Unsupervised Morphology Induction for Part-of-Speech Tagging

Abstract

In this paper we present an unsupervised morphology induction algorithm that uses Alignment Based Learning (ABL) e. g. (Zaanen, 2001) for hypothesis generation. We show how this algorithm can be used to induce a lexicon and morphological rules for a wide range of natural languages. The resulting morphological rules and structures are optimized during the induction process using a constraint satisfaction model which enforces preferences as to the size and statistical properties of the respective grammars. In particular, we use constraints based on MINIMIUM DESCRIPTION LENGTH, RELATIVE ENTROPY, and MAXIMUM AVERAGE MUTUAL INFORMATION. The resulting morphological segmentation reaches approx. 99% precision over all languages to varying levels of recall.

Given the very precise morphological grammar we were able to generate, lexical classification is performed on the basis of the resulting signatures with simple clustering algorithms, resulting in separation of the elements into basic lexical classes, e. g. verbs and nouns. Different algorithms make use of similar morphological classification in Part-of-speech (POS) tagging, cf. (Brants, 2000), (Lee *et al.* , 2002).

To take a particularly suggestive example, of the words in the WSJ section of Penn that end in "able", 98% are adjectives, and only 2% are nouns (e. g. "cable", "variable") (Brants, 2000). This means that the suffix highly predicts the categorization of the word and is therefore a powerful aid to any POS tagger. Samuelsson (1994) introduced an algorithm to utilize these end-of-word substring "suffixes" to categorize words into POSs by taking probabilities of substring word endings of 7 characters or less and smoothing this over by averaging in the probability with one less character each iteration. TnT (Brants, 2000), a statistical n-gram POS tagger, uses an implementation of this algorithm as the primary component of its algorithm to tag words not seen in the training corpus. (Brants, 2000) reports 89.0% accuracy on these unknown words using the Penn Treebank (Marcus *et al.* , 1993) as a corpus.

Lee et al. (2002) performed a similar experiment on Korean. Their approach uses a morpheme pattern database to automatically tag the agglutinative morphology of Korean. After assigning all possible morpheme tags to a morpheme, the text is run through a statistical POS tagger which uses the Viterbi algorithm to assign word categories. This is then run through a correction layer, using a rule-based correction system. Even though 10% of the words were unknown, Lee reports a tagging accuracy of 97% .

Precision and Recall of the morphologic component of TnT was not reported. Lee et al. reported a 94.9% recall and 89.7% precision on the Korean data. Our algorithm's high precision and lower level of reliance on supervised knowledge makes it an attractive replacement for either of these systems. We will present the results of a comparison between the TnT-tagger and our morphology-induction based category guessing on the Brown corpus (Kucera & Francis, 1967) and the Penn Treebank (Marcus *et al.* , 1993). We can show that a smaller training set is enough to reach higher precision with our algorithm.

References

Brants, Thorsten. 2000 (April 29 – May 3). TnT – a statistical part-of-speech tagger. In Proceedings of the 6th Applied NLP Conference, ANLP-2000.

Kucera, Henry, & Francis, W. Nelson. 1967. *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Lee, Gary G., Lee, J.-H., & Cha, J. 2002. Syllable-Pattern-Based Unknown-Morpheme Segmentation and Estimation for Hybrid Part-of-Speech Tagging of Korean. *Computational Linguistics*, **28**(1), 53–70.

Marcus, Mitchell P., Santorini, Beatrice, & Marcinkiewicz, Mary Ann. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

Samuelsson, Christer. 1994 (3–5 June). Morphological Tagging Based Entirely on Bayesian Inference. *In:* Eklund, R. (ed), *Proceedings of the 9th Nordiska Datalingvistikdagarna (NODALIDA 1993)*.

Zaanen, Menno M. van. 2001. *Bootstrapping Structure into Language: Alignment-Based Learning*. Doctoral dissertation, The University of Leeds.