Memory-Based Language Processing

Walter Daelemans and Antal van den Bosch
(University of Antwerp and Tilburg University)

Cambridge: Cambridge University Press, 2005, vii+189 pp; hardbound,
ISBN 0-521-80890-1, $75.00

Reviewed by
*Sandra Kübler**
Indiana University

Machine learning has become the predominant problem-solving strategy for computational linguistics problems in the last decade. Many researchers work on improving algorithms, developing new ones, testing feature representation issues, and so forth. Other researchers, however, apply machine-learning techniques as off-the-shelf implementation, often with little knowledge about the algorithms and intricacies of data representation issues. In this book, Daelemans and van den Bosch provide an in-depth introduction to *Memory-Based Language Processing* (MBLP) that shows for different problems in NLP how the technique is successfully applied. Apart from the more practical issues, the book also explores the suitability of the chosen learning paradigm, *memory-based learning* (Stanfill and Waltz 1986), for NLP problems. Thus the book is a valuable source of information for a wide range of readers from the linguist interested in applying machine-learning techniques or the machine-learning specialist with no prior experience in NLP to the expert in machine learning wanting to learn more about the appropriateness of the MBLP bias for NLP problems.

Memory-based learning is a machine-learning method based on the idea that examples can be re-used directly in processing natural language problems. Training examples are stored without modification or abstraction. During the classification process, the most similar examples from the training data are located, and their class is used to classify the new example.

The book addresses different levels of understanding and working with MBLP: On one level, it explains the theoretical concepts of memory-based learning; on another, it provides more practical information: The implementation of memory-based learning, TiMBL, is described as well as different extensions such as FamBL and MBT. On a different level, the application of these techniques is described for typical problems in natural language processing. The reader learns how to model standard classification problems such as POS tagging, as well as sequence-learning problems, which are more difficult to model as classification problems. Daelemans and van den Bosch also cover critical issues, such as problems that arise in the evaluation of such experiments and the automation of searching for suitable system parameter settings. On a more abstract level, they approach the question of how suitable the bias of MBLP is. In chapter 6, they compare memory-based learning as an instance of lazy learning to an instance of eager learning, rule induction, with regard to their classification accuracy if, for example, more abstraction is introduced. Since MBLP does not abstract over the training data, it is called a lazy learning approach. Rule induction, in contrast, learns rules and does not go back to the actual training data during classification.

---

* A shorter version of this review will be published in German in the journal *Linguistische Berichte*.

The book consists of 7 chapters. Chapter 1 situates memory-based language processing firmly in the domain of empirical approaches to NLP. Empirical approaches became attractive in the early 1990s, replacing knowledge-based approaches to a high degree. Daelemans and van den Bosch argue that in the range of empirical approaches, memory-based learning offers the advantage over statistical approaches that it does not abstract over low-frequency events. Such low-frequency events are necessary in processing natural language problems because they often describe exceptions or subregularities. The chapter also introduces the major concepts of MBLP and provides an intuitive example from linguistics: PP attachment.

Chapter 2 locates central concepts of MBLP in neighboring areas of research: In linguistics, the idea of processing by analogy to previous experience is a well-known concept. Psycholinguistics often uses exemplar-based approaches or, more recently, hybrid approaches that combine rules with exceptions. Applications of memory-based principles can be found in explanation-based machine translation (Nagao 1984) and data-oriented parsing (Bod 1998).

Chapter 3 gives a simultaneous introduction to memory-based learning and TiMBL, the Tilburg implementation of the method. This strategy of combining theory and practice gives the reader an impression of the importance of selecting optimal parameter settings for different problems. The application of TiMBL is demonstrated on the example of plural formation in German. The chapter ends with the introduction of evaluation methodology and TiMBL's built-in evaluation functions.

Chapter 4 describes the application of TiMBL to two more complex linguistic examples: grapheme to phoneme conversion and morphological analysis. In order to find optimal solutions for these problems, two algorithms that deviate from the standard memory-based learning algorithm are introduced: IGTree and TRIBL. IGTree is a decision tree approximation, which bases the comparison of an example to others on a small number of feature comparisons. TRIBL is a hybrid model between the standard memory-based learning algorithm, IB1, and IGTree. Both modifications reduce memory requirements and processing time during classification, but they may also affect classification accuracy. Unfortunately, the presentation of the first example suffers from unreadable phonetic transcriptions throughout the chapter.

