Induction of a Stem Lexicon for Two-level Morphological Analysis

Erika F. de Lima
Institute for Natural Language Processing
Stuttgart University
Azenbergstr. 12
70174 Stuttgart, Germany
delima@ims.uni-stuttgart.de

Abstract

A method is described to automatically acquire from text corpora a Portuguese stem lexicon for two-level morphological analysis. It makes use of a lexical transducer to generate all possible stems for a given unknown inflected word form, and the EM algorithm to rank alternative stems.

1 Motivation

Morphological analysis is the basis for most natural language processing tasks. Hand-coded lists used in morphological processing are expensive to create and maintain. A procedure to automatically induce a stem lexicon from text corpora would enable the creation, verification and update of broad-coverage lexica which reflect evolving usage and are less subject to lexical gaps. Such a procedure would also be applicable to the acquisition of domain-specific vocabularies, given appropriate corpora.

In the following, a method is described to automatically generate a stem lexicon for two-level morphological analysis (Koskenniemi, 1983). The method, which was implemented and tested on a newspaper corpus of Brazilian Portuguese, is applicable to other languages as well.

2 Method

The learning algorithm consists of a procedure which attempts to determine the stem and part of speech for each (unknown) inflected form in its input. For instance, given the inflected form *recristalizações* ('recrystallizations'), the procedure induces that *cristal* ('crystal') is a noun, and adds it to the set of learned stems.

The system makes use of a two-level processor–PC-KIMMO (Antworth, 1990)–to generate a set of putative stems for each inflected form in its input. (For a detailed account of the PC-KIMMO two-level framework, see (Antworth, 1990).) In order to morphologically analyze its input, the processor makes use of a set of two-level rules, a lexicon containing inflectional as well as derivational affixes, and a unification-based word grammar. No stem lexicon is provided to the system. In the word grammar and lexical transducer, a stem is defined to be a nonempty arbitrary sequence of characters.

The current system contains 102 two-level rules, accounting for plural formation, e.g., *cristal* ('crystal') - *cristais* ('crystals'), diminutive and augmentative formation, e.g., *casa* ('house') - *casinha* ('house-DIM'), feminine formation, e.g., *alemão* ('German-MASC') - *alemã* ('German-FEM'), superlative formation *pagão* ('pagan') - *pananíssimo* ('pagan-SUP'), verbal stem alternation, e.g., *dormir* ('to sleep') - *durmo* ('sleep-1P-SG-PRES'), and derivational forms, e.g., *forum* ('forum') - *forense* ('forensic'). The affixes lexicon consists of 511 entries, of which 236 are inflectional and 275 derivational. The unification-based word grammar consists of 14 rules to account for prefixation, suffixation, and inflection.

Each word parse tree produced for an inflected form yields a putative stem and its part of speech through the constraints provided by the grammar and affix lexicon. For instance, given the unknown inflected form *cristalizar* ('crystallize'), and the constraint that the suffix *izar* ('ize') may only be applied to nouns or adjectives to form a verb, the system induces that the string *cristal* ('crystal') is possibly a nominal or adjectival stem.

Since a stem is defined to be an arbitrary nonempty string, a parse forest is usually produced for each inflected form, yielding a set of putative stems, each corresponding to one parse tree. In order to establish the correct stem for an inflected form, the learning procedure attempts to combine the accumulated evidence provided by related word forms, i.e., word forms sharing a common stem. For instance, the word *recristalizações* ('recrystalliza-

Erika F. de Lima (1998) **Induction of a Stem Lexicon for Two-level Morphological Analysis.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning,* ACL, pp 267-268.

tions') shares a common stem with related words such as *cristal* ('crystal'), *cristalino* ('crystalline-MASC'), *cristaliza* ('crystallize-3P-SG-PRES'), etc. The EM algorithm, used to assign a probability to each stem in a set, makes use of this fact to determine the most probable stem.

3 EM Algorithm

The system uses the expectation maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) to assign probabilities to each stem in a set, given all sets obtained for a given corpus of inflected word forms. In the current setting, the algorithm is defined as follows.

