INFORMATICA, 2009, Vol. 20, No. 1, 35–50

Analysis of Automatic Stress Assignment in Slovene

Domen MARINČIČ, Tea TUŠAR, Matjaž GAMS, Tomaž ŠEF
Jozef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
e-mail: {domen.marincic, tea.tusar, matjaz.gams, tomaz.sef}@ijs.si

Received: October 2007; accepted: October 2008

Abstract. We tested the ability of humans and machines (data mining techniques) to assign stress to Slovene words. This is a challenging comparison for machines since humans accomplish the task outstandingly even on unknown words without any context. The goal of finding good machine-made models for stress assignment was set by applying new methods and by making use of a known theory about rules for stress assignment in Slovene. The upgraded data mining methods outperformed expert-defined rules on practically all subtasks, thus showing that data mining can more than compete with humans when constructing formal knowledge about stress assignment is concerned. Unfortunately, compared to humans directly, the data mining methods still failed to achieve as good results as humans on assigning stress to unknown words.

Keywords: stress assignment, data mining, expert knowledge.

1. Introduction

In general, all stress assignment tasks deal with differences between written and spoken words and sentences. Linguistics explains stress as the relative emphasis that can be given to some syllables in a word. How the stress manifests itself in the speech highly depends on the language. Some languages have fixed stress placement inside a word, meaning that stress is placed always on a predetermined syllable. An example of such a language is Finnish where stress is always located on the first syllable. Macedonian is an example of a language with the stress being placed on the third syllable counting backwards. In some other languages, stress is placed on different syllables but determined by certain rules, as in Latin where stress depends on the structure of the penultimate syllable.

In English, stress partly depends on the origin of the word. In words with the anglo-saxon origin the first syllable (not counting possible prefixes) is stressed. On the other hand, words originating from Latin, are stressed differently.

In Slovene, stress can be positioned almost arbitrarily on any syllable. Further examples of such languages are: Romanian, Lithuanian, Ukraininan, etc. In Slovene some stress assignment rules are well known. For decades, the dilemma between (i) learning specific pronunciation of each word and (ii) at least partially relying on stress assignment rules remains unsolved for the Slovene language.

One of our aims was to investigate whether there exist stress assignment rules of appropriate quality. More important, we tried to find out whether new data mining (DM) methods outperform expert-defined rules and humans on the task of assigning stress to unknown Slovene words.

The remainder of this article is organized as follows. In Section 2, we introduce the problem of assigning stress to Slovene words. In Section 3 we present the related work. Thereafter, in Section 4, we describe the approaches used: expert-defined rules and data mining methods. In Sections 5 and 6, we present experiments and their results. Finally, in Section 7, we conclude by summarizing our findings.

2. Stress Assignment in Slovene

In Slovene, one syllable contains one vowel or one reduced vowel. We have defined the stress assignment task as classifying vowels of a word-form into one of the classes presented in Table 1. Because stress position and type are almost arbitrary (Toporišič, 1984), it is hard to find good overall rules for stress assignment in Slovene.

Most Slovene word-forms have only one stressed syllable, but there exist also word-forms with no stress (the so-called *clitics*) and word-forms with more than one stressed syllable. Furthermore, different forms of the same word can be stressed differently. Therefore, the exact vocal representation of a word-form is highly dependent on several known and some unknown parameters. For example, to classify the stress of a syllable, we probably need more information than just the local context of the observed syllable.

We decompose the problem of stress assignment into two subproblems – determining the stress position and, once stressed vowels have been identified, determining the type of stress (narrow or open) for vowels *e* and *o* only.

If vowels *a*, *i* or *u* are stressed, they are marked with the diacritic '´'. Stressed narrow *e* and *o* are marked with the diacritic '´' as well, while stressed open *e* and *o* are marked with the diacritic 'ˆ'. In Slovene, the reduced vowel can appear instead of a proper vowel. When stressed it is marked as '*è*' or '*ŕ*' if it appears before the letter *r*. Note that the diacritics are practically never written in the normal text.

Unlike most languages, in Slovene stress assignment depends on the morphological category of the word-form. Therefore, we use a Slovene pronunciation dictionary where

Table 1

Stress assignments

Vowel	Classes
a	unstressed (*a*), stressed (*á*)
e	unstressed (*e*), narrow stressed (*é*), open stressed (*ê*), stressed red. vowel (*è*)
i	unstressed (*i*), stressed (*í*)
o	unstressed (*o*), narrow stressed (*ó*), open stressed (*ô*)
u	unstressed (*u*), stressed (*ú*)
r	unstressed (*r*), stressed (*ŕ*)

for every word-form we have the following information: word-form (without stress), lemma, stressed word-form, and morphological information of the word-form. The dictionary contains almost 600.000 word-forms with more than 2.000.000 syllables. Many of these word-forms share the same lemmas.

The dictionary covers around 20.000 lemmas, a small, but relatively frequently used part of the Slovene vocabulary. Overall, there are approximately 300.000 lemmas in the Slovene language. A simple experiment on the FIDA corpus of Slovene language (FIDA, 2006), which contains over 100 million word-forms of contemporary Slovene texts, showed that the word-forms from our dictionary cover 85.9% of word-forms from FIDA.

