USING MORPHEME AND SYLLABLE BASED SUB-WORDS FOR POLISH LVCSR

M. Ali Basha Shaik, Amr El-Desoky Mousa, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition – Computer Science Department
RWTH Aachen University, 52056 Aachen, Germany

ABSTRACT

Polish is a synthetic language with a high morpheme-per-word ratio. It makes use of a high degree of inflection leading to high out-of-vocabulary (OOV) rates, and high Language Model (LM) perplexities. This poses a challenge for Large Vocabulary and Continuous Speech Recognition (LVCSR) systems. Here, the use of morpheme and syllable based units is investigated for building sub-lexical LMs. A different type of sub-lexical units is proposed based on combining morphemic or syllabic units with corresponding pronunciations. Thereby, a set of grapheme-phoneme pairs called *graphones* are used for building LMs. A relative reduction of 3.5% in Word Error Rate (WER) is obtained with respect to a traditional system based on full-words.

Index Terms— language model, morpheme, syllable, graphone, Polish

1. INTRODUCTION

Polish is considered as one of the morphologically rich languages. It belongs to the family of Slavic languages like Russian, Czech, and Bulgarian. Polish is characterized by a high degree of inflection, having seven cases and three genders. Declensional endings depend on case, number, gender and animacy. In addition, declension changes if the word is noun or adjective. Moreover, word stems are frequently modified by the addition or absence of endings. This provides huge lexical variety that causes data sparsity and leads to high OOV rates and high LM perplexities. Normally, traditional Polish LVCSR systems use a large recognition lexicon having several hundred thousands of full-words [1]. However, still relatively high OOV rates are obtained. On the other side, the ASR system suffers from high resource requirements. Therefore, sub-words are used instead of full-words in order to reduce the lexical variety. Normally, the number of possible sub-words in a corpus is smaller than that of full-words, giving higher average frequency. This helps to reduce OOV rates and limit the recognition search space.

A possible type of sub-word is the *morpheme* which is the smallest linguistic component of the word that has a semantic meaning. For Slavic languages, morpheme based LMs are proposed [2, 3]. They are based on decomposing words into stems and endings. Moreover, morpheme based LMs are used for other languages as German [4] and Arabic [5].

Another type of sub-word is the *syllable* which is considered as a phonological building block of words. A syllable is usually made up of a nuclear vowel with optional initial and final consonants [6]. Syllable based LMs are successfully used for languages like Chinese [7]. In [8] a syllable based LM is proposed for Polish, where both OOV rate and LM perplexity are reduced but no WERs are provided.

A different approach is to combine the graphemic sub-words with their corresponding pronunciations. This allows different context dependent pronunciations of sub-words to be captured on the level of the LM rather than the lexicon level. In [9], a set of automatically derived morphemes joint with pronunciations augments a normal word model and used for an English LVCSR task. In [10, 4] a set of combined grapheme-phoneme pairs (called graphones) are used. Only OOV words are replaced by sequences of such pairs. Graphones are automatically derived from grapheme-to-phoneme (G2P) conversion [11]. In those experiments, the grapheme part is just a sequence of letters with some length constraints, but without a linguistic relationship. Moreover, those models are mainly used to cope with OOV words, but no attempt is made to use graphones for in-vocabulary words.

In this work, the use of sub-lexical LMs for LVCSR of Polish is investigated. Different types of sub-words, namely morphemes and syllables are compared. In addition, morphemes and syllables are combined with their pronunciations. The resulting graphones are used to partially model in-vocabulary words.

2. METHODOLOGY

2.1. Morpheme based sub-words

We perform Polish morphological decomposition using a statistical tool called *Morfessor* [12]. It is a data driven tool that autonomously discovers the optimum decomposition for the words in unannotated text corpora based on the Minimum Description Length (MDL) principle. Moreover, it is a general model for unsupervised induction of morphology from raw text. It is designed to cope with languages with rich morphology, where the number of morphemes per word is varying so much and not known in advance [13]. Although Morfessor is successfully used for various languages, its application to Polish is not sufficiently investigated.

We train our decomposition model using a vocabulary of unique words that occur more than 5 times in the training data; this gives about 1.1 Million words. We do not include less frequent words in order to avoid irregular words which are harmful to the training process. Nevertheless, the trained model can be used to decompose unseen words. In addition, the resulting decompositions are adapted so as to avoid very short morphemes which are usually difficult to recognize. This is found helpful to improve the final WER.

