Combining symbolic and statistical methods in morphological analysis and unknown word guessing

Attila Novák[*†], Viktor Nagy[*], Csaba Oravecz[*]

[*]Research Institute for Linguistics
Benczúr u. 33., Budapest, Hungary
{novak, nagyv, oravecz}@nytud.hu

[†]Morphologic Ltd.
Orbánhegyi út 5., Budapest, Hungary

Abstract

Highly inflectional/agglutinative languages like Hungarian typically feature possible word forms in such a magnitude that automatic methods that provide morphosyntactic annotation on the basis of some training corpus often face the problem of data sparseness. A possible solution to this problem is to apply a comprehensive morphological analyser, which is able to analyse almost all wordforms alleviating the problem of unseen tokens. However, although in a smaller number, there will still remain forms which are unknown even to the morphological analyzer and should be handled by some guesser mechanism. The paper will describe a hybrid method which combines symbolic and statistical information to provide lemmatization and suffix analyses for unknown word forms.

1. Introduction

The problem of data sparseness is ubiquitous in stochastic morphosyntactic annotation systems trying to cope with highly inflectional/agglutinative languages like Hungarian. The number of word forms can reach such a magnitude that no training corpus thus far constructed contains enough of them for efficient annotation. The problem cannot simply be tackled by independently preparing huge morphological dictionaries (Hajič, 2000); these will grow to sizes unmanageable to any efficient application. A hand-on solution might be to apply a comprehensive morphological analyser (Prószéky and Kis, 1999), 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 annotation environment (Oravecz and Dienes, 2002).

Independently of the type of the source that provides the lexical information, morphological processing of huge corpora inevitably faces the problem of a significant number of word forms missing not only from the training coprus but also from the external knowledge base. If this external knowledge is provided by a symbolic morphological analyzer then this means that the particular base form is not listed in the analyser's lexicon so its derivatives cannot be analyzed. In order to tackle this problem a combined method can be applied utilizing symbolic constraints and statistical information from large unannotated corpora. The paper will describe and empirically investigate how this method can be put into practice to improve on the output of the morphosyntactic annotation. Section 2. will give a brief overview of related work with emphasis on the problems that arise when they are directly applied to Hungarian. Section 3. will describe the symbolic guesser module while section 4. the domain of application of the statistical information. In section 5. we will present an evaluation of the methods with several testing scenario in the context of morphosyntactic disambiguation. Conclusions and suggestion for further work will end the paper in section 6.

2. Related work

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 morphology external information like capitalization (Weischedel et al., 1993). However, the direct application of these models, even when supported by information from very large corpora (Cucerzan and Yarowsky, 2000), can be debatable in the case of Hungarian, given the agglutinative nature of the language and the limited size of available hand annotated corpora.

As for the stochastic algorithms, in Hungarian, fixed and variable length suffix models based on annotated training corpora face the same data sparseness problem as any other pure stochastic NLP method. Models built upon unannotated corpora of potentially unlimited size (Cucerzan and Yarowsky, 2000) introduce a huge search space in our case which might be difficult to manage computationally. In addition, when using these models in a practical application, a limit must be set on the maximal length of the suffixes to be considered. In Hungarian, due to the agglutinating nature of the language, very long inflectional suffix sequences do occur, which might pose a problem for purely stochastic suffix models. Table 1 illustrates the error rate in terms of word form types and tokens with regard to suffix cutback threshold (SCT) on the basis of a 120 million corpus analyzed by the MA. Any threshold set below the specified suffix length will leave no chance whatsoever for a guessing algorithm to correctly analyze word forms having longer suffixes. (It is true, however, that e.g. for a threshold of 6 we would only lose 0.2% of tokens from this corpus but this still is an unwanted handicap.)

Another issue concerns the identification of potential stems and the segmentation of suffixes in a suffix sequence for unknown tokens. For higher level of language process-

SCT	Cumulative error rate			
	Tokens	%	Types	%
0	49778192	41.74%	1774516	81.21%
1	40719741	34.14%	1603942	73.40%
2	25662692	21.52%	1168639	53.48%
3	11356258	9.52%	656921	30.06%
4	4089022	3.43%	335040	15.33%
5	1067129	0.89%	111623	5.11%
6	197002	0.17%	26178	1.20%
7	55263	0.05%	7306	0.33%
8	14625	0.01%	859	0.04%
9	1683	0.00%	147	0.01%
10	39	0.00%	24	0.00%
11	2	0.00%	2	0.00%

Table 1: Cumulative error rate with respect to SCT

ing in an agglutinative language, to make use of essential linguistic information (e.g. subcategorization frames, semantic information), this should be attached to stems and not word forms for efficient processing. Therefore stemming or lemmatization is of great importance even in unknown word guessing. Furthermore, individual suffixes carry important linguistic information so they must be identified as separate elements in a possible suffix sequence attached to a word stem. Current statistical guessers are not prepared to perform these tasks.

