PART-OF-SPEECH TAGGING BASED ON COMBINING MARKOV MODELS AND MACHINE LEARNING

ALEKSANDAR KUPUSINAC[1], MILAN SEČUJSKI[2]

Abstract: The task of part-of-speech (POS) tagging is to mark each word of a text for its part-of-speech and morphological properties. Various techniques have been used, including HMMs and machine learning, and it has been shown that higher accuracy can be achieved for languages with simpler morphology. The paper considers the possibility of combining HMMs and machine learning based on transformation rules. The experiments carried out on a tagged text corpus of Serbian, containing 200.000 words, using a dictionary containing 3.9 million inflected forms (100.000 lexemes). The dictionary and the corpus were built within the AlfaNum project at the Faculty of Technical Sciences in Novi Sad, for the purpose of research and development of speech technologies for Serbian.
The experiments show that combining the two methods increases the accuracy of POS tagging, although it is still notably lower than the one of an expert system previously built within the same project.

Keywords: machine learning, Markov models, POS tagging, Serbian language

1. INTRODUCTION

The task of part-of-speech (POS) tagging is to mark each word of an unknown text for its part-of-speech and values of its morphological categories. As such, this problem plays an important role in almost all applications of language technologies. A number of different techniques has so far been used for automatic POS tagging, including hidden Markov models and machine learning techniques, and it has been shown that, regardless of the technique, significantly higher accuracy can be achieved for languages with simpler morphology. However, the Serbian language belongs to a group of languages with relatively complex morphology, requiring certain modifications to be introduced into standard algorithms in order to increase their accuracy and achieve the performance closer to the one of systems developed for some other languages. The aim of the research described in this paper is to systematically explore the possibility of obtaining a highly-accurate POS tagger for Serbian language.

Section 2 deals with natural language processing in general as well as most common problems related to it. It also introduces the basic terms related to POS tagging and gives the outline of the main approaches to the problem of POS tagging. Special attention is given to additional difficulties that present themselves when tagging languages with relatively complex morphology.

Section 3 contains a description of the language resources needed for the research into POS tagging described in this work. These consist of a module for morphological analysis which makes use of a comprehensive morphological dictionary of Serbian language, with entries also marked for accentuation (about 100,000 lexemes i.e. about 3.9 million inflected word forms) and a morphologically annotated corpus of Serbian language, consisting of texts of various kinds and containing approximately 200,000 words.

Section 4 gives the basics of POS tagging techniques that rely on Markov models and presents the results of experiments in which such a technique was used for POS tagging of Serbian.

Section 5 explains how certain machine learning techniques oriented towards identification and application of error-correcting transformation rules can be applied to POS tagging. This chapter also presents the results of experiments in which a system based on transformation rules was used for POS tagging of Serbian. The basic algorithm was modified to a certain extent in order to be better suited to particular issues related to highly inflected languages. The performance of the algorithm modified in such a way was compared to the performance of the original algorithm, with special attention given to

[1] Aleksandar Kupusinac, Faculty of Technical Sciences, Novi Sad, e-mail: sasak@uns.ac.rs
[2] Milan Sečujski, Faculty of Technical Sciences, Novi Sad, e-mail: sasak@uns.ac.rs

the influence of the size of the training corpus. The possibility of combining HMMs with machine learning techniques has also been investigated.

The paper is concluded by a comparison between the results obtained by several techniques for automatic POS tagging. The general conclusion is that in present conditions (regarding the availability of tagging techniques and language resources) expert systems still outperform fully automatic systems to a certain extent.

2. NATURAL LANGUAGE PROCESSING AND PART-OF-SPEECH TAGGING

Natural language processing (NLP) is the area of computer sciences dealing with human-machine interaction by means of natural[3] language. Natural language processing investigates strategies aimed at enabling computers to understand and process natural language in its written or oral form. Another term, *computational linguistics*, is often used with the same or a similar meaning.

Statistical natural language processing relies on stochastic, probabilistic and statistic methods in order to resolve some of the problems mentioned above, especially in cases of long sentences, which could produce hundreds of millions of different parse trees if analysed based on classical grammar. Methods for determining the correct interpretation typically rely on extremely large text corpora and on methods such as e.g. Markov models (MM). Statistical natural language processing encompasses all quantitative approaches to automatic processing of natural language, including probabilistic modelling, as well as knowledge from areas such as information theory and linear algebra (Manning & Schütze, 1993). Technology of statistical natural language processing largely relies on machine learning and data mining, two areas of artificial intelligence that are related to learning from extensive collections of data.

