Adapting a parser to clinical text by simple pre-processing rules

Maria Skeppstedt
Dept. of Computer and Systems Sciences (DSV)
Stockholm University, Forum 100, 164 40 Kista, Sweden
mariask@dsv.su.se

Abstract

Sentence types typical to Swedish clinical text were extracted by comparing sentence part-of-speech tag sequences in clinical and in standard Swedish text. Parsings by a syntactic dependency parser, trained on standard Swedish, were manually analysed for the 33 sentence types most typical to clinical text. This analysis resulted in the identification of eight error types, and for two of these error types, preprocessing rules were constructed to improve the performance of the parser. For all but one of the ten sentence types affected by these two rules, the parsing was improved by pre-processing.

1 Introduction

Input speed is often prioritised over completeness and grammatical correctness in health record narratives. This has the effect that lower results are achieved when parsers trained on standard text are applied on clinical text (Hassel et al., 2011).

Syntactic annotations to use for training a parser on clinical text are, however, expensive (Albright et al., 2013) and treebanking large clinical corpora is therefore not always an option for smaller languages (Haverinen et al., 2009). There are studies on adaptation of standard parsers to the biomedical domain, focusing on overcoming difficulties due to different vocabulary use (Candito et al., 2011). How to overcome difficulties due to syntactic differences between standard and clinical language is, however, less studied. The aim of this study was therefore to explore syntactic differences between clinical language and standard language and to analyse errors made by the parser on sentence types typical to the clinical domain. To exemplify how this knowledge can be used, two simple preprocessing rules for improving parser performance on these typical sentences were developed.

2 Method

To find sentence types typical to the clinical domain, a comparison to standard text was conducted. The used clinical corpus was: free-text entries from assessment sections, thus mostly containing diagnostic reasoning, that were randomly selected from the Stockholm EPR corpus[1] (Dalianis et al., 2009); and the used standard corpus was: Läkartidningen (Kokkinakis, 2012), a journal from the Swedish Medical Association.

The comparison was carried out on part-of-speech sequences on a sentence level. The part-of-speech tagger Granska (Carlberger and Kann, 1999), having an accuracy of 92% on clinical text (Hassel et al., 2011), was applied on both corpora, and the proportion of each sentence tag sequence was calculated. 'Sentence tag sequence' refers here to the parts-of-speech corresponding to each token in the sentence, combined to one unit, e.g. '*dt nn vb nn mad*' for the sentence '*The patient has headache.*'. Pronouns, nouns and proper names were collapsed into one class, as they often play the same role in the sentence, and as terms specific to the clinical domain are tagged inconsistently as either nouns or proper names (Hassel et al., 2011). As sentences from Läkartidningen not ending with a full stop or a question mark are less likely to be full sentences, they were not included, in order to obtain a more contrasting corpus.

A 95% confidence interval for the proportion of each sentence combination was computed using the Wilson score interval, and the difference between the minimum frequency in the clinical corpus and the maximum frequency in the standard language corpus was calculated. Thereby, statistics for the minimum difference between the two domains was achieved.

[1] This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/834-31/5.

Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 98–101,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

A total of 458 436 sentence types were found in the clinical corpus. Of these, there were 1 736 types significantly more frequent in the clinical corpus than in the standard corpus, not having overlapping confidence interval for the proportions. 33 sentence types, to which 10% of the sentences in the corpus belonged, had more than 0.1 percentage points difference between minimum frequency in the clinical corpus and maximum frequency in the standard language corpus. For each of these 33 sentence types, 30 sentences were randomly extracted and the dependency parser MaltParser (Nivre et al., 2009), pre-trained on Talbanken (Nivre et al., 2006) using the algorithm *stacklazy* (Nivre et al., 2009), was applied to these part-of-speech tagged sentences. Error categories were manually identified, using MaltEval (Nilsson and Nivre, 2008) for visualisation.

Given the identified error categories, two pre-processing rules were constructed. These were then evaluated by applying the same pre-trained parser model on pre-processed sentences as on original sentences. A manual analysis was performed on a subset of the sentences that were differently parsed after pre-processing.

3 Results

Although only one sentence type was a full sentence (*nn vb pp nn mad*), most sentences were correctly parsed. Omitted words could be inferred from context, and therefore also the intended syntax. Eight error types, to which most errors belonged, were identified: 1) Abbreviated words ending with a full stop interpreted as the last word in a sentence, resulting in an incorrect sentence splitting. 2) Abbreviations incorrectly labelled as nouns by Granska, resulting in sentences exclusively containing nouns. 3) Adjectives not recognised as such (often because they were abbreviated), resulting in AT relations being labelled as DT relations. 4) A general adverbial relation incorrectly assigned an adverb of place or time relation or vice versa. 5) The first word in compound expressions parsed as a determiner to the second. 6) *nn pp nn pp nn mad* sentences for which a preposition had been incorrectly attributed. 7) The sentence type *nn jj* (noun adjective), for which most evaluated sentences were incorrectly parsed. 8) An omitted initial subject, resulting in the object incorrectly being parsed as the subject of the sentence.

