Types of errors in the automatic syntactic parsing of Romanian

Verginica Barbu Mititelu, Elena Irimia

Research Institute for Artificial Intelligence "Mihai Drăgănescu"
Romanian Academy
vergi@racai.ro, elena@racai.ro

Abstract

Within the large effort of creating language resources for Romanian, we have been working on the development of a core treebank. This is meant to contain 5000 sentences: 1000 sentences from each of the following domains: juridical, journalistic, belletristic, academic (literary critique and history), medical; the sentences were selected from ROMBAC, a Romanian balanced corpus (Ion et al, 2012) ; the selection criterion was that each sentence should contain one or more of the most frequent verbs in the corpus. This collection of sentences was meant to be completely and correctly annotated with dependency relations and to serve as a gold standard for training a statistical parser and consequently developing a larger treebank for Romanian.

As we do not have access to a language-aware, rule-based parser for Romanian and as the hand annotation of 5000 sentences is a time- and effort-wise overwhelming activity, we took advantage of a statistical parser largely and successfully used by the research community: MaltParser (Nivre and Hall, 2005). Together with a Spanish syntactic statistical model (following a methodology already tested for another Romance language, Catalan, see (Arias et al., 2014)), MaltParser was used for the syntactic annotation of 500 Romanian sentences. Afterwards, they were manually corrected, thus obtaining a gold standard micro-corpus. These sentences constituted the training material for a Romanian syntactic statistical model, which was hereinafter used to annotate the 4500 remaining Romanian sentences, regularly retraining the model after each 500 manually corrected sentences.

In the process of manually correcting the automatically annotated sentences as well as in the automatic evaluations of the progress using the evaluation software provided by the CONLL[1] dependency parsing shared tasks 2006-2007, we were interested in an analysis of the errors produced by the statistical parser in different training stages.

The first and most relevant remark for the methodology we adopted is that the number of errors decreases with each retraining of MaltParser on the enlarged hand-validated corpus. In our presentation we will show concrete data, first for the amount of errors made by the parser trained on Spanish (as compared with the results of the same parser working with Catalan data), and then for several stages of training and retraining with Romanian data.

A second remark is that in the first stages (i.e., when the training material is scarce) the types of errors are very heterogeneous: they affect all kinds of syntactic phenomena and even punctuation.

However, in further stages, the errors tend to reflect the ambiguities and homonymy in language, as well as its characteristics. We provide here some types of errors identified in our sentences, which we will detail in the presentation: (a) structures made up of [preposition + definite noun + adjective] and [preposition + an adjective bearing a definite article + noun] are analysed in the same way, i.e. as [prepositional modifier + prepositional argument + adjectival modifier]; (b) homonymous adjectives and adverbs are treated identically (as adjectival modifiers); (c) the clitic *SE* (which can be a reflexive/reciprocal clitic, a direct object or a passive marker) is misanalysed; (d) a prepositional phrase predicative is misanalysed as a prepositional modifier.

Several other types of errors will be discussed in the presentation, and relevant examples will be provided.

Analysing these types of errors, we can find solutions which, when implemented, can improve the parser results. Two types of solutions can be formulated: on the one hand, morpho-syntactic restrictions and, on the other hand, semantic and syntactic restrictions, which are the most difficult to implement, though not impossible.

A syntactically annotated corpus (a treebank) is a language resource with intrinsic value, but also of utmost importance for speech synthesis: prosody reflects the syntactic relationships between words phrases.

References

Arias, B., Bel, N., Fomicheva, M., Larrea, I., Lorente, M., Marimon, M., Mila, A., Vivaldi, J. and Padro, M., 2014. Boosting the creation of a treebank, *Proceedings of LREC 2014*, Reykjavik, Iceland

Ion, R., Irimia, E., Ştefănescu, D. and Tufiş, D., 2012. ROMBAC: The Romanian Balanced Annotated Corpus, *Procedings of LREC 2012* Istanbul, Turkey.

Nivre, J., Hall, J. 2005. Maltparser: A language-independent system for data-driven dependency parsing, *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, 137-148.

[1] [1] http://ifarm.nl/signll/conll/