Annotating Wall Street Journal Texts Using a Hand-Crafted Deep Linguistic Grammar

Valia Kordoni & **Yi Zhang**
DFKI GmbH and Dept. of Computational Linguistics, Saarland University
66041 Saarbrücken, GERMANY
{kordoni,yzhang}@coli.uni-sb.de

Abstract

This paper presents an on-going effort which aims to annotate the Wall Street Journal sections of the Penn Treebank with the help of a hand-written large-scale and wide-coverage grammar of English. In doing so, we are not only focusing on the various stages of the semi-automated annotation process we have adopted, but we are also showing that rich linguistic annotations, which can apart from syntax also incorporate semantics, ensure that the treebank is guaranteed to be a truly sharable, re-usable and multi-functional linguistic resource[†].

1 Introduction

The linguistic annotation of a corpus is the practice of adding interpretative linguistic information in order to give "added value" to the corpus. Linguistically annotated corpora have been shown to help in many kinds of automatic language processing or analysis. For example, corpora which have been POS-tagged can automatically yield frequency lists or frequency dictionaries with grammatical classification. Another important use for linguistically annotated corpora is in the area of automatic parsing. In terms of re-usability of linguistic annotations, what is to be advocated here is that – as long as the annotation provided is a kind useful to many users - an annotated corpus gives "value added" because it can be readily shared by others, apart from those who originally added the annotation. In short, a linguistically annotated corpus is a sharable resource, an example of the electronic resources increasingly relied on for research and study in the humanities and social sciences.

In this paper, we present an on-going project whose aim is to produce rich syntactic and semantic annotations for the Wall Street Journal (henceforward WSJ) sections of the Penn Treebank (henceforward PTB; Marcus et al. (1993)). The task is being carried out with the help of the English Resource Grammar (henceforward ERG; Flickinger (2002)), which is a hand-written grammar for English in the spirit of the framework of Head-driven Phrase Structure Grammar (henceforward HPSG; Pollard and Sag (1994)).

2 Background & Motivation

The past two decades have seen the development of many syntactically annotated corpora. There is no need to defend the importance of treebanks in the study of corpus linguistics or computational linguistics here. Evidently, the successful development of many statistical parsers is attributed to the development of large treebanks. But for parsing systems based on hand-written grammars, treebanks are also important resources on the base of which statistical parse disambiguation models have been developed.

The early treebanking efforts started with manual annotations which are time-consuming and error-prone procedures. For instance, the WSJ sections of the PTB has taken many person years to get annotated. Similar efforts have been carried out in many more languages, as can be seen in the cases of the German Negra/Tiger Treebank (Brants et al., 2002), the Prague Dependency Treebank (Hajič et al., 2000), TüBa-D/Z[1], etc. Although many of these projects have stimulated research in various sub-fields of computational linguistics where corpus-based empirical methods are used, there are many known shortcomings of the manual corpus annotation approach.

Many of the limitations in the manual treebanking approach have led to the development of several alternative approaches. While annotating linguistically rich structures from scratch is clearly inpractical, it has been shown that the different

[†]We thank Dan Flickinger and Stephan Oepen for their support with the grammar and treebanking software used in this project. The second author is supported by the German Excellence Cluster: Multimodal Computing & Interaction.

[1]http://www.sfs.nphil.uni-tuebingen.de/en_tuebadz.shtml

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 170–173,
Suntec, Singapore, 6-7 August 2009.

structures in various linguistic frameworks can be converted from annotated treebanks to a different format. And the missing rich annotations can be filled in incrementally and semi-automatically. This process usually involves careful design of the conversion program, which is a non-trivial task. In very recent years, based on the treebank conversion approach and existing manually annotated treebanks, various "new" annotations in different grammar frameworks have been produced for the same set of texts. For example, for the WSJ sections of the PTB, annotations in the style of dependency grammar, CCG, LFG and HPSG have become available. Such double annotations have helped the cross-framework development and evaluation of parsing systems. However, it must be noted that the influence of the original PTB annotations and the assumptions implicit in the conversion programs have made the independence of such new treebanks at least questionable. To our knowledge, there is no completely independent annotation of the WSJ texts built without conversion from the original PTB trees.

Another popular alternative way to aid treebank development is to use automatic parsing outputs as guidance. Many state-of-the-art parsers are able to efficiently produce large amount of annotated syntactic structures with relatively high accuracy. This approach has changed the role of human annotation from a labour-intensive task of drawing trees from scratch to a more intelligence-demanding task of correcting parsing errors, or eliminating unwanted ambiguities (cf., the Redwoods Treebank (Oepen et al., 2002)). It is our aim in this on-going project to build a HPSG treebank for the WSJ sections of the PTB based on the hand-written ERG for English.

3 The Annotation Scheme

3.1 Grammars & Tools

The treebank under construction in this project is in line with the so-called dynamic treebanks (Oepen et al., 2002). We rely on the HPSG analyses produced by the ERG, and manually disambiguate the parsing outputs with multiple annotators. The development is heavily based on the DELPH-IN[2] software repository and makes use of the English Resource Grammar (ERG; Flickinger (2002), PET (Callmeier, 2001), an efficient unification-based parser which is used in our project for parsing the WSJ sections of the PTB, and [incr tsdb()] (Oepen, 2001), the grammar performance profiling system we are using, which comes with a complete set of GUI-based tools for treebanking. Version control system also plays an important role in this project.

