COMPARISON BETWEEN HUMANS AND MACHINES ON THE TASK OF ACCENTUATION OF SLOVENE WORDS

Tea Tušar, Andrej Bratko, Matjaž Gams, Tomaž Šef
Department of Intelligent Systems
Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia
tea.tusar@ijs.si, andrej.bratko@ijs.si, matjaz.gams@ijs.si, tomaz.sef@ijs.si

ABSTRACT

The accentuation of Slovene words represents a challenging task for automated solvers since in Slovenian, stress can be located on arbitrary syllables. This paper compares the performance of humans, expert-defined rules and computer methods, such as machine learning methods and n-gram Markov models, on this task. We find that humans tend to accentuate the words correctly, even when they have never heard or seen them before. On the other hand, expert-defined rules for accentuation perform quite poorly, achieving worse results than machines. This indicates that humans are good at accentuating, but very limited when their knowledge has to be formalized. Therefore, machine methods have to be employed for automatic accentuation of Slovene words.

1 INTRODUCTION

The grapheme-to-phoneme conversion can be described as a function mapping the spelling form of words to a string of phonetic symbols, representing the pronunciation of the word. Most work on data-oriented grapheme-to-phoneme conversion has been performed on a few worldwide languages, especially on English [2]. Several highly inflected languages lack large databases that give the correspondence between the spelling and the pronunciation of all word-forms. For example, no database for orthography/phonology mappings for Russian-inflected words is known [6].

While most other languages have difficulties with grapheme-to-phoneme conversion, in the Slovene language this is rather straightforward if the accentuated form of words (i.e. stress assignment) is known. The conversion can be done on the basis of less than 100 context-dependent letter-to-sound rules with over 99% accuracy [5]. However, no good rules exist for stress assignment of Slovene words. This is somehow in contradiction to the general observation that humans can often pronounce words reasonably well, even though they have never seen or heard them before.

Accentuation of Slovene words is a hard problem, since the lexical stress can be located almost arbitrary on any syllable of the word [7]. Most words have only one stressed syllable, but there exist also words with no stress and words with more than one stress. Furthermore, different forms of the same word can be stressed differently (see Table 1).

In this paper, we inspect the performance of humans (human volunteers, human experts and expert-defined rules) and machines (machine learning methods and n-gram Markov models) on the task of stress assignment of Slovene words. We analyze the relation between human knowledge and their rules and compare the results of expert-defined rules and machines to find the best method for automatic accentuation of Slovene words.

Slovenian	English translation
Danes moraš *peljáti* ti.	Today, you have *to drive*.
Rada se *péljeva* po deželi.	We like *to drive* across the country.
Pêlji, prosim.	*Drive*, please.

Table 1: Stress assignment for the verb *peljati* (to drive) in different forms (infinitive, first person in present tense and imperative). The stressed vowels are marked with the phonetic signs ‘´’ and ‘ˆ’, which are usually not written in the normal text.

2 PROBLEM DESCRIPTION AND RESOURCES

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. In the Slovene language, the stressed vowels differ according to *duration* (short and long vowels – all vowels) and *quality* (narrow and wide vowels – only the vowels e and o). In this paper we distinguish only the quality of stress. Besides the vowels a, e, i, o and u in Slovenian there exists also the reduced vowel, which can appear instead of the vowel e and is always present before the consonant r, provided there are no other vowels around it.

The stress assignment task thus consists of classifying all vowels (and reduced vowels) of a word in one of the following classes: *unstressed vowel* (all vowels and reduced vowels), *stressed vowel* (vowels *a*, *i* and *u*), *wide stressed vowel* (vowels *e* and *o*), *narrow stressed vowel* (vowels *e* and *o*), or *stressed reduced vowel*.

Unlike in other languages, in the Slovene language the stress assignment depends on the morphological category of the word. Therefore, we use a Slovene pronunciation dictionary (created in previous work [8]), where for every word we have the following information (see Table 2): the word (without stress), the word's lemma, the stressed word and the word's morphological information. The dictionary contains almost 600.000 words with more than 2.000.000 syllables.

