IMPROVING BRILL'S POS TAGGER FOR AN AGGLUTINATIVE LANGUAGE

Beáta Megyesi
Stockholm University
Department of Linguistics
Computational Linguistics
S-10691 Stockholm, Sweden
bea@ling.su.se

Abstract

In this paper Brill's rule-based PoS tagger is tested and adapted for Hungarian. It is shown that the present system does not obtain as high accuracy for Hungarian as it does for English (and other Germanic languages) because of the structural difference between these languages. Hungarian, unlike English, has rich morphology, is agglutinative with some inflectional characteristics and has fairly free word order. The tagger has the greatest difficulties with parts-of-speech belonging to open classes because of their complicated morphological structure. It is shown that the accuracy of tagging can be increased from approximately 83% to 97% by simply changing the rule generating mechanisms, namely the lexical templates in the lexical training module.

1. Introduction

In 1992 Eric Brill presented a rule-based tagging system which differs from other rule-based systems because it automatically infers rules from a training corpus. The tagger does not use hand-crafted rules or prespecified language information, nor does the tagger use external lexicons. According to Brill (1992) 'there is a very small amount of general linguistic knowledge built into the system, but no language-specific knowledge'. The grammar is induced directly from the training corpus without human intervention or expert knowledge. The only additional component necessary is a small, manually and correctly annotated corpus – the training corpus – which serves as input to the tagger. The system is then able to derive lexical/morphological and contextual information from the training corpus and 'learns' how to deduce the most likely part of speech tag for a word. Once the training is completed, the tagger can be used to annotate new, unannotated corpora based on the tag set of the training corpus. The tagger has been trained for tagging English texts with an accuracy of 97% (Brill, 1994).

In this study Brill's rule-based part of speech (PoS) tagger is tested on Hungarian, a dissimilar language, concerning both morphology and syntax, to English. The main goal is i) to find out if Brill's system is immediately applicable to a language, which greatly differs in structure from English, with a high degree of accuracy and (if not) ii) to improve the training strategies to better fit for agglutinative/inflectional languages with a complex morphological structure.

Hungarian is basically agglutinative, i.e. grammatical relations are expressed by means of affixes. Hungarian is also inflectional; it is difficult to assign morphemes precisely to the different parts of the affixes. The morphotactics of the possible forms is very regular. For example, Hungarian nouns may be analyzed as a stem followed by three positions in which inflectional suffixes (for number, possessor and case) can occur. Additionally, derivational suffixes, which change the PoS of a word, are very common and productive. Verbs, nouns, adjectives and even adverbs can be further derived. Thus, a stem can get one or more derivational and often several inflectional suffixes. For example, the word *találataiknak* 'of their hits' consists of the verb stem *talál* 'find, hit', the deverbal noun suffix *-at*, the possessive singular suffix *-a* 'his', the possessive plural suffix *-i* 'hits', the plural suffix *–k* 'their', and the dative/genitive case suffix *-nak*.

In this study it is shown that Brill's original system does not work as well for Hungarian as it does for English because of the great dissimilarity in characteristics between the two languages. By adding lexical templates, more suitable for complex morphological structure (agglutination and inflection), to the lexical rule generating system, the accuracy can be increased from 82.45% up to 97%.

2. The Tagger

The general framework of Brill's corpus-based learning is so-called Transformation-based Error-driven Learning (TEL). The name reflects the fact that the tagger is based on transformations or rules, and learns by detecting errors.

Roughly, the TEL (see figure 1 below) begins with an unannotated text as input which passes through the 'initial state annotator'. It assigns tags to the input in a heuristic fashion. The output of the initial state annotator is a temporary corpus, which is then compared to a goal corpus, i.e. the correctly annotated training corpus. For each time the temporary corpus is passed through the learner, the learner produces one new rule, the single rule that improves the annotation the most compared with the goal corpus. It replaces the temporary corpus with the analysis that results when this rule is applied to it. By this process the learner produces an ordered list of rules.

