Syntactic Wordclass Tagging

Hans van Halteren (editor)
(University of Nijmegen)

Dordrecht: Kluwer Academic
Publishers (Text, speech and language
technology series, edited by Nancy Ide
and Jean Véronis, volume 9), 1999,
xvii+334 pp; hardbound, ISBN
0-7923-5896-1, $149.00, £93, Dfl 280.00

Reviewed by
Adwait Ratnaparkhi
IBM T.J. Watson Research Center

Part-of-speech (POS) tagging is one of the most popular and thoroughly researched tasks in the field of natural language processing, particularly since it is a prerequisite for a wide variety of more complex tasks. The book *Syntactic Wordclass Tagging* is a multiauthor collection of articles giving advice on how to use and implement a POS tagger. Part I of the book is entitled "The User's View" and is geared towards novices and researchers who are interested in the POS annotation that taggers produce. Part II, entitled "The Implementer's View," is more technical and is written for researchers who want to understand the advantages of the various computational techniques used for POS tagging; it includes an introductory chapter by Hans van Halteren.

After an introductory chapter by Atro Voutilainen, Part I begins with "A Short History of Tagging," also by Voutilainen, which describes some of the notable developments in both data-driven and linguistic approaches to tagging.

Then, in "The Use of Tagging," Geoffrey Leech and Nicholas Smith argue that POS tagging is useful for almost every task in corpus linguistics, and has made its way into a number of practical applications as well, such as information retrieval, spelling correction, and machine-aided translation. Leech and Smith also point out that syntactic parsing is arguably the central task of natural language processing, since it is a prerequisite for any kind of semantic analysis of text, and claim that tagging, by being a prerequisite to parsing, is effectively an "entry to the most central area of corpus processing" (p. 27).

Jan Cloeren then describes tagsets for wordclass annotation, and discusses the different levels of linguistic details—morphological, syntactic, semantic, discoursal—captured by various tagsets. He also discusses how certain phenomena, such as multiunit tokens, multitoken units, wordclass underspecification, and wordclass ambiguity, present challenges to the design of a tagset. He further describes a proposal by the Text Encoding Initiative (TEI) that specifies how to encode wordclass tags with SGML.

"Standards for Tagsets" by Geoffrey Leech and Andrew Wilson discusses the guidelines for POS annotation standards developed by the Expert Advisory Group on Language Engineering Standards, or EAGLES. The chapter describes the obligatory, recommended, and optional attribute-value sets for describing wordclasses across languages and tasks. The intent is that the widely recognized attribute-value pairs (major parts of speech, gender, etc.) would be obligatory or recommended, whereas the task or language-specific attribute-value pairs would be optional.

In "Performance of Taggers," Hans van Halteren discusses the relationship between correctness, ambiguity, precision, and recall, which are the most commonly reported accuracy measures in the POS tagging literature. He also discusses how the overall accuracy is affected when the tagset, lexicon, and method of evaluation are varied. The lesson is that published POS accuracy results for different taggers can often be difficult and sometimes impossible to compare.

"Selection and Operation of Taggers," also by van Halteren, lists some of the factors that are relevant when one is trying to select "the best tagger." Due to the difficulty in weighing all the factors when choosing a tagger, van Halteren makes no specific recommendations, but instead suggests that potential users will have to make their own judgments in the context of the task that they are trying to accomplish.

Part II begins with an introduction by van Halteren and Voutilainen. Then, a chapter by Gregory Grefenstette discusses tokenization—a very important but often dismissed preprocessing step—and quantifies the accuracy gain from incrementally adding certain knowledge sources: regular expressions, a corpus, a lexicon, and an abbreviation list. Grefenstette also gives code for a sample tokenizer.

In "Lexicons for Tagging," Anne Schiller and Lauri Karttunen discuss how a tagging lexicon can be derived from a morphological lexicon. They first discuss the implementation of morphological lexicons with finite-state techniques, and show how to map, merge, and refine morphological classes into classes suitable for a POS tagger. Lastly, they summarize the problems that occur when using a corpus to derive a lexicon.

"Standardization in the Lexicon," by Monica Monachini and Nicoletta Calzolari, discusses the EAGLES effort on lexicon standardization, which is designed to be compatible with the EAGLES effort on POS annotation standards. Monachini and Calzolari argue that a lexicon can be used as an "interface" between two tagsets, and report that an automatic method to map the UPenn tagset to the BNC tagset was successfully implemented with a lexicon as the interface. The authors reiterate that "it is not possible, nor desirable, to arrive at identical tagsets across languages, even for those within the same language family" (p. 153). Like Leech and Wilson, Monachini and Calzolari describe a multitiered standard, where certain widely recognized attribute-value pairs are obligatory or recommended, and where language- and task-specific attribute-value pairs are considered to be optional. It is argued that this multitiered architecture will be flexible enough to accommodate all languages and tasks but also rigorous enough to permit comparison between the lexicons for different languages and tasks.

