A BILINGUAL STUDY ON THE PREDICTION OF MORPH-BASED IMPROVEMENT

Balázs Tarján [1], Tibor Fegyó [1,2], Péter Mihajlik [1,3]

[1] Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary
[2] AITIA International Inc., Hungary
[3] THINKTech Research Center Nonprofit LLC, Hungary

tarjanb@tmit.bme.hu, tfegyo@aitia.ai, mihajlik@thinktech.hu

ABSTRACT

Morph-based language modeling has been efficiently applied in improving the accuracy of Large-Vocabulary Continuous Speech Recognition (LVCSR) systems – especially in morphologically rich languages. However, the rate of improvements varies greatly and the underlying principles have been only superficially studied. Having a method that can predict the expected improvement prior to experimentations would be largely useful. In this paper, we introduce language-independent factors affecting morph-based improvement and show how they can be utilized in estimating the effectiveness of statistical morph-based language modeling. The task was broadcast news transcription in two less investigated languages, Hungarian and Romanian. It was found that in case of under-resourced conditions morph-based models can bring significant improvement – even for a morphologically less rich language like Romanian. In addition, it was shown that non-initial morph tagging can constantly outperform explicit modeling of word-boundaries both in terms of letter and word accuracies.

Index Terms— LVCSR, morphologically rich languages, under-resourced languages, broadcast news transcription, Hungarian, Romanian

1. INTRODUCTION

In the last several decades many efforts have been devoted to the development of LVCSR systems; however there are still many issues that the research community has not been able to overcome. In a recent, comprehensive study [1] the task dependency of the models and lack of noise robustness were identified as the greatest shortcomings of HMM (Hidden Markov Model) framework. This study also recommends some research directions to find remedies for the problems. It is claimed that transcending the limitations of the current modeling techniques is not possible without making use of diagnostic analysis.

Accordingly, in this paper, we investigate the relation between the performance of word- and the morph-based language modeling approaches evaluated on two broadcast news recognition tasks. Some aspects of this work have already been investigated for Hungarian in our previous papers [5]–[7]. In [5], [6] the performance of word- and morph-based language models were compared on various recognition tasks and a strong relationship was found between morphological richness and the morph-based improvement. However, we also found that the benefit of subword lexical models over word-based models was diminishing – until it completely disappeared [7] – as the training text size was being increased.

Here we extend our former work in many directions. Firstly, in this study the morph-based improvement is determined not only as a function of training corpus size but also with various vocabulary sizes, in different acoustic conditions and with a new type of word boundary reconstruction technique (non-initial tagging). Secondly, we make an attempt to estimate the impact of word boundary modeling on the recognition performance. And last but not least, the new results are used to refine our model for the prediction of morph-based improvement.

Our findings can be especially beneficial for the community studying under-resourced languages; hence we performed the analysis not only for Hungarian but Romanian, as well. Based on "Languages in the European Information Society" reports Romanian language can be considered as under-resourced [2]. Although Hungarian is slightly better supported with speech and text resources and categorized as a moderately-resourced language [3], there are no publicly available resources for broadcast speech. As a consequence, in addition to morph-based language modeling other under-resourced language specific modeling approaches (use of the Web, grapheme-based approach) were also utilized in our experiments [4].

Apart from those that were already cited, we know about two papers dealing with LVCSR for Hungarian broadcast speech. The first one [8] compared the performance of deep neural networks and standard HMM acoustic models. Whereas the second one [9] investigates

Table 1. Distribution of the Hungarian test set according to acoustic categories

Categ. name	Meaning	Length [min.]	SNR [dB]
F0	Clean, planned speech	38	20-25
F1	Clean, spontaneous speech	18	20-25
F2	Speech on telephone channels	2	-
F3	Speech with background music	10	8-10
F4	Speech in degraded acoustics	84	10-15
F5	Non-native speaker	3	-

Table 2. Characteristics of training text corpora

Lang.	Corpus	# of words	Vocab. size	Eval PPL	Eval OOV
HUN	TRS	530k	74k	598	10.0%
	WEB 30D-	1.2M	115k	761	8.3%
	WEB 30D+	50.1M	955k	551	1.5%
ROM	TRS	305k	27k	315	8.8%
	WEB 30D-	1.5M	72k	382	3.9%
	WEB 30D+	20.3M	215k	345	1.7%

the potentials in unsupervised training methods. Note that none of these two studies apply morph-based language models. There are only a few papers reporting LVCSR results for Romanian language. In [10] statistical machine translation supported language model adaptation is investigated for the recognition of tourism-specific Romanian read speech. The unsupervised adaptation method presented in that study was later outperformed in [11] with semi-supervised methods, however the recognition task remained the same. In both studies the vocabulary size was restricted to the most frequent 64000 words due to the limitation of Sphinx3 decoder. To the best of our knowledge the only paper presenting Romanian broadcast news results is our former work [12], in which we introduced a web-based, automatic language model updating method for three East-Central European languages including Hungarian, Romanian and Polish.

