Review of the book Memory-based parsing, S. Kuebler, 2004

van den Bosch, A.

Link to publication in Tilburg University Research Portal

Memory-Based Parsing

Sandra Kübler
(University of Tübingen)

This monograph, based on Sandra Kübler's Ph.D. thesis, introduces the reader to research at the intersection of data-driven parsing and memory-based learning. Compared to other approaches to parsing, including knowledge-based methods, memory-based parsing takes the provocative standpoint that new structures can be parsed by analogical reasoning over stored structures, rather than by abstracted rules. All that the approach needs is parsed example sentences stored in memory and a similarity function to find candidate nearest-neighbor sentences that can act as the basis for the analogical-reasoning step.

Kübler begins with a walk past the known approaches in the field to partial and full memory-based parsing. As for partial parsing (constituent chunking and basic grammatical-relation assignment such as subject-verb relations), the work of Daelemans, Veenstra, Buchholz, Tjong Kim Sang, and others using memory-based learning is summarized, as well as Krymolowski, Argamon, and Dagan's work on memory-based sequence learning. Kübler then proceeds to review Streiter, whose memory-based parser is an example of a more holistic, sentence-oriented approach which, in contrast to the aforementioned approaches, needs a more complex similarity metric to compute the distance between a complete new sentence and stored parsed sentences. A separate chapter is devoted to data-oriented parsing, which uses probabilistic machinery and extensive back-off from larger to smaller substructures, instead of a single similarity function; in particular, two nonprobabilistic variants by Bod and De Pauw are close cousins of the other memory-based approaches.

The heart of the book is the TüSBL (Tübingen Similarity-Based Learning) memory-based parser, which implements a similarity-based approach that, analogous to Streiter's approach, attempts to fully parse complete sentences by analogy, as rapidly as possible. Kübler's solution is original. While a naive approach based on a similarity between full sentences would be able to correctly parse only a few sentences very close to sentences in memory, Kübler introduces at least two smart generalization enhancements. One is that the TüSBL parser has more than one similarity metric. When a new sentence is parsed, it is first analyzed at the levels of part-of-speech tags and base constituents. If no reliable nearest neighbors matching on the word level can be found in memory, the other levels act as back-offs on which to measure similarity. The second enhancement is that the search for nearest neighbors is extended by allowing them to have a word or constituent too many or too few or to be longer but contain a good matching subtree. TüSBL's similarity metric, or rather its case-based reasoning function, is actually aware of the internal structure of the nearest-neighbor trees and the partial syntactic structure of new sentences.

TüSBL is put to the test on the NEGRA-formatted TüBa-DS treebank of spontaneous speech in specific domains (hotel reservations, business appointments, and travel scheduling), gathered in the context of the VERBMOBIL project. Kübler makes the credible point that data of this type offer a more interesting challenge to parsing methods than nonspontaneous, professionally authored texts in a (similarly) closed domain, such as the *Wall Street Journal* Penn Treebank. An excellent point is made on the limitations of the standard PARSEVAL evaluation's focus on syntactic chunking and labeling; arguably, the correctness of the parser in assigning functional labels to grammatical relations is at least as interesting as an evaluation metric. From the reported results we learn that TüSBL does a good job; it attains a PARSEVAL F-score of about 85 on the spontaneous speech corpus. We also learn that the back-off part of TüSBL performs as well as the whole TüSBL system in PARSEVAL terms. This underlines the point that parsing based on matching on smaller, local structures, as is done by most other memory-based methods, performs at least on par with more holistic methods. However, TüSBL's holistic memory-based core is more reliable in assigning correct functional tags to correctly identified grammatical relations.

In sum, the book offers a comprehensive
and well-illustrated overview of the area
of memory-based parsing, makes all the
right methodological points, and describes
a system that performs a complex task in a
refreshingly simple and smart way.—*Antal
van den Bosch, Tilburg University*