Precision of Statistical Syllable Segmentation as a Function of Training Data Quality

Jozef Ivanecký[1] and Daniela Majchráková[2]

[1] European Media Laboratory GmbH, Heidelberg, Germany
Jozef.Ivanecky@eml-d.de
[2] Jazykovedný ústav Ľ. Štúra, Slovenská akadémia vied,
Bratislava, Slovakia
danam@korpus.juls.savba.sk

Abstract. The statistical approach for syllabic segmentation in Slovak seems to be easier to implement and gives better or at least similar results when compared with a rule-based system. The performance strongly depends on the quality as well as quantity of the training data. The proper test set is also very important. The paper describes our efforts to achieve the optimal error rate. We give a theoretical overview on training and testing techniques as well as a description of the real experiments with different selections of training and test data sets. The results lead to the conclusion that in case of limited training data, the selection of the data is particularly important.

1 Introduction

Although the automatic determination of the syllabic boundaries does not have many practical applications, it is necessary for automatic transcription in case the methods are based on the production rules instead of statistical approaches [2]. In Slovak language the pronunciation of *de, te ne, le* and *di, ti, ni, li* changes on syllabic boundaries. Therefore the detection of distinct boundaries is necessary for the correct determination of the pronunciation of *de, te, ne, le* and *di, ti, ni, li*.

Determining syllabic boundaries is complicated by the fact that the syllable definition is ambiguous. For our experiments we applied Pauliny's syllable definition which is primarily based on phonological principles [6]. Considering this fact, that syllable definition is not strictly defined, the specification of exact rules for syllabic segmentation is also difficult. Rule-based systems require sets of simple rules. Applying these rules we did not achieve better results than an 80%–85% success rate. Better results can be achieved only by using more complicated rules, but this correlates with an increase in exceptions.

The other problem is that in some words more than one syllable segmentation is possible which is also issued from the absence of strict rules. For example the word *bystrý* can be segmented as *by-strý, bys-trý* and *byst-rý*. In each case the syllabic segmentation is correct and the number of syllables is fixed.

In our experiment we had to reflect two facts:

- There are no exact rules for the determinination of syllabic boundaries in Slovak.
- In some cases, several different, though correct segmentations are possible.

Our first approach combined the rule-based syllabic segmentation together with a new approach which applies language model theory to the syllabic segmentation. Since we believe that a merely statistical approach is promising, we focused only on the improvement of the statistical model.

In the statistical approach we applied well known methods used in language modeling to the syllabic segmentation. In the case of language models the basic unit is a word. In our case these are syllables. Each word is first split into all possible sequences of syllables. For each syllable sequence $\mathbf{S}$, where

$$\mathbf{S} = s_1, s_2, \ldots, s_n \qquad s_i \in \xi \tag{1}$$

and ξ is the set of all possible syllables. Based on the Bayesian criterion, one can define the likelihood for each given sequence of syllables $\mathbf{S}$ as

$$P(\mathbf{S}) = \prod_{i=1}^{n} P(s_i \mid s_1, \ldots, s_{i-1}) \tag{2}$$

where $P(s_i \mid s_1, \ldots, s_{i-1})$ is the likelihood, that the syllable s_i follows after syllables $s_1, \ldots, s_{i-1}$. If we consider just two anterior syllables, the equation (2) can be rewritten as

$$P(\mathbf{S}) = \prod_{i=1}^{n} P(s_i \mid s_{i-2}, s_{i-1}) \tag{3}$$

To estimate the likelihoods $P(s_i \mid s_{i-2}, s_{i-1})$ we used the "syllabic corpus" created from the training set. The selection of the training set is discussed in Section 2. For the estimation we used the counts of syllabic sequences. As we mentioned above, we consider just two anterior syllables and thus (3) can be written as

$$P(s_3 \mid s_1, s_2) = f(s_3 \mid s_1, s_2) \doteq \frac{C(s_1, s_2, s_3)}{C(s_1, s_2)} \tag{4}$$

where $f(\ \mid\ )$ is the occurrence count function.

For two reasons equation (4) is not suitable for the likelihood estimation of a given syllable:

- for Slovak language monosyllabic and bisyllabic words are common
- not all sequences of syllables s_1, s_2, s_3 may occure in the training set

Based on the previous equation it is necessary to consider the likelihood $P(s_3 \mid s_1, s_2)$ as an interpolation of the count occurance for the sequence of three, two and one syllable:

$$P(s_3 \mid s_1, s_2) = \lambda_3 f(s_3 \mid s_1, s_2) + \lambda_2 f(s_3 \mid s_2) + \lambda_1 f(s_3) \tag{5}$$

where $\lambda_1 + \lambda_2 + \lambda_2 = 1$. The detailed description of the training as well as the testing process can be found in [3].