The dictionary holds only the most common words. To find words, which are unknown to an average Slovene-speaking person, we use the word-stock of Slovene language. It contains almost 180.000 rare words (mostly technical terms and foreign words) together with their morphological information. Beside these two resources, we provide several lists of parts of words that bear some information on the stress of the word, for example, prefixes and suffixes that are usually not stressed. These lists were derived from expert-defined rules and together contain 230 entries.

word	lemma	stressed word	morphological information
peljati	*peljati*	*peljáti*	verb, infinitive
peljeva	*peljati*	*péljeva*	verb, present tense, first person, dual
pelji	*peljati*	*pêlji*	verb, imperative, present tense, second person, singular

Table 2: Information, contained in the dictionary for the words *peljati*, *peljeva* and *pelji*.

3 METHODS FOR AUTOMATIC ACCENTUATION

3.1 Expert-defined Rules

The rules for accentuation of Slovene words were created by the best human experts more than 20 years ago [7]. They were written to help foreigners who are studying Slovenian and Slovene people who speak local dialects to learn the correct formal pronunciation of Slovene words. In this paper, a slightly modernized machine-readable version of expert-defined rules [3] is implemented in 68 IF-THEN rules.

The expert-defined rules that predict the position of stress rely mostly on common word prefixes and suffixes mentioned in the previous section. These lists are scanned in a predefined order. If a word's prefix or suffix matches an entry in the current list, the word is stressed accordingly. If the word does not match any of the list's entries, the stress position is set to be the most frequent stress position in other words with the same number of syllables. This happens with approximately 25% of all words.

The stress type has to be defined only for stressed syllables containing the vowels *e* or *o*. For this task, the expert-defined rules make use of the context of the observed vowel in the word as well as the word's morphological information. For example, the following expert-defined rule predicts a wide stressed *o* [7]:

All nouns that contain a stressed o *in the endings* -oba, -oča *or* -ota, *have a wide stressed* o.

3.2 Machine Learning Methods

Machine Learning (ML) methods build a model from a given data set and use this model to classify new instances. We applied many ML methods on this task (decision trees, decision rules, one variety of naive Bayes classifier and meta methods boosting and bagging), all from the WEKA ML toolkit [9].

In the task of stress assignment of Slovene words with ML, we classify the stress on every vowel individually. We first use a ML 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.

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*. Machine learning is performed on each group separately. Each vowel is thus treated as an instance, described with a set of attributes. These attributes contain information on the word in which the vowel appears and the context of the vowel in the word, as well as characteristic prefixes and suffixes contained in the word. When predicting stress position, each vowel is described with 75 attributes (see Figure 1). The class attribute can take on one of the values *stressed* and *unstressed*. The same attributes are employed for predicting stress type, with the exception of the class attribute, which can have one of the following values: *narrow stressed vowel*, *wide stressed vowel* or *stressed reduced vowel*.

3.3 n-gram Markov Model

We also employ a character-level Markov model, called *Prediction by Partial Matching* (PPM), which was originally designed for lossless text compression [1]. The PPM algorithm predicts the next character in a sequence based on preceding text. The text is approximated with a finite-length n-gram Markov model, so that the current symbol is considered independent of all but the previous $n-1$ characters. These characters are called the current *context*, its length $n-1$ is the *order* of the PPM

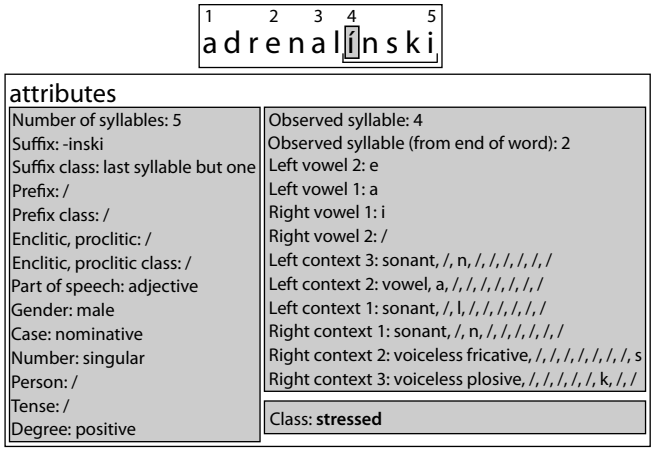

Figure 1: Attributes for the fourth vowel of the adjective *adrenalinski* (adrenaline). The attributes on the left are bound on the word, while the attributes on the right depend on the observed syllable.

model. Many variants of the PPM algorithm exist. We use escape method D in combination with the exclusion principle.