2.2. Syllable based sub-words

The general structure of a Polish syllable consists of the onset, nucleus and coda. The nucleus is usually a vowel sound. while the onset and coda are usually zero or more consonants. The onset is the sound occurring before the nucleus, and the coda is the sound following the nucleus. A syllable without a coda is called a free syllable, while a syllable with a coda is called a closed syllable [14]. We perform decomposition into syllables (syllabification) of Polish words using a phonological rule based tool called *KombiKor v.8.0* [15]. For the same reasons as in case of morphemes, we adapt the syllabification output so as to avoid very short syllables.

2.3. Combining sub-words with pronunciations

For words whose pronunciations are unknown, we use a statistical G2P approach to get the missing pronunciations. Our approach is based on joint-sequence models as shown in [11]. Therein, the aim is to find the most likely pronunciation $\varphi \in \Phi^*$ for a given orthographic form $g \in G^*$, where Φ and G are the sets of phonemes and letters respectively:

$$\varphi(g) = \arg\max_{\acute{\varphi} \in \Phi^*} p(\acute{\varphi}, g) \tag{1}$$

We refer to the joint probability distribution $p(\varphi, g)$ as a "graphonemic" joint sequence model. We assume that for each word, its orthographic form and its pronunciation are generated by a common sequence of graphonemic units called *graphones*. Each graphone is a pair $q = (g, \varphi) \in Q \subseteq G^* \times \Phi^*$ of a letter sequence and a phoneme sequence of possibly different lengths. The joint probability distribution $p(\varphi, g)$ is reduced to a probability distribution over graphone sequences $p(q)$ which are modeled by a standard M-gram:

$$p(q_1^N) = \prod_{i=1}^{N+1} p(q_i | q_{i-1}, ..., q_{i-M+1}) \tag{2}$$

This model has two parameters: the order of the M-gram model, and the allowed size of graphones. The number of letters and phonemes are allowed to vary between zero and an upper limit L. Such a model can be trained using Maximum Likelihood (ML) training via the Expectation Maximization (EM) algorithm as presented in [11]. To produce a pronunciation for a given word, we use the maximum approximation over the set $S(g, \varphi)$ of all joint segmentations of g and φ:

$$p(\varphi, g) \approx \max_{q \in S(g,\varphi)} p(q_1, ..., q_L) \tag{3}$$

In the above model, the inventory of graphones Q is automatically inferred from the training data. The letter and phoneme sequences are grouped into an equal number of segments. The number of letters in each segment depends on the the parameter L. Normally, we choose the optimum L which gives the minimum Phoneme Error Rate (PER) over some test dictionary. This guarantees the best possible pronunciations for some letter sequences. Here, it is worth noting that the letter sequences are not representing any type of linguistic units; rather, they are groups of letters of almost fixed length.

The set of graphones inferred during G2P training constitutes a graphone model that can be integrated with the normal word model. Thus, it is possible to combine lexical vocabulary entries with sub-lexical graphones derived from G2P conversion to form a unified set of recognition units. In our experiments, we estimate a normal graphone model as described above, and then we adapt the letter sequences such that they represent morphemes or syllables of the underlying words. To adapt the initial graphones, we need to do letter-phoneme alignment. For that, we follow the same approach described in [16]. The adapted graphones replace some chosen subset of words of the original full-words vocabulary. The LM training data is re-written accordingly such that it contains full-words with interspersed sequences of graphones.

2.4. Experimental considerations

As shown in our earlier work [5], it is better for sub-word based LMs to not decompose the N most frequent decomposable full-words (words with more than one morpheme or syllable). This prevents those most important words from being mixed-up with other sub-words in the search space. Here, we optimize the value of N over the development corpus. For easy recovery to full-words in the recognition output, we attach a '+' sign to the end of non-boundary sub-words. An example is the word *'niejednokrotnie'* which is decomposed into *'nie+ jedno+ krotnie'*. On the other side, we compute the OOV rate of any corpus such that a word is considered an OOV if and only if it is not found in the vocabulary and it is not possible to compose it using vocabulary sub-words.

3. EXPERIMENTAL SETUP

The basis of the acoustic model is the cross-language unsupervised trained acoustic model described in [17]. This model is originally trained on about 128 hours of untranscribed recordings from the European Parliament Plenary Sessions (EPPS).

Our LM training corpora consist of around 630 Million running full-words including data from EPPS, Kurier Lubelski, Nowosci, in addition to official data provided for the Quaero project (mainly news and blogs). The text corpora are used for vocabulary selection (N most frequent words) and to estimate back-off N-gram LMs by the SRILM toolkit [18].

Our speech recognizer works in 3 passes. In the first pass, across-word acoustic models are used with no speaker adaptation. The second pass applies speaker adaptation based on

Constrained Maximum Likelihood Linear Regression (CMLLR). The third pass adds Maximum Likelihood Linear Regression (MLLR) adaptation. In each pass, either a 3-gram or 5-gram LM is used to construct the search space.

