A Treebank-Driven Approach to Semantic Lexicons Creation

Kiril Simov, Petya Osenova
BulTreeBank Project
http://www.BulTreeBank.org
Linguistic Modelling Laboratory, Bulgarian Academy of Sciences
Acad. G. Bonchev St. 25A, 1113 Sofia, Bulgaria
kivs@bultreebank.org, petya@bultreebank.org

1 Introduction

In this paper we aim at showing how the information from a richly annotated treebank can be used for facilitating the construction of a semantic lexicon when such a lexicon lacks for a certain language. We demonstrate this idea with the Bulgarian treebank (BulTreeBank).

The structure of the paper is as follows: the next section briefly describes the levels of linguistic interpretation in the treebank. In section 3 we present the model of a semantic lexicon which we are using. Section 4 outlines the algorithm for the extraction of the semantic information from the treebank. The last section concludes the paper.

2 The Levels of Linguistic Interpretation in BulTreeBank

Our treebank (200 000 words) is a part of a morphosyntactic corpus (1 000 000 words). It is manually processed and consists of the following analytical levels:

1. Token level - the tokens are divided into common words, names, abbreviations, numericals, symbols, punctuation.

2. Morphosyntactic level - the correct POS tag with the appropriate characteristics is selected among alternatives (if any). At this level we designate different semantic types of adverbials: time, place, manner, quantity, modal and named entities: person, organization, local, other. This classification

helps us to form the verb frames at later stages. In contrast to [Hajič, 2003] we do not exclude them from the 'inner participants' list.

3. HPSG-oriented syntactic level - it combines the constituent representation, grammatical roles assignment and head-dependent distinction. For each phrasal domain we annotated the role of the dependent element(s): complement, subject, adjuncts.

4. Intrasentential co-reference relations level - here we rely on the structure-sharing mechanisms in HPSG and we assume different relations between nominals or nominalized elements that reflect the phenomena binding, pro-drop, control etc.

In NLP literature there are a number of schemes for annotating more complex co-reference relations in treebanks, see [Kunz and Hansen-Schirra 2003], and [Kučová and Hajičová (in press)] among others. For the moment we have annotated the following referential relations: equality, subset-of and member-of (we have not annotated relations like part-of). We capture all main co-references of the following syntactic representations: subject and object relations, reflexivity, possession, clitic-doubled structures, secondary predicated adjectives with the subject or the object. Also we represent co-reference between synonymic expressions, changed referring expressions in direct-indirect speech, nominalizations. Part of the co-reference relations within a sentence are not explicated because they can be easily inferred from the syntactic structure like co-reference between the relative pronoun and the head noun when a relative clause modifies a noun phrase.

HPSG theory implies a lexicon, which in a general way reflects the idea of the 'frame-semantic approach' as stated in [Lowe, Baker and Fillmore, 1997] and [Kingsbury and Palmer, 2003]. For instance, the semantics of the verb *give* will include a representation of the relation 'give' with corresponding arguments:

$$\left[\begin{array}{ll} \text{CONTENT} & \left\{\left[\begin{array}{ll} \text{REL} & \text{give} \\ \text{ARG1} & \text{giver} \\ \text{ARG2} & \text{given} \\ \text{ARG3} & \text{givee} \end{array}\right]\right\} \end{array}\right]$$

In order to supply this type of information and/or to make it more concrete, we use two dictionaries in the semantic annotation of the treebank: the machine-readable Valency lexicon and the Seed Semantic lexicon[1]. On the one hand, these

[1]We call this lexicon a Seed Semantic lexicon because it contains only about 3000 nouns and does not contain other parts of speech. But otherwise it follows the chosen model for the sematic dictionary we want to construct.

lexicons represent the model of argument-predicate and semantic representation. But, on the other hand, they are far from covering all the treebank data.

In the following section we describe the model of the semantic dictionary which we follow in our work.

3 The Semantic Lexicon

Our aim is to show how the construction of a semantic lexicon can be facilitated by using the annotated linguistic relations in the treebank as supplementary to the available, but incomplete Valency and Seed Semantic lexicons.

In our view, an elaborate semantic lexicon has to contain both pieces of information: subcategorization and semantics. Additionally, the argument positions in the subcategorization need to be syntactically and semantically constrained.

As it was mentioned above, semantic information plays a crucial role in the process of parse discrimination on which the construction of our treebank depends. Thus, in order to support the selectional restrictions imposed by the valency dictionary and to facilitate its usage, we decided to compile a semantic lexicon along the guidelines of SIMPLE project — [Lenci, A. et. al., 2000]. Generally, the structure of the lexical items follow the structure of predetermined templates which contain several fields and relations between them. For consistency each template is connected to a concept in the SIMPLE core ontology. It is worth mentioning that we follow an extended variant of the core ontology, namely - with taking into account Pustejovsky's qualia. Also the SIMPLE model of semantic lexicon includes representation of the valency of the words together with constraints over the arguments. Another important advantage of the SIMPLE model is that it is compatible with WordNet model of a semantic lexicon. Thus it is a good model for the creation of a semantic lexicon for HPSG. Our goal is to create such a dictionary for Bulgarian with wide coverage.

In our work we used two lexicons which were at our disposal before the experiment with the extraction of additional information from the treebank:

The Valency Dictionary consists of 1000 most frequent verbs and their valency frames. Each verb has a gloss and one or more frames. Each frame defines the number and the kind of the arguments imposing morphosyntactic and semantic restrictions over them. The original semantic restrictions over the arguments are extracted and matched against the SIMPLE core ontology. The frames of the most frequent verbs are compared to the corpus data (the morphologically annotated corpus) and repaired if necessary (new frames are added, some of the existing frames are deleted or fine-grained). We envisage to enlarge the coverage of this dictionary with the help of some derivational means, such as the verb prefixes.