3. Related Work

It is common to include the stress assignment task as a part of a G2P (grapheme-to-phoneme) system. The stress assignment module, part of a G2P system Plattos (Rojc and Kačič, 2007), uses a CART model for stress prediction. Feature vectors contain linguistic knowledge extracted from a phonetic lexicon. (Gros *et al.*, 2005) use a simpler approach with a dictionary lookup. For out-of-dictionary word-forms, stress is predicted by hand-crafted rules according to (un)stressable affixes, prefixes and suffixes or by a statistical analysis based on the number of syllables in the word-form.

An example of a rule-based approach can be found in (Skripkauskas and Telksnys, 2006) where an algorithm for stress assignment is a part of a G2P transformation for Lithuanian. In (Taylor, 2005) an approach for stress prediction in English using a hidden Markov model is described. Only vowels are represented as symbols of the HMM, while consonants (except for '*y*') are not included in the decision process.

For Romance languages, such as Spanish, French and Romanian the approaches considered range from hidden Markov models (Yarowsky, 1999; Simard and Deslauriers, 2001; Tufiş and Chiţu, 1999), to Bayesian classifiers (Yarowsky, 1999) and decision lists (Yarowsky, 1994). Although these studies deal with the problem of stress assignment, they are fundamentally different to our work. The problem is posed as a disambiguation problem: all possible valid variants of the stress of a word-form are presumed to be given in advance. The stress assignment problem is then reduced to choosing the correct variant based on local context, while our methods work at the phoneme-level similarly as (Lauriciukaite and Lipeika, 2007), with the intent to generalize from one word to another.

Mihalcea and Nastase (Mihalcea and Nastase, 2002) propose the use of an instance-based data mining algorithm for the general task of diacritics restoration. They tested their algorithms on Czech, Hungarian, Polish and Romanian. As in our work, individual examples for the learner are letters within their local context, to which diacritical marks could be attached. However, they use different data mining algorithms. Furthermore, their general problem of diacritics restoration is materially different to the stress assignment task that is considered in this study.

4. Computer Methods

4.1. *Expert-Defined Rules*

Rules for stress assignment in Slovene were first created by human experts more than 20 years ago (Toporišič, 1984). They occupy around 10 pages in their original source and demand memorization of many exceptions. The text is rather declarative and not directly applicable for classification. Sometimes the rules require knowledge which cannot be easily obtained for the purposes of machine stress assignment, especially the recognition of loan words. In this article, a modernized machine-readable version (Šef, 2001) is implemented as 68 IF-THEN rules, which does not include loan word identification because of the lack of resources. However, the list of loan words is unavailable to the DM methods as well, ensuring a fair comparison with the expert-defined rules. Experts estimate that the newly defined rules should achieve practically the same or even higher accuracy compared to the original rule set (at least for non-loan words) due to the absence of human errors at classification.

Some rules are used for determining stress position while the others are applied for finding the type of stress. The stress type has to be defined only for stressed syllables containing the vowels *e* or *o*. Other vowels can only assume a single stress type and are denoted simply as stressed or not stressed. For this task, the expert-defined rules make use of the context of the observed vowel in the word-form as well as the morphological information of the word-form. For example, the following rule predicts an open stressed *o*:

> All adjectives that contain a stressed *o* in the genitive ending *-okega*, have an open stressed *o*.

There are a total of 37 rules for prediction of stress type for the vowel *e* and 12 rules for prediction of stress type for the vowel *o*.

4.2. *Data Mining*

Data mining methods build a model from a given data set and use this model to classify new instances. In our previous work, we considered the use of decision trees and boosted decision trees for machine stress assignment (Šef and Gams, 2001). The results were not encouraging in particular when combining predictions on single syllables to assign stress to whole word-forms. In this paper we describe several new methods and a new mechanism for word stress assignment. We also extended the number of attributes describing learn samples. Most notably, the attributes describing morphological properties of words have been added.

We experimented with a larger selection of DM methods to other approaches like bagging, decision rules, and a method that uses naive-Bayes-like models, all from the WEKA data mining toolkit (Witten and Frank, 2000):

- *J48* is WEKA's reimplementation of Quinlan's C4.5 algorithm for building decision trees (Quinlan, 1993).

- *Bagging* is a meta-classifier method that combines various classifiers with equal weights into a single prediction using votes (Breiman, 1996).
- *AdaBoostM1* is an adaptive boosting algorithm. Subsequent classifiers are built in favor of those instances which were misclassified by previous classifiers (Freund and Schapire, 1996).
- *PART* generates decision rules using partial C4.5 decision trees (Frank and Witten, 1998).
- *AODE* averages the predictions of alternative naive-Bayes-like models that have weaker independence assumptions than naive Bayes (Webb *et al.*, 2002).

Several other DM systems in WEKA, such as *Logistic*, *MultilayerPerceptron*, *RandomForest*, *RandomCommittee*, and *SMO* were omitted due to their difficulties with the large amount of data. Here we use only the widely-known rule generator (*PART* rules) and methods for rule generation like k-optimal rule sets (Li, 2006). Some pattern-based methods, such as PPM (Cleary and Witten, 1984) were tested as well, however are not presented here since their constructed knowledge is not transparent to humans.

