Data-Driven Methods for Building a Swedish Treebank

Beáta Megyesi & Rolf Carlson
Centre for Speech Technology
Royal Institute of Technology
[bea|rolf]@speech.kth.se

1 Introduction

There are two types of techniques that are commonly used to build syntactically annotated corpora: rule-based techniques and data-driven methods. Rule-based systems are manually constructed and often time-consuming to develop in order to achieve high accuracy. Data-driven methods, on the other hand, can be easily adapted to many different natural language processing tasks, such as part-of-speech (PoS) tagging and shallow parsing within a short period of time given some correctly annotated training data. In this paper, we will describe evaluation and comparison of three different data-driven algorithms when applied to shallow parsing of Swedish. Additionally, based on the results, we propose a method for a fast and efficient development of a treebank.

2 Building Data-driven Shallow Parsers

To build data-driven shallow parsers for Swedish, we adapted Abney's idea of approaching parsing by dividing the sentence into syntactically related non-overlapping groups of words, so called chunks (Abney, 1991). A chunk is a major phrase category consisting of the phrasal head and its modifiers on the left hand side. Additionally, the chunks were grouped together in order to represent the hierarchical structure of the sentence by describing the whole constituent structure the words belong to and thereby giving the shallow parse tree of each sentence.

Since there exist efficient machine learning algorithms that have been applied to PoS tagging of a variety of languages with great success, we applied and tested three of the most successive data-driven PoS taggers to shallow parsing of Swedish. These taggers are based on hidden Markov modeling (Brants, 2000), maximum entropy learning (Ratnaparkhi, 1996), and transformation-based learning (Brill, 1994).

Given these algorithms, three different aspects need to be addressed in order to build a data-driven shallow parser: the data used for training and test, the choice and the representation of the target classes that the algorithms have to learn to predict, and the attributes or features included in learning.

Since there exists no treebank for Swedish, we have to construct parsed texts that can serve as training data and benchmark corpus. For this purpose, an Earley Parser, SPARK (Aycock, 1998) is used together with a context-free grammar for Swedish developed by Megyesi. The second version of the Stockholm-Umeå corpus (Ejerhed, *et al.*, 1992) annotated with PAROLE tags served as input to the rule-based parser. The PoS tagged texts were parsed by SPARK using nine phrase categories. Some categories correspond to the chunks, while other categories are designed to handle arguments on the right hand side of the phrasal head. The nine phrase types are: adverb phrase, adjective phrase, noun phrase, numerical expressions, prepositional phrase, verb clusters, infinitive phrase, maximal projection of noun phrase and adjective phrase, see Megyesi (2002) for more details.

Each phrase type is represented with an additional tag marking position information in a manner similar to that proposed by Ramshaw and Marcus (1995) and used in the CoNLL-2000 competition:

XB – the initial word of the phrase X
XI – non-initial word inside the phrase X
O – word outside of any phrase.

Thus, each word and punctuation mark in a sentence is accompanied by a tag which indicates the phrase structure the token belongs to in the parse tree together with the position information. Since a token may belong to several phrases, it can have several tags.

The representation is illustrated in the example below for the Swedish equivalent of the sentence "Everybody should read Pilger's very good books about politics" represented first by parenthesis notation, and second by PoS and phrase tags.

[NP Alla NP] [VC borde läsa VC] [NPMAX [NP Pilgers [AP [ADVP mycket ADVP] bra AP] böcker NP] [PP om [NP politik NP] PP] NPMAX].

Various experiments were carried out in order to ascertain how well the different data-driven algorithms can learn the whole hierarchical constituent structure of the word sequences, and to examine what kinds of linguistic information shall be included in learning to achieve the best result. Particular

WORD	POS + MORPHOLOGY as PAROLE tags	PHRASE TAGS
Alla	PI@0P0@S	NPB
borde	V@IIAS	VCB
läsa	V@N0AS	VCI
Pilgers	NP00G@0S	NPB_NPMAXB
mycket	RGPS	ADVPB_APB_NPI_NPMAXI
bra	AQP00N0S	API_NPI_NPMAXI
böcker	NCUPN@IS	NPI_NPMAXI
om	SPS	PPB_NPMAXI
politik	NCUSN@IS	NPB_PPI_NPMAXI
.	FE	0

attention has been directed to the various types of input features that the taggers learn from, such as words, PoS tags, and a combination of both.

The results show that all three data-driven algorithms can be efficiently used for shallow parsing of texts, given that PoS information only, i.e. without the presence of words, is included in the training data. By excluding the lexical information during learning and testing, all classifiers obtain an accuracy above 92% when the training set contains at least 50k tokens (see Megyesi, 2002 for more details).

Thus, in order to achieve high accuracy in parsing, we need good PoS taggers for Swedish. Currently, four data–driven part–of–speech (PoS) taggers applied to PoS tagging of Swedish exist at our department. These taggers are: hidden Markov modeling (Brants, 2002), maximum entropy learning (Ratnaparkhi, 1996), memory-based learning (Daelemans et al., 1996), and transformation-based learning (Brill, 1994). All taggers have been tested, evaluated and compared, see Megyesi (2001). They have an average accuracy of 95%. However, by using the ensemble technique with standard consensus vote procedure, the error rate can be decreased to 1% (Megyesi, forthcoming).