For further improvements of this technique we focus now on the selection of the training and test data sets. During the last years we found out, that training data selection is particularly important. Since data for the training has to be manually checked before one can use it, we were looking for the right pre-selection to minimize the human work.

The remainder of the paper is organized as follows: In Section 2 we give a brief overview on the data source for the training as well as testing set. In Section 3 we focus on the experiment design, the experiments and provide a brief summary in Section 4.

2 Data selection

Experimental results in [3] showed that the amount of training data is important. Besides the importance of the amount of the data, we believe that the quality of the data is also important. Since our ability to manually check huge amounts of data is limited, we decided to use first the 11 000 most frequent words from the Slovak National Corpus [1]. For the testing we also used the last 500 words from the corpus.

The Slovak National Corpus in version prim-3.0 has approximately 340 million tokens and contains Slovak language text(mainly journalistic) of many language styles. We extracted a list of the most frequent tokens from prim-3.0. As the list contained not only words but also many tokens like punctuation, one-letter words and abbreviations, we first had to modify the list to get a proper sample set for the experiment. Then we changed all characters to lower case and eliminated all duplicate words. The sample of the first 11 000 words from the corpus is representative of the most frequently used words and their case-forms in Slovak language. It also shows that the most frequent words in Slovak language are usually the shortest (mainly pronouns, conjunctions) and many frequent words appear in our list in all case-forms. Therefore the variability of the first 500 words is not very high.

The frequency sorting of the 11 000 words can be useful if we want to achieve the right segmentation of the most frequent words. Unfortunately it does not have to necesserily imply a good syllable coverage. The amount of syllables in the entire corpus is the same. By using just a fraction of the corpus, we wanted to know if it is better to use a corpus sorted by occurrence frequency or by randomly selected words.

Thus we created two lists from the 11 000 words. In the first list the words were sorted as they were in the corpus. The second one contained the same words but in random order. For each list we counted the number of different syllables after each 100 words. The number of unique syllables in our experiment was 2670. The results are shown in Figure 1.

From Figure 1. it is clear that randomly selected data has a better syllabic coverage. This is important information, as the number of words incorrectly

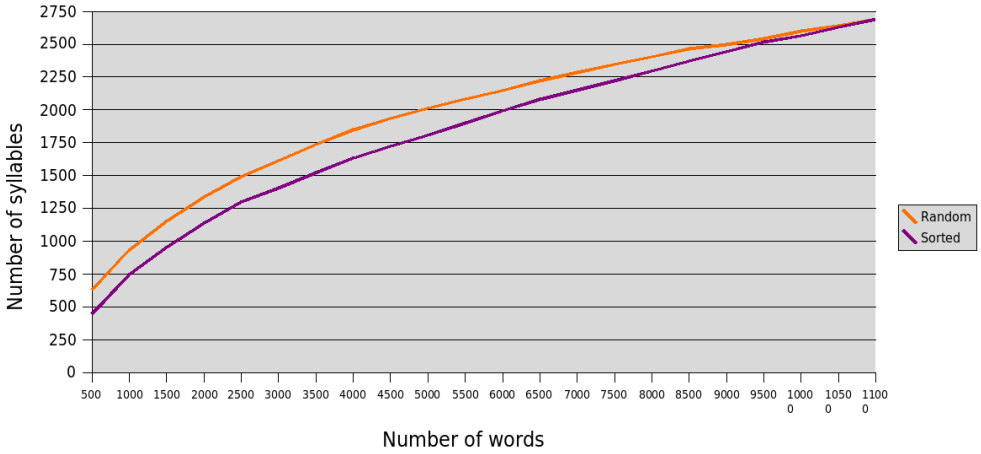

Fig. 1. Number of unique syllables as function of word amounts for a sorted and a random corpus.

segmented due to missing syllables in the training set is relatively high. Results in Section 3 demonstrate this conclusion.

3 Experiments

For the training data selection we designed several scenarios. We always used 10 000 words for the training and 1000 words for testing. In addition we also used the last 500 words from the corpus for testing. We tested all sets against the old system described in [3]. From 11 000 words the first 10 000 were used for the training and last 1000 for the testing. The same was done with the random list of 11 000 words. Therefore we ended up with 3 different systems for the syllabic segmentation as well as 3 different data sets for testing.

During the testing we generated the best sequence of syllables for all the words in the test set. We then compared the best syllable with the segmentation from the data preparation. If they did not match, we manually checked if it was another correct segmentation, an incorrect segmentation or an incorrect segmentation due to a missing syllable in the training set.