In the next section the databases used for the training and evaluation of the broadcast news transcription systems are presented. The subsequent section introduces the applied modeling approaches that is followed by the experimental results and their discussion. In section 5 the potentials in the prediction of morph-based improvement are presented, while in the last section we summarize our findings.

2. TASK AND DATABASES

2.1. Acoustic training and test data

The acoustic training database of our Hungarian and Romanian broadcast speech transcription system consisted of manually transcribed TV news programs. The Hungarian database contained 50 hours, whereas the Romanian was based on 31 hours of transcribed speech.

The Hungarian test set comprises 6 TV news (155 minutes altogether) collected in January of 2012 from various Hungarian TV channels (*TV2, MTV, Duna TV*). The test set was split into two parts: one part used for development tests (**HUN-Dev**, 2 TV news, 50 minutes) and a second part for evaluation (**HUN-Eval**, 4 TV news, 105 minutes). The 125 minutes long Romanian test database dates from the same period of time as the Hungarian and consists of the recordings of 3 TV news. One of the 3 news was selected for development purposes (**ROM-Dev**, 42 minutes), while the other 2 was used for evaluation (**ROM-Eval**, 83 minutes).

During the transcription of the Hungarian test set not only the common acoustic events were noted but speech segments were categorized based on the acoustic conditions (see **Table 1**). Note if more categories occurred in a speech segment, the whole segment was classified into the highest category. As it can be seen in Table 1, the majority of segments were recorded in degraded acoustics conditions, which can be surprising in a broadcast news task. However, in our test database all parts of the news were kept even the reports made on the spot. The approximate signal-to-noise ratios (SNR) were determined by using NIST STNR[1] and WADA SNR [13] algorithms. In the case of F2 and F5 there were not enough data for reliable estimation of SNR.

2.2. Training text corpora

The textual training data were collected from two sources. On the one hand the manual transcriptions (**TRS**) of acoustic training data were utilized, while on the other hand additional texts were gathered from news site on the web. Only those websites were selected that used proper diacritics. More information about the collection, storage and processing of web-based news can be found in [12]. In order to separate the more recent and this way more relevant training data from the older ones, the web-based training corpora were split into two parts. The first part (**WEB 30D-**) contains all the articles from web-based news database, which are newer than 30 days (published between the 1st and 31st December, 2011), whereas the second part (**WEB 30D+**) consists of the articles, which are older than 30 days. For further details see **Table 2**.

3. METHODOLOGY

3.1. Acoustic modeling

The Hungarian speaker independent, cross-word triphone models were trained using decision trees and embedded Baum-Welch re-estimation with the HTK toolkit [14]. Word-to-phoneme mapping was obtained by applying simple grapheme-to-phoneme rules considering consonant assimilations of Hungarian. The final model had 4630 states and 15 Gaussians per state.

[1] http://labrosa.ee.columbia.edu/~dpwe/tmp/nist/doc/stnr.txt

There was no publicly available pronunciation dictionary for Romanian that would suit our task. However, Romanian has a largely phonemic orthography that enabled us to train data-driven, grapheme-based acoustic model called as "trigraphones" [15]. Grapheme-based acoustic models were trained similarly as the Hungarian phoneme-based models. Since in the current state of the research we wanted to avoid the usage of any language-specific knowledge we applied a simplistic singleton clustering technique [16] in decision tree constructions. Number of states in the model was 4672 and 15 Gaussians were used per state.

3.2. Language modeling

The parameter estimation and interpolation of language models were performed by using the SRI Language Modeling Toolkit (SRILM) [17]. The n-gram orders were optimized individually for every model by minimizing word error rate.

