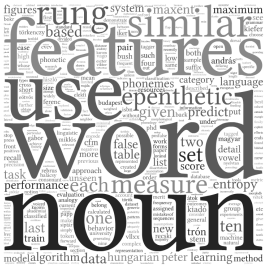

Identifying Epenthetic Nouns using Maximum Entropy Classification

Gábor Recski and András Rung

1 Introduction

This paper will demonstrate the use of a generic machine learning algorithm in predicting a particular morphological behavior exhibited by a group of Hungarian nouns. *Epenthetic nouns* (eg *bokor* 'bush') appear without the last vowel phoneme under certain conditions (*bokr#ok* 'bush-PLUR', *bokr#ot* 'bush-ACC'). Rule-based accounts of this phenomenon, whether generative (Vago 1980, Törkenczy 1994) or traditional (Papp 1975, Elekfi 1994), do not enable us to predict the behavior of new, unseen words. More recently, Rebrus & Törkenczy (2008) have explored the effects that previously attested noun forms may have on the the behavior of new stems, but without providing an algorithm for predicting the behavior of unknown words.

The methods described in this paper make no use of linguistic analyses put forward in the literature. On the contrary, we wish to demonstrate that given a sufficiently large data sample it is possible to predict the behavior of unseen nouns with high accuracy while using minimal linguistic knowledge. In §2 we give a brief account of an analogy-based approach put forward by (Rung 2012). This approach limits phoneme representations to a small set of phonetic features and makes use of a fairly sophisticated measure of word similarity, called the *complex feature measure*. §3 will use the same data and the same representation to train a Maximum Entropy classifier that predicts epentheticity without adopting any notion of phonetic similarity. For both experiments we use a dataset containing ca. 1000 epenthetic and 48 000 non-epenthetic nouns. The sample was compiled using the *morphdb.hu* morphological database (Trón et al. 2006), the annotation was manually improved (see Rung 2012 for details).

2 Analogy-based approaches and the complex feature measure

The first method for predicting epenthetic behavior of nouns is based on similarity of words. Given some training data, ie a database of nouns whose status is known, the algorithm will assume that new, unseen words behave as their closest neighbor in the training set. Details of the algorithm are given in (Rung 2012), we will only summarize the most important characteristics of the system.

Similarity is measured based on the feature representation of phonemes in each word. Vowels are represented by the binary features HIGH, LOW, LONG, SHORT, FRONT, BACK, ROUND, UNROUND, consonants by AFFRICATE, ALVEOLAR, APPROXIMANT, BILABIAL, CONTINUANT, CORONAL, DORSAL, FLAP, FRICATIVE, GLOTTAL, LABIAL, LABIODENTAL, LATERAL, NASAL, OBSTRUENT, PALATAL, PLOSIVE, RHOTIC, SPREAD GLOTTIS, STOP, VELAR, VOICED, VOICELESS. The selection of these features was based on descriptions by Kiefer (1994) and Siptár & Törkenczy (2000).

In order to choose, given some unseen word, the one most similar to it in the training data, an exact metric of similarity must be given. Rung's metric, the *Complex Feature Measure* compares two words by aligning their phonemes, starting by the last one, and measuring differences between phoneme pairs. Each articulatory feature the two phonemes do not share reduces their similarity by a factor of 2: identical phoenemes have a similarity of 1, those differing in one feature have similarity 0.5, etc. Similarity between a vowel and a consonant is zero. Word similarity is calculated as the weighted average of phoneme similarities, where phoneme pairs contribute less as we get further from the right edge of the stem (details are given in Rung 2012). Given some unseen word, its distance from all words in the sample data is calculated and the category of the nearest neighbor is assigned to it. Performance of the algorithm was further improved with one tweak: words whose last four characters do not match the CV pattern of the word under examination are left out of the comparison (unless no word matches).

3 The Maximum Entropy approach

3.1 Maximum Entropy Learning

We will now attempt to use a more general machine learning approach that is widely used in various classification tasks, both within and outside the domain of natural language processing. Maximum Entropy (or *maxent* for

short) is a widely used, general algorithm for supervised learning. Given some set of events (in our case, words), each of which are represented by a finite set of features (eg phonetic ones), and each of which are assigned one of finitely many categories or *outcomes,* eg whether they are epenthetic or not, maximum entropy learning will produce a model that assigns scores to each *(feature, outcome)* pair that describe the contribution of the feature to the likelihood of the word belonging to the category. When confronted with new, unseen words, the model will thus determine the probability of each category based on its features. The mathematical details are discussed in Ratnaparkhi (1998).

Any application of Maximum Entropy learning to some classification task will involve establishing a set of features used to represent the entities that are to be classified, all of which encode information that might be relevant to the classification task. It is then up to the algorithm to determine, by assigning weights, which of these features are more important than others. Applying the maxent method to the identification of epenthetic nouns simply means generating features for each word and feeding the representations to the maxent learner.

3.2 Features

When choosing features to encode words for the current task, we limited our representation to the set of phonetic features used by the algorithm described in the previous section. Since the maxent method can efficiently handle tens of thousands of word features, the majority of which are probably irrelevant to the categorization task, it is feasible to represent each word with thousands of features. Working on the minimal assumption that epentheticity somehow depends on the rightmost phonemes of a word and that it is sensitive to particular sequences of phonetic properties, we decided to create word features by considering a suffix of arbitrary size and generating all *feature sequences* present in the suffix.