"Morphological Analysis," by Kemal Oflazer, discusses the process of decomposing words into their constituents. Oflazer describes two-level morphological analyzers, and shows how they can analyze various examples, including Turkish vowel harmony. He also discusses in detail the design of a two-level morphological analyzer for the Turkish language.

Eric Brill then discusses several techniques for tagging unknown words, and shows how a particular corpus-based learning paradigm, called Transformation-Based Learning (TBL), can learn rules for guessing the POS tags for words that it has not seen before in the corpus. Brill uses the observation that low-frequency words behave like unknown words, and therefore uses low-frequency words (annotated with POS tags) as training material in the learning phase.

Voutilainen, in "Hand-Crafted Rules," gives a very interesting comparison of the linguistic and statistical paradigms for POS tagging, and claims that "contrary to common belief, a useful grammar can be written without spending years on it, given that reasonable attention has been given to the lexicon, tagset and test corpus—in fact, a few months of rule-writing and testing is likely to suffice for making a useful grammar" (p. 219). This chapter describes an experiment in which the author (an expert linguist) writes a nontrivial accurate POS tag disambiguation grammar in a few hours.

The experiment was duplicated with novices, who took considerably longer, but eventually succeeded in writing accurate grammars. In all the experiments, a small training corpus was used to develop the grammars, and a separate test corpus was used to evaluate them.

Brill then, in "Corpus-Based Rules," continues the discussion on Transformation-Based Learning, and focuses on learning rules to tag known words. He describes the TBL algorithm in more detail, and explains an extension of TBL that can provide *n* likely tags for any given word, instead of just one tag per word. An unsupervised version is also presented, in which tagging rules are induced from an unannotated corpus and a lexicon; it exploits the fact that some words only have one permissible tag, and uses rules derived from those instances to resolve the tags of ambiguous words.

Marc El-Beze and Bernard Merialdo discuss hidden Markov models (HMMs) for POS tagging, which use mathematically motivated techniques to tag in either supervised or unsupervised modes. They give an informal description of the Baum-Welch algorithm, which is used for unsupervised training, as well as a description of the Viterbi algorithm, which is used for testing. They report the result that using Baum-Welch with a "bad" initial tagging model increases tagging accuracy, but that using Baum-Welch with a "good" initial tagging model only decreases tagging accuracy.

Walter Daelemans then discusses machine learning approaches for POS tagging, in which each POS assignment is viewed as a classification task. He gives an overview of some popular machine learning techniques for POS tagging, such as case-based learning, decision tree induction, and neural networks. While the approaches differ in their learning characteristics, the accuracies of all these approaches appear to be similar. Daelemans argues that "learning is preferable to programming" because it "provides a solution to the knowledge-acquisition and reusability bottlenecks and to robustness and coverage problems" (p. 303).

The level of detail in the book varies from chapter to chapter; some chapters have copious details of implementation or annotation standards while others are more informal. Part I is excellent for novices, users of tagging software, and researchers who want to rapidly familiarize themselves with POS tagging. Part II gives insight into how and why the various computational techniques for POS tagging work, and will be useful for researchers who want to write their own tagging software. After reading the book, one doesn't get the impression that a particular technique or paradigm stands out as the preferred technology for tagging. Accuracies of the taggers from different chapters are impossible to compare because of mismatches in the tagsets, annotation consistencies, test data, and evaluation techniques. (The book warns the reader of this issue several times.)

The methodological distinction between the rule-based and statistical methods is not as dramatic as the chapters imply. The machine learning and statistical methods mostly use information in a small window (usually ±2 words) around the word being tagged; most rules in the EngCG-2 tagger of Voutilainen's chapter on hand-crafted rules also look in this small window (p. 224). The grammar development process of this chapter uses a corpus in the same spirit of most machine learning techniques—rules are accepted or rejected on the basis of their correctness in a manually annotated benchmark corpus.

The primary value of the book is that it has a nice diversity of approaches to one problem, and gives a sense of the challenges that are faced by those designing wordclass annotation and lexicon standards.

Adwait Ratnaparkhi received his Ph.D. in 1998 from the University of Pennsylvania, where he studied statistical and corpus-based techniques for natural language ambiguity resolution. He is

the author of MXPOST, a freely available POS tagger built with the maximum-entropy learning framework. He is currently a research staff member at the IBM T.J. Watson Research Center. Ratnaparkhi's address is: IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA; e-mail: aratnapa@us.ibm.com; URL: http://www.research.ibm.com/people/a/adwaitr