Something Old, Something New – Applying a Pre-trained Parsing Model to Clinical Swedish

Martin Hassel
DSV, Stockholm University
Kista, Sweden
xmartin@dsv.su.se

Aron Henriksson
DSV, Stockholm University
Kista, Sweden
aronhen@dsv.su.se

Sumithra Velupillai
DSV, Stockholm University
Kista, Sweden
sumithra@dsv.su.se

Abstract

Information access from clinical text is a research area which has gained a large amount of interest in recent years. Automatic syntactic analysis for the creation of deeper language models is potentially very useful for such methods. However, syntactic parsers that are tailored to accommodate for the distinctive properties of clinical language are rare and costly to build. We present an initial study on the applicability of an existing parser, pre-trained on general Swedish, to clinical text in Swedish. We manually evaluate twelve documents and obtain a 92.4% part-of-speech tagging accuracy and a 76.6% labeled attachment score for the syntactic dependency parsing.

1 Introduction

The increasing use of electronic patient records has made it possible to explore this rich source of information by means of natural language processing. In order for the many potential applications to be successful, lexical information is often insufficient; syntactic information, such as dependency structures, is also needed.

The *MaltParser* system (Nivre et al., 2007) may be employed to provide such an analysis: it allows a dependency parser to be induced from a treebank, i.e. a syntactically annotated corpus. By building a treebank one can generate a parser for any language or sublanguage. This is, however, a somewhat demanding task.

The purpose of this paper is to evaluate to which degree a pre-trained model is directly transferable to a new domain, in this case clinical text. Given the linguistic differences between clinical and general Swedish (Dalianis et al., 2009), it may prove necessary to create a model tailored specifically to the clinical domain.

2 Background

There are a number of parsers that have been developed for Swedish, two of which are grammar-based dependency parsers (Knutsson et al., 2003). The *MaltParser*, however, is language-independent and data-driven, allowing parsers to be induced for any language or sublanguage. This approach is advantageous when the linguistic data resources are available (Nivre et al., 2007). The system has been successfully applied to an array of languages, yielding an average labeled attachment score[1] of 80.8%[2] across 13 languages (84.6% for Swedish) (Nivre, 2008).

Clinical text differs from general text in terms of both language and content, making it a discourse of its own. It is written by professionals responsible for patient care and is primarily used for record-keeping and transfer of information between healthcare personnel (Allvin et al., 2010).

In addition to the fragmentary style of clinical language and the prevalence of misspellings and abbreviations, there is also a significant difference in the respective vocabularies. This makes the application of natural language processing methods—including syntactic parsing—to clinical text a potential challenge (Dalianis et al., 2009).

Haverinen et al. (2009) have built a treebank for clinical Finnish[3], which was used to induce a domain-specific dependency parser using the *MaltParser* system. They annotated the corpus using the Stanford Dependency scheme, which was adapted to accommodate for properties of the Finnish language in general and clinical Finnish in particular. Using a standard version of Nivre's arc-eager parsing algorithm (Nivre et al., 2007), they report an overall labeled attachment score of 69.9%.

[1] The proportion of scoring tokens that are assigned both the correct head and the correct dependency relation label.
[2] Using Nivre's algorithm.
[3] 1,019 intensive care unit nursing documents.

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 287–290

A comparative study of intensive care unit documents written in Swedish and Finnish respectively shows that, despite significant linguistic differences, there are many structural and content-related similarities, such as missing predicates, copulas and subjects (Allvin et al., 2010).

3 Method

We apply the pre-trained model for Swedish, *SweMalt*[4], developed for the *MaltParser* system, on a set of Swedish clinical assessment entries[5], where each entry is treated as a document. These entries are written by physicians from an emergency ward. The model is trained on the *Talbanken* section of the *Swedish Treebank*[6]. As input to the system, we need part-of-speech (PoS) tagged data. We use the *Granska Tagger* (Carlberger and Kann, 1999) in this initial step. No cleaning or other preprocessing is performed on the documents prior to applying the PoS tagger; however, the evaluation is performed only on content tokens, i.e. punctuation and formatting issues are ignored.

As there is no morphologically and syntactically annotated corpus of Swedish clinical text that can be used for evaluation, we manually evaluate twelve randomly extracted documents with regards to the following: (1) PoS tagging accuracy and (2) labeled and unlabeled attachment score (LAS and UAS), as well as labeled accuracy score (LA), of the syntactic parses. For the evaluation of the syntactic parses, we use the visualization tool provided by *MaltEval*[7].

The results are evaluated manually by two researchers, both educated in (Swedish) linguistics, but with no formal training in the specific morphological and syntactic schemas used by *MaltParser*. One document is evaluated jointly, while the remaining eleven documents are evaluated individually, after which differences are resolved through discussion.

[4]Available at `http://maltparser.org/mco/swedish_parser/swemalt.html`

[5]This research has been carried out after approval from the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2009/1742-31/5.

