Book Review

Pierre M. Nugues. *An Introduction to Language Processing with Perl and Prolog*. Springer-Verlag. 2006.

Reviewed by:
Ilyas Cicekli
Department of Computer Engineering
Bilkent University

An Introduction to Language Processing with Perl and Prolog is mainly designed to be used as a textbook in computational linguistics (CL) and natural language processing (NLP) courses. The book is comprehensive and covers all the required material for one semester NLP course. In addition to the major NLP concepts such as morphology, part-of-speech tagging, parsing, semantics and discourse, the book also covers the corpus linguistic related concepts such as entropy, n-grams, perplexity, and finding collocations. However the speech related concepts such as speech recognition, speech synthesis and phonetics are not discussed in the book. An NLP course covering the speech related material in depth should consider another NLP text that covers those subjects such as Jurafsky and Martin (2008), and Coleman (2005).

This is a very well written introductory textbook, and it is easy to read. The coverage and order of the concepts in the book are very suitable for an NLP course. In the following, I present the detailed reviews of the chapters of the book. Although I have minor critics in the details of chapters, overall it is a very nice textbook.

After giving a gentle overview of language processing in Chapter 1, the book covers most of the subjects related to corpus linguistics in Chapters 2-4. These chapters discuss simple algorithms such as string matching, and tokenization as well as the important statistical NLP concepts such as N-grams and language models. Chapter 2 introduces finite state automata and regular expressions, and their simple applications in corpora processing. For example, Nugues gives Perl and Prolog programs that find concordances and edit distances. Corpora encoding and annotation schemes are introduced in Chapter 3. Nugues discusses character encoding schemes and the usage of XML in corpora building as well as the theoretical concepts entropy and perplexity. The statistical NLP techniques are gaining importance in many fields of language processing. Nugues present statistical NLP concepts in Chapter 4. After explaining simple tokenization with Perl and Prolog programs, he presents N-gram, language model, and collocation concepts. Although statistical NLP concepts are covered, they are not discussed in depth as much as they are discussed in Manning and Schütze (1999).

In Chapter 5, Nugues explains morphology and morphological parsing clearly by giving a simple Prolog implementation of finite state transducers (FST). He also discusses two-level morphology and the relation between two-level morphological rules and finite state transducers. Like the authors of other NLP textbooks, Nugues also argues that a single FST representing a morphological processor can be created by intersections and compositions of the finite state transducers representing the parts of the morphological analyzer without presenting the details of the algorithms. Although it is possible to find the algorithms that convert two-level morphology rules into FSTs, intersect FSTs and compose FSTs in other resources, the students would like to see them in an NLP textbook.

Nugues explains both rule-based and stochastic part-of-speech tagging techniques in Chapters 6-7. He discusses Brill's tagger by giving its partial Prolog implementation in Chapter 6. Noisy channel model, Viterbi algorithm and hidden Markov models are presented in Chapter 7. Surprisingly, there are no Prolog implementations of the stochastic part-of-speech tagging algorithms in Chapter 7.

Nugues reserves four chapters (Chapters 8-11) for syntactic formalisms and parsing techniques. Nugues covers not only phrase-structure grammars and the well-known parsing techniques, but also partial parsing techniques and dependency parsing methods. I observe that this is the first NLP book which reserves a separate chapter for partial parsing methods. In addition to Earley parsing algorithm, Nugues also presents Cocke-Younger-Kasami (CYK) algorithm clearly by giving its Prolog's implementation. Nugues also pays a special attention to dependency parsing and the implementations of the dependency parsing techniques. He covers even the recent dependency parsing technique of Nivre (2006).

Nugues starts the discussion of the parsing with phrase-structure grammars in Chapter 8. He discusses the usage of Definite Clause Grammar (DCG) notation of Prolog how to easily transcribe phrase-structure rules into Prolog programs. However he directly jumps into parsing without giving a formal definition of context free grammar (CFG) and explaining the concept of parsing. He could have explained context free grammar, derivation, and parse tree concepts in depth before jumping directly into parsing in Chapter 8.