Pre-processing rules were constructed for error types 7) and 8). As a verb in the middle of *nn jj*-sentences (in most cases copula) was left out, the first pre-processing rule added copula in the middle of these sentences. The second rule added the pronoun *I* as the first word in sentences starting with a verb, as this was the most frequently left out subject, along with the slightly less frequent omission, *patient*. The rules were not applied on sentences ending with a question mark.

10 (out of 33) sentence types were affected by the two rules. The proportion of those receiving a different parsing after pre-processing is shown in the column *Changed* in Table 1. A subset of these sentences, for which the parsing was changed, was manually classified as either incorrect (= containing at least one parsing or labelling error) or completely correct.

For sentences classified as incorrect, a more granular comparison between the original and the modified parsing was carried out. For these sentences, the difference in average unlabelled (*UAS*) and labelled (*LAS*) attachment score between the pre-processed and the original parsing was computed. A positive value indicates that although the pre-processing resulted in some incorrectly parsed sentences, these sentences were improved by pre-processing. The sentence types *vb pp nn nn mad* and *vb pp nn pp nn mad* were thus slightly improved by the pre-processing, although they had a low proportion of correctly parsed sentences.

A negative value for attachment score difference, on the other hand, indicates that parsing for the incorrectly parsed sentences was impaired by pre-processing. As these figures only apply to sentences incorrectly parsed after pre-processing, this means that although e.g. the type *vb ab nn mad* has negative UAS and LAS difference, this only applies to the 3 sentences that were incorrectly parsed by the pre-processed version.

With one important exception, sentences modified by pre-processing, were either a) given a completely correct parsing and labelling in between 64% and 100% of the cases, or were b) slightly improved by pre-processing. A reasonable simplification in this case is that there can only be one correct parsing of a sentence, as although there might be occurrences of syntactically ambiguous sentences, it is unlikely that their interpretation is not given by the context in the closed domain of language used for diagnostic reasoning.

Given this simplification, this means that a sentence was transformed from an incorrectly parsed sentence to a correctly parsed sentence in 64% or more of the cases, when pre-processing was applied. The difference in attachment score shows that the parsing is not drastically degraded for the rest of the sentences, although it mostly changed to a worse parsing. The overall effect of applying pre-processing is therefore positive. Sentences of the type *vb nn pp nn mad* is the important exception to this positive effect, important as 54% of the sentences belonging to this type received a different parsing after pre-processing and as 0.39% of the sentences in the corpus belong to this type. Only 61% of the pre-processed sentences of this type had a correct unlabelled parsing and only 32% had a correct labelled parsing. Many of these sentences were similar to *Writes a prescription of Trombyl*, for which *of Trombyl* incorrectly is given the word *write* as the head after pre-processing.

Almost all of the sentences of the type *nn jj mad* were correctly parsed when a copula was inserted between the noun and the adjective. Of the other types of sentences that improved, many improved by an incorrectly labelled subject relation being changed to an object relation. There were, however, also improvements because some adverbs of place and time were correctly labelled after the pre-processing rules had been applied.

4 Discussion

Even if quantitative data is given in Table 1, the core of this study has been to use a qualitative approach: searching for different categories of errors rather than determining accuracy figures, and investigating whether pre-processing has a positive effect, rather than determining the final accuracy.

The next step is to apply the findings of this study for developing a small treebank of clinical text. A possible method for facilitating syntactic annotation is to present pre-annotated data to the annotator (Brants and Plaehn, 2000) for correction or for selection among several alternatives. As the overall effect of applying pre-processing were improved parsings, the pre-annotation could be carried out by applying a model trained on standard language and improve it with the pre-processing rules investigated here. The other identified error types also give suggestions of how to improve the parser, improvements that should be attempted before using a parser trained on standard language for pre-annotation. Error types 1), 2) and partly 3) were due to abbreviations negatively affecting part-of-speech tagging and sentence splitting. Therefore, abbreviation expansion would be a possible way of improving the parser. That available medical vocabularies also could be useful is shown by error type 5), which was due to the parser failing to recognise compound expressions.

Of the sentences in the corpus, only 10% belonged to the analysed sentence types, and even fewer were affected by the evaluated pre-processing rules. It is, however, likely that the two developed pre-processing rules have effects on all sentence types lacking a verb or starting with a verb, thus effecting more sentence type than those included in this study. This is worth studying, as is also syntactic differences for shorter part-of-speech sequences than sentence level sequences.

