Book review

Automatic Ambiguity Resolution in Natural Language, Alexander Franz

Christof Monz
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
The Netherlands
E-mail: christof@wins.uva.nl

BOOK TITLE: Automatic Ambiguity Resolution in Natural Language
AUTHOR: Alexander Franz
TOWN OF PUBLICATION: New York
YEAR OF PUBLICATION: 1996
SERIES: Lecture Notes in Artificial Intelligence 1171
PUBLISHER: Springer-Verlag
PRICE IN $ 36
PAGES: xix + 155
ISBN: 3-540-62004-4

During the last ten years, corpus-based approaches to natural language processing (NLP) have received a lot of attention, resulting in extensive ongoing work in statistical methods in Computational Linguistics. Corpus-based linguistics focuses on language data, for instance as available through large text corpora, whereas approaches in the Chomskian tradition try to give an abstract model of grammatical competence. It is not uncommon in the latter approach to focus on natural language examples that sometimes appear to be rare and hard to judge as grammatical or ungrammatical. Corpus-based linguistics, on the other hand, examines examples that were actually expressed in a spoken or written way. This change in perspective leads to a totally different set of problems and a different methodology to solve them. Often it is difficult for traditional and corpus-based linguists to understand the problems of the other, and it appears to be very hard to work out a unified approach to solve the difficulties in natural language processing. This situation is nicely exemplified by Franz' book.

Franz discusses three kinds of ambiguity from which two, namely classification of unknown words and part-of-speech tagging, are mainly considered in corpus-based linguistics. PP-attachment ambiguity, the third kind of ambiguity discussed in the book, already received extensive attention in traditional linguistics. There are some heuristics that help to resolve ambiguities, but a wide-coverage approach is still missing. Previous approaches on modifier attachment ambiguity failed because they were either too simplistic, for instance considering only the syntactic structure, or they were too ambitious by trying to incorporate world-knowledge. Franz tries to solve this problem by taking an empirical approach which tries to extract resolution strategies from large parsed language data. This seems to be the right track in that he does not only consider the syntactic structure, but also a set of linguistic features assigned to the lexical items, among them lexical association strength. So it is not too simplistic, but not too ambitious either, because only a limited amount of world-knowledge is considered. Approaches that try to integrate world-knowledge in general have to implement it manually, which is very time-consuming and not feasible for a large domain. Lexical association strengths, on the other hand, can be extracted from large language data in a fully automated fashion.

The book is a slightly revised version of the author's dissertation at the Carnegie Mellon University in 1995. It is very well written and contrary to a lot of the literature in statistical linguistics, not overloaded with formulas and diagrams. The theoretical results are well motivated and most of the time their presentation is accompanied by a short prose explanation. Although the book is better suited for people with some background in statistical linguistics, it is not totally inaccessible to students who only have some background in probability theory. Common concepts from statistical linguistics used throughout the book, are shortly explained. The explanations may be not long enough, if one has never heard of these concepts before, but it is not the aim of the book to be a textbook on statistical linguistics. The aim is to show how a particular method, namely the loglinear method, allows for a statistical approach to ambiguity resolution.

The book can be subdivided into four parts. The first part consists of Chapter 1 and Chapter 2, where some preliminary remarks and a brief review of previous work on syntactic ambiguity resolution are given. The second part encompasses only Chapter 3. Here, the main theoretical results are established. Part three, (Chapter 4 – 7) evaluates the theory according to some problems in ambiguity resolution. The six appendices which provide some background information form the fourth and final part.

Chapter 1 provides some preliminaries and an overview on the remaining chapters. Some examples of ambiguities, like part-of-speech (POS), ambiguities are considered and the problem of ambiguity and robust parsing is discussed. In this context the author relates coverage, perplexity, and ambiguity to each other. A grammar with a high coverage and thereby also a high perplexity, has in practise a higher degree of ambiguity. Then he contrasts the idea behind empirical or corpus-based approaches to natural language processing with approaches in the Chomskian framework.