The worst problem of all is that a stochastic tool can rely only on the set of tags present in the training corpus to produce analyses. Unfortunately, the tag coverage of this can be far away from the set of analyses found in large corpora.[1]

We are not arguing here that stochastic algorithms as such cannot be extended to overcome some of these problems. Rather we propose that once a symbolic morphological analyzer is available for a language one can build an efficient guesser around it and not lose useful statistical information either.

3. The symbolic guesser module

The symbolic guesser module is built around 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.[2] The morphological knowledge built into the symbolic guesser is directly derived from the linguistic description used for the creation of the morphological analyser.

It is worth noting, that with the application of a morphological analyzer (MA) in an annotation system, there is an important difference in the nature of the unknown word problem: we have to handle word forms unknown to the morphological analyzer and not word forms not found in training corpora.[3]

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. This resulted in fairly restrictive constraints on the possible forms of open class word classes, particularly verbs. On the other hand, many of the unknown word forms are of foreign origin with an 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. For this reason, a number of constraints, observed when creating the database of the regular morphological analyser (e.g. vowel harmony), had to be relaxed or discarded. Other phonological and orthographic constraints on suffixation which are not violated even by stems of irregular orthography are directly encoded in the data and are checked by the guesser. Figure 1 presents the output of the symbolic guesser module (N=noun, NOM=nominative, PL=plural, DAT=dative, PSe3=third person singular possessive; stem is separated from the morphosyntactic tags by a '\').

```
guesser>Ginának
Ginának\[N][NOM]
Ginán\[N][PL][NOM]
Giná\[N][DAT]
Gina\[N][DAT]
Gin\[N][PSe3][DAT]
```

Figure 1: Sample output of the symbolic guesser module

4. Stochastic filters

The hypothesis space of the symbolic guesser is pruned using statistical information concerning word form and suffix sequence distribution gathered from a 120 million word corpus analyzed by the morphological analyzer. To associate weights to the outputs of the partial analyzer and to exclude improbable analyses several models were developed based on the statistical information from the corpus. In Novák et al. (2003), we evaluated various measures ranging 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. Here we selected the two best performing models for evaluation in a POS disambiguation task. In Model 1., the selection of a particular analysis for a word form was driven by the corpus frequency of the form the guesser proposed as a stem in the given analysis. That is, the analysis whose stem appeared the highest number of times as an independent token in the corpus was chosen as the preferred reading of the word form. This was augmented with a filter which worked as follows: if the MA did

[1]Our 270k word training corpus contains around 800 tags in contrast to the 3100 manifest in the large 120 million word corpus.

[2]For brevity's sake, we skip much of a detailed description, not only here but in the next section as well. Details can be found in (Novák et al., 2003).

[3]Assuming that these unknown forms, when provided with possible analyses by the guesser, will be handled the same way in the annotation system as the forms analyzed by the MA.

Test		Model	Performance (accuracy)		
			Overall	Known tokens	Unknown tokens
0.	Baseline	TnT suffix guess	97.35%	97.75%	81.19%
1.	Uniform	Model 1.	97.26%	97.75%	77.38%
		Model 2.	97.27%	97.75%	77.57%
2.	Weighted	Model 1.	97.41%	97.75%	83.65%
		Model 2.	97.42%	97.75%	83.90%

Table 2: The performance of the guesser module.

manage to assign analyses to the guessed stem, but none of these analyses was compatible with the proposed stem category (e.g. the stem had an analysis as a verb form and the proposed stem was a noun), the analysis was discarded.

In Model 2., the analyses produced by the MA were consulted like in Model 1. But in addition to filtering out incompatible analyses, the stem category tag for compatible analyses was changed to that proposed by the MA, and the measure used for these modified analyses was not the plain stem form frequency, but the frequency of all analyses produced by the MA containing the proposed stem. For stems left unanalyzed by the MA, word form frequency is used instead of stem frequency as in the previous models. A sample output is illustrated in Figure 2.

```
Gina\[N][DAT] (794)
Ginának\[N][NOM] (49)
Gin\[N][PSe3][DAT] (48)
Ginán\[N][PL][NOM] (2)
Giná\[N][DAT] (0)
```

Figure 2: Output of the guesser in Model 2

5. Evaluation

Evaluation is carried out with respect to the induction of possible analyses and their respective lexical probabilities for unknown word forms in a part-of-speech tagging system developed especially for morphological processing of unconstrained Hungarian language data (Oravecz and Dienes, 2002). In the original test bed, there were no unknown tokens allowed, those not seen in the training corpus were all either present in the lexicon of the MA plugged into the system or were added to its user dictionary. Here, this artificial constraint was relaxed and tokens unknown to the MA are also presented to the system and handled in two ways described below.