In our experiments we found it optimal to consider suffixes of length 5 and sequences of at most 4 features. For example, the word *kereskedelem* will receive features such as -4_0_VOWEL_CONSONANT_VOWEL_CONSONANT, which simply indicates that the word ends in a VCVC sequence, but also every possible combination of phonetic features that these last four phonemes have, such as -4_0_UNROUND_CONSONANT_FRONT_VOICED. In fact, the word *kereskedelem* will be represented by 2430 distinct features.

4 Evaluation

4.1 Methodology

In order to evaluate the system, we used the standard method of *tenfold cross-validation.* We divided the entire data (49 466 nouns) into ten random parts of equal size and performed ten independent experiments: in each run, we used nine sets to train a model which we then used to predict the categories of words in the last tenth of the data. We then average the ten scores to obtain overall figures of merit. The standard method for measuring classifier performance are the three figures known as *precision, recall,* and *F-score*: given a dataset of N words, n of which belong to a category under evaluation, K of which were classified as belonging to this category, k of which were correctly classified, precision and recall will be calculated as k/K and k/n respectively. Neither of these two figures alone can adequately measure the performance of a system: a model that predicts every noun in a corpus to be epenthetic will achieve 100% recall, while the converse model that tags no noun as epenthetic will achieve 100% precision. It is therefore customary to calculate the harmonic mean of the two figures (2PR/(P+R)), the F-score, to measure general performance.

4.2 Results

Table 1 lists figures of merit for the best model along with the figures obtained using the complex feature measure approach described in the previous section. Both methods have been evaluated using tenfold cross-validation, standard deviation figures are also available for the current experiment. Since most nouns that have been incorrectly tagged by either system have relatively low frequency, we also calculated, based on frequency information from the Szószablya Webcorpus (Halácsy et al. 2004), precision and recall figures based on token frequency (table 2).

table 1: Performance measured on types

	#epenthetic	false pos	false neg	precision	recall	F-score
CFM	107.8	3.1	1.4	97.17%	98.70%	97.79
maxent	107.5	3.3±1.56	2.1±4.2	96.96%±1.46%	98.05%±1.26%	97.50%±1.17

table 2: Performance measured on tokens

	#epenthetic	false pos	false neg	precision	recall	F-score
CFM	10.87m	7.41k	.71k	99.932%	99.994%	99.956
maxent	10.87m	20.16k±26.25k	2.23k±1.9k	99.82%±9.27%	99.98%±1.41%	99.90±6.63

Finally in tables 3 and 4 we also list the most frequent errors made by the system, divided into two groups: false positives (nouns incorrectly tagged as epenthetic) and false negatives (epenthetic nouns that weren't recognized as such).

table 3: Frequent mistakes of the maxent tagger

false positives	
meleg	110182
török	59266
menedzser	13012
észak	6468
üstökös	2505
false negatives	
átok	7225
vétek	2635
karom	2568
iker	1675
járom	1602

table 4: Frequent mistakes of the CFM tagger

false positives	
török	59266
donor	2328
sógor	1594
pacal	1342
vigyor	1222
false negatives	
iker	1675
lator	1257
kölök	1068
takony	642
berek	537

5 Conclusion

Hungarian epenthetic nouns have been studied in great detail. Recently, they have served to demonstrate the role of analogy in the production of individual word forms. Previous work has shown that automatic prediction of epenthesis is possible based on similarity of phonemes in noun stems. We created a system that uses a generic machine learning algorithm, has no explicit notion of similarity, and relies only on the most basic assumptions about what is relevant to the phenomenon. The system's performance is comparable to previous, more task-specific algorithms, demonstrating that simple machine learning tools with no specific linguistic component can treat the vast majority of cases adequately.

REFERENCES

Elekfi, László. 1994. *Magyar ragozási szótár* [Dictionary of Hungarian Inflections]. Budapest: MTA Nyelvtudományi Intézet.

Halácsy, Péter, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. 2004. Creating open language resources for Hungarian. In: *Proceedings of Language Resources and Evaluation Conference (LREC04).* European Language Resources Association.

Kiefer, Ferenc (ed.). 1994. *Strukturális magyar nyelvtan 2. Fonológia* [Structural Grammar of Hungarian 2. Phonology]. Budapest: Akadémiai Kiadó.

Papp, Ferenc. 1975. *A magyar főnév paradigmatikus rendszere* [The Hungarian Noun Paradigm]. Budapest: Akadémiai Kiadó.

Ratnaparkhi, Adwait. 1998. Maximum Entropy Models for Natural Language Ambiguity Resolution. PhD dissertation, University of Pennsylvania.

Rebrus, Péter and Miklós Törkenczy. 2008. Morfofonológia és a lexikon [Morphophonology and the Lexicon]. In: Ferenc Kiefer (ed.), *Strukturális Magyar Nyelvtan 4. A szótár szerkezete.* Budapest: Akadémiai kiadó. 683–786.

Rung, András. 2012. Magyar főnévi alaktani jelenségek analógiás megközelítésben. [An Analogy-based Approach to Hungarian Noun Morphology]. PhD dissertation, Eötvös Loránd University, Budapest.

Siptár, Péter and Miklós Törkenczy. 2000. *The Phonology of Hungarian.* Oxford: Oxford University Press.

Törkenczy, Miklós. 1994. A szótag [The Syllable]. Kiefer 1994 : 273–392.

Trón, Viktor, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon. 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In: *Proceedings of 5th International Conference on Language Resources and Evaluation.* European Language Resources Association.

Vago, Robert M. 1980. *The Sound Pattern of Hungarian.* Georgetown University Press. Washington.

Gábor Recski and András Rung
recski@sztaki.hu
Research Group for Human Language Technologies, Computer and Automation Research Institute, Hungarian Academy of Sciences