In the task of stress assignment of Slovene words with DM, we classify the stress on every vowel individually. We first use a DM method to predict whether the vowels are stressed, after which we apply the same method to predict the stress type. The predictions made on the vowels are combined to produce the final stress assignment of the whole word-form using a heuristics, which we describe later in this section.

4.2.1. *Assigning Stress to Vowels*

To evaluate prediction of stress position and type for every vowel, we divide all vowels from the dictionary into six groups: *a*, *e*, *i*, *o*, *u*, and *r*. Data mining is performed on each group separately. This yields eight DM tasks: six for predicting stress position and two for stress type (for vowels *e* and *o*). Each vowel is thus treated as an instance, described with a set of attributes. These attributes contain information on the word-form in which the vowel appears and the context of the vowel in the word-form, as well as whether certain characteristic prefixes and suffixes are contained in the word-form. When predicting stress position, each vowel is described with the following 75 attributes:

- number of syllables in the word-form (1 attribute),
- clitics, prefixes, and suffixes that influence the stress of the word-form (6 attributes),
- morphological information of the word-form (7 attributes),
- position of the observed syllable counted from the beginning and the end of the word-form (2 attributes),
- two vowels that precede and two vowels that follow the observed vowel (4 attributes),
- three graphemes that precede and three graphemes that follow the observed vowel (9 attributes for each grapheme = 54 attributes),
- the class attribute (1 attribute with values *stressed* and *unstressed*).

The same attributes are employed for predicting stress types, with the exception of the class attribute, which can take one of the following values: narrow stressed vowel, open stressed vowel, or stressed reduced vowel.

4.2.2. *Assigning Stress to Words*

A new method for word stress assignment was introduced, namely a heuristic probability model which was used to predict the stress assignment of the whole word-form. Unlike sophisticated ways of combining different classifiers (Bell *et al.*, 2005), we combine predictions of the same classifier, thus the following simple heuristic function seems sufficient. The heuristics chooses the most probable pattern of stress assignment among patterns with sufficient occurrences in the training set. For example, for word-forms with three syllables, the possible patterns are: 'Ynn' (only the first vowel is stressed), 'nYn' (only the second vowel is stressed), 'nnY' (only the third vowel is stressed) and 'YYn' (the first and second vowel are stressed). The patterns "nYY" and 'YYY' never appear in the dictionary, while 'nnn' and 'YnY' occur in less than 0.1% of cases. Among the remaining patterns, the heuristic algorithm chooses the pattern with the highest probability, which is calculated as the product of probabilities for each vowel. For example, the probability of the pattern 'YYn' is computed as $P(\text{first vowel stressed}) \times P(\text{second vowel stressed}) \times P(\text{third vowel not stressed})$. The individual probabilities are estimated by the confidence for prediction returned by the different DM methods. Combining stress assignments to individual vowels is a specialized mechanism intended to upgrade DM techniques for stress assignment tasks.

5. Dictionary Experiments with Computer Methods

5.1. *Experiments*

The first set of experiments was performed on the word-forms contained in the pronunciation dictionary described in Section 2. We first test machine stress assignment methods on the dictionary using cross validation.

We divided the word-forms from the dictionary into three corpora of similar size in such a manner that word-forms with the same lemma were always placed in the same corpus. In this way, the entries in different corpora were not too similar. We then evaluated the methods with a 3-fold cross validation. One corpus was used for testing the method, while the remaining two were used for training, so that each word-form in the dictionary appeared in the test set exactly once. The expert-defined rules do not require training, but need statistics of the most frequent stress assignments for word-forms with different number of syllables. When predicting the stress of word-forms contained in one corpus, the remaining two corpora were used to calculate these statistics.

All the DM methods were run with WEKA'a default parameter settings (Witten and Frank, 2000). The meta-methods bagging and boosting were run using ten *J48* decision trees as basic classifiers (the pruning factor was set to 2 minimum examples in leaves).

Two measurements were carried out. The first measurement was designed to measure the accuracy of different methods. With the second measurement we tried to discover simple and accurate models for stress assignment and compare them to the expert-defined rules. The tests of statistical significance were carried out using the resampled t-test proposed by (Witten and Frank, 2000). The results were calculated for the 95% confidence level.

5.2. *DM Methods*

5.2.1. *Stress Position*

Table 2 presents the accuracy of different methods for the task of assigning stress position on vowels. The methods are sorted based on their performance over all vowels in the dictionary. For the DM methods, the values in Table 2 show the accuracy achieved after the use of the heuristic probability model. The model decreased the performance of *J48* and *PART* on individual vowels by less than 1%, while improving the performance of the other three DM methods by 1 to 1.5%.

The results show that all the machine methods outperform the expert-defined rules. Among the data mining methods, the ensemble methods achieve the best results as expected. The best method is boosting. All the results are statistically significantly different except when comparing *J48* trees, *PART* rules and *AODE* algorithm among each other. As an example, we present two detailed measurements: the minimal statistically significant difference at the 95% confidence level between the algorithm AdaboostM1 and the bagging algorithm amounts to 0.53%; the minimal statistically significant difference comparing the algorithm AdaboostM1 and the expert-defined rules equals 0.70%.