To evaluate the recognition performance, we use the Quaero 2010 development and evaluation corpora (dev10: 3.2h; eval10: 3.5h). Each corpus consists of audio material from Broadcast News (BN) and podcast sources.

4. EXPERIMENTS

4.1. Baseline recognition

In Table 1, we summarize the results of our baseline recognition experiments using traditional LMs based on full-words.

Table 1. *Baseline word error rates [%] using 5-gram LMs based on full-words (voc: vocabulary).*

voc size	Dev10		Eval10	
	OOV [%]	**WER** [%]	**OOV** [%]	**WER** [%]
300k	1.70	22.70	1.88	26.84
500k	1.08	22.13	1.21	25.63

4.2. Morpheme and syllable based LMs

In Table 2, we summarize the results of our recognition experiments using morpheme based LMs. The vocabulary size is fixed to 300k. The LM is either 3-gram or 5-gram. We optimize the number of full-words over the dev10 corpus (see Section 2.4). We get the best results using a vocabulary of *70k full-words + 230k morphemes* and a 5-gram LM. We achieve WER reductions of 2.31% relative (0.62% absolute) for the eval10 corpus compared to the 300k baseline in Table 1. We got no improvement for the dev10 corpus.

Table 2. *Word error rates [%] using morpheme based LMs (mrfs: morphemes, wrds: words).*

corpus	#full wrds	# mrfs	OOV [%]	WER [%] 3-gram	WER [%] 5-gram
Dev10	30k	270k	1.47	23.05	-
	50k	250k	1.51	22.77	-
	70k	230k	1.57	**22.73**	**22.71**
	90k	210k	1.67	22.82	-
	100k	200k	1.69	22.85	-
Eval10	70k	230k	1.77	**28.95**	**26.22**

In Table 3, we record the recognition results using syllable based LMs. Similar to Table 2, we use 300k vocabularies, and either 3-gram or 5-gram LM. We got the best results using a vocabulary of *130k full-words + 170k syllables*. We achieve WER reductions of [dev10: 1.19% relative (0.27% absolute); eval10: 2.76% relative (0.74% absolute)] compared to the 300k baseline in Table 1.

4.3. Morphemic and syllabic graphone based LMs

In Table 4, we record the recognition results using morphemic and syllabic graphone based LMs as described in Section 2.3.

Table 3. *Word error rates [%] using syllable based LMs (slbs: syllables).*

corpus	#full wrds	# slbs	OOV [%]	WER [%] 3-gram	WER [%] 5-gram
Dev10	50k	250k	0.50	23.09	-
	70k	230k	0.52	22.67	-
	90k	210k	0.56	22.65	-
	110k	190k	0.59	22.47	-
	130k	170k	0.62	**22.43**	**22.33**
	150k	150k	0.73	22.50	-
Eval10	130k	170k	0.82	**26.32**	**26.10**

The initial graphone model is based on $L = 2$; this achieves a PER of 0.22% on a held-out test dictionary. The graphone model is further adapted so that letter sequences represent morphemes or syllables. The number of decomposable full-words retained without decomposition is now fixed after the optimization performed in Tables 2 and 3. It is worth noting that exactly the same morphemes or syllables which give the best results before are extended here into graphones by considering context dependent pronunciations. This increases the overall vocabulary size, but retains the same OOV rate as in the case of normal morphemes or syllables. Finally, we can see that morphemic graphones perform better than syllabic graphones. We can achieve WER reductions of [dev10: 2.38% relative (0.54% absolute); eval10: 3.54% relative (0.95% absolute)] compared to the 300k baseline in Table 1.

Table 4. *Word error rates [%] using morphemic and syllabic graphone based 5-gram LMs (grfs: graphones).*

grfs type	#full wrds	# grfs	Dev10 OOV [%]	Dev10 WER [%]	Eval10 OOV [%]	Eval10 WER [%]
mrfs	70k	277k	1.57	**22.16**	1.77	**25.89**
slbs	130k	173k	0.62	22.75	0.82	26.57

4.4. Impact on OOV words

In Table 5, we record the percentage of correctly recognized OOVs (with respect to the 300k full-words vocabulary) for the eval10 corpus.