In predicting stress position (and type), the PPM model is built on the training portion of the pronunciation dictionary. To predict the accentuation of words in the test set, we generate all plausible stress assignments for each word. We then compute the probability of each such solution using the trained model and predict the solution that is deemed most probable. When predicting the position of stressed vowels, only combinations that contain up to three stressed vowels are considered. An example for predicting stress position for the word *relief* is given in Figure 2.

The probability of a word is computed as the product of character probabilities, as predicted by the PPM

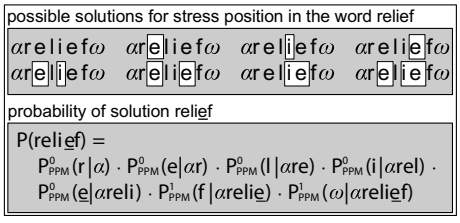

Figure 2: Predicting stress position for the word *relief* (same word in English and Slovenian). All solutions considered by the method are listed in the top part of the figure. The special characters α and ω mark the beginning and the end of a word. The evaluation of the correct solution *reliéf* is depicted in the bottom part of the figure. P^0_{PPM} denotes the character probability of the first PPM model. After encountering the stressed vowel *e*, PPM switches to the second model P^1_{PPM}.

model. For every word, we trained two models, one for prediction from left to right and a second model for prediction from right to left. The final probability assigned to a stressed word is simply the average of both models.

4 EXPERIMENTS AND RESULTS

4.1 Experimental Setup

We perform three different experiments. In the first, we compare the expert-defined rules to machine methods. To this end we use the dictionary, described in Section 2. The words from the dictionary are divided into three corpora of similar size in such way that words with the same lemma are always placed in the same corpus. All applied methods are thus evaluated with 3-fold cross validation, where two corpora are used for training and the remaining corpus for testing.

In the second experiment, the whole dictionary is used as the training set and 100 random words from the same dictionary are selected to represent the test set. These words are common Slovene words and are "known" to machine methods as well as to humans. For the third experiment, we use the dictionary as the training set and 100 unknown words as the test set. The "unknown" words were obtained from the word-stock (see Section 2) in the following manner. First, we randomly selected 200 words from the word-stock. Words, which might have been known to an average Slovene-speaking person, were manually eliminated. This yielded 100 unknown words that are used in the third experiment.

For space limitations, among all ML methods we report only the results of the best ML method – boosting [4] (called *AdaBoostM1* in WEKA). Boosting is run using ten C4.5 decision trees as basic classifiers. All other parameter settings are the same as the default settings in WEKA [9]. The PPM method accepts a single parameter – the order of the PPM model. An order-4 model was found to perform best or near-best in all tasks and is used as the default setting in all comparisons.

In experiments with 100 known/unknown words we asked ten Slovene-speaking people and a human expert to accentuate these words. All volunteers have at least a university degree in technical sciences. The results of volunteers are averaged. Humans first stressed the known/unknown words by marking the stress on words written on paper, and second, by reading the known/unknown words aloud. The sound records were later analyzed by the human expert, which annotated the spoken stress assignment. In this way, mistakes made due to difficulties with phonetic signs were avoided.

4.2 Results

The results of all three experiments are presented in Figure 3. On the words from the dictionary, boosting

