Development of a morphological guesser for Hungarian

Attila Novák[12], Viktor Nagy[2], and Csaba Oravecz[2]

[1] MorphoLogic Ltd., Budapest
[2] Research Institute for Linguistics, Budapest
{novak,nagyv,oravecz}@nytud.hu

Keywords: guesser, morphological analysis, similarity of distributions, disambiguation

Highly inflectional/agglutinative languages like Hungarian typically feature possible word forms in such a magnitude that machine learning methods which are most often used in NLP and rely on training data almost always face the problem of data sparseness. This problem cannot simply be tackled by independently preparing huge morphological dictionaries; these will grow to sizes unmanageable to any efficient application. A hand-on solution to this problem is to apply a comprehensive morphological analyser, which works in tandem with a base form lexicon and has the capability of analysing all inflectional, productive derivational and compounding phenomena and is also capable of doing base form reduction. Although essentially being a symbolic tool, such an analyser can be efficiently utilized even in a stochastic NLP environment.

Independently of the type of the source that provides the lexical information, morphological processing of huge corpora inevitably faces the problem of a large number of unknown word forms. For the symbolic analyser, this means that the particular base form is not listed in the analyser's lexicon so its derivatives cannot be analysed. (Many of these are of foreign origin with irregular orthography, which poses a special problem in Hungarian where suffixation is primarily determined by the phonological shape of the stem which is not reflected by the orthographic form of these words in any consistent way.) In order to cope with the problem of unknown words in unconstrained corpora, generally some stochastic method is used based on suffix models built from training corpora and aided by some external information like capitalization. However, the direct application of these models, even when supported by information from very large corpora, is debatable in the case of languages like Hungarian.

As far as the external information is concerned, its use does not contribute significantly to the knowledge of the guesser model. In Hungarian, almost any type of special tokens can get suffixation very productively, so e.g. the information that some form is capitalized so it is probably a proper noun is uninformative for the system: it does not deliver the essential information on the base form and the details of suffixation attached to it.

As for the stochastic suffix models, it is worth noting, that with the application of a morphological analyser there is an important difference in the nature of the unknown word problem: we have to handle word forms unknown to the morphological analyser and not word forms not found in training corpora. Fixed and variable length suffix models are based on annotated training corpora and in Hungarian face the same data sparseness problem as any other pure stochastic NLP method. Models built upon unannotated corpora of potentially unlimited in size introduce a huge search space in our case which is difficult to manage computationally. In addition, when using these models in a practical application, a fairly strict limit must be set on the maximal length of the suffixes to be considered. But in Hungarian, due to the agglutinating nature of the language, very long inflectional suffix sequences do occur, which poses an inherent problem for all purely stochastic suffix models.

The paper present a combined method for unknown word guessing featuring symbolic constraints and statistical information. The former is embodied in a partial word form analyser (guesser) which generates hypotheses on possible lemma-plus-suffix sequences along with properties which can be inferred for the lemma from the suffix sequence. This hypothesis space is then pruned using statistical information concerning word form and suffix sequence distribution gathered from a 150 million word corpus analysed by the morphological analyser. The morphological knowledge built into the symbolic guesser is directly derived from the linguistic description used for the creation of the morphological analyser.

Since unknown words in general tend to belong to productive inflectional and derivational paradigms the hypothesis space can effectively be reduced in the first place by considering only these paradigms in the partial analysis. To associate weights to the outputs of the partial analyser and to exclude improbable analyses several models are developed based on the statistical information from the corpus. Measures evaluated range from simple relative stem frequency to similarity measures like L1 norm between stem/suffix distribution proposed for the unknown forms and stem/suffix distribution of the known word forms. Evaluation is carried out from two perspectives: with respect to the extension of the lexicon of the morphological analyser with base forms gained from the unknown word analysis, and with respect to the induction of lexical probabilities for the unknown forms; these probabilities are used by a part-of-speech tagging system especially well suited and robust for morphological processing of unconstrained Hungarian language data.