Diagnosing Diagnoses in Swedish Clinical Records

Sumithra Velupillai, Hercules Dalianis and Martin Hassel

DSV, KTH-Stockholm University, 164 40 Kista, Sweden
{sumithra,hercules,xmartin}@dsv.su.se

1 Introduction

Electronic clinical record systems are becoming the standard for many hospitals, providing an extensive amount of valuable information which could be used for important research in different research areas. In our project, we have access to a large set of de-identified clinical records from several departments in one of the largest hospitals in Sweden: Karolinska University Hospital. To our knowledge, this set is unique in at least two ways; it is the first set of clinical records written in Swedish, and it is the first set covering several medical departments, thus providing an invaluable data set for many research areas.

Clinical records contain both structured and unstructured entries, such as measurement values and sections of free text. However, the free text sections of clinical records have not, until recently, been used for further research. Such sections hold great potential for inventive text mining and computational linguistics research.

The language use in clinical records is very specific and noisy, containing domain-specific vocabulary, and often ad-hoc abbreviations and misspellings. Moreover, these types of text contain a potentially large amount of speculation, uncertainty and negation together with certainty and confirmation. This property is significant for the diagnosis and documentation procedure, and is very important to extract. For many text mining and information extraction tools, such issues are seldom taken into account, which we believe is problematic. These aspects have gained a lot of interest recently, and many methods for handling such parts in text sets have been proposed. However, most experiments have been performed on text sets in English, and mostly on similar contents. We plan to apply and evaluate existing state-of-the-art methods on Swedish clinical records. Moreover, we plan to develop these methods further with the goal of being as language independent as possible and generic for different medical specializations.

2 Related Work

Research on speculative language, or identification of both negations, uncertainties and other hedging cues in text has, in the (bio)medical domain, been performed both on biomedical scientific literature (full articles as well as abstracts) and clinical records. Many methods have been developed using handcrafted rule-based negation and uncertainty detection modules (see for instance Szarvas et al. (2008)). In Kilicoglu & Bergler (2008), the hedge classification dataset developed by Medlock & Briscoe (2007) was used, utilizing existing lexical resources with an extension of syntactic patterns and weighting schemes. In Szarvas et al. (2008), the creation of the BioScope corpus is described, a project with the aim of creating an annotated text set which can be used for developing and evaluating automatic classification systems for this specific phenomenon. The created corpus consists of biomedical scientific full papers and abstracts as well as medical free texts. A corpus consisting only of clinical free text has been used in a shared task on multi-label classification described in Pestian et al. (2007).

3 Proposed Work

The set of Swedish clinical records that we have access to is, as stated above, unique in many ways. The risk for accessing information that may be used to identify patients is an important aspect that has to be taken into account. The records we have obtained have been automatically de-identified, but many records may still have information in the free-text sections that may be used for identification, such as phone numbers, family member names, specific occupations etc. We intend to use and evaluate de-identification methods such as named-entity recognizers, in order to remove the risk of accessing private data. For the project described here, we propose to extract a small (fully de-identified) balanced subset for annotation of negation, speculation and certainty, based on the guidelines described in Szarvas et al. (2008). This set of annotated data will be classified applying current state-of-the-art methods, such as the ones described in Kilicoglu & Bergler (2008), and evaluated on our data set, which differs both in language and in text type.

From a small amount of clinical records from the Rheumatology clinic at the Karolinska University Hospital, we have identified several examples of both uncertain and certain diagnoses:

(1) Patient med oklar myalgi, muskelsvaghet med 10 mg Prednisolon, har ingen CK stegring, inga säkra förändringar på muskelbiopsi. Negativ EMG. Oklar diagnos. Statinutlöst myopati?
Patient with unclear myalgia, muscle weakness with 10 mg Prednisolon, no CK increase, no certain changes in muscle biopsy. Negative EMG. Unclear diagnosis. Statin triggered myopathia?

(2) Tydlig effekt av Methotrexate o Remicade i händerna, dock resterande sjd-aktivitet. Kan absolut inte avstå från NSAID.
Clear effect of Methotrexate and (abbr) Remicade in the hands, yet remaining signs of active disease (abbr). Still in need of NSAID (formulated with negation in Swedish).

As we can see in the examples above, conclusions (uncertain or certain) may span over sentence boundaries. Therefore, we will extend the annotations to span over sequences of sentences covering a diagnosis. In these guidelines, speculative elements are marked by angled brackets (<>), and negative elements are marked by square brackets ([]). We will also extend the guidelines to annotate elements that indicate certainty, with curly brackets ({}). The certain diagnosis in Example 2 for instance, would be annotated the following way:

(3) ({Tydlig} effekt av Methotrexate o Remicade i händerna), dock resterande sjd-aktivitet. Kan absolut ([inte] avstå från NSAID).

We propose the following work plan:

- Annotate a (fully de-identified) subset of the data set
- Apply existing state-of-the-art tools for classification of speculative sequences
- Analyze and evaluate the results, especially with regards to differences in the following aspects: language, medical specialization, and style and tradition in writing clinical records
- Develop methods for improving performance, primarily using word space models (and possibly extensions to constructions that can be modeled as words)

As many natural language processing tools will be needed in preprocessing steps, especially the ones used for de-identifying the full data set, current tools may not work optimally for these types of texts in Swedish. Evaluation and fine-tuning of such preprocessing steps will also have to be made. However, by using word space models, thus utlizing distributional patterns and relations in the text sets, heavy lexical and linguistic resources will not be needed once the records are de-identified.

Acknowledgements

Our thanks go to Vinnova (Swedish Governmental Agency for Innovation Systems) for funding the KEA-project (grant p29670-1), and to Jussi Karlgren (SICS) for initializing the idea and commenting the proposal.

References

H. Kilicoglu and S. Bergler: Recognizing Speculative Language in Biomedical Research Articles: A Linguistically Motivated Perspective. In BioNLP 2008: Current Trends in Biomedical Natural Language Processing, Columbus, Ohio, Association for Computational Linguistics, 46–53, June 2008.

B. Medlock and T. Briscoe: Weakly Supervised Learning for Hedge Classification in Scientific Literature. In Proceedings of the 45th Meeting of the Association for Computational Linguistics, 647–656, Prague, Czech Republic, June 2007.

J. P. Pestian, C. Brew, P. Matykiewicz, DJ Hovermale, N. Johnson, K. B. Cohen and W. Dutch: A Shared Task Involving Multi-label Classification of Clinical Free Text. In Biological, Translational, and Clinical Language Processing, Prague, Czech Republic, Association for Computational Linguistics, 97–104, June 2007.

G. Szarvas, V. Vincze, R. Farkas and J. Csirik: The BioScope Corpus: Annotation for Negation, Uncertainty and their Scope in Biomedical Texts. In BioNLP 2008: Current Trends in Biomedical Natural Language Processing, Columbus, Ohio, Association for Computational Linguistics, 38–45, June 2008.