Another possible method for domain adaptation would be to adapt the training data to construct a model more suitable for parsing clinical text. Instead of applying pre-processing, sentences in the training data could be modified to more closely resemble sentences in clinical text, e.g. by removing words in the treebank corpus to achieve the incomplete sentences typical to clinical text. Differences in vocabulary has not been included in this study, but methods from previous studies for bridging differences in vocabulary between the general and medical domain could also be applied for improving parser performance.

For supplementing a treebank to also include sentences typical to clinical text, some of the methods investigated here for extracting such sentence types, could be employed

5 Conclusion

Sentence types typical to clinical text were extracted, and eight categories of error types were identified. For two of these error types, pre-processing rules were devised and evaluated. For four additional error types, techniques for text-normalisation were suggested. As the pre-processing rules had an overall positive effect on the parser performance, it was suggested that a model for syntactic pre-annotation of clinical text should employ the evaluated text pre-processing.

Acknowledgements

Many thanks to Joakim Nivre and to the four reviewers for their many valuable comments.

Sentence type	# In test	% Changed	# Manually classified	% Correct unlabelled (labelled)	# Incorrect unlabelled (labelled)	pp UAS(LAS) difference among incorrect
a) vb nn mad	1181	30%	40	100 (100)%	0 (0)	
vb jj nn mad	317	13%	32	100 (94) %	0 (2)	
nn jj mad	316	100%	200	94 (94) %	12 (12)	
vb ab nn mad	256	33%	31	90 (90) %	3 (3)	-25 (-25) pp
vb pp nn mad	674	5%	27	100 (85) %	0 (4)	(-19) pp
vb ab pp nn mad	222	21%	30	100 (70) %	0 (9)	(+7) pp
vb pp jj nn mad	207	7%	14	100 (64) %	0 (5)	(-16) pp
b) vb pp nn nn mad	197	5%	9	22 (11) %	7 (8)	0 (+10) pp
vb pp nn pp nn mad	232	5%	12	75 (4) %	3 (12)	0 (+2) pp
c) vb nn pp nn mad	813	54%	28	61 (32) %	11 (19)	-20 (-15) pp

Table 1: *In test*: Number of sentences in test set of this type. *Changed*: Proportion of sentences that received a different parsing after pre-processing had been applied. *Manually classified*: Number of manually classified sentences. *Correct*: Proportion of sentences that were correctly parsed (and labelled) after pre-processing had been applied. *# Incorrect:* Number of incorrectly parsed (and labelled) sentences after pre-processing. *UAS (LAS) difference*: For these incorrect sentences: The difference in UAS, unlabelled attachment score, (and LAS, labelled attachment score) before and after pre-processing. (For sentence types with more than 90% correct sentences, this difference was not calculated.)

References

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler, 4th, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, Wayne Ward, Martha Palmer, and Guergana K Savova. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J Am Med Inform Assoc*, Jan.

Thorsten Brants and Oliver Plaehn. 2000. Interactive corpus annotation. In *LREC*. European Language Resources Association.

Marie Candito, Enrique H. Anguiano, and Djamé Seddah. 2011. A Word Clustering Approach to Domain Adaptation: Effective Parsing of Biomedical Texts. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 37–42, Dublin, Ireland, October. Association for Computational Linguistics.

Johan Carlberger and Viggo Kann. 1999. Implementing an efficient part-of-speech tagger. *Software–Practice and Experience*, 29:815–832.

Hercules Dalianis, Martin Hassel, and Sumithra Velupillai. 2009. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of ISHIMR 2009, Evaluation and implementation of e-health and health information initiatives: international perspectives. 14th International Symposium for Health Information Management Research, Kalmar, Sweden*, pages 243–249.

Martin Hassel, Aron Henriksson, and Sumithra Velupillai. 2011. Something Old, Something New - Applying a Pre-trained Parsing Model to Clinical Swedish. In *Proceedings of NODALIDA'11 - 18th Nordic Conference on Computational Linguistics*, Riga, Latvia, May 11-13.

Katri Haverinen, Filip Ginter, Veronika Laippala, and Tapio Salakoski. 2009. Parsing Clinical Finnish: Experiments with Rule-Based and Statistical Dependency Parsers. In Kristiina Jokinen and Eckhard Bick, editors, *Proceedings of NODALIDA'09, Odense, Denmark*, pages 65–72.

Dimitrios Kokkinakis. 2012. The journal of the Swedish medical association - a corpus resource for biomedical text mining in Swedish. In *The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), an LREC Workshop. Turkey.*

Jens Nilsson and Joakim Nivre. 2008. Malteval: An evaluation and visualization tool for dependency parsing. In *Proceedings of the Sixth International Language Resources and Evaluation. LREC*, pages 161–166.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, pages 24–26.

Joakim Nivre, Marco Kuhlmann, and Johan Hall. 2009. An improved oracle for dependency parsing with online reordering. In *Proceedings of the 11th International Conference on Parsing Technologies*, IWPT '09, pages 73–76, Stroudsburg, PA, USA. Association for Computational Linguistics.