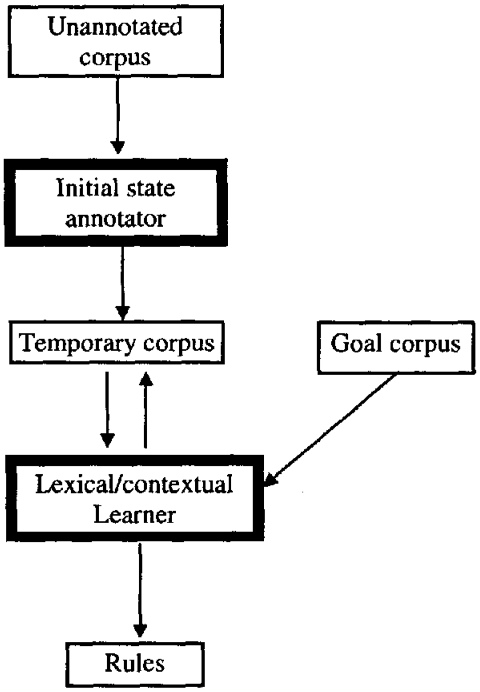

Figure 1. Error-driven learning module in Brill's tagger (data marked by thin lines)

The tagger uses TEL twice: once in a lexical module deriving rules for tagging unknown words, and once in a contextual module for deriving rules that improve the accuracy.

A rule consists of two parts: a condition (the trigger and possibly a current tag), and a resulting tag. The rules are instantiated from a set of predefined transformation templates. They contain uninstantiated variables and are of the form 'if trigger, change the tag X to the tag Y' or 'if trigger, change the tag to the tag Y (regardless of the current tag)'. The triggers in the lexical module depend on the character(s), the 'affixes', i.e. the first or last one to four characters of a word and on the following/preceding word. For example, the lexical rule

kus hassuf 3 MN

means that if the last three characters (hassuf 3) of the word are 'kus', annotate the word with tag MN (as an adjective). The triggers in the contextual module, on the other hand, depend on the current word itself, the tags or the words in the context of the current word. For example, the contextual rule

DET FN NEXTTAG DET

means that 'change the tag DET (determiner) to the tag FN (noun) if the following tag is DET'.

The ideal goal of the lexical module is to find rules that can produce the most likely tag for any word in the given language, i.e. the most frequent tag for the word in question considering all texts in that language. The problem is to determine the most likely tags for unknown words, given the most likely tag for each word in a comparatively small set of words. This is done by TEL using three different lists: a list consisting of Word Tag Frequency - triples derived from the first half of the training corpus, a list of all words that are available sorted by decreasing frequency, and a list of all word pairs, i.e. bigrams. Thus, the lexical learner module does not use running texts.

Once the tagger has learned the most likely tag for each word found in the annotated training corpus and the rules for predicting the most likely tag for unknown words, contextual rules are learned for disambiguation. The learner discovers rules on the basis of the particular environments (or the context) of word tokens.

The contextual learning process needs an initially annotated text. The input to the initial state annotator is an untagged corpus, a running text, which is the second half of the annotated corpus where the tagging information of the words has been removed. The initial state annotator also uses a list, consisting of words with a number of tags attached to each word, found in the first half of the annotated corpus. The first tag is the most likely tag for the word in question and the rest of the tags are in no particular order. With the help of this list, a list of bigrams (the same as used in the lexical learning module, se above) and the lexical rules, the initial state annotator assigns to every word in the untagged corpus the most likely tag. In other words, it tags the known words with the most frequent tag for the word in question. The tags for the unknown words are computed using the lexical rules: each unknown word is first tagged with a default tag and then the lexical rules are applied in order.

There is one difference between this module and the lexical learning module, namely that the application of the rules is restricted in the following way: if the current word occurs in the lexicon but the new tag given by the rule is not one of the tags associated to the word in the lexicon, then the rule does not change the tag of this word.

When tagging new text, an initial state annotator first applies the predefined default tags to the unknown words (i.e. words not being in the lexicon). Then, the ordered lexical rules are applied to these words. The known words are tagged with the most likely tag. Finally, the ordered contextual rules are applied to all words.

3. Testing Brill's Original System on Hungarian

3.1 Corpora and Tag Set

Two different Hungarian corpora[1] were used for training and testing Brill's tagger. The corpus used for training is the novel *1984* written by George Orwell. It consists of 14,034 sentences: 99,860 tokens including punctuation marks, 80,668 words excluding punctuation marks. The corpus has been annotated for part of speech (PoS) including inflectional properties (subtags).