[6]For a description of the treebank, see `http://stp.ling.uu.se/~nivre/swedish_treebank/`

[7]Available at `http://w3.msi.vxu.se/~jni/malteval/`

4 Results

In Table 1 we present overall information about the twelve documents used in the experiment. Sentences vary greatly in length, ranging from two to 36 content tokens, but consist of around ten words per sentence on average. The documents are very short (5.6 sentences on average), ranging from only one sentence to ten.

	#	min - max	avg ± std
Sentences	68	1 - 10 (/d)	5.6±2.0 (/d)
Tokens	676	2 - 36 (/s)	9.9±6.4 (/s)
		4 - 128 (/d)	57.6±30.5 (/d)

Table 1: General statistics for the twelve documents. Minimum, maximum, average and standard deviation for sentences per document and tokens per sentence and document. d = document, s = sentence.

Analysis and measure	# errors	% score
PoS, accuracy	51	92.4
Syntactic parse, LAS	142	76.6
Syntactic parse, UAS	117	80.7
Syntactic parse, LA	133	78.1

Table 2: Results of part-of-speech (PoS) tagging and syntactic parses. LAS = Labeled Attachment Score, UAS = Unlabeled Attachment Score, LA = Label Accuracy Score.

4.1 Part-of-Speech Tagging

In general, the PoS tagging results were very high (see Table 2). However, some mistakes frequently recur. One of the most common words, *pat* (”patient”, abbreviated), was consistently assigned the class proper name (PM), which is erroneous. Moreover, other abbreviations such as *ua* (*utan anmärkning*, ”without remarks”) and *ca* (”approx.”) were, in many cases, broken up in the morphological analysis, or split in the original text (with a space or a colon inserted in between), which led to errors in the PoS assignments.

Moreover, the documents contain many clinical terms, such as disease names, medications, wards, etc. These were sometimes tagged as nouns (NN), sometimes as proper names (PM), although not always consistently. These cases are not easy to discriminate, as they, in context, work either way. For instance, the disease *dvt* (*djup ventrombos*, ”deep

venous thrombosis", abbreviated) is tagged as a proper name (PM), while *myocardit* (*myocarditis*) has been tagged as a noun (NN). These conflicting analyses are not, however, problematic for the syntactic analysis.

4.2 Syntactic Parsing

The overall results of the syntactic parsing are presented in Table 2. The most common errors are related to conjunctions, adverbials and prepositional constructions. These are also among the most common dependency relation types, with PA (*Complement of preposition*) being the most frequent (77 instances).

Many sentences lack a main predicate (32%), which is known to be a problematic issue for syntactic parsers (Haverinen et al., 2009). Moreover, main subjects are often omitted (43%), which further complicates the analysis. This feature is not unique to Swedish in the clinical domain (Allvin et al., 2010). In Figure 1, we see an example of a rather typical sentence, where there is no predicate or subject, with several errors in the syntactic analysis as a result.

In general, shorter sentences such as *således dvt* ("thus dvt") are analyzed correctly (ROOT → *dvt* → CA → *således*), while longer sentences contain several errors, in particular sentences with complicated conjunctional, conditional and prepositional constructions.

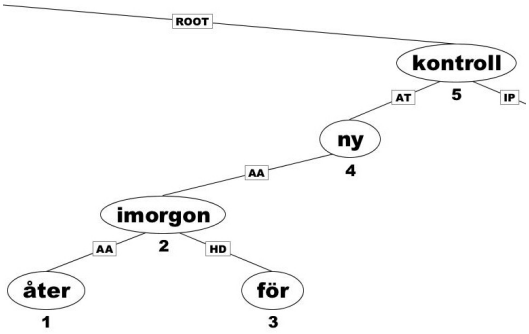

Figure 1: Example parse tree for the sentence *åter imorgon för ny kontroll* ("back tomorrow for new check-up")

5 Discussion

Although this study is performed on a very small set of documents, the general results could be interpreted as indicators. The twelve documents were randomly extracted from a total of 150 parsed documents. A comparison of the general characteristics—the average number of tokens per sentence, common tokens and dependency relations—of the sample set with that of all the parsed documents shows that the evaluation was conducted on a data set that is at least fairly representative of Swedish emergency ward documentation.

5.1 Part-of-Speech Tagging

The pre-trained *SweMalt* model presupposes input text that has been morphosyntactically disambiguated using the Stockholm-Umeå Corpus (SUC) tag set (Ejerhed et al., 1992). It should thus be noted that there has not been any tailoring to resolve differences between morphosyntactic categories in the *Granska Tagger* system compared to the SUC categories. This fact could possibly influence results in the syntactic parses, as the parser might encounter tags that it does not recognize[8] and the lack of information on past participles, for instance, might be harmful for the parser.