Part-of-speech tagging denotes marking each word of an unknown text for its part-of-speech, or, in a more general sense, determining its morphological properties. It is clearly a problem largely dependent on language, having in mind the fact that the degree of complexity of morphology of different languages can be quite different as well. Based on the complexity of morphology, the information that should be determined for each word can vary. In case of inflective languages, such as Serbian, the process of part-of-speech tagging can include determining the values of morphological categories of certain parts-of-speech. For instance, morphological categories of nouns in Serbian language are *number* and *case*, and part-of-speech tagging requires that their values be determined for each noun. However, besides the aforementioned categories, nouns also possess the grammatical category of *gender*, which, although not morphological but classificational, can be of great importance for any subsequent processing of the text (e.g. syntax analysis). Besides, it is often useful to know whether the noun is proper, common, countable etc., and thus it would be beneficial to determine such its properties as well. It is, therefore, clear that part-of-speech tagging in a wide sense can include annotation for various other information beside part-of-speech and the values of morphological categories themselves, i.e. that the morphological descriptor (*tag*) assigned to each word can be as detailed and extensive as required by specific application of annotated text. Complexity of the tag set can, thus, also vary depending on the purpose of the system, but it is generally greater in case of languages with complex morphology. For that reason, it can be expected that part-of-speech tagging for such languages will not be as accurate as it would be for languages with simpler morphology.

For instance, several standard tagsets have been defined for English, such as the Brown tagset, which consists of 179 unique descriptors used to annotate the Brown corpus of American English (Francis & Kučera, 1967) and the Penn Treebank tagset, containing 45 tags, representing a simplified version of the Brown tagset used to annotate the Penn Treebank corpus (Marcus et al, 1993). The fact that the size of these tagsets is relatively small is due to the relative simplicity of English morphology. For instance, nouns in English have only two forms (*house/houses*), and verbs have five (*write*/*writes*/*wrote*/*written*/*writing*). There are many more wordforms in languages with more complex morphology, and thus, e.g., the tagset used for annotation of the Prague Dependency Treebank (Hajič, 1998), as well as the Czech national corpus (Hajič & Hladká, 1998) theoretically contains as much as 3,030

[3] The term „natural“ in this context denotes a language used by humans in their communication, in contrast to formal or computer languages.

tags, although only a third of this number actually appear in the corpora. It should be noted that some of the very frequent function words have their own tags because of their very specific behaviour and use. For instance, in the Brown corpus the word *do* is not assigned the tag regularly used for base forms of verbs (VB) but a specialised tag (DO). This principle has been used by many other tagset designers because it has a positive effect on the accuracy of POS tagging as well as the usability of the results.

In tagging inflective languages, such as Serbian and Czech, *positional* tagsets are used, which means that morphological descriptors actually represent strings of characters denoting parts-of-speech as well as values of particular morphological categories and other categories of interest. This contributes to the readability of tags and they are also easier to operate.

2.1. Basic approaches to part-of-speech tagging

Two basic group of contemporary systems for automatic part-of-speech tagging are (1) *expert systems*, oriented on application of rules obtained by expert linguists, and (2) *automatic systems*, which, through analysis of large text corpora, attempt to extract knowledge that will subsequently be used for annotation of unknown text (van Guilder, 1995).

Expert systems are completely language-dependent, and a system developed for one language cannot be as successfully used for part-of-speech tagging of another, even in case of very similar languages, and even in case certain modifications are carried out. On the other hand, these systems are the ones that achieve the highest accuracy. As an example of a system achieving extremely high accuracy, the EngCG (*English Constraint Grammar*) system is often cited (Karlsson et al, 1995), (Samuelsson et al, 1997). This system is based on application of finite state automata (FSA) realised based on hand-written grammatical rules, and depending on the type of text to be annotated, its accuracy can exceed 99%.