3.2.1. Morph segmentation

In the case of morphologically rich languages the usage of word-based language models (**WORD**) often results in very large vocabularies, high Out Of Vocabulary (OOV) rate, and inaccurate language model parameter estimation due to large number of distinct word forms. These issues can be handled by changing the base units from words to sub-word lexical units in the language model [4]. In our experiments the morphological segmentation was performed by applying an unsupervised vocabulary decomposition technique, called Morfessor Baseline [18]. As a result of the decomposition a "word-to-morph" dictionary is obtained, which can be used to replace a word by the corresponding morph sequence. The term "morph" stands for statistically derived sub-word lexical units.

3.2.2. Word boundary modeling

As morph-based recognition systems create morph sequences as an output, reconstruction of word boundaries is essential. Two approaches are compared in our paper. The first places word boundary (**WB**) symbols into the training text between each word and considered as a separate morph [19]. Word boundaries are reconstructed in the ASR output by merging every morph between WB symbols. According to the other approach only those morphs are tagged, that are not the initial morphs of a given word. This technique is called non-initial (**NI**) tagging [20]. Word reconstruction here is solved by merging non-initial morphs with the preceding morphs in the output. A sample training sentence for both approaches can be found in **Fig. 1**.

3.3. Decoding and evaluation

The knowledge sources were integrated into a triphone level Weighted Finite State Transducer (WFST) [21] recognition network. Standard MFCC feature extraction with Blind Equalization [22] was applied on all audio data, including first and second derivatives and energy, resulting in a total of 39 dimensional vectors. The one-pass recognition tests were performed using the WFST decoder called VOXerver [7] on a standard PC with a Core i7 processor at 3.5 GHz. The Real Time Factor (RTF) for a morph- and the corresponding word-based system were adjusted to be close to equal using standard pruning techniques. The performance of the broadcast news transcription system is characterized by Word Error Rate (WER) in all of our experiments, except for one case in section 4.4 where Letter Error Rate (LER) is given.

Word:
jó napot kívánok (Good afternoon)
NI tagging:
jó nap –ot kíván –ok
WB morph:
jó <WB> nap ot <WB> kíván ok

Figure 1. Sample sentence from the training corpus for demonstrating word boundary modeling

4. EXPERIMENTAL RESULTS

4.1. Morph-based improvement as a function of training corpus size

The aim of our first experiment was to investigate the relation between the morph-based improvement and the amount of text data used for training the language models. Previously we found that the benefit of morph-based lexical models was diminishing [5], [6] – until it completely disappeared [7] – as the training text size was being increased. However, in our former papers only the WB reconstruction technique was investigated.

In the current comparison 3 measuring point was applied. The first measurement was taken with models trained on the in-domain manual transcriptions (**TRS**). In the second case transcriptions were extended with WEB 30D- web corpora (**TRS+WEB 30**), whereas in the last case all available training data (TRS, WEB 30D-, WEB 30D+) were utilized (**TRS+WEB ALL**). Training data were integrated in a common language model by using linear interpolation. Interpolation weights were optimized on the development test sets.

If Hungarian (**Table 3**) and Romanian (**Table 4**) recognition results are compared, we find that WERs of the Hungarian system is much lower. It can be explained by the fact that notably more training data were available in Hungarian (see section 2). We can also get an impression about the difference in the morphological richness of the two languages, if we compare the size of the vocabularies. This difference implied that morph-based language modeling was only successful for Hungarian and in the case of Romanian degradations were measured (see **Fig. 3**). The

Table 3. Hungarian recognition results with various training corpus sizes and lexical modeling

Corpus	Lexical model	N-gram order	Vocab. size	Dev WER	Eval WER
TRS	WORD	3	74k	38.2%	41.4%
	WB	4	12k	35.2%	39.6%
	NI	4	16k	34.8%	38.3%
TRS+ WEB 30	WORD	3	151k	32.4%	36.1%
	WB	4	24k	31.1%	35.7%
	NI	4	31k	30.9%	34.2%
TRS+ WEB ALL	WORD	4	978k	23.0%	25.6%
	WB	5	175k	23.1%	25.6%
	NI	4	204k	22.3%	24.7%

Table 4. Romanian recognition results with various training corpus sizes and lexical modeling

Corpus	Lexical model	N-gram order	Vocab. size	Dev WER	Eval WER
TRS	WORD	3	27k	45.8%	45.1%
	WB	5	7k	48.6%	48.0%
	NI	4	9k	47.5%	47.2%
TRS+ WEB 30	WORD	4	80k	38.9%	39.6%
	WB	6	17k	41.6%	42.8%
	NI	5	21k	40.4%	41.4%
TRS+ WEB ALL	WORD	4	233k	34.6%	35.2%
	WB	6	42k	36.8%	38.1%
	NI	5	53k	36.2%	36.6%