3.2 Preprocessing

The sentences from the Wall Street Journal Sections of the Penn Treebank are extracted with their original tokenization, with each word paired with a part-of-speech tag. Each sentence is given a unique ID which can be used to easily look up its origin in the PTB.

3.3 Annotation Cycles

The annotation is organised into iterations of parsing, treebanking, error analysis and grammar/treebank update cycles.

Parsing Sentences from the WSJ are first parsed with the PET parser using the ERG. Up to 500 top readings are recorded for each sentence. The exact best-first parsing mode guarantees that these recorded readings are the ones that have "achieved" highest disambiguation scores according to the current parse selection model, without enumerating through all possible analyses.

Treebanking The parsing results are then manually disambiguated by the annotators. However, instead of looking at individual trees, the annotators spend most of their effort making binary decisions on either accepting or rejecting constructions. Each of these decisions, called discriminants, reduces the number of the trees satisfying the constraints (see Figure 1). Every time a decision is made, the remaining set of trees and discriminants are updated simultaneously. This continues until one of the following conditions is met: i) if there is only one remaining tree and it represents a correct analysis of the sentence, the tree is marked as gold; ii) if none of the remaining trees represents a valid analysis, the sentence will be marked as "rejected", indicating an error in the grammar[3]; iii) if the annotator is not sure about any further decision, a "low confidence"

[2] http://www.delph-in.net/

[3] In some cases, the grammar does produce a valid reading, but the disambiguation model fails to rank it among the top 500 recorded candidates. In practice, we find such errors occuring frequently during the first annotation circle, but they diminish quickly when the disambiguation model gets updated.

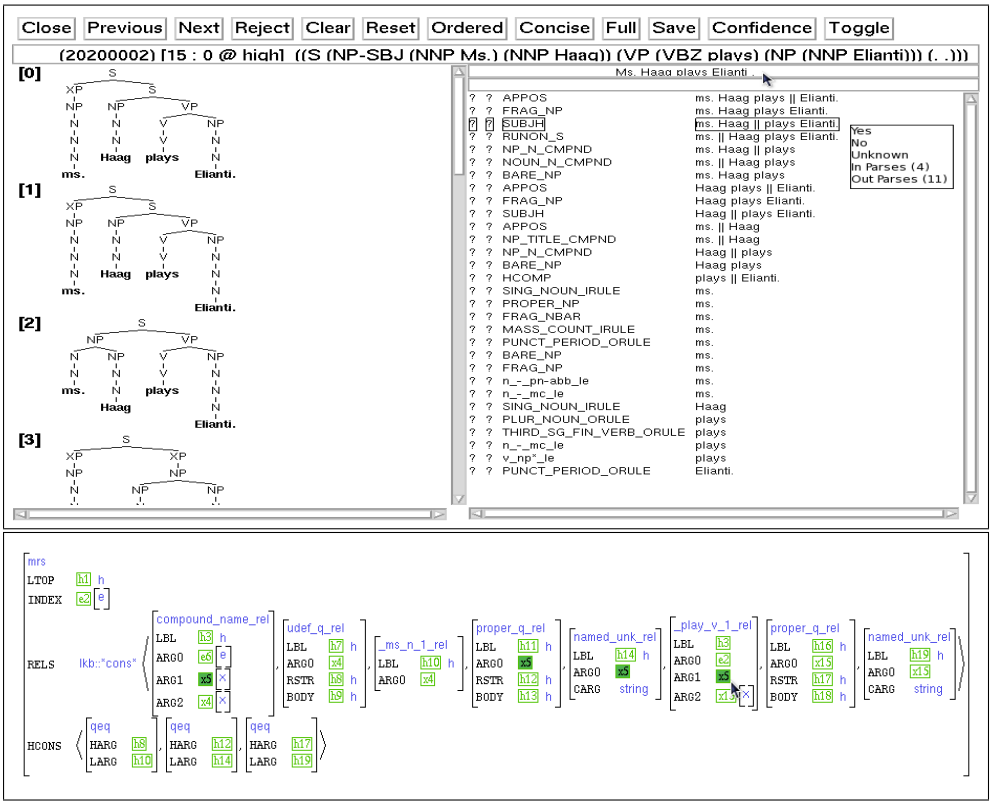

Figure 1: Treebanking Interface with an example sentence, candidate readings, discriminants and the MRS. The top row of the interface is occupied by a list of functional buttons, followed by a line indicating the sentence ID, number of remaining readings, number of eliminated readings, annotator confidence level, and the original PTB bracket annotation. The left part displays the candidate readings, and their corresponding IDs (ranked by the disambiguation model). The right part lists all the discriminants among the remaining readings. The lower part shows the MRS of one candicate reading.

state will be marked on the sentence, saved together with the partial disambiguation decisions. Generally speaking, given n candidate trees, on average $\log_2 n$ decisions are needed in order to fully disambiguate. Given that we set a limit of 500 candidate readings per sentence, the whole process should require no more than 9 decisions. If both the syntactic and the MRS analyses look valid, the tree will be recorded as the gold reading for the sentence. It should be noted here that the tree displayed in the treebanking window is an abbreviated representation of the actual HPSG analysis, which is much more informative than the phrase-structure tree shown here.