On the other side, automatic systems have significantly lower accuracy, but their realisation does not require engagement of expert linguists. The text used for training can be annotated or not, which divides such techniques into two main classes – *supervised* and *unsupervised* ones. Supervised techniques use text that has been previously annotated as the basis for extracting information to be used later for annotation of unknown text, such as relative frequencies of words or tags, relative frequencies of sequences of words or tags, as well as automatically identified grammatical rules. In contrast, unsupervised techniques use far more sophisticated mathematical algorithms in order to discover similarities in the behaviour of particular words and divide them into groups based on their relatedness, thus defining the tagset itself automatically, and subsequently being able to classify words of unknown texts into these groups. Unsupervised techniques are much more portable than supervised ones and can be used when annotated corpora and other such linguistic resources are not available. However, it has been shown that higher accuracy in general can be achieved using supervised techniques, i.e. that in case that an annotated corpus of sufficient size is available, it is always preferable to use supervised techniques (Merialdo, 1994).

3. RESOURCES FOR AUTOMATIC PART-OF-SPEECH TAGGING IN SERBIAN

The first step in part-of-speech tagging of a particular word is determining a list of tags that could theoretically be assigned to it. Depending on the language, a morphological dictionary could be required for that. For example, in case of languages with relatively simple morphology, the dictionary need not exist as a separate linguistic resource, because if a sufficiently large training corpus is available, a relatively comprehensive dictionary could be made by simply listing all the words in the corpus and their corresponding tags. On the other hand, there are languages with extremely complex morphology (such as Turkish), where not only the dictionary must exist as a separate linguistic resource, but certain algorithms for morphological analysis are also required, since in case of such languages there is a much larger number of out-of-dictionary words in an arbitrary text (Hakkani-Tür et al, 2002).

Within this research, by using a software tool created for that purpose, the AlfaNum morphological dictionary was created, containing approximately 100.000 lexemes at this moment, i.e. approximately 3.9 million inflected forms. This research also required that an extensive part-of-speech

tagged text corpus be built. Within this research, by using another software tool created for that purpose, the AlfaNum Text Corpus (ATC) was created and part-of-speech tagged, containing approximately 11.000 sentences with approximately 200.000 words in total.

3.1. Morphological dictionary of Serbian

AlfaNum morphological dictionary was developed within the project dealing with research and development of speech technologies, and its original purpose was to support the part-of-speech process within the system for speech synthesis in Serbian (Sečujski et al, 2007). For that reason, each entry in the dictionary, besides the morphological descriptor, also contains the data related to the accentuation of the word, as well as the lemma (base form), which is useful for lemmatisation. The term *entry* denotes a particular inflected form of a word, together with the corresponding lemma, values of part-of-speech and morphological categories, as well as its accent structure (a string of characters denoting accent type associated to each syllable). An example of an entry would be:

Vb-p-1-- izgubićemo (izgubiti) [0\000].

Morphological categories that are marked are dependent on the part-of-speech, and thus e.g. verbs are marked for tense/mood, gender, number and person, but only in case a particular category is applicable to the tense/mood in question. The example above represents a verb (V) in 1st person (1) plural (p) of the future tense (b), whose surface form is *izgubićemo* and whose base form is *izgubiti*. The data related to accentuation are given in square brackets.

The latest version of the dictionary, the one which was used in the experiments described in the paper, contains 3.888.407 entries (100.517 lexemes). There are 748 morphological descriptors (tags) in the dictionary.

3.2. Part-of-speech tagged corpus of Serbian

Development of morphologically annotated text corpora are an extremely expensive and demanding task, and the perpetual need for them is a problem present in all language communities. Most of the existing corpora of Serbian language are not part-of-speech tagged. Notable exceptions are the *Corpus of Serbian Language*, developed at the Institute for Experimental Phonetics and Speech Pathology and Faculty of Philosophy in Belgrade (Kostić, 2001), containing texts from the period from 12th to 20th century, with a total of approximately 11 million words, as well as the Serbian translation of the novel "1984" by George Orwell, which represents the most valuable element of MULTEXT-East linguistic resources for Serbian language (Krstev et al, 2004), containing 108.805 words. However, due to the incompatibility of these corpora with the AlfaNum dictionary, as well as the need for texts of different types in the corpus (including contemporary daily press), it was decided that it would be a better solution to develop an entirely new corpus than to adapt any of the existing corpora to the dictionary.

