

# Using Morphological and Distributional Cues for Inductive Part-of-Speech Tagging

Damir Čavar, Paul Rodrigues and Giancarlo Schrementi

Indiana University, Linguistics Dept.  
{dcavar,prrodrig,gischrem}@indiana.edu

## Abstract

In this paper we evaluate the role of morphological and distributional cues in PoS induction, using an incremental and unsupervised learning algorithm with clustering on a vector space.

## 1 Introduction

Morphological properties are intuitively understood to be cues for PoS. Together with inherent lexical and prosodic properties (e. g. length and frequency), positional cues, i. e. distributional properties of a word in the utterance context seem to offer additional cues about its PoS. Classical taggers exploit this information in either handcrafted rule-sets, or via e. g. n-gram and/or entropy models. All these strategies require either language experts, and time and effort, or large collections of annotated corpora to train statistical models. For most of human languages these prerequisites are not given.

## 2 Our Approach

Our approach is an induction approach, where we aim at inductive PoS and mapping to deductive PoS for bootstrapping of NL resources and algorithms. We have shown in previous work that an Alignment Based Learning (ABL) algorithm (Zaanen, 2001) can be used in combination with Minimum Description Length (Grünwald, 1998) and Entropy-based measures to incrementally induce regularities of unknown natural languages. This algorithm induces co-occurrence rules (affixes, infixes and templates),

such that for each morpheme a set of possible morpheme patterns is generated. We have shown that the resulting patterns as such are sufficient to derive the distinction between stems and affixes with simple binary clustering based on K-means or Expectation Maximization (EM), given only the number of morphological co-occurrence patterns (signatures) and the respective morpheme frequencies as the feature vectors. Given two clusters from the set of morphemes, we select the cluster that contains morphemes with the shortest, but most frequent signatures to represent stems (henceforth types). The morphemes from the second cluster represent affixes (henceforth non-types). In evaluations on the basis of Indo-European languages, the resulting morphological segmentation reaches approx. 99% precision over all, to varying levels of recall, depending on the amount of learning or input utterances and the richness of the morphological paradigms found in the respective language. Further detailed evaluations on agglutinative and Semitic languages will be available in the next weeks.

### 2.1 Morphological Cues for PoS Induction

We take an approach as suggested in (Mintz et al., 2002) for token clustering, where lexical co-occurrences are mapped on a vector space.<sup>1</sup> However, in order to take the induced morphology into account, we do not take tokens, but rather the in-

<sup>1</sup>Lexical tokens are represented as vectors where each other word in its context represents a column with the relative frequency as its content. (Mintz et al., 2002) restrict their experiments to the most frequent 1000 and 2000 words, we extend the experiment to all context words, with a window size of word to the left and right.

duced types as the row vectors, and the types of their direct neighbors for the relative frequency cell content. The morphological co-occurrence patterns are added to the distributional information in the vector space by representing all non-types as columns and the respective cell content as the relative frequency of co-occurrence with the corresponding types.

The vector space is used subsequently to induce part-of-speech information for each morpheme using clustering based on similarity metric like Euclidean distance and Cosine similarity. We use agglomerative and fuzzy clustering algorithms based on Soft-K-Means, Density and Expectation Maximization clustering. A specific lexical PoS information is assumed to correspond to the resulting clusters. The success rate is measured in terms of the purity of each resulting cluster, with respect to a specific PoS that we can identify.

### 3 Results

As has been shown elsewhere, simple category information can be derived from basic inherent and distributional properties of tokens. As for morphological cues, various algorithms make use of “quasi” morphological information in PoS tagging, cf. (Brants, 2000), (Gary Geunbae Lee and Lee, 2002). For 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”). 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” for tagging by taking probabilities of substring word endings of 7 characters or less and smoothing this by averaging in the probability with one less character each iteration. TnT (Brants, 2000), uses an implementation of this algorithm as a central. (Brants, 2000) reports 89.0% accuracy on unknown words using the Penn Treebank (Marcus et al., 1993). (Gary Geunbae Lee and Lee, 2002) report a similar experiment on Korean, where they use a morpheme pattern database to tag the agglutinative morphology of Korean. After assigning all possible morpheme tags, the text is run through a statistical PoS tagger which uses the Viterbi algorithm to assign word categories. This is 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 algorithms 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 and 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

- Thorsten Brants. 2000. Tnt – a statistical part-of-speech tagger. In *Proceedings of the 6<sup>th</sup> Applied NLP Conference (ANLP-2000)*, Seattle, WA, April 29 – May 3.
- Jeongwon Cha Gary Geunbae Lee and Jong-Hyeok Lee. 2002. Syllable-pattern-based unknown-morpheme segmentation and estimation for hybrid part-of-speech tagging of Korean. *Computational Linguistics*, 28(1):53–70.
- Peter Grünwald. 1998. *The Minimum Description Length Principle and Reasoning under Uncertainty*. doctoral dissertation, Universiteit van Amsterdam.
- Henry Kucera and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Toben H. Mintz, Elissa L. Newport, and Thomas G Bever. 2002. The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26:393–424.
- Christer Samuelsson. 1994. Morphological tagging based entirely on bayesian inference. In R. Eklund, editor, *Proceedings of the 9th Nordic Conference on Computational Linguistics, 9th Nordiska Datalingvistikdagarna (NODALIDA 1993)*, Stockholm, Sweden, June.
- Menno M. van Zaanen. 2001. *Bootstrapping Structure into Language: Alignment-Based Learning*. Doctoral dissertation, The University of Leeds.