The corpus used for testing the tagger consisted of two texts that were extracted from the Hungarian 'Hand' corpus: a poem and a fairy tale, both modern literary pieces without archaic words. The test corpus contains approximately 2,500 word tokens.

The tag set of the training corpus consists of 452 PoS tags including inflectional properties of 31 different parts of speech.

3.2 Training Process and Rules

The tagger was trained on the same material twice: once with PoS and subtags and once with only PoS tags.

The threshold value, required by the lexical learning module, was set to 300, meaning that the lexical learner only used bigram contexts among the 300 most frequent words. Two non-terminal tags were used for annotating unknown words initially, depending on whether the initial letter was a capital or not.

The lexical learner, used to tag unknown words, has derived 326 rules based on 31 PoS tags while it has derived 457 rules based on the much larger tag set, consisting of 452 PoS and subtag combinations. Note that if the tag set consists of a large number of frequently occurring tags, the lexical learner necessarily generates more rules simply to be able to produce all these tags. On the other hand, if only PoS tags (excluding subtags) are used the first rules score very high, in comparison with the scores of the first rules based on PoS and subtags. Another difference is that the score decreases faster in the beginning and slower in the end, compared to the rules based on PoS and subtags, resulting in a larger amount of rules, relative to the size of the tag set.

The contextual learner, used to improve the accuracy, derived approximately three times more rules based on 31 PoS tags than it derived from the text annotated with both PoS and subtags. This is somewhat harder to interpret since the output of the contextual learner does not contain scores. It seems reasonable that the contextual rule learner easier find 'globally good' rules, i.e. rules that are better in the long run, since the subtags contain important extra information, for instance about agreement.

The conclusion that can be drawn from these facts together with the fact that the test on the training corpus achieved slightly higher precision using subtags, is that it is probably more difficult to derive information from words, which are annotated with only PoS tags, than from words whose tags include information about the inflectional categories.

3.3 Results and Evaluation of Brill's Original Tagger

The tagger was tested both on new test texts with approximately 2500 words and on the training corpus. Precision was calculated for all test texts, and recall and precision for specific part of speech tags. Testing on the training set, i.e. using the same corpus for training and testing, gave the best result

[1] The corpora were annotated by the Research Institute for Linguistics at the Hungarian Academy of Sciences (Pajzs, 1996).

(98.6% and 98.8%), as would be expected. Due to the fact that the tagger learned rules on the same corpus as the test corpus, the outcome of the testing is much better than it is for the other types of test texts. The results do not give a valid statement about the performance of the system, but indicate how good or bad the rules the system derived from the training set are. These results mean that the tagger could not correctly or completely annotate approximately one in every hundred words.

In order to get a picture of the tagger's performance the tagger was tested on two different samples other than the training set. The accuracy (i.e. precision) of the test texts was 85.12% for PoS tags only, 82.45% for PoS tags with correct and complete subtags, and 84.44% for PoS tags with correct but not necessarily complete subtags, see Table 1 below.

Since one of the test texts contained three frequently occurring foreign proper names divergent from Hungarian morpho-phonological structure, the tagger's performance was also tested by preannotation[2] of these proper names as nouns before the tagging. Hence, the tagging performance increased: 86.48% for PoS tags only, 85.98% for PoS tags with correct and complete subtags, and 88.06% for PoS tags with correct but not necessarily complete subtags. The reason for the higher accuracy in this case is that these words are unknown and have atypical Hungarian morpho-phonological structure why the tagger can not guess their correct PoS tag by the application of the rules, derived from Hungarian words. Therefore, for achieving higher accuracy it is a good idea to handle foreign proper names before the tagging occurs, either by preannotation or by listing the words in the lexicon together with their correct tag.

The accuracy can be further increased if we do not consider the correctness of the subtags but only the annotation of the PoS tags in the evaluation. The accuracy in this case is 90.61%.

Table 1. Precision for the test corpora with and without the preannotation of foreign proper names in the tests.