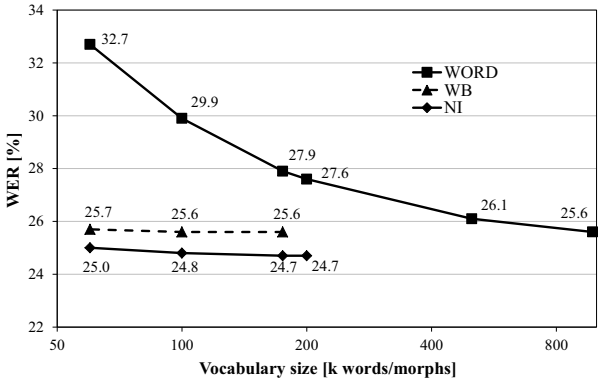

Figure 2. Hungarian recognition results as a function of vocabulary size

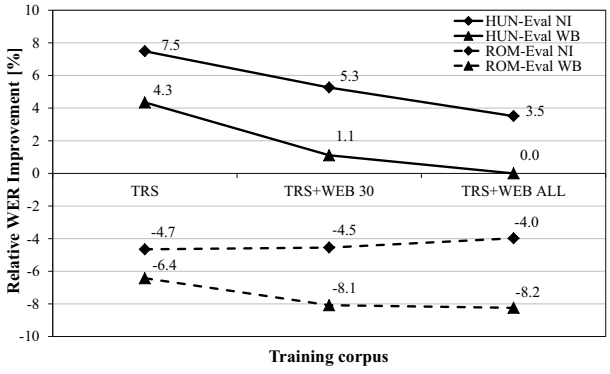

Figure 3. Morph-based WER improvement as a function of training corpus size

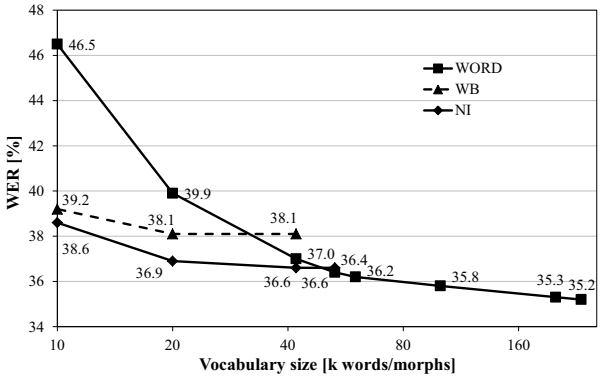

Figure 4. Romanian recognition results as a function of vocabulary size

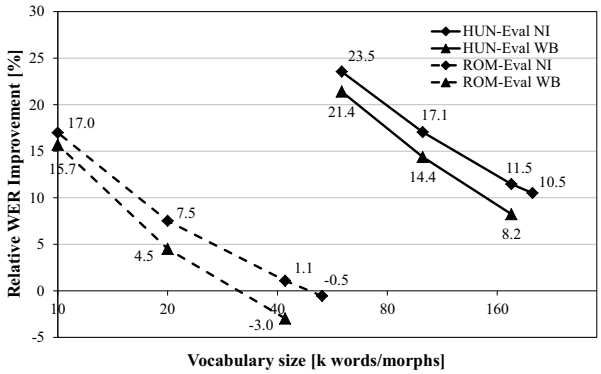

Figure 5. Morph-based WER improvement as a function of vocabulary size

results in Fig. 3 (apart from ROM-NI) are in accordance with our earlier finding, namely, the more textual training data is available; the lower is the morph-based improvement. Note that NI word boundary modeling technique outperformed WB at all measuring points.

4.2. Morph-based improvement as a function of vocabulary size

The results presented in the previous section were got by including all lexical units found in the training texts into the vocabulary. However, the utilizable vocabulary size is often restricted due to built-in limitations of the framework or resource efficiency reasons. Hence, we decided to investigate the effect of restricted vocabulary on morph-based improvement. We took the word- and morph-based language models trained on TRS+WEB ALL dataset and limited their vocabulary to 60, 100, 175, 204, *500* thousand of most frequent items for Hungarian (see **Fig. 2**), and 10, 20, 42, 53, *60*, *100*, *200* thousand items for Romanian (see **Fig. 4**). The vocabulary limited language models were created with *-limit-vocab* switch of SRILM tools. Note that for vocabulary sizes indicated with italic letters there are only word-based results available, since they exceed the full size of morph-based vocabularies.

