ata, citation and similar papers at core.ac.uk

brought to yo

provided by DSpace at Tartu U

Bootstrapping a Swedish Treebank Using Cross-Corpus Harmonization and Annotation Projection

Joakim Nivre and Beáta Bandmann Megyesi
Uppsala University
Department of Linguistics and Philology

Abstract

In this paper, we describe an ongoing project with the aim of bootstrapping a large Swedish treebank, ultimately with a size of about 1.5 million tokens, by reusing two previously existing annotated corpora: an old treebank of about 350,000 tokens and a more recently developed part-of-speech-tagged corpus of about 1,2 million words. A key component in the bootstrapping methodology is the use of cross-corpus harmonization and annotation projection.

1 Introduction

Given the high cost of manual annotation and post-editing in treebank development, the possibility to reuse existing annotated resources is of great importance. Often the efficient reuse of such resources is hampered by the fact that different resources, even for the same language, have been developed with different annotation guidelines or encoding standards. In many cases, however, it is possible to overcome these obstacles through a process of cross-corpus harmonization and annotation projection.

In this paper, we describe an ongoing project with the aim of bootstrapping a larger Swedish treebank, ultimately with a size of about 1.5 million tokens, by reusing two previously existing annotated corpora: an old treebank of about 350,000 tokens (Einarsson, 1976a; Einarsson, 1976b) and a more recently developed part-of-speech-tagged corpus of about 1,2 million words (Ejerhed and Källgren, 1997). Although many of the details involved in the process are dependent on properties of the pre-existing corpora, and therefore specific to this particular project, we believe that there are general points about methodology that are of relevance to the community.

We first give an overview of the project and the different steps needed to develop a new treebank from existing resources and then focus on the two most interesting steps: the harmonization of tokenization and sentence

De Smedt, K., Hajič, J. and Kübler, S. (Eds.)
Proceedings of the Sixth International Workshop
on Treebanks and Linguistic Theories (2007)

segmentation, and the projection of annotation from one corpus to the other using data-driven taggers and parsers. In addition, we briefly discuss how much is gained by reusing existing corpora and annotation, as opposed to creating a new treebank from scratch.

2 Project Overview

The final goal of the project is to produce a treebank of Swedish containing 1.5 million words by reusing two existing annotated corpora. We begin by describing these two corpora and some of their important properties.

Talbanken (Einarsson, 1976a; Einarsson, 1976b) is a syntactically annotated corpus, containing both written and spoken Swedish, produced in the 1970s at the Department of Scandinavian Languages, Lund University, by a group led by Ulf Teleman. In total, the corpus contains about 350,000 tokens, divided into 200,000 tokens of written Swedish (professional prose and high school essays) and 150,000 tokens of spoken Swedish (interviews, debates, and informal conversations). The annotation consists of two layers: a lexical layer, with parts of speech and morphosyntactic features, and a syntactic layer, with a relatively flat phrase structure and grammatical functions (or dependencies). The annotation scheme, known as MAMBA, is described in Teleman (1974).

The main asset of Talbanken, from our point of view, is the syntactic annotation, which contains enough information to support the derivation of both phrase structure and dependency structure representations, as shown in Nilsson, Hall, and Nivre (2005) and Nivre, Nilsson, and Hall (2006), and therefore provides a good base representation for a treebank. Moreover, since Talbanken is by far the largest available corpus of Swedish with manually validated syntactic annotation, including it in the new treebank not only lets us reuse a manually validated syntactic annotation of 350,000 tokens, but also gives us a good basis for training parsers that can be used in the annotation of additional data.

The Stockholm-Umeå Corpus (SUC) (Ejerhed and Källgren, 1997) is a balanced corpus of written Swedish, modeled after the Brown Corpus and similar corpora for English, developed at Stockholm University and at Umeå University in a project led by Gunnel Källgren and Eva Ejerhed. The corpus consists of some 1.2 million tokens of text from a variety of different genres, the corpus encoding follows the guidelines of the Text Encoding Initiative (TEI), and the annotation includes lemmatization, parts of speech, morphosyntactic features, and named entities. Since SUC was first released in the 1990s, its annotation scheme has become a de facto standard for Swedish, especially in research on part-of-speech tagging, where SUC data is standardly used for training and evaluation, e.g., in Nivre (2000) and Megyesi (2002).

