ABSTRACTION IS HARMFUL IN LANGUAGE LEARNING

Walter Daelemans
DILK (Induction of Linguistic Knowledge)
Computational Linguistics
Tilburg University, The Netherlands
and
CNTS (Center for Dutch Language and Speech)
Linguistics
University of Antwerp, Belgium
Walter.Daelemans@kub.nl

1. Abstract

The usual approach to learning language processing tasks such as tagging, parsing, grapheme-to-phoneme conversion, pp-attachment, etc., is to extract regularities from training data in the form of decision trees, rules, probabilities or other abstractions. These representations of regularities are then used to solve new cases of the task. The individual training examples on which the abstractions were based are discarded (forgotten). While this approach seems to work well for other application areas of Machine Learning, I will show that there is evidence that it is not the best way to learn language processing tasks.

I will briefly review empirical work in our groups in Antwerp and Tilburg on lazy language learning. In this approach (also called, instance-based, case-based, memory-based, and example-based learning), generalization happens at processing time by means of extrapolation from the most similar items in memory to the new item being processed. Lazy Learning with a simple similarity metric based on information entropy (IB1-IG, Daelemans & van den Bosch, 1992, 1997) consistently outperforms abstracting (greedy) learning techniques such as C5.0 or backprop learning on a broad selection of natural language processing tasks ranging from phonology to semantics. Our intuitive explanation for this result is that lazy learning techniques keep all training items, whereas greedy approaches lose useful information by forgetting low-frequency or exceptional instances of the task, not covered by the extracted rules or models (Daelemans, 1996). Apart from the empirical work in Tilburg and Antwerp, a number of recent studies on statistical natural language processing (e.g. Dagan & Lee, 1997; Collins & Brooks, 1995) also suggest that, contrary to common wisdom, forgetting specific training items, even when they represent extremely low-frequency events, is harmful to generalization accuracy.

After reviewing this empirical work briefly, I will report on new results (work in progress in collaboration with van den Bosch and Zavrel), systematically comparing greedy and lazy learning techniques on a number of benchmark natural language processing tasks: tagging, grapheme-to-phoneme conversion, and pp-attachment. The results show that forgetting individual training items, however `improbable' they may be, is indeed harmful. Furthermore, they show that combining lazy learning with training set editing techniques (based on typicality and other regularity criteria) also leads to worse generalization results.

I will conclude that forgetting, either by abstracting from the training data or by editing exceptional training items in lazy learning is harmful to generalization accuracy, and will attempt to provide an explanation for these unexpected results.

2. References

Collins, M. and J. Brooks. `Prepositional Phrase Attachment through a Backed-off Model. Proceedings Third Workshop on Very Large Corpora, MIT, 1995.

Daelemans, W. and A. van den Bosch. `Generalization Performance of Backpropagation Learning on a Syllabification Task.' In: M.F.J. Drossaers and A. Nijholt (eds.) Connectionism and Natural Language Processing. Proceedings Third Twente Workshop on Language Technology, 27-38, 1992.

Daelemans, W., Van den Bosch, A., & Weijters, A. `IGTree: Using Trees for Compression and Classification in Lazy Learning Algorithms.' Artificial Intelligence Review 11, 407--423, 1997.

Daelemans, W. `Abstraction Considered Harmful: Lazy Learning of Language Processing.' In: van den Herik, J. and T. Weijters (eds.) Benelearn-96. Proceedings of the 6th Belgian-Dutch Conference on Machine Learning. MATRIKS: Maastricht, The Netherlands, 3--12, 1996.

Dagan, I., L. Lee, F. Pereira. `Similarity-Based methods for Word Sense Disambiguation.' Proceedings 35th ACL - 8th EACL, Madrid, 1997.

Walter Daelemans (1998) **Abstraction is Harmful in Language Learning.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning*, ACL, pp 1-1.