Natural parsing: a psycholinguistically motivated computational language processing model

Gábor Prószéky and Balázs Indig

MTA–PPKE Hungarian Language Technology Research Group, Budapest, Hungary
Pázmány Péter Catholic University, Faculty of Information Technology and Bionics,
50/a Práter street, Budapest, 1083, Hungary
MorphoLogic Ltd., Budapest, Hungary
{proszeky.gabor,indig.balazs}@itk.ppke.hu
http://nlpg.itk.ppke.hu/

We present a new paradigm and framework for syntactic and semantic analysis, which is based on the following principles: (1) *Psycholiguistically* motivated means, our model hold on to the inner algorithms of human language processing as much as possible. (2) As a *performance-based* system the model tries to process 2-3 sentences long coherent texts, where our focus is not the handling of theoretically existing but practically rather rare phenomena. Instead, we consider and try to interpret any – no matter how badly formed or agrammatic – text that appears as a natural language utterance[1]. (3) In our model the parser *processes the text strictly left-to-right incrementally* and does not utilize or reference the parts that succeed the current position. (4) The general architecture of the parser framework is naturally *parallel* from the beginning, in contrast to the traditional approach, where the analysis is generated at the end of a pipeline of modules. Here the program processes the actual word using parallel threads (a morphological analyser thread, a corpus statistics thread, etc.). These threads analyse each word together at the same time and communicate to correct each other's errors and to make a final decision in the analysis. (5) The framework's processing and representational units are not individual sentences, rather *utterances* consisting of one or more sentences. This enables the unified handling of intra- and intersentential anaphoric relations, as there is no psycholinguistic evidence that they are handled differently [2]. (6) In accordance with the principles described so far, in order to be able to handle all the different phenomena at the same time the *representation* is not necessarily a tree, but a connected graph containing different types of (colored) edges. Besides the used resources, we needed a certain description of the main phenomena of the language, namely such a grammar which enumerates all possible roles for linguistic units (e.g. a noun in nominative case or a comma) as in [3]. Therefore we developed the correct handling of these structures in Hungarian by parallel threads. Basically, two basic thread types seemed necessary: an *offer* type thread provides information on the current element (e.g. this element is in nominative case), and a *demand*

[1] This does not mean that there are no utterances which "circumvent" the most likely analysis, sometimes forcing even the human parser to backtrack, but in everyday communication humans seem to follow the maxims of Grice [1], and avoid these constructions.

type is looking for a required element with a specific property (e.g. a possessed noun looks for its possessor, a determiner seeks the NP head, a transitive verb needs its object etc.). The output of our parser contains syntactic and semantic information to identify both participants and events, building a representation of the whole utterance, and formulating statements about who does what, where and when. We illustrate the basic principles and "perform" some of the analysis steps conforming to the theory.

References

1. Grice, H.P., Harman, G.: Logic and conversation. Encino:Dickenson (1975)
2. Pléh, C.: Formal connexity and pragmatic cohesion in anaphora interpretation. In: Text and discourse connectedness. Proceedings of the Conference on Connexity and Coherance, Urbino. pp. 137–52 (1989)
3. Small, S.L.: Word expert parsing. In: Proceedings of the 17th Annual Meeting on Association for Computational Linguistics. pp. 9–13. ACL '79, Association for Computational Linguistics, Stroudsburg, PA, USA (1979), http://dx.doi.org/10.3115/982163.982167