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

Given that SUC is a larger and more recently developed corpus, which has been extensively used to train taggers and other tools for Swedish, it makes sense to use SUC as a model for the new treebank wherever possible, thus minimizing the need for (new) manual validation and maximizing the conformance with current practice in Swedish language technology. This means, among other things, that principles of tokenization and sentence segmentation should be kept intact in SUC but modified for Talbanken in cases of conflict. We will refer to this as the *harmonization* of tokenization and sentence segmentation. The same holds for the annotation of parts of speech and morphosyntactic features, where the kind of annotation used in SUC has to be *projected* to Talbanken, which unfortunately uses a different scheme. [1] Since no simple mapping exists from the Talbanken scheme to the SUC scheme (nor in the other direction), this projection will have to be induced by training a tagger on the SUC corpus, using it to reannotate Talbanken, and finally correcting the errors performed by the tagger in a manual post-editing phase.

Given the considerations so far, we propose the following overall plan for the production of a new treebank based on Talbanken and SUC:

1. Convert both corpora with their existing annotation into a common standard for corpus encoding.[2]

2. Harmonize tokenization and sentence segmentation in Talbanken, applying as far as possible the principles adopted in SUC.

3. Project part-of-speech tags and morphosyntactic features from SUC to Talbanken, using a data-driven tagger trained on SUC with manual post-editing.

4. Project syntactic annotation from Talbanken to SUC, using a data-driven parser trained on Talbanken with manual post-editing.

In the following two sections, we describe the problems involved in harmonization and annotation projection in a little more detail.

3 Harmonization

To harmonize the two corpora, we convert the tokenization and sentence segmentation of Talbanken according to the principles of SUC. In the tokenization of SUC, abbreviations are always represented as single tokens.

[1] Other kinds of annotation found in SUC, such as lemmatization and named entities, are outside the scope of the current project but should in principle be projected in the same way from SUC to Talbanken.

[2] The exact standard used is not important in this context, but we plan to use the XML-based Corpus Encoding Standard (XCES) with standoff annotation.

This means that when abbreviations in the original text consist of several tokens, these are concatenated into one token where the space character in the original are represented by an underscore. In Talbanken, on the other hand, abbreviations, as any other multi-word expressions, consist of several tokens, independently of how they appear in the original text. Each token in the abbreviation is shown on a separate line where the first token is annotated with the part-of-speech tag, while the other tokens in the expression receive an `ID` tag. To find the abbreviations in Talbanken, we automatically extract tokens annotated with `ID` tags together with the preceding head token, and convert these into one single token while separating the included tokens with an underscore. Then, we manually extract the abbreviations from the multi-word expression list. Lastly, we remove the underscore in cases where the internal tokens in the abbreviation are separated by a period, according to SUC's tokenization standards.

The sentence segmentation also differs in the two corpora. Above all, lists have a different structural annotation. In SUC, items in lists are handled as different sentence units, while in Talbanken the entire list consisting of several items can be treated as one sentence. This might lead to parser annotation errors as the sentences in Talbanken that will serve as training data to a data-driven parser will have a different structure compared to the sentences in SUC that need to be parsed. Therefore, we treat each item in a list in Talbanken as a sentence unit as far as possible.

4 Annotation Projection

In order to harmonize the morphological annotation of the two corpora, we project the part-of-speech tags and morphological features from SUC to Talbanken. We do this by training the data-driven TnT tagger (Brants, 2000) on SUC, bootstrapping the tagger by training it on a considerably larger automatically tagged corpus (Forsbom, 2005), and then applying the trained model to Talbanken. Finally, we correct the automatic annotation manually following SUC's annotation principles. The result is a corpus with consistent morphological annotation.