Test corpus ----- correct tags in per cent	PoS tags only	PoS tags with correct and complete subtags	PoS tags with correct but not necessarily complete subtags	Without conside-ration of the correctness of subtags
Original test	85.12%	82.45%	84.44%	87.55%
Test with pre-annotated names	86.48%	85.98%	88.06%	90.61%

In order to know which categories the tagger failed to identify, precision and recall were calculated for each part of speech category of the test corpus (Megyesi, 1998). The results are given in the table below.

Table 2. Precision (correct_found/retrieved_total) and recall (correct_found/intended_total) for PoS categories of both test texts

PoS tags	**Precision**	**Recall**
DET (Determiner)	1.0	1.0
NM (Pronoun)	0.94	0.80
FN (Noun)	0.83	0.78
MN (Adjective)	0.70	0.75
IGE (Verb)	0.70	0.83
INF (Infinitive)	0.90	0.96
IK (Verbal Particle)	0.74	0.84
HA (Adverb)	0.85	0.74
SZN (Numeral)	0.73	0.89
NU (Postposition)	0.83	0.97
KOT (Conjunction)	0.91	0.96
ISZ (Interjection)	1.0	0.20

To sum up the results, the tagger has greatest difficulties with categories belonging to the open classes because of

[2] The preannotation was done by placing two slashes (//) between the word and its tag (instead of one slash), meaning that the tagger does not change the specific tag.

their morphological structure and homonymy, while grammatical categories are easier to detect and correctly annotate. Complex and highly developed morphological structure and fairly free word order, i.e. making positional relationships less important, lead to lower accuracy compared to English when using Brill's tagger on Hungarian.

These results are not very promising when compared with Brill's results of English test corpora which have an accuracy of 96.5% trained on 88200 words (Brill, 1995:11). The difference in accuracy might depend on i) the type of the training corpus, ii) the type and the size of the test corpus, and iii) the type of language structure, such as morphology and syntax. The corpus which was used to train the tagger on Hungarian consisted of only one text, a fiction with 'inventive' language, while Brill used a training corpus consisting of several types of texts (Brill, 1995). Also, there is a difference between the types and the sizes of the test corpora. In this work, small samples, which greatly differ in type from the training corpus, have been used while Brill's test corpus was larger and probably did not differ from the training corpus as much as in this study. Nevertheless, the most significant difference between the results lies in the type of the language structure, as will be shown later in this paper.

I argue that the low tagging accuracy for Hungarian mostly depends on the fact that the templates of the learner modules of the tagger are predefined in such a way that they include strong language specific information which does not fit Hungarian or other agglutinative/inflectional languages with complex morphology. The predefined templates are principally based on the structure of English and, perhaps, other Germanic languages.

The contextual templates are not as important for Hungarian as for English since Hungarian has free, pragmatically oriented word order. Also, Hungarian is a pro-drop language, i.e., the subject position of the verb can be left empty, which implies a larger number of contextual rules for Hungarian than for English because of the paradigmatic and/or syntagmatic difference between personal pronouns and nouns. The contextual templates however are necessary and fit as well for Hungarian as for English.

The lexical templates are, on the other hand, of greater importance for Hungarian than for English because of the structural differences on the word level between these languages. In Hungarian, the number of forms that a word can have is much greater than in English. Hungarian has a great number of common and productive derivational and inflectional suffixes that can be combined in many ways. The major problem is that Hungarian is partly inflective, i.e. one suffix can have several analyses depending on the type of grammatical relation it expresses. Sometimes the PoS tag of the stem indicates which properties the particular suffix has and sometimes the surrounding suffixes does the same. When a particular suffix is combined with the stem together with other suffixes there are often no alternate analysis, i.e. tag combinations for the word. For instance, in the training corpus only 1.78% of the words have more than one possible PoS tag, and 1.98% of the words have more than one possible PoS and subtag. On the other hand, according to Pajzs' examination (1996), more than 30% of the Hungarian lexical morphemes are homographs.

For the above mentioned reasons the lexical templates are much more important for Hungarian than the contextual templates.

Those lexical templates whose triggers depend on the affixes of a word examines only the first or last four characters of a word. In other words, defining that a lexical trigger is 'delete/add the suffix x where |x| < 5'

is to assert that it is only important to look at the last or first four letters in a word which is often not enough for correct annotation in Hungarian. For example, the word *siessu2nk*[3] 'hurry up' was annotated by the tagger as IGE_t1, i.e. as a verb in present indicative first person plural with indefinite object. The correct annotation should be IGE_Pt1, i.e. as a verb in the imperative (P) first person plural with the indefinite object. Because the tagger was only looking at the last four characters *-u2nk*, it missed the necessary information about the imperative -*s*-.