The AlfaNum Text Corpus of Serbian language is part-of-speech tagged and contains a variety of texts from different sources (Sečujski & Delić, 2008). Most of it (between 80% and 90%) are texts of different topics and styles from daily press. The rest of the corpus is composed of fiction (by various authors) as well as texts taken from encyclopaedias. The corpus was designed with special attention to its representativeness, not only as regards topics and functional styles, but also as regards grammatical content (e.g. presence of various verb forms).

In its basic form, the corpus consists of sequences of dictionary entries corresponding to correct part-of-speech tagging of sentences. The corpus is annotated at sentence level, which means that sentences were considered in isolation rather than as parts of a text. Anaphoric references are ignored, which increased the number of cases where tagging remained ambiguous. More detailed information on the content of the dictionary and the corpus can be found in (Sečujski, 2009).

4. AUTOMATIC PART-OF-SPEECH TAGGING USING HIDDEN MARKOV MODELS

Hidden Markov Models (HMM) represent probabilistic functions of a Markov process (Markov chain). The basic assumption upon which they are based is that for predicting the future state of the system it is sufficient to know its current state.

In case of HMMs, the state sequence is not known, and the only thing that is known is some probabilistic function of the state sequence. The model crosses from one state to another, and in each state it emits a *symbol* according to a known probabilistic function. The sequence of symbols is visible and known as the *observation sequence*, while the state sequence itself is not visible. The aforementioned probabilistic function defines the probabilities of emitting particular symbols in a given state. HMMs are thus very useful when it is necessary to model a certain invisible sequence of events generating a sequence of visible events according to a certain probabilistic function.

A very good example is, indeed, the problem of part-of-speech tagging. The sequence of tokens in a text to be annotated is visible, while the sequence of tags that correspond to these tokens is not visible. Owing to grammatical rules that define the interior structure of the sentence, not all sequences of tags are equally probable. Moreover, since not all words are equally probable, particular tokens will have different emission probabilities in a given state. HMM based part-of-speech tagging is carried out in two stages – in the first stage the probabilities of moving from each state to each state are to be estimated on a large corpus, as well as the probabilities of emitting each particular token in a given state, and in the second stage this knowledge is to be used for analysis of an unknown text (sequence of tokens) and determining the most probable state sequence given the observation sequence.

4.1. The experiment

For the results of all experiments to be evaluated in an objective way, it has been established that, if part-of-speech tagging is carried out in a completely random way (by random choice of one of the possible tags for each token), the tagging error rate is 45.7% while the accentuation error rate is 15.0%. The results of the experiments should be evaluated keeping in mind these values as a reference, i.e. baseline.

Within this research a comparison between bigram and trigram HMMs was made. In order to determine the influence of training corpus size on tagging accuracy, the system was trained on segments of ATC whose size varied from 10,000 to 190,000, while the size of the test corpus was kept at a constant value of 10,000 in each round of the experiment.

The results of the experiments are shown in Fig. 1. The trigram model performs consistently worse than the bigram model, which was expected in view of the relatively small size of the ATC and the fact that the problem of data sparsity becomes much more acute when higher order HMMs are used. The tagging error decreases with the increase in the training corpus size, and in both cases it approaches saturation. A similar tendency can be noted in case accentuation error is analysed, and its lowest values are 2.65% (bigram model) and 3.05% (trigram model).

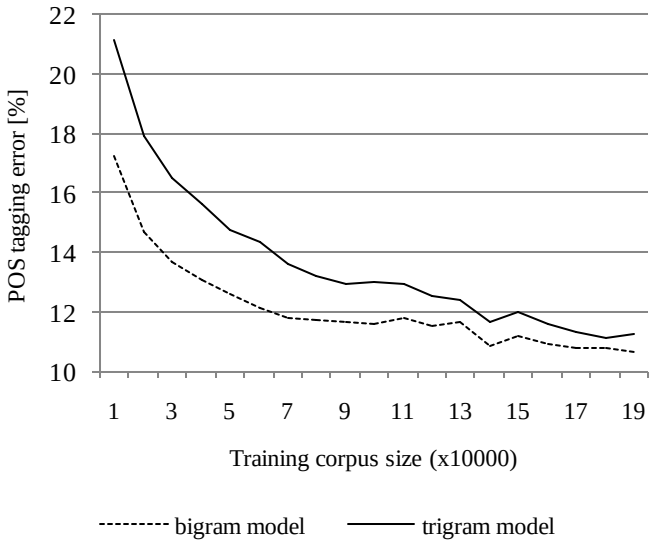

Fig 1. Dependency of POS tagging error on training corpus size
(POS tagging based on bigram and trigram HMMs)