Algorithm. Let S be a set of *stems*. Further, let $\mathcal{S}$ be a finite set of nonempty subsets of $\wp(S)$, and let $S_0 = \bigcup_{X \in \mathcal{S}} X$. For each stem x in S_0:

Initialization:

$$c_0(x) = \sum_{X \in \mathcal{S}} (I(x, X) \cdot g_{\mathcal{C}}(X))$$

Step $k+1$:

$$c_{k+1}(x) = c_k(x) + \sum_{X \in \mathcal{S}} (P_k(x, X) \cdot g_{\mathcal{C}}(X))$$

Where $g_{\mathcal{C}}$ is a function from $\mathcal{S}$ to the natural numbers mapping a set X to the number of times it was produced for a given corpus $\mathcal{C}$ of inflected forms, and I, P_k, and p_k are functions defined as follows:

$$\begin{array}{lll} I: & S \times \wp(S) & \to [0,1] \\ & (x, X) & \mapsto \begin{cases} \frac{1}{|X|} & \text{if } x \in X \\ 0 & \text{else} \end{cases} \end{array}$$

$$\begin{array}{lll} P_k: & S \times \wp(S) & \to [0,1] \\ & (x, X) & \mapsto \begin{cases} \frac{p_k(x)}{\sum\limits_{\bar{x} \in X} p_k(\bar{x})} & \text{if } x \in X \text{ and } |X| > 1 \\ 0 & \text{else} \end{cases} \end{array}$$

$$\begin{array}{lll} p_k: & S & \to [0,1] \\ & x & \mapsto \frac{c_k(x)}{\sum\limits_{\bar{x} \in S_0} c_k(\bar{x})} \end{array}$$

A stem x is considered to be *best in the set X at the iteration k* if $x \in X$ and $p_k(x)$ is an absolute maximum in $\bigcup_{\bar{x} \in X} p_k(\bar{x})$.

In the experiment described in the next section, a set of stems was considered disambiguated if it contained a best set at the final iteration; the final number of iterations was set empirically.

4 Results

The method described in the previous sections was applied to a newspaper corpus of Brazilian Portuguese containing 50,099 inflected word types. The system produced a total of 2,333,969 analysis (putative stems) for these words. Of the 50,099 stem sets, 33,683 contained a best stem.

In order to measure the recall rate of the learning algorithm, a random set of 1,000 inflected word types used as input to the system was obtained, and their stems manually computed. The recall rate is given by the number of stems learned by the system, divided by the total number of stems, or 42,3%. The low recall rate is due partially to the fact that not all sets produced by the system contained a best stem. The system produced partial disambiguation for 15,814 of the original 50,099 sets, e.g., after the final iteration, there was a proper subset of stems with maximal probability, but no absolute maximum. A large number of partial disambiguations involved sets containing a stem considered to be both an adjective and a noun, e.g., {AJ *stem*, N *stem*}. This reflects the fact that very often Portuguese words are noun-adjective homographs, and assignment to one category cannot be made based on the morphological evidence alone. If the system were to consider partial disambiguation as well, the recall rate could be significantly improved.

In order to evaluate the precision of the learning algorithm, a random set of 1,000 stems produced by the system was compared to the judgements of a single judge. The precision of the system is given by the number of correct learned stems divided by the total number of learned stems, or 70.4%. A small percentage of errors was due to the fact that closed-class words were assigned open-class word categories. A closed-class word lexicon would eliminate these errors. Spelling errors are another source of errors. Taking frequency of occurrence into account would alleviate this problem. By far the largest percentage of errors was due to the fact that the system was not able to correctly segment stems, mostly due to incorrect prefixation. In order to improve precision, the system should make use of not only of the stem provided by each parse tree, but take the structure itself into account in order to correctly determine the stem boundaries.

References

Antworth, Evan L. 1990. *PC-KIMMO: a two-level processor for morphological analysis.* Summer Institute of Linguistics, Dallas.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from inclomplete data via the EM algorithm. *J.R.Statis. Soc. B*, 39:1–38.

Koskenniemi, Kimmo. 1983. *Two-level morphology: a general computational model for word-form recognition and production.* University of Helsinki Department of General Linguistics, Helsinki.