Book Review

Language Processing with Perl and Prolog: Theories, Implemetation, and Application

Pierre M. Nugues
(Lund University, Sweden)

Springer (Cognitive technologies series, edited by A. Bundy et. al), 2014,
Second Edition, xxv+662 pp; hardcover, ISBN 978-3-642-41463-3, $89.99;
ebook, ISBN 978-3-642-41464-0, $69.99; doi 10.1007/978-3-642-41464-0

Reviewed by
Wei Lu
Singapore University of Technology and Design

Although Prolog was intended to be used in building natural language processing (NLP) applications, the language has often been neglected in many modern NLP introductory courses and texts. Because of the popularity of machine learning and statistical approaches to language processing problems over the past decades, researchers and practitioners tend to use other modern programming languages such as C++ and Java for developing NLP applications. However, with a growing interest in semantic processing and knowledge base construction in recent years, researchers have found Prolog to be an indispensable tool for tasks such as searching databases and performing symbolic reasoning. We have already seen recent successful deployment of Prolog-based NLP systems in industry. An example would be the use of Prolog in IBM's DeepQA project to express pattern-matching rules.[1] Pierre M. Nugues' comprehensive textbook *Language Processing with Perl and Prolog* provides a timely, in-depth introduction to the field of NLP using Prolog and, to a lesser extent, Perl. This book is suitable for anyone wanting to enter NLP through learning to prototype NLP modules in Prolog.

I approve of how Nugues has structured his textbook. It starts from first principles, moves on to concepts such as words, syntax, and semantics, and concludes with a discussion on the more advanced and open topics of discourse and dialogues.

Chapter 1 gives an overview of the field of natural language processing. Chapters 2 and 3 then provide a technical discussion of corpus processing and data representations.

Chapter 4 focuses on information theory and machine learning. The important concepts of entropy and perplexity are introduced first. They are followed by a discussion on the learning of a decision tree model. One minor quibble is that general machine learning concepts such as supervised and unsupervised learning are presented under the subsection "Machine Learning" within "Entropy and Decision Trees." A brief introduction to "Linear Models" is given before moving on to regression and classification models, including perceptrons, support vector machines, and logistic regression. I liked the way that such concepts were elucidated—I thought the use of automatic language detection as an example to illustrate how the different models worked was a nice

[1] `http://www.cs.nmsu.edu/ALP/2011/03/natural-language-processing-with-prolog-in-the-ibm-watson-system/`.

doi:10.1162/COLI_r_00238

touch. Although advanced readers might find this section unsatisfying since it does not cover more sophisticated models and algorithms, I believe this section is ideal for beginners, as it covers the most fundamental and essential topics for computational linguistics.

Chapters 5 to 8 discuss word-level concepts. Chapter 5 touches on some fundamental issues related to word-level processing, including how to count the words in sentences, tokenization issues, and how to model word sequences. Next, the concepts of parts of speech (POS), lexicon, and morphology are introduced in Chapter 6. An in-depth treatment of morphology is also given in this chapter, which includes a review of finite state automata and finite state transducers. The chapter mainly focuses on European languages, but I believe additional discussions on other language families would be helpful. Chapter 7 presents rule-based techniques for performing POS tagging. In this chapter, Nugues describes Brill's tagger, one of the earliest successful approaches to POS tagging. Recent developments in multilingual POS tag sets are also introduced. In Chapter 8, statistical approaches to POS tagging are discussed. Methods based on noisy-channel models, hidden Markov models (HMMs), perceptrons, and decision trees are presented. HMMs, in particular, are illustrated with detailed inference and decoding algorithms and well-chosen examples. Although HMMs are influential graphical models, I feel that its discriminative counterpart, linear-chain conditional random fields (CRFs), which are now widely used in the NLP community, could also be covered in the chapter. At the end of the chapter, two applications of the noisy-channel model are given. One is on spellchecking, and the other on machine translation. The treatment of machine translation in this example constitutes most of the discussions of this topic in the text.