The annotation system is the same as in Oravecz and Dienes (2002): a symbolic morphological analyzer together with TnT (Brants, 2000) as the POS tagger. As a baseline model for the analysis of unknown words (i.e. those not present in the training corpus nor in the lexicon of the MA), the suffix guessing algorithm built in TnT was applied, while the guesser module was tested in 2 different scenarios. In Test 1., uniform distribution was assumed over the analyses proposed by the guesser, i.e. practically no statistical information was used in this scenario. Test 2. applied the weighted distribution as output from the stochastic filters. Each test was run on both models.[4]

The main figures for the test data are as follows: for training the tagger we used a manually disambiguated corpus of 270.000 tokens (56000 types). The test corpus consisted of 68100 tokens (18500 types). The percentage of unknowns is 2.4%. The results of the experiments are presented in Table 2.

The best performance was achieved by Model 2., when the distribution output from the filter was used. Model 1. performed comparably in this scenario. The performance of the baseline TNT suffix guesser was slightly worse than those of our guesser models. However, both models performed much worse when the stem distribution information output from them was ignored.

We also performed tenfold cross validation on the training corpus. Overall results were somewhat worse than those obtained on the independent test set and the standard deviation was large. This might be due to the fact that the ratio of unknown word forms was much lower in this corpus than in other corpora of comparable size, since the training corpus was assembled with some care to keep the number of unknown words low to save human labor when annotating/disambiguating the corpus. For this reason, idiosyncratic unknown words seemed to play a statistically significant role in the tests resulting in a great deviation in the results. Therefore we considered the results presented here more reliable.

6. Conclusion and further work

In this paper we compared the performance of some variations of a combined symbolic/stochastic guesser model with that of the purely stochastic suffix guessing model built into the TNT tagger in a POS disambiguation task. The suffix guesser built into TNT uses conditional tag probabilities given word ending learnt from a manually annotated training corpus. It does not provide lemma or suffix segmentation information, at least the former of which seems indispensable (and the latter is also often useful) for further linguistic analysis of the corpus.

Our guesser model is based on stem statistics gathered form a huge unannotated corpus, and it does produce the lemma and suffix analysis missing from the TNT guesser.

[4]Several other scenarios were tested such as the linear combination of the ambiguity class distribution from the training corpus and from the guesser but these did not result in any improvement on the tests presented here.

In our experiments, the two models yield comparable performance in the POS disambiguation task. We expect that a combination of the two models: using statistical information concerning both stem and ending would yield even better results.

The main conclusion of Oravecz and Dienes (2002) still holds: in the case of languages like Hungarian, symbolic morphological analysis seems indispensable in order to achieve acceptable POS disambiguation performance (when using a manually disambiguated training corpus of limited size), and the most effective way of boosting performance is improving the coverage of the morphological analyzer. Our next goal will thus be the adaptation of our guesser models to the task of learning lexical information from the corpus for inclusion in the stem database of the MA. This task amounts to the identification of stems to be added along with all their unpredictable (irregular) morphological/morphosyntactic properties which can be inferred from the corpus. We plan to enrich the symbolic guesser to provide this kind of information (e.g. the proposed analysis presupposes that the word is irregular with regard to vowel harmony.) This type of information could then also be used to refine the ranking of hypothetical analyses.

7. References

Thorsten Brants. 2000. Tnt - a statistical part-of-speech tagger. In *6th Applied Natural Language Processing (ANLP '00), April 29 - May 4*, pages 224–231, Seattle, USA. Association for Computational Lingusitics.

Silviu Cucerzan and David Yarowsky. 2000. Language independent minimally supervised induction of lexical probabilities. In *Proceedings of the 38th Annual Meeting of the Associations for Computational Linguistics*, pages 270–277, Hong Kong.

Jan Hajič. 2000. Morphological tagging: Data vs. Dictionaries. In *Proceedings of ANLP-NAACL Conference*, pages 94–101, Seattle, Washington, USA.

Attila Novák, Viktor Nagy, and Csaba Oravecz. 2003. Corpus assisted development of a Hungarian morphological analyser and guesser. In Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, editors, *Proceedings of the Corpus Linguistics 2003 Conference*, number 16 in UCREL Technical Paper, pages 583–590. Lancaster University.

Csaba Oravecz and Peter Dienes. 2002. Efficient stochastic part-of-speech tagging for Hungarian. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC2002), May 29-31*, pages 710–717, Las Palmas, Canary Islands, Spain.

Gábor Prószéky and Balázs Kis. 1999. Morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 261–268, College Park, Maryland, USA.

Ralph Weischedel, Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeff Palmucci. 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, 19(2):359–382.