Two-level probabilistic grammars for natural language parsing
Infante Lopez, G.G.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 8

Conclusions

In this thesis we have applied an abstract model to learn more about natural language parsing. We have surveyed three state-of-the-art probabilistic parsers and we have identified their characteristic features. We have abstracted from these particular features and models to produce and formalize a general, and abstract, language model. This abstract language model provides us with a suitable framework to carry out principled investigations of new directions into parsing based on different parameterizations of the general model.

While reviewing state-of-the-art parsers, we focused on crucial issues like the role of probabilities in the context of parsing, their importance and their possible uses. We have shown that, when used as a filtering mechanism, probabilities can add expressive power to context free grammars, defining a class of tree languages beyond the expressivity of context free grammars. We have shown that it is not possible to decide whether probabilities solve all ambiguities in a language. In addition, we have shown that probabilities can be used for purposes other than the mere filtering of unwanted trees. We have illustrated this claim with some examples, like using probabilities to evaluate the quality of PCFGs and to boost the performance of parsers.

The general language model we have presented is based on W-grammars. We have introduced constrained W-grammars and have augmented them with probabilities. The resulting formalism, probabilistic constrained W-grammars, is the backbone formalism for all our experiments. Like every general model, the suitability of PCW-grammars is mainly given by two aspects: that they can capture multiple existing formalisms in a single formalism, and that they provide a structured framework where new directions of research can be identified and pursued in a principled way.
8.1 PCW-Grammars as a General Model

We have shown that PCW-grammars provide an encompassing formalism for explaining three state-of-the-art language models. PCW-grammars are based on a well-known grammatical framework, and their computational properties are well-understood. It is true that the expressive power of PCW-grammars is significantly lower than that of W-grammars, but their expressiveness is perfectly adequate to capture grammatical formalisms underlying state-of-the-art parsers.

For example, probabilistic constrained W-grammars are capable of capturing bilexical grammars, Markovian context free rules, and stochastic tree substitution grammars. We have described the expressive power of these three formalisms, together with some conditions under which grammars inferred from treebanks are consistent. Despite the similarities between PCW-grammars and PCFGs, there is a fundamental difference between the two: the two-level mechanism of PCW-grammars. This mechanism allowed us to capture these three state-of-the-art natural language models mentioned above, which cannot be done using standard PCFGs only.

The suitability of the general language model provided by PCW-grammars is that it allowed us to compare three apparently different formalisms within the same formal perspective. We have shown that the essence of bilexical grammars and Markovian context free grammars is quite comparable: both are based on approximating bodies of rules using Markov models. We also found similarities between STSGs and Markov rules: both suppose that rule bodies are obtained by collapsing hidden derivations. More concretely, for Markovian rules, a rule body is a regular expression (or a Markov chain, which is equivalent) and STSGs take this idea to the extreme by considering the whole sentence as the yield of a hidden derivation.

8.2 PCW-Grammars as a New Parsing Paradigm

PCW-grammars are not only useful for capturing the formalisms underlying state-of-the-art parsers, but also for suggesting new research directions. These come as a consequence of different instantiations of the parameters of the general model, or by rethinking the set of assumptions the particular instances have made. A brief description of the directions explored in this thesis follows.

Explicit Use of Probabilistic Automata: PCW-grammars allowed us to use general methods for inducing regular languages instead of the usual n-based algorithms. Our experiments along this line lead to two types of conclusions. First, that modeling rules with algorithms other than n-grams does not only produce smaller grammars, but also better performing ones. Second, that the procedure used for
optimizing the parameters of the parser reveals that some POS behave almost deterministically for selecting their dependents, while others do not. This conclusion suggests that splitting classes that behave non-deterministically into homogeneous ones could improve the quality of the inferred automata. We argued that lexicalization and head-annotation seem to take advantage of the properties of splitting.

**Splitting the Training Material:** We have presented an approach that aims at finding an optimal splitting of the material before inducing a PCW-grammar. The splitting was aimed at improving parsing performance. For this purpose, we defined a quality measure to quantify the quality of different partitions of the material. Using this measure, we searched among a subset of all possible partitions for the one maximizing the proposed quality measure. This measure is a combination of a quality measure defined for each component in a partition. For each component, we built an automaton and computed the automaton's missed samples and perplexity. The measure we presented combines the values of perplexity and missed samples for all resulting automata. We used the resulting automata to build grammars that were subsequently used for parsing the Penn Treebank. We have shown that the quality measure we defined can be used for comparing two grammars' parsing scores if the grammars are built from partitions having a similar number of components. Since our measure $q$ is a good indicator of the parsing performance, the process of inferring grammars can be treated as an optimization task. This implies that this procedure spares us the need to assess the performance of a particular grammar by parsing.

**Sequences as Features:** The usual perspective on parsing is that all features that improve parsing performance are used for parsing, without a clear study of how these features improve parsing. Our approach is aimed at changing this perspective; we have designed grammars and experiments for isolating, testing and explaining the value of two particular features that are known to improve parsing performance: sequences of POS tags and sequences of GRs. We have shown that sequences of POS tags are fundamental for parsing performance, because they provide a reliable source for predicting and detecting dependents. Our experiments have also shown that sequences of GRs are not as reliable as sequences of POS tags. We think this is the case because the training material for GRs is small compared to the training material for sequences of POS.

PCW-grammars are versatile enough to allow us to address the variety of experimental questions and aspects we tried out in this thesis. For all of them we used only one parsing algorithm and a unified mathematical theory. PCW-grammars have reduced
the design cycle for all of our experiments; we only need to focus on very specific aspects of parsing, and we could leave aside all conditions on expressive power, complexity of parsing, and parsing algorithms. We think that PCW-grammars have proven their suitability: the formalism is abstract enough to capture the formalisms underlying state-of-the-art parsers and to suggest new research directions in parsing.

8.3 Two Roads Ahead

This thesis presents a rather unusual perspective on parsing. The usual perspective aims at designing and building parsers that produce better scores on parsing the Penn treebank. In contrast, we presented measures, grammars, tasks and experiments that were designed for testing particular aspects of syntax, language modeling and parsing.

In my opinion, these two approaches are exponents of two different research directions. The first one focuses on understanding why particular features improve parsing performance, while the second focuses on finding new features that can improve parsing performance. The second direction has reached a plateau; different approaches do not differ substantially in terms of their parsing scores and it is hard to identify the key features that may produce a future jump in performance scores. It is also difficult to determine which differences in performance are statistically significant. I think that in the forthcoming years the research focus will shift from the second line of research to the first one. I think that this shift will result in deeper, more detailed knowledge of the structure of human language, and its impact on parsing performance.

The ultimate aim of parsing is twofold: understanding human language and producing parsers that process naturally occurring sentences with an acceptable error rate. Depending on the scientific discipline, one of these interests might be more important than the other: linguistics aims at understanding human language, while statistical parsing aims at producing computational models that process natural language with an acceptable level of performance. If there is any flow of ideas here, it seems to go mostly from the linguistic side to the statistical one. The reverse direction is blocked to the point that some linguists claim that the knowledge that can be inferred by statistical methods cannot be considered as a reliable model of language (Andor, 2004). I hope this situation will change in the years to come. I think that the parsing community has to focus on two main research areas. One area will focus on identifying, understanding and testing particular aspects of features used for parsing which will provide linguists with interesting data about language as they understand it. The second stream will try to incorporate insights suggested by the first line of research in parsers. I hope that this thesis has suggested new directions; translating them into improvements in state-of-the-art parsing performance is the next step.