Chapter 2 surveys some previous approaches to ambiguity resolution and summarizes the pros and cons of them. Four problems are considered (i) unknown words, (ii) lexical syntactic ambiguity or POS-tagging, (iii) structural ambiguity, and (iv) prepositional phrase attachment ambiguity. For each problem, syntactic or semantic approaches are compared to a corpus-based one. Most attention is paid to PP-attachment, where the work of Hindle and Rooth (1993) is reviewed. The author concludes that syntactic factors are important for ambiguity resolution, but they also heavily rely on the way the grammar is set up. To give a fully satisfying account of ambiguity resolution, semantics has to be considered. However, in the state of the art it is still a problem how to encode world-knowledge in an efficient way. Corpus-based approaches, on the other hand, approximate ambiguity resolution of human speakers pretty well, but previous work on resolution is still too simplistic.

Chapter 3 is the heart of the book; the author explains how ambiguity resolution can be regarded as a classification problem. Let $I = \{i_1, \ldots, i_n\}$ be a set of possible unambiguous interpretations of some expression, then the set C of classifications is the power set of I modulo the empty set. Classifications then still leave some space for uncertainty or underspecification. Given an expression with a vector $\mathbf{v}$ of linguistic features, the decision for a classification is guided by a strategy to minimize the loss.

$$\text{Expected Conditional Loss}(c_{\text{proposed}}|\mathbf{v}) = \sum_{i \in I} P(i|\mathbf{v})\lambda(c_{\text{proposed}}|i)$$

where $\lambda(c_{\text{proposed}}|i)$ returns 0 if $i \in c_{\text{proposed}}$ and 1 otherwise. The 'zero-one' loss function λ checks whether a given interpretation is contained in a proposed classification. The 'Expected Conditional Loss' function computes the classification with the minimal error rate. To extract $P(i|\mathbf{v})$ from a set of data, contingency tables are used, where the dimensions correspond to the set of linguistic features in $\mathbf{v}$. To deal with the problem of sparse data and interdependence of features, loglinear models are introduced as an effective means. Loglinear models meet several essential requirements: (i) they go beyond one-dimensional probabilistic models, (ii) they allow to describe interdependence, (iii) they are widely applicable, and (iv) they can be applied to data that has not been covered well during the training phase. The strength of the dependence between features is described by interaction terms. At the end an iterative procedure for building a smoothed contingency table is presented.

Although the material is presented rather perspicuously, more detail might have been appropriate. It is basically the only chapter where the author's theory is established and therefore it deserved more attention.

Chapter 4 describes the application of the disambiguation procedure to the problem of classifying unknown words. To classify the part-of-speech (POS) of an unknown word, it is useful to consider its linguistic features: its affixation, whether it is sentence-initial, whether it includes a period, and so on. But considering all these features results in an impossibly large contingency table. Therefore, the author extracts some essential discriminating features on an experimental basis, where features that are independent of each other are stored in different contingency tables. Experimental results show that classifying unknown words with loglinear models performs significantly better in accuracy (8% – 17%) than using models which assume independence of features. As was stated earlier, classifications may still be ambiguous, i.e., an unknown word may be classified with a set of POS tags, and an answer by a model is judged as correct, if the correct POS tag is an element of the classification. So there is a trade-off between accuracy and ambiguity, and the choice between these two depends on the area of application.