As it can be observed in both Fig. 2 and 4, word-based language models are far more sensitive to the limitation of vocabulary then morph-based approaches. Consequently, the more limited is the vocabulary the higher is the morph-based improvement (see **Fig. 5**). For instance the morph-based models have around 20% lower WER than word-based models at 60k vocabulary size in the case of Hungarian. Although morph-based language models did not suit the Romanian task with full vocabulary (see section 4.1), if the vocabulary has to be restricted (e.g. built in limitations of modeling framework) morph-based modeling can be a good choice.

Table 5. Word error rates [%] measured according to the different acoustic categories of HUN-Dev+Eval

Corpus	Lexical model	F0	F1	F2	F3	F4	F5
TRS	WORD	33.7	42.1	76.0	46.3	42.3	55.4
	WB	29.2	40.2	67.4	42.9	39.8	54.7
	NI	29.6	41.0	67.1	43.0	39.5	52.2
TRS+ WEB 30	WORD	26.3	39.1	67.1	40.0	37.0	49.5
	WB	24.8	38.5	60.5	38.5	35.8	48.7
	NI	24.7	37.2	64.7	37.3	35.4	47.3
TRS+ WEB ALL	WORD	15.6	31.5	56.2	28.7	27.5	38.2
	WB	15.9	29.7	53.5	30.7	27.4	35.8
	NI	15.6	29.5	56.6	28.1	26.6	37.6

4.3. Morph-based improvement as a function of speech type and signal-to-noise ratio

As the Hungarian test set was labeled according to the acoustic conditions (see section 2.1), we also had the opportunity to investigate the effect of speech type and SNR on morph-based improvement (see **Table 5**). The best recognition results can be obtained on clean, planned speech (F0), where WER ranges between 16-30% depending on the size of training corpus. However, as the SNR degrades from 20-25dB (F0, F1) to 10-15dB (F3, F4), the average WER increases with around 10-15%. The worst results (56-76%) were got for telephone speech (F2), which can be explained by the mismatch in the bandwidth of the task and the acoustic model. Although the SNR of F0 and F1 tasks are roughly equal, the WER of F1 is 8-15% higher than of F0 due to the sloppy articulation and higher perplexity in spontaneous speech [5].

In the comparison of morph-based improvements F2 and F5 categories were not taken into consideration due to their small proportion in the test set (see Table 1). If there are only limited resources available for language modeling (TRS, TRS+WEB30 in **Fig. 6 and 7**), the highest morph-based improvement can be observed on F0 test category.

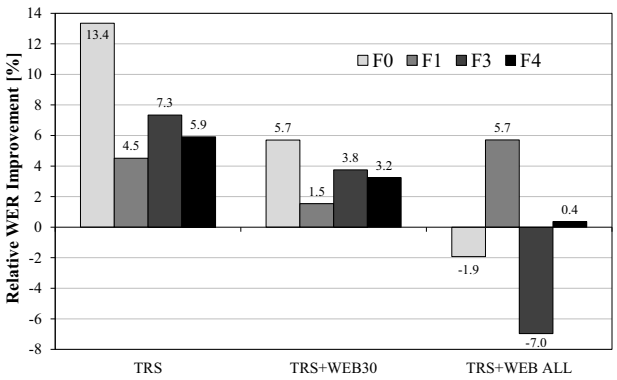

Figure 6. Morph-based improvement as a function of acoustic categories in the HUN test set if WB word boundary modeling is applied

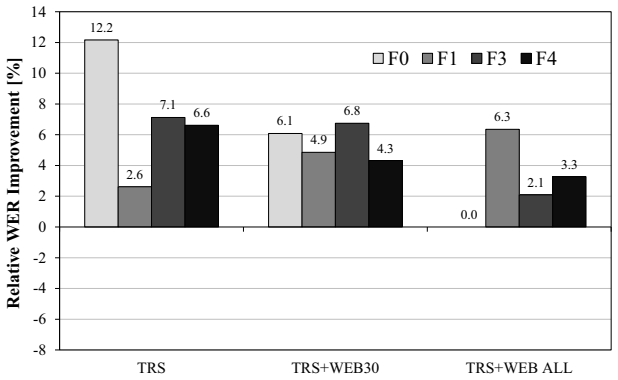

Figure 7. Morph-based improvement as a function of acoustic categories in the HUN test set if NI word boundary modeling is applied