5. AUTOMATIC PART-OF-SPEECH TAGGING USING TRANSFORMATION RULES

Part-of-speech tagging based on transformation rules as an supervised learning technique is one of the most popular approaches to automatic part-of-speech tagging. The basic idea was first introduced in (Brill, 1992), and further elaborated in (Brill, 1995). An algorithm based on machine learning functions in a way that enables it to detect its own weaknesses and remedy them, thus improving its performance.

The algorithm, namely, firstly assigns a certain initial tag to each word regardless of its context. In practice, this means that each word is assigned its most probable tag, according to word and tag frequencies in the training corpus. Then, based on the analysis of the same corpus, using a set of predefined rule-generating templates, a number of transformation rules are identified. These rules are designed so as to transform one tag into another in a particular context, thus eliminating a number of initial tagging errors. The algorithm is based on expectation that the same rules would eliminate a significant number of errors if another, unknown corpus was tagged instead.

Such a tagging strategy overcomes the common shortcomings of classical rule based approaches to natural language processing: it is robust and does not require (almost) any expert knowledge of grammar. Furthermore, it operates with a relatively small set of rules as opposed to a large amount of statistic data required by stochastic taggers to capture contextual information. Besides a significant reduction in stored information required, rules used by such taggers are easy to interpret by humans, unlike large tables of contextual probabilities. Such taggers are also easily fine-tuned manually and used for improving performance of expert systems (Kupusinac & Sečujski, 2007). They are also easily portable to another tagset or even another language.

5.1. The basic algorithm

Brill's algorithm for part-of-speech tagging based on transformation rules is trained in two steps. The first one consists of initial annotation of the training corpus, and the second one consists of identifying transformation rules.

During initial annotation each token is assigned its initial tag. Having in mind that the corpus is already part-of-speech tagged, the correct tag of each token is actually known, but that information is withheld in this phase. In the original version of the algorithm relative frequency of each tag for a given token was analysed in the training corpus and initial tags were assigned with respect to that. For example, the word *run* was initially annotated as a verb in both these cases:

*We **run** three miles every day.* (1)

*The **run** lasted thirty minutes.* (2)

where in (1) it was a correct decision and in (2) it was not.

The second phase consists of acquisition of transformation rules, based on rule templates such as the following:

Change the tag t_i into t_j:

1. If the previous (following) word has the tag t_k.
2. If the word two places to the left (right) has the tag t_k.
3. If any of the two preceding (following) words has the tag t_k.
4. If any of the three preceding (following) words has the tag t_k.
5. If the previous word has the tag t_k, and the following word has the tag t_l.
6. If the previous (following) word is L.

Transformation rule is identified by its originating template and a set of particular values of t_i, t_j, t_k, t_l, and/or L, its *triggering environment*. For each rule template by going through the training corpus a number of rules representing instances of the given template are identified, and the number of errors that would be corrected if the detected rule was applied (minus the number of new errors that the application of that rule would cause) is noted. For instance, after the initial annotation based on a segment of Brown corpus whose size was 900,000, for the rule template "*Change the tag t_i into t_j if any of the two preceding words has the tag t_k*", by going through a corpus segment containing 50.000

words it was established that the rule instantiated for (t_i, t_j, t_k) = (VB, NN, AT)[4] corrects 98 out of 159 existing errors in the initially annotated corpus and introduces 18 new ones (Brill, 1995). The total efficiency of a particular rule can be expressed as the difference between the number of corrected errors and the number of ones newly introduced into the training corpus. In the original version of the algorithm the most efficient transformation rule is identified and added to the interim list of transformation rules, the initially annotated corpus is re-annotated using that rule, and the whole procedure is repeated until a previously established number of transformation rules is established, or until the efficiency of newly identified transformation rules falls below a predefined threshold. Such a procedure is called *validation* of transformation rules.

It is interesting to note that most of the rules discovered are actually quite sensible from a linguistic point of view, especially those discovered earlier. For example, if an English word can be annotated as a verb in its base form (VB) and as a noun (NN), it is more probable that it is a noun if one of the previous words is an article (AT) as was the case in the example (1).