We have not evaluated the full morphosyntactic tag, but rather focused on the part-of-speech. The results of the PoS tagging were very high (92.4% accuracy), which is in line with state-of-the-art performance of Swedish PoS taggers. Carlberger and Kann (1999) report an accuracy of 92.0% for unknown words and 96.3% for all words when evaluated on a part of SUC. Since the clinical domain is also new to the tagger, it is in this study exposed to a higher degree of unknown words in the form of medical brands, tests and diseases, as well as ad-hoc abbreviations of much of the aforementioned terminology (Allvin et al., 2010). These prove to be a challenge already in the tokenisation step of the analysis.

5.2 Syntactic Parsing

The syntactic parsing results (see Table 2) are lower than state-of-the-art results of (general) Swedish (LAS: 84.6%, UAS: 89.5%, LA: 87.4% (Nivre, 2008)). However, the results are still within the range of average overall parser performances accross languages (LAS: 80.8%). Compared to other parsing results for clinical language, we observe higher LAS scores than those presented in Haverinen et al. (2009) (LAS: 69.9% for

[8]The Granska Tagger has, apart from removing 14 tags/features, introduced 5 new tags.

a statistical parser, LAS: 75.2% for a rule-based parser). Although the results are not directly comparable to the mentioned previous studies (e.g. in both Nivre (2008) and Haverinen et al. (2009), the parsing model is not evaluated on a new domain; there are, of course, language differences; different evaluation methods are used; etc.), we believe the general trends are comparable.

The distribution of the most common dependency relation types can—with such reservations as stated above—be compared to those reported for general Swedish in (Nivre et al., 2007). Despite differences in dependency relation schemes, we observe some similarities in the distribution patterns. For example, prepositional dependency relations (PA, *Complement of preposition*, and PR, *Preposition dependent*) and conjunctional relations (CJ, *Conjunct* and CC, *Coordination*) are among the most common (approx. 10% respectively). One important difference is, however, the number of adverbial dependency relation types in the different schemes. In Nivre et al. (2007) only one adverbial relation is used, while in this setting there are ten. Since there are so many different adverbial types in our setting, and the adverbial relations are one of the larger sources of errors in our evaluation, one possible explanation might be the low frequency per adverbial type.

Other than that, although we have not quantified the amount of errors per dependency relation type, similar tendencies are apparent. For instance, among error types categorized as the ”medium-accuracy set” in Nivre et al. (2007), we find error types linked to incorrect attachment, e.g. modifier attachment ambiguities and attachment ambiguities. These are common error types in our experiment as well. The general indications are thus that the error types found in this evaluation are not necessarily domain-dependent; however, modeling syntactic analyses of sentences lacking predicates and/or subjects would probably be needed in order to improve results. This particular characteristic seems to be typical for clinical language (see e.g. Haverinen et al. (2009)).

5.3 Conclusion

The main finding is that the morphological characteristics of Swedish clinical language do not differ greatly from general language and that existing tools can be used successfully when it comes to PoS information. Syntactic parsing also works well in most cases, but the errors that are produced are relatively severe. One solution to this would be to enrich an existing treebank with only these types of sentences. Along these lines, we plan to use this evaluated set as a small gold standard for further development of parsing Swedish clinical documentation, as well as for studying domain adaptation for other professional languages.

References

Helen Allvin, Elin Carlsson, Hercules Dalianis, Riita Danielsson-Ojala, Vidas Daudaravičius, Martin Hassel, Dimitrios Kokkinakis, Heljä Lundgren-Laine, Gunnar Nilsson, Ø ystein Nytrø, Maria Skeppstedt, Hanna Souminen, and Sumithra Velupillai. 2010. Characteristics and Analysis of Finnish and Swedish Clinical Intensive Care Nursing Narratives. In *Second Louhi Workshop on Text and Data Mining of Health Documents (Louhi-10)*, number 2, pages 53–60, Los Angeles, U.S. Association for Computational Linguistics (ACL).

Johan Carlberger and Viggo Kann. 1999. Implementing an efficient part-of-speech tagger. *Software Practice and Experience*, 29(9):815–832, July.

Hercules Dalianis, Martin Hassel, and Sumithra Velupillai. 2009. The Stockholm EPR corpus: Characteristics and some initial findings. In *Proc. 14th ISHIMR*, volume 219, pages 243–249, Kalmar, Sweden.

Eva Ejerhed, Gunnel Källgren, Ola Wennstedt, and Magnus Åström. 1992. The linguistic annotation system of the The Stockholm-Umeå corpus project. Technical Report DGLUUM-R-33, Department of General Linguistics, University of Umeå.

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, *Proc. 17th NODALIDA 2009*, number 1998, pages 65–72.

Ola Knutsson, Johnny Bigert, and Viggo Kann. 2003. A robust shallow parser for Swedish. In *Proc. NODALIDA 2003*, Reykavik, Iceland.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135, January.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553, December.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955