Whereas Chapter 4 analyzes linguistic problems, which are easily described in terms of classification, chapter 5 approaches a problem of sequence learning: partial parsing. For this task, phrase and clause boundaries must be found. In order to apply classification methods to sequence learning, the problem must be redefined as assigning tags to words or word combinations, so-called IOB tagging (Ramshaw and Marcus 1995). This tagging provides information as to whether a word constitutes a boundary or not. One advantage of using MBLP for such problems lies in the fact that different types of information, including long-distance information, can be included without modification of the original algorithm.

In Chapter 6, Daelemans and van den Bosch investigate the difference between lazy and eager learning. As noted earlier, TiMBL is a typical example of lazy learning since it does not abstract from the training data. RIPPER (Cohen 1995), the other classifier used in this chapter, is a typical eager learning approach: It is a rule-induction algorithm, which displays the opposite behavior to TiMBL: a complex learning strategy and simple, efficient classification. The results presented in this chapter show that deleting examples from the training data is harmful for classification, supporting the hypothesis that lazy learning has a fitting bias for natural language problems. However, this seems to be a little too straightforward. Here, one would expect a reference to the findings of Daelemans and Hoste (2002), which show that parameter and feature

optimization often result in larger differences in accuracy than differences between learning algorithms.

Chapter 7 returns to more practical problems: parameter optimization and more advanced types of classifier combinations to handle sequence learning. Parameter optimization is one of the more tedious aspects of the work in machine learning. Researchers often spend a significant amount of time on finding the optimal parameter settings for a specific algorithm and problem. Wrapped progressive sampling is one method to automate this process. An implementation is included in the TiMBL package, making it an even more attractive tool for machine learning in NLP.

The fact that this book provides a very well written and easy-to-follow overview of different aspects of how to apply memory-based learning to problems in NLP makes it a valuable reference for researchers without previous experience in machine learning or natural language processing as well as for experienced researchers. The book provides a well-balanced combination of theoretical aspects of MBLP with an introduction to a software package that allows an easy transformation of the ideas presented in this book into practice. Daelemans and van den Bosch allow the reader to profit from their vast experience with MBLP when they describe different problems in NLP and how they are approached in TiMBL. This is also obvious in the extensive sections on further reading, which inspire the reader to get deeper into the subject. Since the chapters on example problems are combined with sections on evaluation, the book provides a thorough overview of the methodology in machine learning, which researchers new to the area often have to learn the hard way because most reports on results in machine learning tend to gloss over such issues. More experienced researchers who are already familiar with memory-based learning will especially be interested in the comparison of lazy and eager learning. It is very interesting to see that the inclusion of low-frequency examples as well as some high-frequency examples, which may be redundant, always leads to a decrease in accuracy. This shows very clearly that the memory-based strategy of keeping all training examples is a successful strategy for natural language problems.

References

Bod, Rens. 1998. *Beyond Grammar: An Experience-Based Theory of Language*. Stanford, CA: CSLI Publications.

Cohen, William. 1995. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, Tahoe City, CA.

Daelemans, Walter and Véronique Hoste. 2002. Evaluation of machine learning methods for natural language processing tasks. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Gran Canaria.

Nagao, Makoto. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Banerji, editors, *Artificial and Human Intelligence*. Amsterdam: North-Holland.

Ramshaw, Lance and Mitchell Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the SIGDAT/ACL Third Workshop on Very Large Corpora*. Cambridge, MA.

Stanfill, Craig and David L. Waltz. 1986. Towards memory-based reasoning. *Communications of the ACM*, 29(12), 1213–1228.

Sandra Kübler is an assistant professor of computational linguistics at Indiana University. Her research interests include machine learning methods in morpho-syntactic and syntactic annotation. Kübler's address is Indiana University, Linguistics Department, Memorial Hall, 1021 E. Third Street, Bloomington, IN 47405; e-mail: Skuebler@indiana.edu.