Chapter 5 is devoted to part-of-speech tagging in general, i.e., also including unknown words. Standard techniques, like tagging with trigrams, where given a sequence of words W, finding the corresponding sequence of tags T can be approximated by considering the preceding two tags of each tag, only. To maximize T, the Viterbi algorithm can be employed which computes the maximum in polynomial time. Next, several smoothing techniques, like Jeffrey's estimate, linear interpolation, and deleted interpolation, are compared. Smoothing deals with the problem of assigning frequencies to word-tag pairs that did not occur during training. Several experiments are discussed and the author shows that lexical probabilities play a dominant role in POS tagging, because the proportion of unknown words and the overall accuracy are clearly correlated. This is especially important for taggers which have to be applicable to texts stemming from different domains. Using loglinear models instead of regular trigram models, reduces the error rate significantly. On the other hand, the error rate for known words is not significantly reduced by using a loglinear model. It is remarkable of this and the following chapter, that error rates are displayed by boxplots, which also indicate the variance of a method. Therefore, instead of presenting a particular test and its error rate, a series of tests is given and it is shown how the error rate varies among them.

Chapter 6 discusses a particularly difficult instance of structural ambiguity, namely PP-attachment. In a sequence of the form *verb* $noun_1$ [$_{\text{PP}}$ *preposition* $noun_2$], there are two attachment sites, namely *verb* and $noun_1$. Similar to (Hindle and Rooth, 1993), the author employs lexical association strength measured by the Mutual Information between the noun or verb, and the preposition. Besides the lexical association strength also features like definiteness of the noun, or kind of preposition (with, to, over, etc.) are modeled. For cases considering only two attachment sites, the loglinear method does not perform significantly better than the method by Hindle and Rooth. But if sequences of the form *verb* $noun_1$ $noun_2$ [PP *preposition* $noun_3$] are considered, where three attachment sites for the preposition are possible, the loglinear method gains a much higher accuracy.

Chapter 7 summarizes the experimental results and provides some conclusions and prospects for future research. Modeling unknown words can be improved by using a loglinear model, because it does not assume independence among the set of features. Although loglinear models do not significantly improve part-of-speech tagging in general, depending on the percentage of unknown words, they allow for a significant reduction of the tagging error (19%) for proper nouns. If only the problems of PP-attachment analyzed in (Hindle and Rooth, 1993) are considered, there is no significant enhancement, but for more complex cases of PP-attachment the loglinear method gains an improvement. Combining these results, this way of ambiguity resolution might provide a new technique that helps to achieve the goal of robust parsing of unrestricted text. The advantages over previous approaches are that there is no assumption of feature independence in loglinear models and that observed frequencies can be nicely smoothed via the marginal totals of the contingency table. Besides the problems discussed in the book, the author considers it possible to apply the loglinear method to other problems like coordinated constructions, word-sense ambiguity, other modifier-attachment ambiguities, and spelling correction.

The book contains a number of appendices. Appendix A explains the concept of entropy. A list of the Penn Treebank tags is given in Appendix B. Appendix C shows how random samples can be extracted from the Penn Treebank corpus. To this end it is necessary to convert the Penn Treebank files into a LISP-readable format, and Appendix D provides a Common LISP implementation that accomplishes this. Appendix E explains confusion matrices for POS tagging. Appendix F describes an interface between Common LISP and S-PLUS, a statistical software package.

The 13-page bibliography is very comprehensive and it contains almost all relevant references.

After this rather long summary of the contents of the book, I would like to point out again, that this very well written book describes a statistical approach to ambiguity resolution, that might not only be interesting for statistical linguists, but also for computational linguists in general. It would have been nice if more cases of structural ambiguity had been covered, because most of the book is still concerned with the application of the loglinear method to POS tagging and the classification of unknown words. Both are classical problems in statistical linguistics, whereas stochastic approaches to structural ambiguity resolution, a classical problem in traditional linguistics, are still in their infancies. An analysis of a broader spectrum of phenomena of syntactic ambiguity might have attracted even those who are still suspicious about the applicability of stochastic methods in linguistics. But, as the author says in his conclusions, he is convinced that the loglinear method can be applied to several kinds of structural ambiguity resolution. A broader analysis is therefore not prohibited, but rather remains a task of future research.

References

Hindle, D. and Rooth, M.: 1993, Structural ambiguity and lexical relations, *Computational Linguistics* **Vol. 19 no. 1**, pp. 103–120