Chapter 9 presents a detailed discussion on partial parsing that is not covered by most of the NLP textbooks. It is nice to see that shallow parsing and chunk parsing are covered in an NLP textbook since they are gaining importance in many application areas of language processing such as information extraction.

Chapter 10 presents two syntactic formalisms, namely constituent-based and dependency based formalisms. After the discussion of Chomsky's constituent-based grammars, unification-based grammars are introduced. Nugues also discusses dependency grammars as an alternative to constituent-based formalisms.

Chapter 11 is the main parsing section, and it covers most of the efficient parsing algorithms. Nugues starts with a basic shift-reduce parsing algorithm. Then, he

introduces Earley parser which is an efficient bottom-up chart parser. In addition to Earley parser, he discusses the details of Cocke-Younger-Kasami (CYK) algorithm which is another efficient bottom-up chart parser. He also discusses dependency parsing by giving Prolog implementation of Nivre's dependency parsing technique (Nivre 2006).

The usage of predicate logic in the representation of meanings of sentences is discussed in Chapter 12. It is explained how the phrases are mapped into their corresponding logic formulas of first order predicate calculus (FOPC), however there is no formal introduction of FOPC. FOPC introduction would be useful for the students who are unfamiliar with FOPC. Also, for some phrases, only Prolog representations of logical formulas are given, which makes difficult to see the corresponding logical formulas for those phrases.

Chapter 13 discusses the lexical semantics and word sense disambiguation. Although there are some discussions about WordNet, there should have been more discussions in depth about WordNet and its usage in lexical semantics. Nugues also discusses case grammars in this chapter.

The last two chapters introduce discourse and dialogue related issues. Chapter 13 starts with a gentle introduction of discourse concept. Nugues continues with the discussion of anaphora resolution. In Chapter 13, rhetoric structure theory that explains discourse coherence is also discussed. In Chapter 14, Nugues presents the issues in dialogue modeling. The speech act theory and its application to human-machine dialogue are also discussed in this chapter.

The book ends with a comprehensive appendix that covers an introduction to Prolog programming language. The appendix should be a starting point for the readers who do not know Prolog, because most of the algorithms are given as Prolog programs.

Nugues presents the algorithms using Prolog or Perl instead of pseudo-code. Most of the algorithms are given as Prolog programs in the book. Perl is only used in Chapter 2 and Chapter 4 in the implementation of simple corpus linguistics related algorithms such as string matching. Since Prolog is more widely used in the book, and Perl is used only in few sections, the title of the book is kind of misleading. In addition, there is a long appendix for the introduction to Prolog, but there is no introduction to Perl for the readers who do not know Perl. In my opinion, there was no need to use a second programming language (Perl) in order to present algorithms.

Giving the algorithms as Prolog programs makes it easy to understand for the readers who know Prolog or willing to learn Prolog. On the other hand, the readers who do not know Prolog very well will have difficulty to understand the algorithms. The readers of this textbook should be willing to learn Prolog in order to comprehend the language processing issues that are discussed in the book.

Nugues gives the examples in three different languages, namely English, French and German. Although this exposes the reader language phenomena in different

languages, it also sometimes makes it difficult to understand the concepts. In some cases, giving the examples in three languages does not gain any advantage if they do not demonstrate different language phenomena.

In conclusion, this is a very well-written textbook and covers all the required concepts for an NLP course, and I recommend it to the students and the researchers who are willing to learn Prolog. I also plan to use it as a reference book at my NLP course so that my students can benefit from it.

References

Coleman, C. (2005) *Introducing Speech and Language Processing*. Cambridge, UK: Cambridge University Press.

Jurafsky, D. & Martin, J. H. (2008) *Speech and Language Processing, 2nd Edition*. Upper Saddle River, NJ: Prentice Hall.

Manning, C. D. & Schütze H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

Nivre, J. (2006) *Inductive Dependency Parsing*. Dordrecht, The Netherlands: Springer.