The second lexicon contains 3000 of the most frequent nouns. They are classified with respect to the ontological hierarchy without specifying the synonymic relations between them. Also, the named entities and the adjectives have been classified with respect to the same ontology. We call this dictionary Seed Semantic lexicon.

In order to extend both lexicons we use the information encoded in the treebank. First, we annotated all the words in the treebank with the information available in the lexicons. Then we used the syntactic and co-referential information encoded within the treebank in order to disambiguate the annotated words. Afterwards, we collected the new information and inspected it manually.

4 The Algorithm

In order to extend the coverage of the semantic information, we decided to rely on the following corpus-based 'scratch' method along with the classification of the words against the SIMPLE ontology:

1. Verb annotation.

   Each verb in a sentence of the treebank is annotated with the frame descriptions from the Valency dictionary (if there is a lexical entry for the verb). Each of the arguments in a frame of the verb is connected to some of the verb dependents in the syntactic annotation. This is possible for the subject, the direct object and the indirect object. Note that sometimes there is a mapping from an indirect object in the Valency dictionary to an adjunct role in the annotation of the treebank.

2. Noun annotation.

   Each noun in the treebank is annotated with all the semantic classes in the semantic lexicon (if there is a lexical entry). On the one hand, this information is important for the verbs to select the appropriate arguments. On the other hand, it helps to classify named entities with better accuracy.

3. Disambiguation.

   This step is based on the idea of lexical chains: a set of coherently interrelated words in the text as presented in [Hirst and St-Onge 1998]. The connection between the words is defined on the basis of lexical relations like synonymy, hyperonymy, meronymy etc, which are classified as extra-strong, strong, medium, etc. The words in a lexical chain are connected with relations that represent different degrees of ontological similarity. We focused on

extra-strong (i.e. literal repetition) and some of the strong relations, namely - the first type: when there is a synset (a set of synonyms) common to two different words, such as human and person, and the third type, namely when there is some kind of link between a synset associated with each word if one word is a compound word or a phrase that includes the other, such as *school* and *private school*. The second type (when there is a horizontal link between synsets associated with two different words, such as pre- cursor and successor) as well as medium strong relations are not considered, because apart from the upper part, the rest of the hierarchy is rather flat and therefore - unreliable. In the treebank we define lexical chains on the basis of co-referential relations and apply the idea of ontological similarity between the co-referent elements.

For each verb annotated with more than one frame we check whether some of the arguments in some of the frames disagree with the morphological and/or semantic information of the head noun of the corresponding element in the syntactic structure. If such a disagreement exists we delete the frame from the annotation of the verb.

For each noun annotated with more than one semantic class we check two things: (1) whether some of the semantic classes disagree with the selectional restrictions of some frame of the verb in the sentence (if the noun is a head noun mapped to some of the arguments in the frame). In this case we remove the class from the annotation of the noun; (2) we are using the coreferential relations with nouns or pronouns[2] to rule out more semantic classes.

These disambiguation rules are applied only when there are sure indicators for them, otherwise we leave the ambiguity in the annotation unresolved.

4. Classification.

   We classify the nouns in the text in equivalent classes on the basis of their participation in a coreferential relation or their headedness towards the same argument for different occurrences of the same verb. If there is an ambiguity, several equivalent classes are constructed.

5. Manual validation.

   An expert manually checks over the equivalent classes and creates appropriate lexical entries. For example, in the phrase ‘to write an application’, ‘application’ is added to the semantic class of the word ‘letter’.

[2]We assume that two semantic classes agree with each other if they are the same or one is a superclass of the other.

5 Conclusion

Thus, a semantic lexicon can be built in a bootstrapping manner. It is unordered (i.e. most of the hypernymic, synonymic, meronymic relations are hidden), but lexically rich. Later, gradually, the lexical relations will be added to this lexicon.

Note that the Treebank contains implicitly other predicate-argument patterns, which are extracted and processed as well. Here we have in mind not only all the cases of type verb-dependent, but also some fixed phrases: idioms, parenthetical expressions, verbs of saying which uniquely determine the semantic classes of their syntactic context (dependent elements or heads).

References

[Hajič, 2003] Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová and Petr Pajas. 2003. *PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation.* In the Proceedings of The Second Workshop on Treebanks and Linguistic Theories. Sweden, pp. 57–68.

[Hirst and St-Onge 1998] Graeme Hirst and David St-Onge. *Lexical chains as representations of context for the detection and correction of malapropisms.* In: Christiane Fellbaum (editor), WordNet: An electronic lexical database, Cambridge, MA: The MIT Press, 1998.

[Kingsbury and Palmer, 2003] Paul Kingsbury and Martha Palmer. 2003. *PropBank: the Next Level of TreeBank*. In the Proceedings of The Second Workshop on Treebanks and Linguistic Theories. Sweden.

[Kučová and Hajičová (in press)] Lucie Kučová and Eva Hajičová*Coreferential relations in the Prague Dependency Treebank.* In the Proceedings from FDSL5, Leipzig, 26–28 November 2003.

[Kunz and Hansen-Schirra 2003] Kerstin Kunz and Silvia Hansen-Schirra. 2003. *Coreference Annotation of the TIGER Treebank*. In the Proceedings of The Second Workshop on Treebanks and Linguistic Theories. Sweden, pp. 221–224.

[Lenci, A. et. al., 2000] Alessandro Lenci et. al. 2000. *SIMPLE Work Package 2 — Linguistic Specifications, Deliverable D2.1*, ILC-CNR, Pisa.

[Lowe, Baker and Fillmore, 1997] John B. Lowe, Coilin F. Baker, Charles J. Fillmore. 1997. *A Frame-Semantic Approach to Semantic Annotation*. In Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?. USA.