Chapters 9 to 13 focus on syntactic parsing. Chapter 9 begins with the discussion of the definite clause grammar (DCG) used in Prolog to define grammars. DCG is then linked to phrase-structure rules used for syntactic parsing of natural languages. Extensive Prolog code examples on parsing based on DCG are presented in this chapter. A small section in Chapter 9 is devoted to the interesting topic of semantic parsing, where the λ-calculus as a semantic formalism is introduced and its implementation discussed. The topics of the next few chapters include partial parsing, constituency parsing, and dependency parsing.

Chapter 10 discusses partial parsing (also known as shallow parsing or chunking). In this chapter, Nugues explains the many concepts related to information extraction, with a focus on named entities and multiword expressions. Real examples taken from the CONLL shared tasks are used when introducing such concepts.

Chapter 11 gives an overview of syntactic formalisms. The chapter starts with an introduction to Chomsky's ideas on constituency parsing, followed by unification-based grammars. Next, dependency grammars are introduced in detail. At the end of the chapter, the connections between constituency grammars and dependency grammars are highlighted.

Chapter 12 focuses on constituency parsing: Classic PCFG parsing algorithms such as the CYK algorithm and the Earley algorithm are discussed. These algorithms are treated in a non-probabilistic manner and are illustrated extensively in Prolog. However, not much is written about probabilistic constituency parsing apart from the subsection "Adding Probabilities to the CYK Parser," which contains no code examples. Nugues also writes about lexicalized PCFG parsing algorithms, with special attention given to Charniak's parser. In his discussion of dependency parsing in Chapter 13, Nugues covers Nivre's shift-reduce style parser, Covington's non-projective parser, as well as Eisner's cubic-time projective parser.

After discussing syntactic processing, Nugues moves on to semantics in Chapters 14 and 15. The former focuses on formal semantics and the latter on lexical semantics. Chapter 14 starts with predicate logic, which can be naturally implemented with Prolog. The λ-calculus expressions are then revisited, followed by a wide range of technical issues related to representing words and phrases with logical forms. The issue of using semantic representations to interact with databases for question-answering is also covered in this chapter. The area of semantic parsing—mapping natural languages to their semantic representations or denotations—has recently received a lot of attention. Although some recent developments in the field are not covered in these sections, the discussion should already serve as a good introduction to readers interested in this growing area.

Chapter 15 presents lexical semantics. Concepts such as ontology and case grammar are discussed. Some useful resources such as WordNet, FrameNet, and Propositional Bank are also introduced here.

In the last two chapters, Nugues takes one step further by moving language processing beyond the sentence level. Chapter 16 focuses on discourse analysis. Topics under this theme include coreference, rhetoric, as well as event and temporal information processing. The final chapter focuses on the more advanced topic of dialogues. This can be regarded as a shift from a single, static monologue to dynamic, interacting dialogues. Many open issues remain to be addressed on such a topic, but some discussions on building simple dialogue systems are given in this chapter.

The Appendix explains some of the concepts of Prolog. Since the information contained in the Appendix is necessary to understand the examples within the text, it might be better to have it at the front of the book.

Overall, Nugues presents us with a very useful and comprehensive textbook with adequate depth and breadth. The book presents a nice synthesis of actionable code and theory and its well-considered structure allows different concepts in NLP to be strung together. This text would serve as a solid stepping-stone for NLP beginners, and is therefore suitable for graduate or senior undergraduate students who are interested in entering into the field. It would also serve as an excellent reference for researchers as well as a good guide for practitioners who are interested in building NLP applications.

Wei Lu is an assistant professor at Singapore University of Technology and Design (SUTD). His primary research interest is in the application of machine learning techniques to natural language processing problems. He is particularly interested in the modeling of semantics of natural language texts. Lu's email address is `luwei@sutd.edu.sg`.