5.2.2. *Stress Type*

The results for this task are given in Table 3. This time the DM methods do not require the use of the heuristic probability model, since it has no effect on predicting stress type.

Table 2

Stress position accuracy

Method	Vowel
Expert-defined rules	65.44%
Heuristics + *J48*	88.83%
Heuristics + *PART*	88.98%
Heuristics + *AODE*	89.53%
Heuristics + *Bagging*	91.74%
Heuristics + *AdaBoostM1*	92.99%

Table 3

Stress type accuracy

Method	Vowels *e* and *o*
Expert-defined rules	69.37%
AODE	85.21%
PART	88.34%
J48	89.97%
Bagging	90.40%
AdaBoostM1	91.15%

Again, the DM methods outperform the expert-defined rules when no limit is set on complexity. The best results are achieved by boosting. The differences are statistically significant with the exception of the difference between the algorithm AdaboostM1 and bagging and the difference between bagging and *J48* trees. The smallest statistically significant difference between the algorithm AdaboostM1 and the bagging algorithm is 1.01%; the smallest statistically significant difference comparing the algorithm AdaboostM1 and the expert-defined rules equals 1.36%.

5.2.3. *Assigning Stress to Words*

To evaluate the methods on the whole task of stress assignment at the word level, we need to combine the predictions on the two subtasks. The results in Table 4 present the accuracy achieved by the methods on the whole word-forms. The accuracy at word-level is usually much lower than the accuracy at vowel-level due to the fact that only one erroneously stressed vowel in the word-form usually causes an error in the stress assignment on the whole word-form. Here, the use of the heuristic probability model on the predictions of DM methods shows its advantage. On average, the predictions of DM methods alone reached accuracy that was 8.7% lower than when the predictions were combined with the probability model. The results show once again that the DM methods considerably outperform the expert-defined rules.

The following differences are not statistically significant: (i) the difference between *PART* rules and *J48* decision trees, (ii) the difference between the algorithm *AdaboostM1* and the algorithm *AODE* using the heuristics and (iii) the difference between *J48* trees joined with the heuristics and bagging with the heuristics. All the other differences are statistically significant. For example, the minimal statistically significant difference between the algorithm Heuristics + *AdaBoostM1* and the algorithm Heuristics + *Bagging* is 0.73%. The minimal statistically significant difference between the algorithm Heuristics + *AdaBoostM1* and the expert-defined rules is 1,71%.

Table 4
Accuracy of assigning stress to the whole words

Method	Word-form
Expert-defined rules	31.37%
AODE	59.38%
PART	65.40%
J48	65.96%
Bagging	68.10%
AdaBoostM1	70.85%
Heuristics + *AODE*	71.06%
Heuristics + *PART*	72.05%
Heuristics + *J48*	72.21%
Heuristics + *Bagging*	77.66%
Heuristics + *AdaBoostM1*	80.34%

5.3. *DM Methods vs. Expert-Defined Rules on Words*

In the next experiment, we wanted to find small and comprehensible machine-made rules and/or trees that would outperform the expert-defined rules of similar complexity. For this purpose, the *J48* decision trees and *PART* decision rules were subject to pruning with the pruning factor set to 0.01%, 0.02%, 0.05%, 0.1%, 0.2%, 0.5%, 1%, 2%, 5%, 10%, 20%, and 50% minimum examples in leaves/rules. The predictions of these methods were not combined using the heuristic probability model, since we wanted the rules/trees to remain comprehensible as much as possible. The two DM methods and the expert-defined rules were compared regarding complexity and accuracy. *Complexity* was measured in two different ways for each method:

C1 – as the number of rules compared to the number of leaves,

C2 – as the number of clauses in the rules compared to the number of the nodes in trees.

To evaluate the methods on words, we need to combine the predictions on the two subtasks, on stress position and stress type. The results are presented in Table 5 and Fig. 1, logarithmic scale was chosen for the x-axis to make the difference clearly visible. When pruning is high, the rules/trees reach the predictions of the majority class. Since the majority class predicts all unstressed vowels, except for the reduced vowel preceding the consonant *r*, the word-level results for heavily pruned trees approach 0%.

The DM trees/rules with comparable complexity to the expert-defined rules perform similarly or better even though we sacrificed the combining mechanism for the sake of transparency. If the simple probabilistic calculation is allowed, the complexity increases only slightly, while accuracy increases by 7%. The results are encouraging from another

Table 5

Complexity vs. accuracy of *J48* decision trees, *PART* decision rules and the expert-defined rules on assigning stress position and type for word-forms. In the first two columns of each table the number of nodes/clauses and leaves/rules are shown, respectively

C2/nodes	C1/leaves	Accuracy
11251	5614	64.19%
7523	3777	63.43%
4099	2069	61.47%
2443	1230	59.68%
1433	722	57.34%
655	330	53.72%
359	182	51.17%
183	94	47.45%
101	53	39.79%
61	33	27.89%
39	22	14.76%
11	8	1.67%

