Toward Memory-Based Syntactic Processing

Walter Daelemans (WALTER@KUB.NL)
Jakub Zavrel (ZAVREL@KUB.NL)
Jorn Veenstra (VEENSTRA@KUB.NL)
Computational Linguistics, Tilburg University
PO Box 90153
5000 LE Tilburg
The Netherlands

In **Memory-Based reasoning and learning** (Stanfill & Waltz, 1986) it is claimed that reasoning is the result of direct use of stored experiences rather than of the application of abstractions (e.g. rules) extracted from those experiences.

In Memory-Based language processing, if a new utterance is analysed, the most similar cases are retrieved from memory, and used for extrapolation. In this performance-oriented paradigm learning is a straightforward process of storage of cases and adaptation of similarity metrics to a linguistic domain. Processing directly exploits the implicit linguistic knowledge in the stored cases.

In our implementation of Memory-Based reasoning, the cases are represented as feature-value vectors of fixed dimensionality, labeled with a class chosen from a fixed set. Many linguistic phenomena show: i) a large number of cases needed to cover language's (ir)regularity, ii) symbolic feature values, and iii) diverse features of varying importance. In previous research, we have presented a number of methods to deal with these issues, i.e. efficient memory-indexing, using distributional similarity metrics for linguistic symbols, and weighting features by their Information Gain (Daelemans, 1995).

So far, these methods have mostly been used for phonetic and morphological tasks. In order to apply Memory-Based modeling to full-scale syntactic processing, parsing must be decomposed into an ensemble of classification tasks. **We argue that parsing can be rephrased in terms of segmentation** (i.e. which boundary class to insert where in a sentence), **disambiguation** (i.e. how to label a word or constituent in context), **and attachment resolution** (i.e. where to attach a constituent). We have constructed Memory-Based models for two representative aspects of syntactic processing: morpho-syntactic disambiguation, and disambiguation of Prepositional Phrase (PP) attachment.

Morpho-syntactic disambiguation. The task is to assign a morpho-syntactic category (part-of-speech) to words in sentence context. Penn Treebank Wall Street Journal (WSJ) material is used for training and testing. As features we use the disambiguated categories of two words to the left, and the ambiguous categories of the words itself and its direct right neighbor. For out-of-vocabulary words, the first letter of the word and the last three letters of the word were used for a lexical representation. The cases were labeled with their annotated Penn Treebank category. The used metric weights features by their Information Gain. The model was trained on 2 million words, and tested on a held-out set of 200,000 words. The accuracy on this test set was 96.7% for known words, 90.6% for out of vocabulary words, totaling to a competitive 96.4%.

Disambiguating structural attachment. For this task, sentences with ambiguous PP-attachments, extracted from the WSJ corpus, are used as training and test material. All the sentences consisted of a `V NP P NP` sequence, and the features used are its head words, i.e. `V N1 P N2`. The patterns are labeled with the correct attachment decision, i.e. whether the PP attaches to V or to N1. Training was done on 20801 cases and testing on 3097 separate cases (the same train-test partition as in Collins & Brooks (1995)). A Memory-Based model allows us to study the effect on task performance of varying representations (words, syntactic categories, and continuous distributional representations called LexSpace (Zavrel & Veenstra, 1995)) and similarity metrics (weighted and unweighted). The results (83.7% correct for an unweighted word matching metric, 84.1% weighted word matching, 84.4% weighted LexSpace representations) show that the Memory-Based model generalizes to new test material with an accuracy that is equal to or even better than the best method known for this task in the literature (Collins & Brooks, 1995: 84.1%; cf. Human judges 88.2%, syntactic category matching 59.0%).

Conclusions from this work are that: i) Memory-Based models of aspects of syntactic processing are efficient and accurate in performance without using explicit rules; ii) Domain-independent feature-weighting metrics adapt the notion of similarity to the task at hand; iii) The flexibility of the Memory-Based approach supports the use of diverse lexical representations; and iv) LexSpace representations outperform simple word-matching, suggesting that taking word-similarity into account can be beneficial.

References

Collins, M. & Brooks, J. (1995). Prepositional Phrase Attachment through a Backed-Off Model in: *Proc. of WVLC-3*, Cambridge, MA.

Daelemans, W. (1995). Memory-Based Lexical Acquisition and Processing, in: P. Steffens (ed.) *Machine Translation and the Lexicon*, Springer Lecture Notes in AI 898.

Stanfill, C. & Waltz, D.L. (1986). Toward Memory-Based Reasoning. *Communications of the ACM* 29: 1213-1228.

Zavrel, J. & Veenstra, J.B. (1995). The Language Environment and Syntactic Category Acquisition. in: C.Koster & F.Wijnen (eds.), *Proc. of GALA95*, Groningen.