Table 5. *Effect of sub-word based LMs on OOV recognition.*

sub-word type	OOV recognition accuracy [%]
mrfs	4.7
slbs	26.7
graphone mrfs	15.3
graphone slbs	24.5

5. CONCLUSIONS

We investigated four types of sub-words for Polish LMs, namely morphemes, syllables in addition to morphemic and

syllabic graphones. We achieved the best results using morphemic graphones with a vocabulary of *70k full-words + 277k graphones*. This gives improvements in WER of [dev10: 2.38% relative (0.54% absolute); eval10: 3.54% relative (0.95% absolute)] over a 300k full-words baseline. Moreover, these WERs are comparable to the 500k full-words baseline. While normal syllables outperformed normal morphemes, the morpheme based graphones achieved the best results at the end. Given that the average morpheme length is around 6 letters while the average syllable length is around 4 letters, we see that the pronunciation variance becomes less in the case of syllables due to short lengths. This is clear from the total number of syllabic graphones (170k syllables give only 173k graphones). This normally leads to less powerful graphones resulting in increased WERs compared to the normal syllables. On contrary, adding variant pronunciations to morphemes creates powerful graphones that capture context dependent pronunciations leading to decreased WERs.

6. ACKNOWLEDGEMENTS

This work was partly funded by the European Community's 7^{th} Framework Programme under the project SCALE (FP7-213850), and partly realized under the Quaero Programme, funded by OSEO, French State agency for innovation.

7. REFERENCES

[1] D. Falavigna, D. Giuliani, R. Gretter, J. Lööf, C. Gollan, R. Schlüter, and H. Ney, "Automatic transcription of courtroom recordings in the JUMAS project," in *2nd International Conference on ICT Solutions for Justice*, Skopje, Macedonia, Sept. 2009, pp. 65 – 72.

[2] W. Byrne, J. Hajič, P. Ircing, P. Krbec, and J. Psutka, "Morpheme based language models for speech recognition of Czech," in *Text, Speech and Dialogue*, vol. 1902 of *Lecture Notes in Computer Science*, pp. 139 –162. 2000.

[3] T. Rotovnik, M. S. Maučec, and Z. Kačič, "Large vocabulary continuous speech recognition of an inflected language using stems and endings," *Speech Communication*, vol. 49, no. 6, pp. 537 – 452, June 2007.

[4] A. El-Desoky, M. Shaik, R. Schlüter, and H. Ney, "Sublexical language models for German LVCSR," in *IEEE Workshop on Spoken Language Technology*, Berkeley, CA, USA, Dec. 2010, pp. 159 – 164.

[5] A. El-Desoky, C. Gollan, D. Rybach, R. Schlüter, and H. Ney, "Investigating the use of morphological decomposition and diacritization for improving Arabic LVCSR," in *Interspeech*, Brighton, UK, Sept. 2009, pp. 2679 – 2682.

[6] J. Rubach and G. Booij, "Syllable structure assignment in Polish," *Phonology*, vol. 7, pp. 121 – 158, Oct. 1990.

[7] B. Xu, B. Ma, S. Zhang, F. Qu, and T. Huang, "Speaker-independent dictation of Chinese speech with 32K vocabulary," Philadelphia, PA , USA, Oct. 1996, vol. 4, pp. 2320 – 2323.

[8] M. Piotr, "Syllable based language model for large vocabulary continuous speech recognition of polish," in *Text, Speech and Dialogue*, vol. 5246 of *Lecture Notes in Computer Science*, pp. 397 – 401. 2008.

[9] L. Galescu, "Recognition of out-of-vocabulary words with sub-lexical language models," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 249 – 252.

[10] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models," in *Interspeech*, Lisbon, Portugal, Sept. 2005, pp. 725 – 728.

[11] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434 – 451, May 2008.

[12] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0," Tech. Rep., Computer and Information Science Helsinki University of Technology, Finland, Mar. 2005.

[13] M. Creutz, *Induction of the morphology of natural language: Unsupervised morpheme segmentation with application to automatic speech recognition*, Ph.D. thesis, Helsinki University of Technology, Finland, 2006.

[14] E. Czaykowska-Higgins and C. Y. Bethin, "Polish syllables: the role of prosody in phonology and morphology," *Phonology*, vol. 11, no. 2, pp. 354 – 360, 1994.

[15] "Kombikor v.8.0, 3n company," `http://www.3n.com.pl/kombi.php`.

[16] R. I. Damper, Y. Marchand, J. D. Marsters, and A. Bazin, "Aligning letters and phonemes for speech synthesis," in *5th ISCA Speech Synthesis Workshop*, Pittsburg, PA, USA, June 2004, pp. 209 – 214.

[17] J. Lööf, C. Gollan, and H. Ney, "Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a Polish speech recognition system," in *Interspeech*, Brighton, UK, Sept. 2009, pp. 88 – 91.

[18] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. Int. Conf. on Spoken Language Processing*, Denver, Colorado, USA, Sept. 2002, vol. 2, pp. 901 – 904.