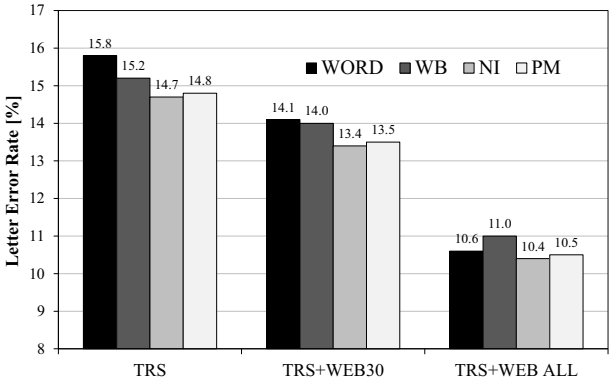

Figure 8. Letter error rates measured on HUN-Eval test set with word-based and various types of morph-based language models

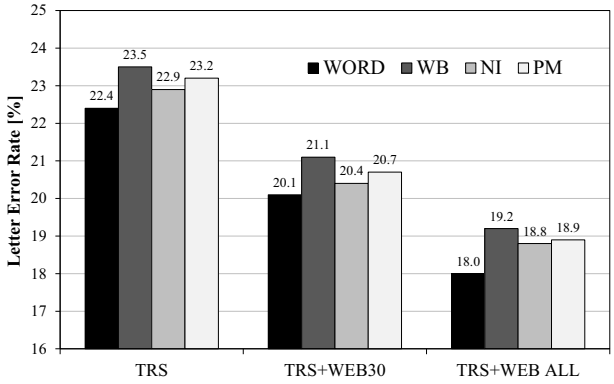

Figure 9. Letter error rates measured on ROM-Eval test set with word-based and various types of morph-based language models

This result is in accordance with our former findings [6] that clean acoustic conditions (good acoustical match) can raise the improvement rate owing to the smaller acoustic confusability of lexical elements. In degraded acoustic conditions (F3, F4) the WB modeling approach performs much worse than NI that can explain the difference in their overall performance. The results based on TRS+WEB ALL corpus show that if there is efficient amount of training texts are available, the advantage of morph-based models disappears on F0.

4.4. The effect of word boundary modeling

In the last experiment we had an interest in the recognition performance of a morph-based system without using any type of word boundary modeling. According to our hypothesis word boundary markings reduce the efficiency of morph-based language models, but we must use them for the reconstruction of word boundaries in output of the transcription system. Hence it follows that the comparison of word-based models, morph-based models with boundary markings (WB, NI) and the morph-based models without any marking (Pure Morph - **PM**) could only be done by using LER as a metric. As it can be seen on **Fig. 8 and 9,** the results refuted our prior assumption. Although the PM approach outperformed both the word- and WB morph-base models, it was not able to provide lower error rate than NI morph approach. Consequently, the additional information coded in the NI models by making difference between initial and non-initial morphs is useful for improving accuracy of language modeling.

5. PREDICTION OF MORPH-BASED IMPROVEMENT

Here we introduce two empirically derived predictors that can be used to estimate the expected morph-based improvement. These predictors are based on the characteristics of the training corpora; hence they are language-independent and can be calculated prior to morphological processing. In our former study [5] we showed the relation between morphological richness of a recognition task (average number of word types in training corpora) and the morph-based improvement. However, as it was presented in the previous section, the improvement has a connection with training corpus size (see Fig. 3) and vocabulary size (see Fig. 5), as well. This way our new predictors take these factors also into account.