(a) Pruned dec. trees

C2/clauses	C1/rules	Accuracy
13005	6029	64.35%
9950	4029	63.00%
5704	2396	59.86%
2997	1398	58.32%
1915	868	55.92%
1045	425	52.08%
570	234	48.45%
312	138	40.92%
135	61	35.02%
73	33	27.03%
40	16	4.30%
17	8	1.67%

(b) Pruned dec. rules

C2/clauses	C1/rules	Accuracy
/	/	
/	/	
/	/	
/	/	
/	/	
/	/	
/	/	
/	/	
205	68	31.37%
/	/	
/	/	
/	/	

(c) Expert-def. rules

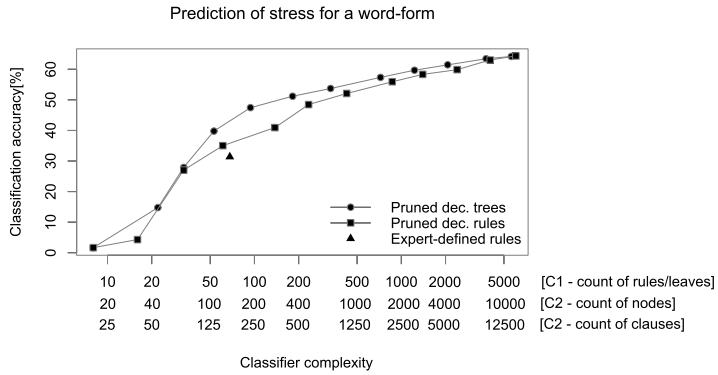

Fig. 1. Complexity vs. accuracy of expert-defined rules, *J48* decision trees and *PART* decision rules on assigning stress position and type for word-forms. The counts of nodes and clauses on the horizontal axis were estimated.

aspect: the DM methods automatically constructed formal knowledge being nearly half the size of the expert-defined rules, while still retaining similar accuracy. The smallest statistically significant difference between *J48* trees and the expert-defined rules is 1.91% while between *PART* rules and the expert-defined rules it equals 1.82%

and the experiment with the algorithm Heuristics + *Bagging* is 0.73%. The minimal statistically significant difference between the algorithms Heuristics + *AdaBoostM1* and the expert-defined rules is 1,71%.

6. Humans Versus Computers on Words

6.1. *Experiments*

In the second set of experiments, the six DM methods and the expert-defined rules were compared to the ability of humans on the task of assigning stress to 100 known and 100 unknown word-forms.

The 100 known word-forms were randomly selected from the dictionary, described in Section 1, while unknown word-forms were extracted from the word-stock of Slovene language, containing almost 180.000 rare Slovene word-forms. They consist mostly of technical terms and foreign words, together with their morphological information. Out of these, 200 word-forms were randomly selected. The extracted word-forms were then inspected by the authors to further eliminate word-forms, which might be known to an average Slovene-speaking person. This yielded 100 unknown word-forms that were used in the experiments.

Since the known word-forms are included in the dictionary and the unknown word-forms are not, the known word-forms are actually all “known” to the machine predictor and usually most of them to humans also, while the unknown word-forms are actually “unknown” to humans and machines.

In the experiments with 100 known and 100 unknown word-forms, we applied the same methodology as in the previous section. Here, the methods for machine

stress assignment used the whole dictionary from Section 1 as the training set and the known/unknown word-forms as the test set. The DM methods used were the following:

- *J48* decision trees with leaves containing at most two examples.
- Heuristics + *J48* decision trees with leaves containing at most two samples.
- Default boosting: boosting using ten *J48* decision trees as basic classifiers. Leaves of the decision trees contained at most two examples.
- Heuristics + default boosting.
- Optimal boosting: boosting using ten *J48* decision trees as basic classifiers. The decision trees were pruned to contain at most 0.01% of examples in leaves.
- Heuristics + optimal boosting.

In addition, we asked an expert and 10 other Slovene-speaking people (non-experts) to assign stress to the same word-forms. All non-experts had at least a university degree in technical sciences but no formal education in linguistics. The results of the non-experts were averaged. Humans first stressed the known/unknown word-forms by marking the stress on word-forms written on paper, and second, by reading the known/unknown word-forms aloud. The sound records were later analyzed by an expert, which annotated the spoken stress. In this way, mistakes made due to the difficulties people usually have with diacritics were avoided.

6.2. *Measurements*

6.2.1. *Stress Position*

When humans stress word-forms, they predict stress position and type simultaneously. To enable the comparison between humans and machines, we extract the information on prediction of stress position out of the results.

On the known word-forms, the expert-defined rules with the accuracy of 74% achieved the worst result. Boosting (default and optimal) alone accurately assigned the stress position on all vowels, but the use of the heuristic probability model slightly deteriorated its performance thus reaching the accuracy of 99.5%, being equal to the accuracy of the expert when dictating. A slightly worse result, 98.8%, was achieved by the expert writing, followed by the non-expert speaking (98.1%) and non-experts writing (95.8%). These and all subsequent results show that humans achieve better results when dictating. Also, the expert in both cases outperformed an average non-expert. Even worse results, 90.5% and 89.1% were achieved by heuristics + *J48* decision tress and *J48* decision trees without heuristics, respectively.

