillai. 2009. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of ISHIMR*.

FASS. 2012. Sök läkemedel. http://www.fass.se/LIF/produktfakta/sok_lakemedel.jsp. Accessed 2012-08-27.

Martin Gellerstam, Yvonne Cederholm, and Torgny Rasmark. 2000. The bank of Swedish. In *Proceeding of LREC*, pages 329–333.

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

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.

Dimitrios Kokkinakis and Anders Thurin. 2007. Identification of entity references in hospital discharge letters. In *Proceedings of NODALIDA*.

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.

NPL. 2011. NPL (Nationellt produktregister för läkemedel) Review and verify product information. https://npl.mpa.se/mpa.npl.services/home2.aspx. Accessed 2011-10-28.

Philip V. Ogren. 2006. Knowtator: A Protégé plug-in for annotated corpus construction. In *Proceedings of HLT-NAACL*.

Maria Skeppstedt, Maria Kvist, and Hercules Dalianis. 2012. Rule-based entity recognition and coverage of snomed ct in swedish clinical text. In *Proceedings of LREC*.

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.