Towards Natural Natural Language Processing
A Late Night Brainstorming

Petr Sojka

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
sojka@fi.muni.cz

Abstract. An essay about mimicking some aspects of language processing in our heads, using information fusion and competing patterns.

1 Introduction

Usual approach to Natural Language Processing separates language processing into word form, morphological, syntactic, semantic and pragmatic levels. Most often processing of these levels are independent, and result of one level is communicated to the other unnaturally disambiguated to cut off less probable (but often linguistically valid) intermediate results (e.g. sentence syntactical parse trees) just to simplify things. Even though ungrammatical sentences are often used for communication between people (English as the second language), they are banned by NLP software. Considerable effort is given to the balancing general purpose corpora to choosen only such text examples, aiming at handling only [syntactically] correct language parts. Given that, for purposes of handling non-polished texts, blogs or even speech, these data resources fail badly, as the tools are trained and fine-tuned to the different type of input than used when processing real [speech] data.

As simple as possible, but not simpler. – Albert Einstein

2 Levels of Processing, Level Interaction and Importance of Complexity

Most of today's available texts is processed on *word form* level only (Google), with PageRank optimizing access to the most credible ones. Texts sharing the same forms are collected together and only the most credible picked up and shown. This suits most, but not all purposes.

Good *morphological* tools allows handling of *all* possible morphological categories, allowing their pruning in further stages of processing (syntactic analysis, etc.). The disambiguation should not be driven by pure statistics in applications like guesser.

Syntactic analysis aiming at only one (best) parsing tree, independently of sentence context, document type and other information is simply wrong.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, pp. 98–100, 2008.

Analysis of potentially billions of possible trees of long sentences is waste of computer resources. Most probable partial parse trees are collected together and shown as the parsed result for further processing. Much better approach is to collect possible sentence segmentations of main building blocks (phrases) and not limiting the analysis outcome to the correct full sentence parses only.

Another bottleneck of today NLP processing is *semantics* handling. Bubble of semantic net starts to blow out, as there is not single semantic representation suitable for all purposes and applications. Linguistic resources are scarce, and wordnets lack many important aspects as deduction and thematic folding (specific domain adaptation and usage, with exception of framenet). Promising formalisms like TIL need necessary language resources.

Little attention is given to *pragmatics* in NLP, as a starter of disambiguation process. Disambiguation, at all levels, should be driven by the final application, deriving from the purpose, classification type of communicated text, intertwisting and backtracking between all levels of linguistic processing. The tools should not be trivialized and should handle multiple lemmata, parses, meanings. Language handling may be as complex as the life it describes, not simpler.

Be as elegant as the situation will allow.

3 Information Fusion and Patterns

The suggested remedy to the current status quo is the design of a modular NLP system for parallel language processing at different levels, allowing mutual interactions and processing between data structures and intermediate results at all levels. The data structures may be not only grammar chunks, framenets and wordnets, but also empirical evidence of language usage (text corpora processed), allowing pattern matching of linguistic data and knowledge representation at various, but interlinked levels.

For several purposes in this scenario, *competing patterns* [1,2] may be used: sentence or phrase segmentation (alphabet is word forms or lemmas), morphological disambiguation patterns (alphabet is gramatical categories and lemmata) [3], and even pragmatics patterns (alphabet being events in time and meaning terms). Same terms in pattern alphabets used will allow for connecting information on different level of language processing – the patterns may be derived from available text and dialogue corporas [4,5]. Pattern storage in the packed digital trie is very compact and allow blindingly fast language data retrieval at the constant time (limited by the pattern length only, e.g. by width of [local] context covered).

4 Conclusion

In this paper, we have presented several thoughts about current state of the art of natural language processing approaches, and have outlined several

directions of improvement towards 'natural' way of text processing, grabbing some metaphors from what is known about language processing in our brains.

Acknowledgments. This work has been partially supported by the Academy of Sciences of Czech Republic under the projects 1ET208050401, 1ET200190513 and by the Ministry of Education of CR within the Centre of basic research LC536 and National Research Programme 2C06009.

References

1. Sojka, P.: Competing Patterns for Language Engineering. In: Sojka, P., Kopeček, I., Pala, K., (Eds.): Proceedings of the Third International Workshop on Text, Speech and Dialogue—TSD 2000. Lecture Notes in Artificial Intelligence LNCS/LNAI 1902, Brno, Czech Republic, Springer-Verlag (2000), pp. 157–162.
2. Sojka, P.: Competing Patterns in Language Engineering and Computer Typesetting. Ph.D. thesis, Masaryk University, Brno (2005).
3. Macháček, D.: Přebíjející vzory ve zpracování přirozeného jazyka (Competing Patterns in Natural Language Processing). Master's thesis, Masaryk University, Brno, Faculty of Informatics, Brno, Czech Republic (2003).
4. Antoš, D., Sojka, P.: Pattern Generation Revisited. In: Pepping, S., (Ed.): Proceedings of the 16th European TEX Conference, Kerkrade, 2001, Kerkrade, The Netherlands, NTG (2001) pp. 7–17.
5. Sojka, P., Antoš, D.: Context Sensitive Pattern Based Segmentation: A Thai Challenge. In: Hall, P., Rao, D.D., (Eds.): Proceedings of EACL 2003 Workshop on Computational Linguistics for South Asian Languages – Expanding Synergies with Europe, Budapest (2003) pp. 65–72.