A New Algorithm for Unsupervised Induction of Concatenative Morphology

Harald Hammarström

Chalmers University of Technology
412 96 Gothenburg
Sweden
harald2@cs.chalmers.se

1 Introduction

This paper sketches a new algorithm for unsupervised induction of concatenative morphology. The algorithm differs markedly from previous approaches in both segmentation and paradigm induction. It is illustrated here with the respect to suffixes, using the following notation:

- W: the set (not bag) of words in the corpus
- $s \triangleleft w$: s is a suffix of the word w i.e there exists a (possibly empty) string x such that $w = xs$
- $Stems(s) = \{x | xs \in W\}$: the set of all strings ("stems") that make a word in the corpus if appended with s
- $f(s) = |\{w \in W | s \triangleleft w\}|$: the number of words with suffix s (equals $|Stems(s)|$)
- $s_i(w)$: the suffix of w that begins at position $0 \leq i \leq |w|$
- $Q(w) = \{s_i(w) | i < |w|\}$: the set of (non-empty) suffixes of s
- $S = \bigcup_{w \in W} Q(w)$: all suffixes in the corpus

2 Segmentation

The segmentation takes a corpus as input and output a ranked list of (all) suffixes. The ranking is meant to say how salient a suffix is for the language of the corpus, and is computed in three steps:

1. **Relative Frequency Increases:** Define $Z : S \times W \rightarrow \mathbf{Q}^+ \cup \{\mathbf{0}\}$:

$$Z(s,w) = \begin{cases} 0 & \text{if not } s \triangleleft w \\ 1 & \text{if } s = s_0(w) \\ \frac{f(s_i)}{f(s_{i-1})} & \text{if } s = s_i(w) \text{ for some } 0 < i < |w| \end{cases} \quad (1)$$

   Note that f, and hence Z, depends on W.
2. **Accumulation:** Calculate $Z^W : S \rightarrow \mathbf{Q}^+$:

$$Z^W(s) = \sum_{w \in W} Z(s,w) \quad (2)$$

3. **Re-scale:** Scale on suffix-length by a parametre $p = 2$: $\overline{Z^W}(s) = |s|^p \cdot Z^W(s)$

A. Yli-Jyrä, L. Karttunen, and J. Karhumäki (Eds.): FSMNLP 2005, LNAI 4002, pp. 288–289, 2006.
© Springer-Verlag Berlin Heidelberg 2006

3 Paradigm Induction

The paradigm induction phase outputs a ranked list of paradigms given a ranked list of segmented suffixes. By paradigm we simply mean a non-empty set of suffixes, and the ranking is meant to convey how salient a declension pattern is for the language in question. At first glance, the task of finding paradigms looks exceedingly difficult since the number of theoretically possible paradigms is exponential in the number of suffixes, and paradigms in real languages are often not mutually disjoint. Moreover, in a corpus of a real language we cannot expect to rely on there existing words that occur in all its forms.

1. **Testing Paradigm Heuristic:** Suppose we have a hypothesis of a paradigm P. We give a test metric using the idea that suffixes of P ought to show up on the "same stems". First, for each suffix $x \in S$, define its quotient function $H_x(y) : S \to [0, 1]$ as:

$$H_x(y) = \frac{|\{z|\exists z(z \in Stems(x) \wedge zy \in W)\}|}{|Stems(x)|} \tag{3}$$

   Construct a rank by summing the quotient functions of the members of P:

$$V_P(y) = \sum_{x \neq y \in P} H_x(y) \tag{4}$$

   The $Rank_P(x) : S \to \mathbf{N}$ is then simply $|\{y|V_P(y) > V_P(x)\}|$.
   Now, the test $VI(P)$ is a measure of how "high up" the sum of ranks of the members of P are, compared to the optimal sum (which depends on $|P|$ and is $0 + \ldots + |P| - 1$):

$$VI(P) = \frac{|P|(|P| - 1)}{2\sum_{x \in P} rank_P(x)} \tag{5}$$

2. **Gradient Search:** It is intractable to list all hypotheses of paradigms P, thus we suggest a way to "grow" paradigms. Start with a one member paradigm and greedily improve the VI-score, by successively adding or taking away one suffix at a time (until the score doesn't improve by a one-member change):

$$G(P) = argmax_{p \in \{P\} \cup \{P \text{ xor } s | s \in S\}} VI(p) \tag{6}$$

$$G^*(P) = \begin{cases} P & \text{if } G(P) = P \\ G^*(G(P)) & \text{if } G(P) \neq P \end{cases} \tag{7}$$

   Where P xor s means $P \setminus \{s\}$ if $s \in P$ and $P \cup \{s\}$ if $s \notin P$.

Naturally, the induced paradigms $A = \{G(\{s\})|s \in S\}$ are all those that can be grown from (at least one) suffix in the corpus, and finally we rank them by their VI-score and average "suffixness" of its members:

$$R(P) : A \to \mathbf{Q}^+ \cup \mathbf{0}$$

$$R(P) = \frac{VI(P)}{|P|} \sum_{s \in P} \overline{Z^W}(s) \tag{8}$$