Another example concerns derivational suffixes giving important information about the PoS tag because they often change the category of the word. They follow the stem of the word and may be followed by different inflectional suffixes. For example, the word ártatlanságát, in English something like 'his harmlessness'

a1rt:atlan:sa1g:a1t
harm:less:Deadjectival_noun:ACC

is wrongly annotated by the tagger because information about the two derivational suffixes is missed if the word *a1rtatlansa1g* does not exist in the lexicon. Thus, if the tagger had looked at more than four characters, it would have been possible to reduce the total number of words in the lexicon. Also, it would have been able to create better and more efficient rules concerning the morphological structure of Hungarian words. This is especially true in the case of the corpora used in this work, since the encoding of accentuation of the vowels is done with extra characters (numbers) which reduces the effective length of the affixes. In the example above, *siessu2nk* (siessünk), at most three of the last letters are actually examined.

For Hungarian, the triggers of templates seem to be unsuccessful because of the Hungarian suffix structure of the open classes, such as the categories noun, verb and adjective. A possible solution is therefore to change the predefined language specific templates to more suitable ones for the particular language.

4. Testing Brill's System with Extended Lexical Templates

To get higher performance, lexical templates have been added to the lexical learner module. These templates look at the six first or last letters in a word. Thus, the maximum length of |x| has been changed from four to six. The lexical templates, which have been used for Hungarian, are the following:

* Change the most likely tag (from tag X) to Y if the character Z appears anywhere in the word.
* Change the most likely tag (from tag X) to Y if the current word has prefix/suffix x, |x| < 7.
* Change the most likely tag (from tag X) to Y if deleting/adding the prefix/suffix x, |x| < 7, results in a word.
* Change the most likely tag (from tag X) to Y if word W ever appears immediately to the left/right of the word.

4.1 Results and Evaluation of System Efficiency

After the changes of the lexical templates the tagger was trained and tested on the same corpus and with the same tag set in the same way as before the changes were done. Thus, all test corpora were annotated with both PoS tags, and PoS together with subtags. The performance of the whole

[3] The character 2 in the word annotates the accentuation of a preceding vowel in the corpus.

system has been evaluated against a total of three types of texts from different domains. Precision was calculated for the entire texts, both for PoS tags and PoS with subtags, based on all the tags and the individual PoS tags.

Testing on the training corpus gave the best result as could be expected. The precision rate increased from 98.6% to 98.9% in the case of PoS annotation only, while the result with PoS and subtags was unchanged (98.8% correct) compared to the original test.

In the case of the test corpus, where foreign proper names were preannotated as nouns, the accuracy increased considerably; from 86.48% to 95.53% for PoS tags only, from 85.98% to 91.94% for PoS tags with correct and complete subtags, and from 88.06% to 94.32% for PoS tags with correct but not complete subtags. Note that the precision is highest (97%), when not considering the correctness of the subtags in the evaluation. The results are also given in Table 3 below.

Table 3. Precision for the test corpora before and after the addition of the extra lexical templates with foreign proper names preannotated as nouns.

Test corpus ----- correct tags in per cent	PoS tags only	PoS tags with correct and complete subtags	PoS tags with correct but not necessarily complete subtags	Without consideration of the correctness of subtags
Original test	86.48%	85.98%	88.06%	90.61%
Extended lexical templates	95.53%	91.94%	94.32%	97%

Thus, we have shown that by changing the lexical templates in the lexical training module, specifically the maximum length of the first and last characters of a word that the tagger examines, the tagging performance is greatly improved.

We can assume that the extended lexical templates, used in this study, also fit for other highly agglutinative languages, such as Turkish, Finnish, Estonian, Japanese and Swahili. In these languages, words are built up of a long sequence of affixes similarly to Hungarian. The maximum length of the characters in the lexical templates should be changed for these languages, too, in order to handle the chain of grammatical morphemes.