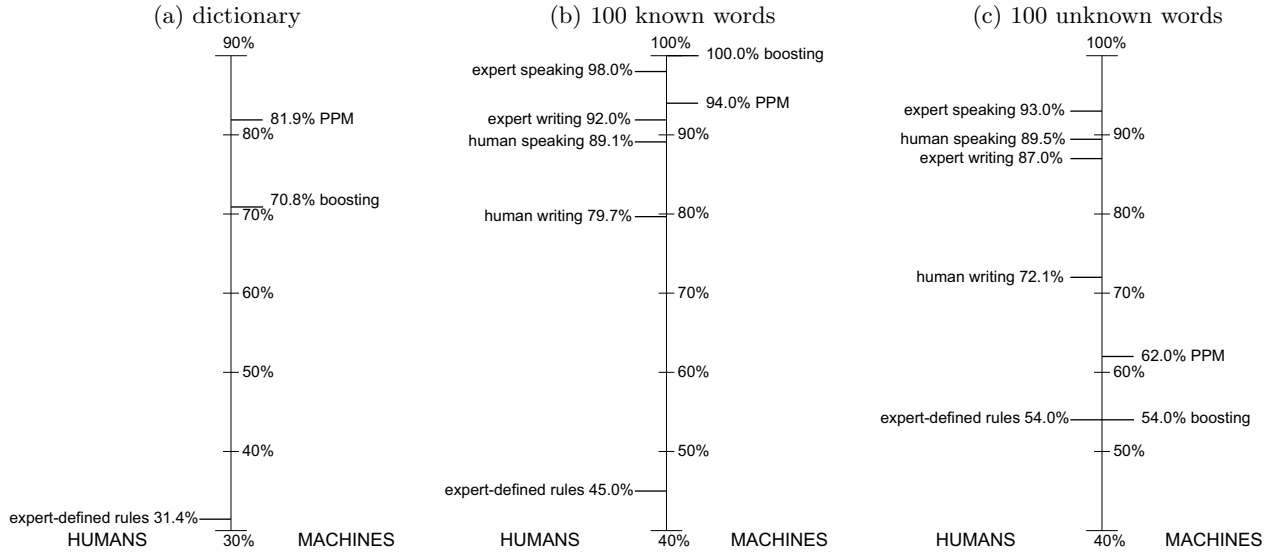

Figure 3: Comparison of accuracy achieved by humans and "machines" on the problem of stress assignment of (a) words from the dictionary, (b) 100 known words and (c) 100 unknown words.

outperformed the expert-defined rules by a 40% improvement in accuracy. Even better results were achieved by PPM. Similar relation between machine methods and expert-defined rules can be observed also on known words, while on unknown words, expert-defined rules equal boosting.

The human expert always achieves better results than the average volunteer and people accentuate more accurately when speaking than writing. On known words, boosting and PPM achieve better accuracy than humans because the test words were also used for training. On unknown words, this changes and humans are better than all artificial methods.

5 CONCLUSION

While humans accentuate Slovene words correctly, they have only a limited ability to formulate their knowledge. This has been shown in two ways. Firstly, they achieve better results when speaking than when writing down the stressed words. And secondly, the expert-defined rules, which should incorporate the human ability for correct accentuation, achieve very bad results.

Machine methods obtained good results on the words from the dictionary and are therefore more suitable for automatic accentuation of Slovene words than the expert-defined rules.

Acknowledgment

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

References

[1] John G. Cleary and Ian H. Witten. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, COM-32(4):396–402, 1984.

[2] Walter Daelemans and Antal van den Bosch. Language-independent data-oriented grapheme-to-phoneme conversion. In *Progress in Speech Synthesis*, pages 77–90. Springer Verlag, New York, 1996.

[3] Matjaž Erpič. Conversion of text into phonemes (Pretvorba besedil v foneme). B.Sc. thesis, University of Ljubljana, Slovenia, 1995.

[4] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of ICML-96*, pages 148–1556, Bari, 1996.

[5] SAZU. *Slovene Orthography - 1. Rules (Slovenski pravopis - 1. Pravila).* Državna založba Slovenije, Ljubljana, 1990.

[6] Richard Sproat, editor. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.* Kluwer Academic Publisher, Boston, 1998.

[7] Jože Toporišič. *Slovene Grammar (Slovenska slovnica).* Založba Obzorja, Maribor, 1984.

[8] Tomaž Šef and Matjaž Gams. Data mining for creating accentuation rules. *Applied Artificial Intelligence*, 18(5):395–410, 2004.

[9] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools with Java implementations.* Morgan Kaufmann, San Francisco, 2000. `http://www.cs.waikato.ac.nz/~ml/index.html`.