For the syntactic annotation, we project the syntactic analysis of Talbanken to SUC as SUC lacks syntactic annotation. The phrase structure and dependency annotation in Talbanken is projected to SUC by training the data-driven MaltParser (Nivre and Hall, 2005) on Talbanken. In the near future, we are going to experiment with various data-driven parsers to model the constituent and dependency structures, which will enable us to use ensemble of classifiers to facilitate the manual correction of the automatic annotation.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

5 What Is Gained?

A reasonable question to ask is how much is actually gained by reusing existing corpora, as opposed to building a new treebank from scratch, given the considerable amount of work involved in the harmonization and projection processes. Let us therefore make an attempt at quantifying the gains and balancing them against the disadvantages.

By reusing all the annotation in SUC and the syntactic annotation in Talbanken, we save all the work needed to manually correct tokenization, sentence segmentation, and morphological annotation of 1.2 million tokens, and syntactic annotation of 350,000 tokens. In addition, we save the work needed to check tokenization and sentence segmentation for 350,000 tokens in Talbanken, minus a few person weeks spent on harmonization. Finally, although the morphological annotation of 350,000 tokens in Talbanken still has to be checked manually, both the efficiency and the accuracy of this process can be improved by making use of the old morphological annotation for consistency checking.

To give just one illustrative example, the string *men* in Swedish can be either a coordinating conjunction (but) or a noun (injury). After projecting the new morphological annotation from SUC to Talbanken, it was found that one occurrence of *men* was tagged as a noun in the old annotation and as a conjunction in the new annotation, whereas the remaining 366 occurrences were tagged as conjunctions in both cases. Unsurprisingly, the single occurrence with inconsistent annotation turned out to be a tagging error, which in this way could be detected and corrected. With very high probability, the remaining 366 occurrences are correctly tagged as conjunctions (since the old annotation has been checked manually) and therefore do not need to be checked.

To sum up, we see that cross-corpus harmonization and annotation projection can lead to substantial gains in the manual work needed to validate segmentation and annotation. This of course has to be weighed against a number of other factors, in particular that the new treebank has to be based on old data (in the case of Talbanken, texts from the 1970s) and that the annotation schemes have to be inherited from at least one of the old corpora. Still, in situations where manual effort has to be minimized, the approach taken appears to be a viable methodology for producing a large-scale treebank from existing resources.

6 Conclusion

In this paper, we have presented ongoing work to produce a large treebank of Swedish by reusing existing annotated corpora. A key component in the bootstrapping methodology is the use of cross-corpus harmonization and

annotation projection. Although many of the details involved in the process are dependent on properties of the pre-existing corpora, we believe that the general approach can be applied to other reusable resources.

References

Brants, Thorsten. 2000. TnT - a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP-00)*, Seattle, Washington, USA.

Einarsson, Jan. 1976a. Talbankens skriftspråkskonkordans. Lund University, Department of Scandinavian Languages.

Einarsson, Jan. 1976b. Talbankens talspråkskonkordans. Lund University, Department of Scandinavian Languages.

Ejerhed, Eva and Gunnel Källgren. 1997. Stockholm Umeå Corpus. Version 1.0. Produced by Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University. ISBN 91-7191-348-3.

Forsbom, Eva. 2005. Big is beautiful: Bootstrapping a pos tagger for swedish. Poster presentation at the GSLT retreat. Gullmarsstrand, January 27-29.

Megyesi, Beata. 2002. *Data-Driven Syntactic Analysis: Methods and Applications for Swedish.* Ph.D. thesis, KTH: Department of Speech, Music and Hearing.

Nilsson, Jens, Johan Hall, and Joakim Nivre. 2005. MAMBA meets TIGER: Reconstructing a Swedish treebank from Antiquity. In Peter Juel Henrichsen, editor, *Proceedings of the NODALIDA Special Session on Treebanks.*

Nivre, Joakim. 2000. Sparse data and smoothing in statistical part-of-speech tagging. *Journal of Quantitative Linguistics*, 7:1–17.

Nivre, Joakim and Johan Hall. 2005. MaltParser. http://www.msi.vxu.se/users/ nivre/research/MaltParser.html.

Nivre, Joakim, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1392–1395.

Teleman, Ulf. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska.* Studentlitteratur.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476