When all the transformation rules have been identified, tagging of an unknown text proceeds in much the same way. Firstly initial tagging is carried out based on frequencies of words and tags in the training corpus, and the identified transformation rules are subsequently applied, one by one, in order to reduce the number of errors. Transformation rules are applied only if the triggering environment fits, and only in case that the tag to be assigned is actually one of the possible tags for the given word. If a word appears in an environment that triggers more than one transformation rule, the rule with the greatest efficiency (the one that caused the greatest error reduction) on the training corpus will be the one to be applied.

5.2. Modifications of the basic algorithm aimed at its application to inflective languages

Due to a significant difference between the number of tags in the Brown and the AlfaNum corpus, as well as in the size of the corpora themselves, it was necessary to introduce certain modifications into the original Brill's algorithm in order to make it more suitable for application to the Serbian language.

Firstly, instead of a morphological dictionary, the Brill algorithm uses a corpus of 900.000 words as a basis for initial part-of-speech tagging. If the token to be annotated is present in this corpus, a most frequent tag for that token is assigned to each of its instances. Due to a much greater number of different tokens in a Serbian corpus of the same size, even 900.000 words would be a far less reliable source for initial tagging than in case of English. For that reason, initial tagging is not carried out based on the most frequent tag *for a given word*, but based on the most frequent tag altogether. For example, the word *knjiga* (book), that can be tagged in two ways, as NNfs1--- and NNfp2---, will be initially tagged as NNfs1---, but not because it is a more frequent tag for that word in the training corpus, but because it is in general a more frequent tag in the training corpus.

In the basic version of the algorithm transformation rules are obtained one by one and the training corpus is reannotated each time. Having in mind the fact that the number of obtained transformation rules in the basic version was 72 and that the difference in tagset sizes suggests that this number would be far higher for Serbian (which was confirmed by the experiments, which have shown that this number is measured in thousands) a somewhat different rule acquisition strategy was adopted, based on evaluation of rules in groups.

For further improvement of performance one can use the fact that in case of inflective languages many of the identified transformation rules are actually instances of more general grammatical rules. For instance, the rule:

"Change the tag of an adjective to genitive plural feminine in case the following word is a genitive plural feminine noun",

is an instance of a (hypothetical) general transformation rule:

"If a noun follows an adjective, change the values of the morphological categories case, number, and gender of the adjective to those assigned to the noun".

[4] VB = verb (base form), NN = noun, AT = article.

A strategy able to infer general rules from a sample of their instances would be of great use for tagging highly inflected languages, since it would enable the tagger to perform correctly even in situations not explicitly present in the training corpus.

Defining such a strategy was made easier by the fact that a positional tag system is used. The first step consists of grouping rules based on the originating template, source tag, target tag as well as the tag(s) defining triggering environment. For example, all rules stating that an adjective tag is to be replaced by an adjective tag with different values of morphological categories if followed by a noun tag ($A \rightarrow A[N_{+1}]$) are examined together, under the hypothesis that all these rules, or most of them, are instances of a general rule. A detailed presentation of this modification, aimed at *generalisation* of transformation rules, is given in (Sečujski, 2009).

5.3. The experiment

Within this research experiments related to POS tagging based on transformation rules have been carried out, with special attention given to the influence of the proposed method for rule generalisation on tagging accuracy. In order to determine the influence of training corpus size on tagging accuracy, the system was trained on segments of ATC whose size varied from 20,000 to 190,000, while the size of the test corpus was kept at a constant value of 10,000 in each round of the experiment. A section of the corpus containing 10,000 words was withheld and used for rule validation in each round of the experiment.

Besides the version of the algorithm presented in 5.2, another version was examined – the one that uses HMM as the method for initial tagging. In this way, transformation rules acted as correctors of the HMM-based tagger, which led to an increase in tagging accuracy. This idea was motivated by the remark that the two automatic systems analysed within this research (HMMs and transformation rules) both attempt to capture the regularities in the training data, but they do it in fundamentally different ways, and thus have different drawbacks. The HMM system lacks flexibility, which is one of the greatest advantages of the system based on transformation rules, but, on the other hand, the system based on transformation rules is capable of capturing very complex regularities in the training data, but lacks the possibility of expressing how reliable a rule is, i.e. a rule will either be applied always or never. The proposed method of initialisation attempts to use the advantages of both techniques.