Grammar & Treebank Update While the grammar development is independent to the treebanking progress, we periodically incorporate the recent changes of the grammar into the treebank annotation cycle. When a grammar update is incorporated, the treebank will be updated accordingly by i) parsing all the sentences with the new grammar; ii) re-applying the recorded annotation decisions; iii) re-annotating those sentences which are not fully disambiguated after step ii. The extra manual annotation effort in treebank update is usually small when compared to the first round annotation.

Another type of update happens more frequently without extra annotation cost. When a new portion of the corpus is annotated, this is used to retrain the parse disambiguation model. This improves the parse selection accuracy and reduces the annotation workload.

3.4 Grammar coverage & robust parsing

Not having been specifically tuned for the newspaper texts, the ERG achieved out-of-box coverage of over 80% on the WSJ dataset. While this is a respectably high coverage for a hand-written precision grammar, the remaining 20% of the data is not covered by the first round of annotation. We plan to parse the remaining data using a less-restrictive probabilistic context-free grammar extracted from the annotated part of the treebank. The PCFG parser will produce a pseudo-derivation tree, with which robust unifications can be applied to construct the semantic structures (Zhang and Kordoni,

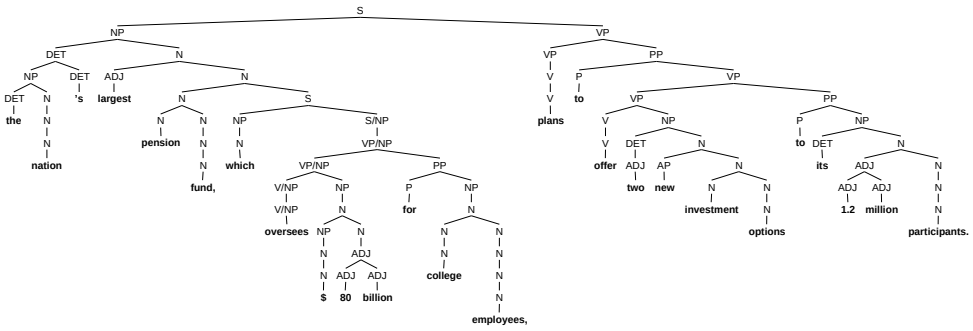

Figure 2: An example tree including a ”heavy” NP-subject, a relative clause, and noun-noun compounds

2008).

3.5 Multiple annotations

To speed up the annotation, the project employs three annotators. They are assigned with slightly overlapping sections of the WSJ dataset. The overlapping part allows us to measure the inter-annotator agreement for the purpose of quality control. To estimate the agreement level, the WSJ Section 02 has been completely annotated by all three annotators. Analysis shows that the annotators reach exact match agreement for around 50% of the sentences. Many disagreements are related to subtle variations in the linguistic analyses. The agreement level shows improvement after several treebanker meetings. For future development, a more fine-grained disagreement assessment is planned.

4 Discussion

The WSJ section of the PTB is not only a challenging corpus to parse with a hand-written grammar. It also contains various interesting and challenging linguistic phenomena. Figure 2, for instance, shows the syntactic analysis that the ERG produces for a sentence which includes a “heavy” NP (noun phrase) containing a relative clause introduced by *which* in the subject position, as well as many interesting compound nouns whose interpretations are missing from the PTB annotation.

The newly annotated data will be also very important for the cross-framework parser development and evaluation. While almost all of the state-of-the-art statistical parsers for English use PTB annotations for training and testing, it would be interesting to see whether a comparable level of parsing accuracy can be reproduced on the same texts when re-annotated independently.

References

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The tiger treebank. In *Proceedings of the workshop on treebanks and linguistic theories*, pages 24–41.

Ulrich Callmeier. 2001. Efficient parsing with large-scale unification grammars. Master’s thesis, Universität des Saarlandes, Saarbrücken, Germany.

Dan Flickinger. 2002. On building a more efficient grammar by exploiting types. In Stephan Oepen, Dan Flickinger, Jun’ichi Tsujii, and Hans Uszkoreit, editors, *Collaborative Language Engineering*, pages 1–17. CSLI Publications.

Jan Hajič, Alena Böhmová, Eva Hajičová, and Barbora Vidová-Hladká. 2000. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In A. Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 103–127. Amsterdam:Kluwer.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002. The LinGO Redwoods treebank: motivation and preliminary applications. In *Proceedings of COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*, Taipei, Taiwan.

Stephan Oepen. 2001. [incr tsdb()] — competence and performance laboratory. User manual. Technical report, Computational Linguistics, Saarland University, Saarbrücken, Germany.

Carl J. Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, USA.

Yi Zhang and Valia Kordoni. 2008. Robust Parsing with a Large HPSG Grammar. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco.