Theory and Applications
of Natural Language Processing

Aims and Scope

The field of Natural Language Processing (NLP) has expanded explosively over the past decade: growing bodies of available data, novel fields of applications, emerging areas and new connections to neighboring fields have all led to increasing output and to diversification of research.

"Theory and Applications of Natural Language Processing" is a series of volumes dedicated to selected topics in NLP and Language Technology. It focuses on the most recent advances in all areas of the computational modeling and processing of speech and text across languages and domains. Due to the rapid pace of development, the diversity of approaches and application scenarios are scattered in an ever-growing mass of conference proceedings, making entry into the field difficult for both students and potential users. Volumes in the series facilitate this first step and can be used as a teaching aid, advanced-level information resource or a point of reference.

The series encourages the submission of research monographs, contributed volumes and surveys, lecture notes and textbooks covering research frontiers on all relevant topics, offering a platform for the rapid publication of cutting-edge research as well as for comprehensive monographs that cover the full range of research on specific problem areas.

The topics include applications of NLP techniques to gain insights into the use and functioning of language, as well as the use of language technology in applications that enable communication, knowledge management and discovery such as natural language generation, information retrieval, question-answering, machine translation, localization and related fields.

The books are available in printed and electronic (e-book) form:

* Downloadable on your PC, e-reader or iPad
* Enhanced by Electronic Supplementary Material, such as algorithms, demonstrations, software, images and videos
* Available online within an extensive network of academic and corporate R&D libraries worldwide
* Never out of print thanks to innovative print-on-demand services
* Competitively priced print editions for eBook customers thanks to MyCopy service http://www.springer.com/librarians/e-content/mycopy

For other titles published in this series, go to
www.springer.com/series/8899

Slav Petrov

Coarse-to-Fine Natural Language Processing

Foreword by Eugene Charniak

Springer

Slav Petrov
Google
slav@petrovi.de

ISSN 2192-032X e-ISSN 2192-0338
ISBN 978-3-642-22742-4 e-ISBN 978-3-642-22743-1
DOI 10.1007/978-3-642-22743-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011939484

Cover design: deblik, Berlin

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my family

Foreword

Grammars for natural languages show how sentences (and their meaning) are built up out of smaller pieces. Syntactic parsing is the task of applying a grammar to a string of words (a sentence) in order to reconstruct this structure. For example, “The dog thought there was day-old food in his dish” has a sub-structure “there was day-old food in his dish” which in turn contains structures like “day-old food.” Before we can build the meaning of the whole we must at least identify the parts from which it is built. This is what parsing gives us.

As with most all areas of natural-language processing (NLP) parsing research has greatly benefited from the statistical revolution — the process of absorbing statistical learning techniques into NLP that began about twenty five years ago. Prior to that time we had no parser that could, say, assign a plausible structure for every sentence in your local newspaper. Now you can download several good ones on the web.

From the outside the result has looked sort of like a Moore’s law scenario. Every few years parsers got more accurate, or much more efficient, or both. From inside, however, things looked quite different. At more than one occasion we in the community had no idea where the next improvement would come from and some thought that we had, perhaps, reached the end of the road. The last time the improvement came from Slav Petrov and the ideas in this monograph. The embodiment of these ideas is the “Berkeley Parser.”

The best parsers models are all “supervised,” e.g., we have a corpus of sentences, in the case here the so-called “Penn tree-bank” where sentences have been analyzed by people so for each sentence has been broken down into a tree structure of components. A computer learns to parse new sentences by collecting statics from the training data that (we hope) reflect generalizations about a particular language, in this case English. We then recast the parsing problem as one of applied statistics and probability — find the most probable parse for the sentences according the the probabilities already obtained from the corpus.

To over simplify, until Slav’s work, the best parsers could be thought of as word-based — rules should be based upon the words found in their examples. A paradigmatic case wold be, say, the use of the prepositional phrase “out of ...”

when talking about removing something by "washing", but not by, say, "sanding." Unfortunately the number of words in English is very large (really unbounded), so this data would be missing many crucial word-grammar combinations. In this cases the parser would "back off" and look from grammar rules ignoring the particular words in question.

The Berkeley parser, however bases rules not on words, but on sets of words. The "coarse to fine' of the title refers to the graularity of these sets. So the counter claim would be that "washing" is not unique here, but is rather one of a group of words that also include "scrubbing" and in some cases "flooding" (I flooded the cinder out of my eye). Unfortunately such groups can be quite idiosyncratic, so it might be that we are still better off at the word level. Indeed, the two methods can be thought of as two ends of a continuum, and perhaps future work can now combine the approaches. But until the Berkeley parser we did have a good concrete example of this second approach.