The experiment results are shown in Fig. 2. One can note a certain improvement with respect to the results obtained by using the bigram HMM model in case of sufficiently large training corpus. The tagging error of 9.97% is the best result obtained by using a fully automatic technique for POS tagging.

6. CONCLUSION

The comparison of the results is given in Table 1. For the sake of simplicity, if an experiment was carried out in several variants, only the best result is displayed in the table. Among automatic techniques, the greatest accuracy was obtained by using the system based on transformation rules (including the generalised ones) and initialised by bigram HMMs. This system achieves the accuracy of 90%, which is comparable to the results given for other languages with a similarly complex morphology (87%-91%) (Hajič, 1998), (Džeroski et al, 2000). However, one must keep in mind the extreme dependence of the results on the language, type of text and the tagset used.

On the other hand, a significantly greater accuracy of morphological annotation, of more than 93%, was achieved using the expert system whose internal structure is described in more detail in (Sečujski, 2009). The superiority of this system is even more evident in terms of accentuation error. The most successful automatic accentuation technique (HMM, bigram model) achieves the error rate of 2,65%, while the expert system achieves 1,26% . Such a difference is partly a consequence of the fact that, having in mind the specific purpose of this system within a speech synthesiser for Serbian, special attention during its development was given to grammatical rules that are related to accentuation.

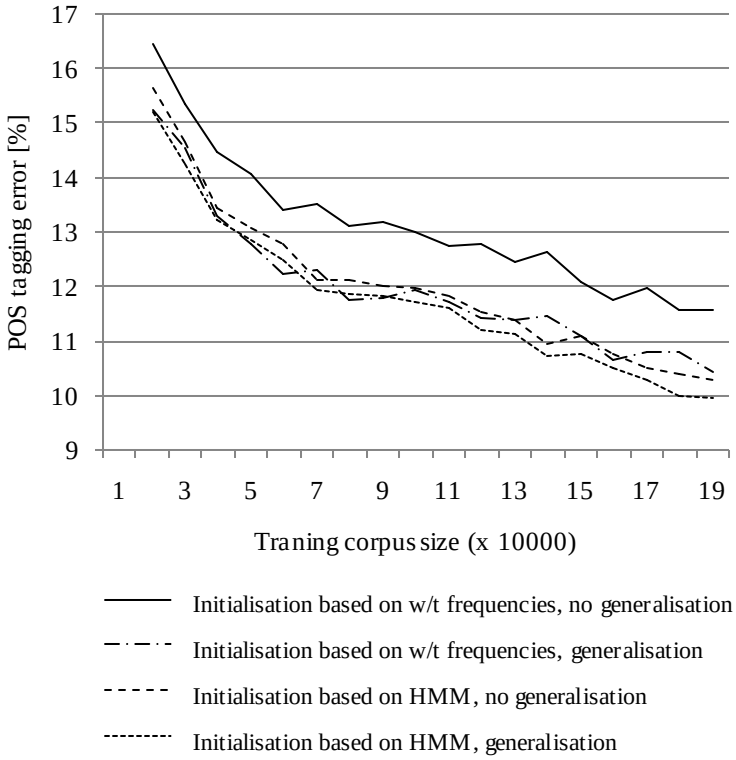

Fig 2. Dependency of POS tagging error on training corpus size (POS tagging based on transformation rules)

Table 1: Comparison of experiment results

Brief description	***Tagging error***	***Accentuation error***
HMM (bigram)	10,65%	2,65%
HMM (trigram)	11,23%	3,05%
transf. rules, initial tagging based on word/tag frequencies, with rule generalisation	10,42%	2,71%
transf. rules, initial tagging based on HMM (bigram model), with rule generalisation	9,97%	3,51%
expert system	6,75%	1,26%

REFERENCES

Brill, E. (1992). A simple rule-based part of speech tagger. *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento, Italy, 152-155.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging, *Computational Linguistics*, 21(4), 543-566.

Džeroski, S., Erjavec, T., Zavrel. J. (2000). Morphosyntactic tagging of Slovene: evaluating taggers and tagsets, *Proc. of the 2nd International Conf. on Lang. Resources and Evaluation*, Athens, Greece, 1099-1104.