In the more interesting test of predicting stress position on unknown word-forms, boosting with the heuristic probability model again performed best of all artificial methods. This time the use of the probability model corrected the predictions of boosting. The results were the following: 89.0% for heuristics + optimal boosting, 87.5% for optimal boosting without heuristics, 86.9% for heuristics + default boosting and 85.7% for default boosting without heuristics. Again, the worst results of DM methods were achieved by *J48* decision trees. The method including heuristics achieved the accuracy of 82.7%, without heuristics the result was 80.7%. Performance of the expert-defined rules was the worst with the accuracy of 76.9%.

The results for humans were as follows: the expert speaking 96.2%, non-experts speaking 95.6%, the expert writing 95.1% and finally, non-experts writing 90.5%.

The humans, the expert and non-experts, outperformed all artificial methods predicting stress on unknown word-forms. All predictions on unknown word-forms were worse than the corresponding predictions on known word-forms with the exception of the expert-defined rules.

Although the use of the heuristic probability model worsened the performance of boosting alone on known word-forms, we are more interested in the accuracy of stressing the unknown words, where the probability model shows its advantage.

6.2.2. *Stress Type*

Since stress type is defined only for the stressed vowels *e* and *o*, the test sets of known and unknown word-forms contained only 41 and 27 instances respectively. Because stress type was predicted, no heuristic model was used with DM methods. For known word-forms, boosting (default and optimal) succeeded in reaching 100% accuracy. The expert speaking assigned stress type with the accuracy of 97.6%, followed by the expert writing with 92.7%, *J48* decision trees with 87.8%, non-experts speaking with 79.0% and non-experts writing with 75.6% of accuracy. The lowest accuracy was achieved by the expert-defined rules: 63.4%.

For Slovene-speaking people, the assignment of stress type is in general more difficult than the localization of stress position. This was confirmed by the results, where the expert achieved nearly 20% better accuracy on this task than the average non-expert.

On unknown word-forms, default and optimal boosting (81.5%) performed better than the expert-defined rules (66.7%), even better than non-experts when writing stress type (74.8%) and also better than *J48* decision trees (74.1%). The expert speaking (100.0%) achieved the best results, while the expert writing (92.6%) and non-experts speaking (91.9%) were considerably worse. Again, the humans pronouncing the words significantly outperform computers on unknown word-forms.

6.2.3. *Assigning Stress to Words*

As in previous tests, the predictions have to be combined in order to evaluate the methods on the full task of stress assignment. The results are presented in Fig. 2 and Fig. 3. On known word-forms, the expert-defined rules achieved by far the lowest accuracy being worse than non-experts writing by 34.7 percentage points. Boosting (optimal and default) alone again reached 100% accuracy on known word-forms, while the use of the heuristics lowered its performance to 98%. Nevertheless, it is still the best machine method, better than all non-experts and the expert.

On unknown word-forms, all machine methods and humans achieved worse results than on known word-forms with one exception – the expert-defined rules. The too-general short rules seem to perform quite well on unknown words. All the DM methods were again worse than humans but better than their rules. The use of the heuristic probability model brought again a major advantage for boosting on this task.

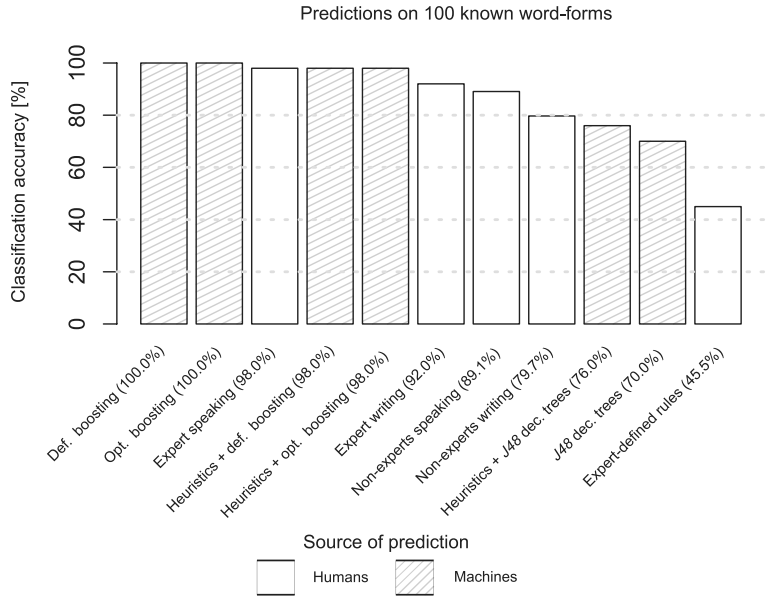

Fig. 2. Accuracy achieved by humans and computer methods on the problem of assigning stress to 100 known word-forms.