For the calculation of the predictors first the training corpus (T) is split into n, disjoint parts (T_i) that contains k tokens. In our experiments k = 160000.

$$\bigcap_{i=1}^{n} T_i = \emptyset \qquad Tokens(T_i, 0 \le i \le n) = k$$

The empirical predictors are constructed to be in direct proportion to the morph-based improvement, thus in the numerator the average amount of word types are placed that relates to the morphological richness of the task. Note that this is our original predictor from [5]. This predictor is extended with the logarithm of training corpus (*predictor 1*) and vocabulary size (*predictor 2*) in the denominator with respect to the relations we found in this paper. Besides both predictors utilize tuning parameters a and b.

$$Empirical\ predictor\ 1 = \frac{\frac{1}{n}\sum_{i=1}^{n} Types(T_i)}{\log_{10} Tokens(T) + a}$$

$$Empirical\ predictor\ 2 = \frac{\frac{1}{n}\sum_{i=1}^{n} Types(T_i)}{\log_{10} Types(T) + b}$$

As it can be seen in **Fig. 10**, after tuning free parameter (a = 10), a very high correlation can be observed between predictor 1 and morph-based improvement measured with various training corpus sizes on the Hungarian and Romanian test set. On the other hand, in **Fig. 11** the vocabulary limited improvement rates are plotted as a

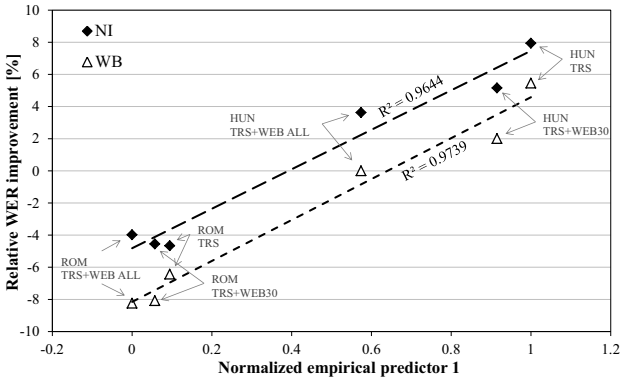

Figure 10. Relationship between the morph-based improvement and empirical predictor 1

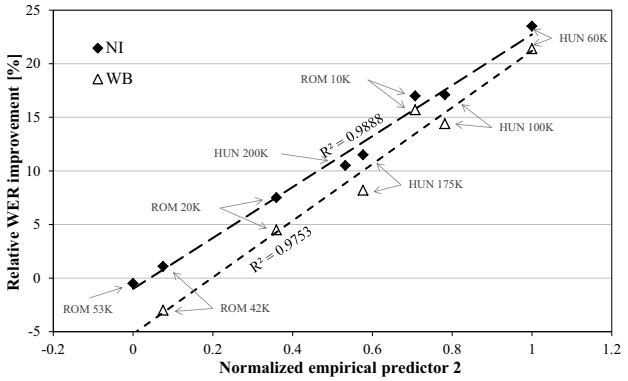

Figure 11. Relationship between the morph-based improvement and empirical predictor 2

function of predictor 2 (b = -2). The correlation here is also very high. In the future these features can be efficiently used to create a more complex model that is able to predict the morph-based improvement considering all the important factors at the same time.

6. CONCLUSIONS

In this paper, we made an attempt to reveal the factors affecting the improvement that can be achieved by using morph-based language model instead of word-based. Our experiments were carried out on a Hungarian and a Romanian broadcast news recognition task utilizing two types of word boundary reconstruction approach. The results showed that the more textual data is available the less morph-based improvement can be expected probably due to the reduction of data sparseness.

The relation between the vocabulary size and the morph-based improvement was also investigated. It was found that the morph-based language models much less sensitive for the limitation of vocabulary than word-based models. This way even higher improvement (~20%) can be measured for a morphologically rich language (e.g. Hungarian) and degradation can be turned to improvement for a morphologically less rich language (e.g. Romanian). Therefore morph-based language models can be especially beneficial for the processing of under-resourced languages, where small training corpus and vocabulary sizes are very common.

The high improvement rate observed on the clean, planned part of Hungarian test set indicate that reduction of acoustic ambiguity of lexical units can further improve the efficiency of morph-based models. The applied word boundary reconstruction method turned out to be another important factor, as non-initial tagging outperformed word boundary morphs in every experimental setup. This difference can be probably assigned to the fact that NI approach differentiates initial and non-initial morphs.

Building morph-based speech recognition systems entails extra work compared to word-based systems, thus it would be helpful to know the expected improvement. Therefore in the last section of our paper, we presented empirical predictors based on language independent parameters. Currently these predictors take morphological richness, training corpus size and vocabulary size into account; however, we plan to extend it with a metric for the match between the acoustic model and the recognition task. In the future, we would like to create a model that is able to predict the morph-based improvement considering all the important factors and evaluate it on an independent recognition task.

