Automatic Morphosyntactic Analysis of Clinical Texts

Filip Antomonov and Beáta Megyesi
Department of Linguistics and Philology
Uppsala University
antomonov@inbox.ru, beata.megyesi@lingfil.uu.se

Abstract

Electronical health records, also called clinical texts, have their own linguistic characteristics and have been shown to deviate from standard language. Therefore, computational linguistics tools trained on standard language presumably do not achieve the same accuracy when applied to clinical data. In this paper, we describe a pipeline of tools for the automatic processing of clinical texts in Swedish from tokenization through part-of-speech tagging and dependency parsing. The evaluation of the components of the pipeline shows that existing NLP tools can be used, but performance drops greatly when models trained on standard language are applied to clinical data. We also present a small, syntactically annotated data set of clinical text to serve as gold standard.

1. Introduction

The increased computerization in society has led to the possibility of making patient's health records electronically accessible. For many people, to access one's health records from home is rather convenient. However, the clinical language used in health records, for example in doctors' daily notes, discharge summaries, or radiology reports, has proven for many to be complicated to understand, as it is mainly meant for communication between medical staff.

Clinical language deviates from standard language in terms of word and sentence composition, lexical complexity and sentence structure. For example, we find more frequent use of technical terms, abbreviations, and omission of words in clinical texts (Smith, 2014).

Studies have shown that patients often have difficulties in understanding their health records. This creates the need for a simplification system, which should be able to automatically change and adjust the text in health records in order to make them easier for layman to understand. Before such a system can be created, a number of pieces must fall into place, including morphosyntactic analysis of clinical texts.

Some attempts have been made to syntactically analyze clinical data. Hassel et al. (2011) presents an initial study on parser applicability, pretrained on standard Swedish, to clinical text, with labeled attachment score of 76.6. And Skeppstedt (2013) introduces pre-processing rules in order to achieve better parsing performance, resulting in improved parsing in 9 out of 10 identified sentence types.

The aim of this study is to build a pipeline of natural language processing tools, adapted to clinical texts including tokenization, PoS tagging and parsing.

2. NLP tools for Clinical Texts

Morphosyntactic analysis is carried out through a pipeline consisting of existing, freely available tools. The idea is to test whether the tools trained on standard Swedish can be applied and adapted to analyze clinical texts with high accuracy. The pipeline consists of 4 main parts as illustrated in Figure 1:

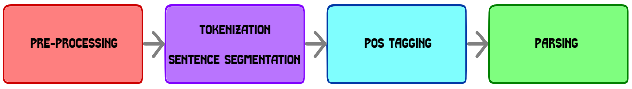

Figure 1. The 4 stages of the pipeline.

The text to be analyzed needs to be first pre-processed, where, for instance, relevant parts of the text are separated from unwanted ones, like personal data or dates. During pre-processing, other practical problems, like required file format, are also taken care of.

After pre-processing, the text needs to go through tokenization and sentence segmentation, essentially meaning that the text is automatically edited on word and sentence level with each word on a new line and empty lines separating sentences. This is performed by the Svannotate tool, developed for the Swedish Treebank (Nivre et al., 2006).

The next step of the automatic processing is the morphological analysis where the text is analyzed at word level with individual tokens being annotated with their appropriate part-of-speech tag, and possibly other morphological information. The part-of-speech tagger HunPos (Halacsy et al., 2007) with the Stockholm Umeå Corpus tagset (Gustafson-Capkova & Hartmann, 2006) was used to annotate the clinical data.

After the text has been cleaned up, tokenized and morphologically analyzed, the last step of the pipeline is the syntactic analysis, or parsing. We choose dependency annotation and test two state-of-the-art freely available dependency parsers, MaltParser (Nivre et al., 2007) and MSTParser (McDonald et al., 2005) for the syntactic annotation of clinical text.

3. Results

To test the performance of the components of the pipeline, especially the performance of the parsers, a small gold standard test set was created, consisting of around 400 sentences of different length, taken randomly from the Stockholm EPR Corpus (Dalianis et al., 2012) with ethical approval by the Regional Ethical Review Board in Stockholm (2012/2028-31/5). The sentences were syntactically annotated by two annotators following the guidelines of the dependency annotation of the Swedish Treebank.

MaltParser and MSTParser, trained on Talbanken of the Swedish Treebank, were used for the syntactic annotation. Table 1 shows the best achieved Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS) for MaltParser and MSTParser in comparison to the results for standard Swedish (Nivre & McDonald, 2008):

	Health Records	**Standard Swedish**
	LAS (UAS)	**LAS**
MaltParser	68.8 (75.5)	84.5
MSTParser	61.6 (72.3)	82.5

Table 1. The results of the parsers.

The results show that it is possible to use existing tools developed for standard language for morphosyntactic analysis of clinical texts, although room for improvement is considerable. Labels which appeared to be especially difficult for the parsers were coordination at main clause level (+F), predicative attribute (PT) and time adverbial (TA). Labels which the parsers passed well were ones like agent (AG), adjectival pre-modifier (AT) and determiner (DT).

To improve the results, a necessary step is to develop a larger syntactically annotated gold standard of clinical texts to train a tagger and a parser in order to adapt models specifically created for clinical domain. Regarding clinical language simplification, what that actually incorporates remains to be seen through user studies.

Acknowledgements

We would like to thank Joakim Nivre and Matilda Bengtsson for their help with the syntactic annotation of the gold standard, and Maria Kvist and Martin Duneld for providing the data sets.

References

H. Dalianis, M. Hassel, A. Henriksson, M. Skeppstedt. *Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care.* Proceedings of SLTC, 2012.

S. Gustafson-Capkova & B. Hartmann. *SUC 2.0 (ed.),* 2006.

P. Halacsy, A. Kornai, C. Oravecz. *Hunpos – an open source trigram tagger.* Association for Computational Linguistics, 2007.

M. Hassel, A. Henriksson, S. Velupillai. *Something Old, Something New – Applying a Pre-trained Parsing Model to Clinical Swedish.* Proceedings of NODALIDA, 2011.

R. McDonald, F. Pereira, K. Ribarov, J. Hajič. *Non-projective dependency parsing using spanning tree algorithms.* Human Language Technology and Empirical Methods in Natural Language Processing, 2005.

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryiğit, S. Kübler, S. Marinov, E. Marsi. *MaltParser: A language-independent system for data-driven dependency parsing.* Natural Language Engineering, 2007.

J. Nivre & R. McDonald. *Integrating Graph-Based and Transition-Based Dependency Parsers.* Proceedings of ACL, 2008.

J. Nivre, J. Nilsson, J. Hall. *Talbanken05: A Swedish treebank with phrase structure and dependency annotation.* Proceedings of LREC, 2006.

M. Skeppstedt. *Adapting a parser to clinical text by simple pre-processing rules.* Proceedings of BioNLP, 2013.

K. Smith. *Treating a case of the mumbo jumbos: What linguistic features characterize Swedish electronic health records?* Master thesis; Department of Linguistics and Philology; Uppsala University, 2014.