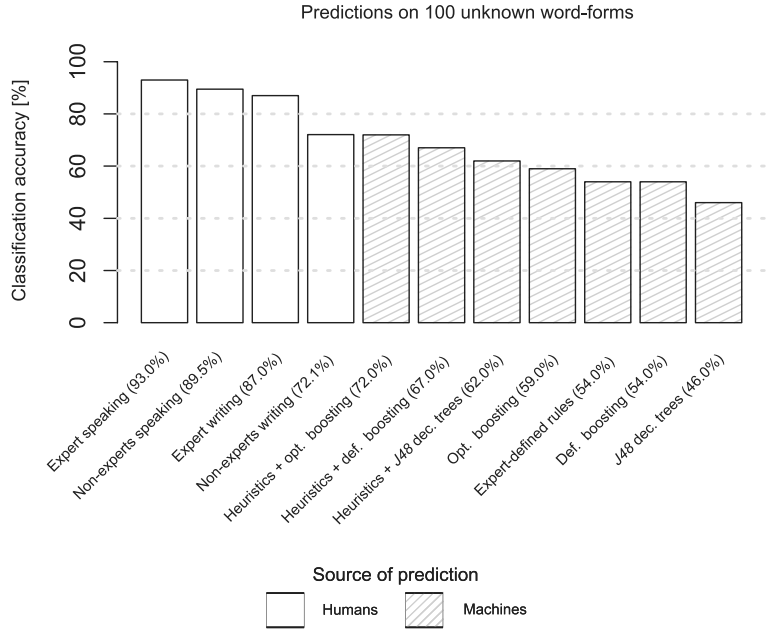

Fig. 3. Accuracy achieved by humans and computer methods on the problem of assigning stress to 100 unknown word-forms.

7. Conclusion

After extensive experiments using DM methods and the expert-defined rules on a dictionary of Slovene word-forms, and additionally also non-experts and an expert on the two sets of known and unknown word-forms, the following conclusions can be drawn.

The morphological information, which is thought to be necessary for predicting stress on word-forms, helped to produce significantly shorter DM models with similar or higher accuracy than the expert-defined rules or significantly shorter rules with similar accuracy (see Table 4). The constructed knowledge differs from the expert-defined rules and therefore represents new, original knowledge. The generated simple rules already cover around half of all the words, thus representing useful information for foreigners learning Slovene. However, it seems that the problem of stress assignment bears no short and accurate rules for all Slovene words.

As expected, on the known word-forms both humans and machines performed well. The expert came close to 100% accuracy by incorrectly stressing only 2 out of 100 word-forms, which can be attributed to simple human error. Obviously, good performance on known word-forms is trivial to achieve with machine methods. The results also showed that humans always achieve better results when dictating the word-forms than when writing them down, and both are always better than the expert-defined rules (see Fig. 2 and Fig. 3). This indicates the humans have only a limited ability to formulate their knowledge.

Regarding the comparison between humans and machines on assigning stress to unknown words, the results indicate that humans stress differently than the applied machine methods: had machine methods copied human knowledge and procedures well, there would not exist a 20% to 30% difference in accuracy between them. Moreover, the "human dictionary" is several times smaller than the dictionary available to computer methods in experiments. Surprisingly, machine methods were not able to cope with unknown words well, although the training data contained around 600.000 word-forms.

The presumption that humans are in general able to correctly stress Slovene words even when they have never seen or heard them before has proved to be mainly true. Humans, and especially experts, can indeed assign stress to Slovene words quite well, although they still have more difficulties with unknown word-forms than with known ones (see Fig. 2 and Fig. 3). They obviously posses some knowledge of pronunciation, otherwise they would not be able to pronounce well the unknown words.

Despite the differences between humans and machines, machine methods proved to be usefully applicable in various scenarios e.g. for speech synthesis. Although machine methods do not reach the performance of native speakers, they are better than the expert-defined rules even when assigning stress to unknown word-forms. Therefore, the DM methods are undoubtedly already very useful in practical tasks where most words are known.

The experiments described in this paper clearly show that there must be some stress assignment rules for Slovene words. Even the short expert-defined rules performed relatively well assigning stress to unknown words. Why is then that the DM methods were

still not able to "mine" these rules even after at least a decade of intensive research of this particular problem?

Acknowledgments

This work was supported by the Slovenian Ministry of Higher Education, Science and Technology. The authors wish to thank the volunteers for participating in the study.

References

Bell, D.A., J.W. Guan and Y. Bi (2005). On combining classifier mass functions for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, **17**(10), 1307–1319.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **2**(24), 123–140.

Cleary, J.G., and I.H. Witten (1984). Data compression using adaptive coding and partial string matching. *IEEE Trans Communications Communications*, **COM-32**(4), 396–402.

FIDA (2006). *Corpus of Slovene Language*. `http://www.fida.net/eng/index.html`

Frank, E., and I.H. Witten (1998). Generating accurate rule sets without global optimization. In *Proceedings of ICML-98*, Madison. pp. 144–151.

Freund, Y., and R.E. Schapire (1996). Experiments with a new boosting algorithm. In *Proceedings of ICML-96*, Bari. pp. 148–156.

Gros, J., A. Mihelic, N. Pavesic, M. Zganec and S. Gruden (2005). Slovenian text-to-speech synthesis for speech user interfaces. In *Proceedings of the Third World Enformatika Conference*, Istanbul. pp. 216–220.

Lauriciukaite, S., and A. Lipeika (2007). Framework for choosing a set of syllables and phonemes for lithuanian speech recognition. *Informatica*, **18**(3), 395–406.

Li, J. (2006). On optimal rule discovery. *IEEE Transactions on Knowledge and Data Engineering*, **18**(4), 460–471.