7. ACKNOWLEDGEMENTS

Our research was partially funded by the TÁMOP-4.2.2.C-11/1/KONV-2012-0013 (FuturICT.hu), GOP-1.1.1-11-2012-0377 (WEBRA TIME SAVE), PIAC_13-1-2013-0234 (Patimedia), KMR_12-1-2012-0207 (DIANA), AAL-08-1-2011-0001 (PAELIFE) projects.

8. REFERENCES

[1] N. Morgan, J. Cohen, S. H. Krishnan, S. Chang, and S. Wegmann, “Final Report : OUCH Project (Outing Unfortunate Characteristics of HMMs),” 2013.

[2] D. Trandabăţ, E. Irimia, V. Barbu Mititelu, D. Cristea, and D. Tufiş, *Limba română în era digitală -- The Romanian Language in the Digital Age*. Springer, 2012.

[3] E. Simon, P. Lendvai, G. Németh, G. Olaszy, and K. Vicsi, *A magyar nyelv a digitális korban -- The Hungarian Language in the Digital Age*. Springer, 2012.

[4] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, “Automatic speech recognition for under-resourced languages: A survey,” *Speech Commun.*, vol. 56, pp. 85–100, Jan. 2014.

[5] B. Tarján and P. Mihajlik, “On morph-based LVCSR improvements,” in *Spoken Language Technologies for Under-Resourced Languages (SLTU-2010)*, 2010, pp. 10–16.

[6] L. Tóth, B. Tarján, G. Sárosi, and P. Mihajlik, "Speech Recognition Experiments with Audiobooks," *Acta Cybern.*, vol. 19, no. 4, pp. 695–713, 2010.

[7] B. Tarján, P. Mihajlik, A. Balog, and T. Fegyó, "Evaluation of lexical models for Hungarian Broadcast speech transcription and spoken term detection," in *2nd International Conference on Cognitive Infocommunications (CogInfoCom)*, 2011, pp. 1–5.

[8] L. Tóth and T. Grósz, "A Comparison of Deep Neural Network Training Methods for Large Vocabulary Speech Recognition," in *Text, Speech, and Dialogue*, vol. 8082, 2013, pp. 36–43.

[9] A. Roy, L. Lamel, T. Fraga, J. Gauvain, and I. Oparin, "Some Issues affecting the Transcription of Hungarian Broadcast Audio," in *14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, 2013, no. August, pp. 3102–3106.

[10] H. Cucu, L. Besacier, C. Burileanu, and A. Buzo, "Enhancing automatic speech recognition for romanian by using machine translated and Web-based text corpora," in *Proc. of SPECOM 2011*, 2011, pp. 81–88.

[11] H. Cucu, L. Besacier, C. Burileanu, and A. Buzo, "Investigating the role of machine translated text in ASR domain adaptation: Unsupervised and semi-supervised methods," *2011 IEEE Work. Autom. Speech Recognit. Underst.*, pp. 260–265, Dec. 2011.

[12] B. Tarján, T. Mozsolics, A. Balog, D. Halmos, T. Fegyó, and P. Mihajlik, "Broadcast news transcription in Central-East European languages," in *IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 2012, pp. 59–64.

[13] C. Kim and R. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *INTERSPEECH*, 2008, pp. 2598–2601.

[14] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The {HTK} Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[15] S. Kanthak and H. Ney, "Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2002, pp. I–845–I–848.

[16] M. Killer, S. Stüker, and T. Schultz, "Grapheme based speech recognition," in *Proceeding of the Eurospeech*, 2003, pp. 3141–3144.

[17] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proceedings International Conference on Spoken Language Processing*, 2002, pp. 901–904.

[18] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0," in *Publications in Computer and Information Science, Report A81*, 2005.

[19] T. Hirsimäki and M. Kurimo, "Decoder issues in unlimited Finnish speech recognition," in *Proc. of the 6th Nordic Signal Processing Symposium (Norsig)*, 2004.

[20] E. Arisoy, D. Can, S. Parlak, H. Sak, and M. Saraclar, "Turkish Broadcast News Transcription and Retrieval," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 17, no. 5, pp. 874–883, Jul. 2009.

[21] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Comput. Speech Lang.*, vol. 16, no. 1, pp. 69–88, 2002.

[22] L. Mauuary, "Blind equalization for robust telephone based speech recognition," in *Proc. European Signal Processing Conference*, 1996, pp. 359–363.