Since Hungarian also has highly inflectional characteristics[4], it can be assumed, that Brill's tagger together with the extended lexical templates and a large tag set would be applicable for inflectional languages with a higher degree of accuracy, too. For example, in Hungarian the grammatical morpheme *-k* may express first person singular present tense of the verb or plural of the noun. In order to know which tag the word should get it is essential to look at the surrounding morphemes.

However, concerning the results, it has to be pointed out that they are based on a small test corpus consisting of approximately 2500 running words. Therefore, it would be necessary to test the tagger on a larger and more balanced corpus with different types of texts, including fiction, poetry, non-fiction, articles from different newspapers, trade journals, etc.

Additionally, since the training and the test corpus are of different text types, it would be very interesting to find out the accuracy results when the tagger is evaluated on the same text type as the training corpus.

Furthermore, for a higher accuracy it would be necessary to train the tagger on a larger corpus with different types of texts or even on several corpora because the

[4] Grammatical relationships are expressed by changing the internal structure of the words by use of inflectional suffixes which express several grammatical meanings at once.

likelihood of higher accuracy increases with the size of the training corpus. It is however still difficult to find correctly annotated balanced Hungarian corpora.

5. Further Development of the Tagger

For higher tagging performance it would also be advantageous to create a very large dictionary of the type *Word Tag1 Tag2... TagN*, (where the first tag is the most frequent tag for that word), listing all possible tags for each word. By using this lexicon, accuracy would be improved in two ways. First, the number of unknown words, i.e. words not in the training corpus, would be reduced. However, no matter how much text the tagger looks at there will always be a number of words that appear only a few times, according to Zipf's law (frequency is roughly proportional to inverse rank).

Secondly, the large dictionary would give more accurate knowledge about the set of possible part of speech tags for a particular word. For example, the template of the type 'Change the most likely tag from X to Y, if...' the template would only change tag X to tag Y, if tag Y exists with a particular word in the training corpus. Thus, a large dictionary would reduce the errors of the annotation by applying better rules and increase the speed of the contextual learning.

6. Conclusion

This work has shown how Eric Brill's rule-based PoS tagger can be applied for highly agglutinative languages with a high degree of accuracy. The results presented in this work show that tagging performance for languages with complex morphological structure can be greatly improved by changing the maximum length of the first/last character of a word from four to six in the lexical templates of the lexical learner module.

Also, it is shown that using a large tag set marking inflectional properties of a word in the training and tagging process improves the accuracy, when not considering the correctness of the subtags at the evaluation.

Acknowledgements

Above all, I would like to thank my advisor, Prof. Gunnel Källgren, who always helped and encouraged me. Unfortunately, she was not able to see the fruit of this work because of her tragic death.

Also, a thousand of thanks to Nikolaj Lindberg for his encouragement and for his many valuable suggestions concerning the preparation of this work and to Sara Rydin for reading and comments. Thanks also to the Hungarian Academy of Sciences for providing me with the Hungarian corpora.

References

Brill, E. 1992. A Simple Rule-Based Part of Speech Tagger. In *Proceedings of the DARPA Speech and Natural Language Workshop.* pp. 112-116. Morgan Kauffman. San Mateo, California.

Brill, E. 1994. A Report of Recent Progress in Transformation-Based Error-Driven Learning. *ARPA-94.*

Brill, E. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of speech Tagging. In *Computational Linguistics. 21:4.*

Brill, E. & Marcus, M. 1992. Tagging an Unfamiliar Text with Minimal Human Supervision. In *Proceedings of the Fall Symposium on Probabilistic Approaches to Natural Language*. 1992.

Megyesi, B. 1998. *Brill's Rule-Based PoS Tagger for Hungarian.* Master's Degree Thesis in Computational Linguistics. Department of Linguistics, Stockholm University, Sweden.

Megyesi, B. 1999. Brill's Rule-Based PoS Tagger with Extended Lexical Templates for Hungarian. To Appear in Technical Report of ACAI'99

Pajzs, J. 1996. Disambiguation of suffixal structure of Hungarian words using information about part of speech and suffixal structure of words in the context. *COPERNICUS Project 621 GRAMLEX, Work package 3 – Task 3E2*. Research Institute for Linguistics, Hungarian Academy of Sciences.