Mihalcea, R., and V. Nastase (2002). Letter level learning for language independent diacritics restoration. In *Proceedings of CoNLL'02*, Taiwan. pp. 105–111.

Quinlan, R. (1993). *C4.5*: *Programs for Machine Learning*. Morgan Kaufmann, San Mateo.

Rojc, M., and Z. Kačič (2007). A unified approach to grapheme-to-phoneme conversion for the plattos Slovenian text-to-speech system. *Applied Artificial Intelligence*, **21**(6), 563–603.

Simard, M., and A. Deslauriers (2001). Real-time automatic insertion of accents in French text. *Natural Language Engineering*, **7**(2), 143–165.

Skripkauskas, M., and L. Telksnys (2006). Automatic transcription of Lithuanian text using dictionary. *Informatica*, **17**(4), 587–600.

Šef, T., and M. Gams (2001). Data mining for creating accentuation rules. *Applied Artificial Intelligence*, **18**(5), 395–410.

Šef, T. (2001). Analiza besedila v postopku sinteze slovenskega govora. *PhD Dissertation*. Ljubljana.

Taylor, P. (2005). Hidden markov models for grapheme to phoneme conversion. In *Proceedings of Interspeech*, Lisbon. pp. 1973–1976.

Toporišič, J. (1984). *Slovene Grammar* (*Slovenska slovnica*). Založba Obzorja, Maribor.

Tufiş, D., and A. Chiţu (1999). Automatic diacritics insertion in Romanian texts. In *Proceedings of COMPLEX*, Pecs.

Webb, G.I., J. Boughton and Z. Wang (2002). Averaged one-dependence estimators: Preliminary results. In *Proceedings of ADM-2002*, Canberra. pp. 65–73.

Witten, I.H., and E. Frank (2000). *Data Mining*: *Practical Machine Learning Tools and Techniques*, 2 edition. Morgan Kaufmann, San Francisco. `http://www.cs.waikato.ac.nz/ ml/index.html`

Yarowsky, D. (1994). Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French. In *Proceedings of ACL 1994*, Las Cruces. pp. 88–95.

Yarowsky, D. (1999). A comparison of corpus-based techniques for restoring accents in Spanish and French text. In *Natural Language Processing Using Very Large Corpora*. Kluwer Academic Publishers. pp. 99–120.

D. Marinčič is a PhD student and a research assistant at the Department of Intelligent Systems, Jozef Stefan Institute, Ljubljana, Slovenia. He received his MSc in computer science at the University of Ljubljana in 2004. His major research interests include language technologies, dependency parsing of unrestricted Slovene text and computer game playing.

T. Tušar is a research assistant at the Department of Intelligent Systems, Jozef Stefan Institute, Ljubljana, Slovenia. She received her BSc degree in applied mathematics at the University of Ljubljana in 2002 and her MSc degree in computer science at the same university in 2007. Her research interests include machine learning methods for text processing and outlier detection and evolutionary algorithms for singleobjective and multiobjective optimization.

M. Gams is an associate professor of computer and information science at the University of Ljubljana and a senior researcher at the Jozef Stefan Institute, Ljubljana, Slovenia. He teaches several courses in computer science at graduate and postgraduate levels at Faculties of Computer and Information Science, Economics, etc. His research interests include artificial intelligence, intelligent systems, intelligent agents, machine learning, cognitive sciences, and information society. He has headed several major artificial intelligence applications in Slovenia, including the major national employment agent on the Internet, and the Slovenian text-to-speech system "Govorec" (Speaker) donated to several thousand users.

T. Šef is a senior researcher at the Department of Intelligent Systems at the Jozef Stefan Institute, Ljubljana, Slovenia. He received his PhD degree in computer science (language and speech technologies) in 2001 at the University of Ljubljana. His research interest include artificial intelligence, intelligent systems, natural language processing, speech processing, Slovenian text-to-speech synthesis and forensic speaker identification. He developed the Slovenian text-to-speech system "Govorec" (Speaker) donated to several thousand users.

Slovėnų kalbos automatinio kirčiavimo analizė

Domen MARINČIČ, Tea TUŠAR, Matjaž GAMS, Tomaž ŠEF

Buvo tikrinamas žmonių ir kompiuterių sugebėjimas kirčiuoti slovėnų kalbos žodžius, naudojant duomenų gavybos metodus. Tai yra iššūkis kompiuteriams, kadangi žmonės atlieka šią užduotį puikiai netgi nežinomiems žodžiams ir nenaudodami konteksto. Tikslas surasti gerus automatinio kirčiavimo modelius buvo siekiamas taikant naujus metodus ir žinomą teoriją apie slovėnų kalbos kirčiavimo taisykles. Patobulinti duomenų gavybos metodai aplenkė ekspertų nustatytas taisykles praktiškai visais atvejais pademonstruodami, kad duomenų gavyba gali daugiau negu varžytis su žmonėmis konstruojant formalias žinias apie kirčio priskyrimą. Tačiau kirčiuojant nežinomus žodžius duomenų gavybos metodais iki šiol nesigauna tokie geri rezultatai palyginus su gaunamais rezultatais kai kirčiuoja žmonės.