3 Elimination of Errors

Since the rule-based parser SPARK introduced some errors in both the training data and benchmark, the presence of the noise can be assumed to influence the result of the classifiers. The noise can be assumed to involve a

simplification of grammatical structures since it was introduced by a rule-based system. To investigate the noise effect on the classifiers some parts of the training and benchmark data were manually corrected and the algorithms were retrained on the noise free data set. The results show that the error rate increases when the classifiers are trained on noise free data. Most of the errors are due to that the rule-based parser wrongly attached the prepositional phrase to a noun phrase, thereby overgenerating maximal projections of noun phrases. If we eliminate the NPMAX target class, final results can be greatly improved. The maximal projections of noun phrases could be taken care of by another, separate classifier at a later stage.

4 Building a Treebank

Based on the results reported in this paper we can conclude that data-driven methods can be efficiently used for shallow parsing Swedish, in particular when we have a part-of-speech tagged text. Thus, we can let the best PoS tagger, or combination of taggers available annotate the text with PoS tags. The next step is to extract the PoS labels from the text but keep the sentence division, and let the parser annotate the PoS sequences. The only thing then remaining to do would be to put the words back into the parsed PoS sequences. The advantage of the development of several classifiers is the possibility to use those in error detection and reduction by using ensemble methods with voting procedure. The classifiers often make different errors, especially in cases when the data contains inconsistencies or vague target classes. By letting human judges inspect those positions where the classifiers disagree, the vague grammatical cases can be detected and corrected. Also, the vote of the human judges and the classifiers together can be assumed to lead to error reduction.

5 Ongoing Work

The grammatical analysis given by the data-driven classifiers presented in this study represent neither clauses, nor syntactic functions. Data-driven classifiers handling these types of analysis are currently under development at our department by using the existing MAMBA (Teleman, 1974) corpus as basis for training and benchmark corpus with promising results. Additionally, building models for spoken Swedish is also of great interest to us in order to be able to model the structuring of speech in various communicative situations. We are, in particular, interested in the relationship between

prosody and linguistic structure with a special attention directed to syntax and discourse (see Carlson et al., 2002). We strongly believe that speech research would greatly benefit from a treebank containing both text and speech. Therefore, we aim to improve our existing models to speech as well.

6 References

Abney, S. 1991. Parsing by Chunks. In *Prin-ciple-Based Parsing.* Kluwer Academic Publ.

Aycock, J. 1998. Compiling Little Languages in Python. In *Proceedings of 7th International Python Conference.*

Brants, T. 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference.* Seattle, Washington, USA.

Brill, E. 1994. Some Advances in Rule-Based Part of Speech Tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94).* Seattle, Washington.

Carlson, R., Granström, B., Heldner, M., House, D., Megyesi, B., Strangert, E., Swerts, M. 2002. Boundaries and groupings - the structuring of speech in different communicative situations: a description of the GROG project. In *Proceedings of Fonetik 2002*, TMH-QPSR Vol 44, pp. 65-69.

Daelemans, W., Zavrel, J., Berck, P., and Gillis, S.E. 1996. MBT: a Memory-Based Part of Speech Tagger-Generator. In *Proceedings of Fourth Workshop on Very Large Corpora (VLC-96).* pp. 14-27. Copenhagen, Denmark.

Ejerhed, E., Källgren, G., Wennstedt, O., & Åström, M. 1992. *The Linguistic Annotation System of the Stockholm-Umeå Project.* Dept. of General Linguistics, University of Umeå.

Megyesi, B. 2001. Comparing Data-Driven Learning Algorithms for PoS Tagging of Swedish. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2001).* pp. 151-158, Carnegie Mellon University, Pittsburgh, PA, USA, June 3 and 4 2001.

Megyesi, B. 2002. Shallow Parsing with PoS Taggers and Linguistic Features. *Journal of Machine Learning Research: Special Issue on Shallow Parsing*, JMLR (2): 639-668. MIT Press

Megyesi, forthcoming. Data-Driven Morpho-Syntactic Analysis of Swedish. Ph.D. thesis. Manuscript.

Ramshaw, L. A. and Marcus, M. P. 1995. Text Chunking Using Transformation-Based Learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora.* ACL.

Ratnaparkhi, A. 1996. A Maximum Entropy Model for Part-of-Speech Tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-96).* Philadelphia, PA, USA.

Teleman, U. 1974. *Manual fr grammatisk beskrivning av talad och skriven svenska.* Lund: Studentlitteratur.

Zavrel, J. and Daelemans, W. 1999. Recent Advances in Memory-Based Part-of-Speech Tagging. In *Proceedings of the VI Simposio Internacional de Comunicacion Social.* Cuba.