Francis W. N., Kučera, H. (1967). *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Hajič, J. (1998). Building a syntactically annotated corpus: the Prague dependency treebank. *Issues of Valency and Meaning*. Karolinum, Prague, Czech Republic, 106-132.

Hajič, J., Hladká, B. (1998). Czech language processing – POS tagging. *Proceedings of the 1st International Conference on Language Resources and Evaluation,* Granada, Spain, 931-936.

Hakkani-Tür, D., Oflazer, K., Tür, G. (2002). Statistical morphological disambiguation for agglutinative languages. *Journal of Computers and the Humanities*, 36(4), Kluwer Acad. Publ., The Netherlands, 381-410.

Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A. (1995). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin, Germany.

Kostić, Đ. (2001). *Kvantitativni opis strukture srpskog jezika: Korpus srpskog jezika*. Institut za eksperimentalnu fonetiku i patologiju govora, Filozofski fakultet, Beograd.

Krstev, C., Vitas, D., Erjavec, T. (2004). Morpho-syntactic descriptions in MULTEXT-East – the case of Serbian. *Informatica*, No. 28, The Slovene Society Informatika, Ljubljana, Slovenija, 431-436.

Kupusinac, A., Sečujski, M. (2007). Povećanje tačnosti poluautomatske morfološke anotacije primenom transformacionih pravila. *TELFOR*, Beograd, 604-606.

Manning, H., Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcus, M. P., Santorini B., Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19, 313-330.

Merialdo, B., 1994. Tagging English text with a probabilistic model. *Computational Ling*., 20, pp. 155-172.

Samuelsson, C., Voutilainen, A. (1997). Comparing a linguistic and a stochastic tagger. *Proc. of the 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, 246-253.

Sečujski, M., Delić, V., Pekar, D., Obradović, R., Knežević, D. (2007). An overview of the AlfaNum text-to-speech synthesis system. *Proceedings of SPECOM*, Moscow, Russia, 3-7 (Addenda Volume).

Sečujski, M, Delić, V. (2008). A software tool for automatic part-of-speech tagging in Serbian language. *Primenjena lingvistika*, 9(1), Društvo za primenjenu lingvistiku, Beograd, 97-103.

Sečujski, M. (2009). *Automatska morfološka anotacija tekstova na srpskom jeziku* (doktorska disertacija), Fakultet tehničkih nauka, Univerzitet u Novom Sadu.

van Guilder, L. (1995). Automated part of spech tagging: a brief overview. *Handout for LING361*, Georgetown University, Georgetown, Washington DC.

MORFOLOŠKA ANOTACIJA ZASNOVANA NA KOMBINOVANJU MARKOVLJEVIH MODELA SA MAŠINSKIM UČENJEM

APSTRAKT

Zadatak automatske morfološke anotacije jeste da za svaku reč u tekstu odredi vrstu reči, kao i vrednosti odgovarajućih morfoloških kategorija. Za rešavanje ovog problema korišćene su razne metode, po¬put skrivenih Markovljevih modela i tehnika mašinskog učenja, pri čemu je znatno veća tačnost postig¬nuta za jezike sa jednostavnijom morfologijom.

U radu se razmatra mogućnost kombinovanja skrivenih Markovljevih modela sa tehnikama mašinskog uče-nja zasnovanim na transformacionim pravilima. Eksperimenti su izvršeni na morfološki anotiranom tek-stualnom korpusu srpskog jezika koji sadrži oko 200.000 reči, uz oslanjanje na morfološki rečnik koji obuhvata oko 3,9 miliona izvedenih oblika reči (100.000 leksema). I rečnik i korpus realizovani su u okviru projekta AlfaNum na Fakultetu tehničkih nauka u Novom Sadu, za potrebe istraživanja i razvoja govornih tehnologija na srpskom jeziku.

Eksperimenti potvrđuju da kombinovanje navedenih pristupa doprinosi tačnosti morfološke anotacije, mada je tačnost i dalje niža u odnosu na tačnost ekspertskog sistema realizovanog u okviru istog projekta.

Aleksandar Kupusinac, Milan Sečujski
Fakultet tehničkih nauka, Novi Sad, Srbija