Furthermore, for anyone with a good machine learning background, once you see how this parser works, it makes immediate sense. Thus for people like me, at least, Slav's work is very easy to read. Perhaps I am not a "typical" person, but take it from me, there are a lot of papers in my research area that I do not find so easy.

Thus I strongly recommend Slav's work to you. It is major advance in the area of syntactic parsing, and a great advertisement for the superiority of the machine-learning approach to the field.

Brown University *Eugene Charniak*

Preface

This book is based on my homonymous PhD thesis filed at the University of California, Berkeley in 2009. It has been updated to reference new work that has happened since then. It has also been reformatted to fit this paper size.

Acknowledgements

This book would not have been possible without the support of many wonderful people.

First and foremost, I would like to thank my PhD advisor Dan Klein for his guidance throughout graduate school and for being a never ending source of support and energy. Dan's sense of aesthetics has shaped the way I see research and will hopefully stay ingrained in me throughout my career. Dan is unique in too many ways to list here, and I will always be indebted to him. Dan was the best advisor I could have ever asked for.

Graduate school would not have been the same without the Berkeley Natural Language Processing (NLP) group. Initially there were four members: Aria Haghighi, John DeNero, Percy Liang and Alexandre Bouchard-Cote. Adam Pauls, David Burkett, John Blitzer and Mohit Bansal joined the group while I was still there and many new faces have joined since I left, but the amazing spirit seems to have remained. Thank you all for a great time, be it at conferences or during our not so productive NLP lunches. I always enjoyed coming to the office and chatting with all of you, though I usually stayed home when I actually wanted to get work done. My plan was to work on a project and write a publication with each one of you, and we almost succeeded. I hope that we will stay in touch and continue our collaborations no matter how scattered around the world we are once we graduate.

I spent two great summers as an intern, working first with Mark Johnson, Chris Quirk and Bob Moore at Microsoft and then with Ryan McDonald and Gideon Mann at Google. I enjoyed my summer in New York so much that I joined Google after graduating.

I would also like to thank the NLP community at large and Eugene Charniak, David Chiang, Hal Daume, Jason Eisner, Tom Griffiths, Mary Harper, Michael Jordan, Dan Jurafsky, Kevin Knight, Chris Manning, Daniel Marcu, David McAllester, Fernando Pereira and Ben Taskar in particular. I enjoyed our numerous conversations so far, and look forward to many more in the future.

Finally, I would like to thank Carlo Tomasi for giving me the opportunity to work with him while I was an exchange student at Duke University and introducing me to research for the first time. Not only did I learn a tremendous amount from him during that project, but it is also in part because of our work that I decided to pursue a PhD degree in the US.

And of course, thank you, dear reader. I feel honored and I hope you will find something useful in it. Besides my academic friends and colleagues, I would also like to thank my friends and family for helping me stay sane (at least to some extent) and providing balance in my life.

A big thank you is due to the two "fellas," Juan Sebastian Lleras and Pascal Michaillat. Living with them was a blast, especially after we survived the "cold war." Graduate school would not have been the same without the two of them. Thank you JuanSe for being my best friend in Berkeley. I am grateful for the numerous trips that we did together (especially Colombia, Hawaii and Brazil), the uncountable soccer games that we played or watched together, and especially the many great conversations we had during those years. Thank you Pascal for literally being there with me from day one, when we met during the orientation for international students. I am grateful for the numerous ski trips, cooking sessions, and lots more. Whenever I make gallettes, I will be thinking about you.

Sports were a big part of my graduate school life and I would like to thank all the members of the Convex Optimizers and Invisible Hands. There were too many to list all, but Brad Howells and Ali Memarsdaeghi deserve special mention. I will not forget our titles and the many games that we played together.

Daniel Thalhammer, Victor Victorson, Arnaud Grunwald and Konstantinos Daskalakis were always there to explore restaurants, bars and clubs in the city and we had a lot of fun together. Thank you for dragging me out of Berkeley when I was feeling lazy, and for exploring the best places to eat good food, drink good (red) wine and listen to good electronic music.

Thanks also to my friends in Berlin, who always made me feel at home when I was there during the summer and over Christmas. We have known each other since high school and I hope we will always stay in touch.

Many thanks also to Natalie, from whom I learnt a lot about (and for) life. I grew a lot as a person during our relationship and I am grateful for having had you in my life.

My brother Anton deserves many thanks for being my best friend. It would be impossible to list all the things that I am grateful for, and I won't even attempt it. I know we will stay always close and that we have many good times ahead of us.

Last but not least, I would like to thank my parents Abi and Orlin for their infinite support and encouragement. I will always be grateful for the opportunities you gave me and Anton by moving from Bulgaria to Berlin. Thank you for raising us with a never ending quest for perfection, and teaching us to believe in ourselves and that we can achieve everything we want. Thank you for your love and thank you for making me who I am.

Contents

List of Figures

List of Tables