In contrast to the test scenario described in [3] we examined just the first (the best) segmentation and did not look at any other generated syllable sequences. The reason for this approach was the fact that if the tool for the segmentation is used in automatized process, it is not possible to determine if only the first or also the second segmentation is correct, or if the first one is not correct and the second one is. We manually checked all incorrect results to ensure proper classification. The initial results are in following table.

	1000 Sorted	1000 Random	Last 500
Sorted training set	17.30 %	7.20 %	30.00 %
Random training set	8.00 %	13.80 %	27.00 %
Old training set	19.20 %	14.60 %	23.20 %

From the table it is clear that results for the last 500 words have a much lower variance than the results for the first two test sets. The explanation is very simple, and also explains why there are 2 results which are much better then the rest. The combination of the sorted training and random test as well as the random training and sorted test should not be part of the test set. In these two particular combinations the test set contains part of the training data. This is the reason why these two results need to be excluded from the results table. These two results point us indeed to the syllable coverage in the test set. This is the main reason why two excluded results are much better then the others. In the following table the number of syllables from the test set is not covered by the training set as well as the percentage of words influenced by missing syllable. The sorted training set contained 2424 unique syllables, the random training set 2459 and finally the old training set 3009 syllables. It is necessary to point out that in the old training set foreign words were not excluded.

	1000 Sorted	1000 Random	Last 500
Sorted training set	116 (10.1%)	9 (0.6%)	139 (20.8%)
Random training set	8 (0.5%)	81 (5.8%)	133 (19.9%)
Old training set	127 (11.7%)	107 (8.3%)	68 (11.8%)

As we can see, the best syllabic coverage is for the "excluded" combinations. In both cases just less then 1% of syllables from the test set are not covered by the training set. From the syllable coverage table it is also clear that there is some relation between word occurrence frequency and syllabic occurrence frequency. 20% of syllables from the last 500 words are not covered by the training set. On average it is 2 times more than for words from the beginning of the corpus.

When we excluded all words containing syllables from the test set which are not covered by the training set, the results of the syllabic segmentation were as follows:

	1000 Sorted	1000 Random	Last 500
Sorted training set	8.00 %	6.63 %	11.61 %
Random training set	7.53 %	8.49 %	8.75 %
Old training set	8.49 %	6.87 %	12.92 %

We can see that the result variance for each test set is lower than in the first results table. But unallowed combinations (sorted–random, random–sorted) are also better than allowed combinations here. From the achieved results we can derive the following conclusions:

– Less frequent syllables appear more often in less frequent words than in common words. The last 500 words test continually showed worse accuracy and had the biggest number of syllables not covered.

- Sorted training data give better performance for more frequent words, but the words from the end of the corpus are significantly worse.
- The amount of training data is still not sufficient. With the current training set almost every second error is caused by syllables not covered by the training data.

To improve the performance of the system the most important thing seems to be to increase the amount of training data. The random selection of the data seems to be more suitable, but this advantage may disappear when the amount of data used for the training is increased several times.

4 Summary

In this paper we described the influence of training data selection for statistical syllabic segmentation to the overall performance of the syllabic segmentation process. We showed that to have better coverage for words from the entire corpus it is better to use random selection of the training words rather than words selected by their occurrence frequency. More important than data selection seems to still be the amount of data. Our experiments confirmed initial assumptions that to achieve acceptable results, the amount of training data has to be increased a few times.

Acknowledgement The authors wish to thank the Slovak National Corpus team for access to the Slovak National Corpus database.

References

1. Garabík, R. Gianitsová, L., Horák, A., Šimková, M., Šmotlák, M.: *Slovak National Corpus*. In: Proceedings of the conference TSD 2004. Brno, Czech Republic: Springer-Verlag, 2004
2. Ivanecký, J.: *Analysis of the Rule Based Phonetic Transcription Technique Applied to the Slovak Language*, Slovko 2005, Bratislava, 2005
3. Ivanecký, J.: *Štatistický prístup pri určovaní slabičných hraníc*, Slovko 2003, Bratislava, 2003
4. Jelinek F.: *Statistical Methods for Speech Recognition*, The MIT Press, Cambridge, Massachusetts, London 1998, 283 s.
5. Kráľ A.: *Pravidlá slovenskej výslovnosti*, Slovenské pedagogické nakladateľstvo, Bratislava 1983, 632 s.
6. Pauliny E.: *Fonológia spisovnej slovenčiny*, Slovenské pedagogické nakladateľstvo, Bratislava 1968
7. Rabiner L., Juang B.-H.:*Fundamental of Speech Recognition*, Prentice Hall, New Jersey, 1993, 507 s.