Investigation of Maximum Entropy Hybrid Language Models for Open Vocabulary German and 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
{shaik,desoky,schlueter,ney}@cs.rwth-aachen.de

Abstract

For languages like German and Polish, higher numbers of word inflections lead to high out-of-vocabulary (OOV) rates and high language model (LM) perplexities. Thus, one of the main challenges in large vocabulary continuous speech recognition (LVCSR) is recognizing an open vocabulary. In this paper, we investigate the use of mixed type of sub-word units in the same recognition lexicon. Namely, morphemic or syllabic units combined with pronunciations called graphones, normal graphemic morphemes or syllables, along with full-words. In addition, we investigate the suitability of hybrid mixed-unit N-grams as features for Maximum Entropy LM along with adaptation. We achieve significant improvements in recognizing OOVs and word error rate reductions for German and Polish LVCSR compared to the conventional full-word approach and state-of-the-art N-gram mixed type hybrid LM.

Index Terms: open vocabulary, maximum entropy

1. Introduction

German and Polish are morphologically rich languages. German has a high degree of inflection, derivation and compounding leading to high lexical variety of words. Polish is characterized by a high degree of inflection, conjugation and numerous declensional endings. This leads to data sparseness, high OOV rates and high LM perplexities. Moreover, OOVs lead to neighboring word errors during recognition. Most of the modern LVCSR systems operate with a hybrid vocabulary containing full-words and sub-word units. The sub-words can be properly combined to produce a wide range of words achieving better lexical coverage using hybrid language models (HLM).

One of the main problems in HLM is the proper choice of sub-words. A possible type of sub-word is the *morpheme* which is the smallest linguistic component of the word that has a semantic meaning. Morphemes could be derived using a data-driven approach [1] based on the Minimum Description Length principle (MDL). The other type of sub-word is the *syllable* [2], a phonological building block of words. Syllables are normally derived using a set of rules. A different type of sub-word unit is a *graphone* [3], where a graphemic sub-word unit is augmented with its context dependent pronunciation.

On the other hand, the state-of-the-art Maximum Entropy (MaxEnt) LM techniques are acquiring more attention for being successful in LVCSR [4, 5, 6]. MaxEnt LMs provide the flexibility to incorporate various features, but consume many resources depending on vocabulary size. For LVCSR, in general, the LM data is obtained from multiple domains like Broadcast News (BN), Podcasts and Web data. It is often unrealistic to significantly reduce word error rate (WER) without adapting the LM to in-domain data [7].

In the literature, for an open vocabulary LVCSR system N-gram HLMs are successfully applied for various languages using different sub-words. For example, morphemic LMs are used for German [8] and Arabic [9]. Syllable based LMs are applied for Chinese [10] and Polish [11]. Graphones are used for languages like English [3], German [8] and Polish [12]. To the knowledge of the authors, MaxEnt techniques and adaptation have not been applied to hybrid LMs, yet.

In our previous work, we applied mixed types of sub-word units HLM to German LVCSR [13] to recognize OOVs. But, similar to N-gram backoff LM, N-gram backoff HLM lacks the ability to incorporate the features during language modeling. For mixed types of word forms, as there are full-words, graphemic sub-words and sub-word graphones, it is complex to extract suitable features as the combinatorial complexity of possible choices of features is high. For example, not all the morphemes obtained from a data-driven segmentation for German are meaningful. Similarly, a syllable itself is a representative phonological feature of the word. For graphones, the pronunciation itself is a feature to its context dependent sub-word. Hence, we assume mixed sub-word units themselves as features for creating N-gram MaxEnt HLMs. The goals of this work are described as follows:

- For Polish, we investigate the use of mixed-unit HLMs to recognize OOVs
- For German and Polish, we investigate:
  - MaxEnt HLMs using higher order N-grams as features for mixed sub-word units

- MaxEnt HLM vs. N-gram HLM
- Adapted HLM vs. MaxEnt HLM

2. Methodology

2.1. Grapheme-to-Phoneme Training

A *graphone* is a sub-word derived from augmenting a grapheme with its corresponding phoneme sequence. To generate sub-word graphones, we train a statistical joint-sequence grapheme-to-phoneme (G2P) model [14]. We seek 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 as:

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

A graphone is represented as a pair $q = (g, \varphi) \in Q \subseteq G^* \times \Phi^*$. The set of co-segmentations of g and φ is represented by $S(g,\varphi)$. The $p(\varphi, g)$ is reduced to a probability distribution over graphone sequences $p(q)$ as:

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

If the number of letters and phonemes ranges from zero to an upper limit *'L'*, the M-gram model is trained using Maximum Likelihood (ML) training using the Expectation Maximization (EM) algorithm as :

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

2.2. Mixed-unit Sub-words

Alternatively, we decompose the words using Morfessor [1]. For both languages, we train word decomposition model using unique words that occur more than 5 times in the LM training corpora. We exclude low frequency words to avoid noise (mis-spelled words e.t.c.,) that are harmful during training. The model is capable of decomposing unseen words. We also perform syllabification of words using a phonological rule based tool [15]. The decomposed words are post-processed to produce a cleaner set of sub-words and to avoid very short sub-words which are usually difficult to recognize.

As described in our previous work for German LVCSR [13], we repeat similar steps for Polish LVCSR to obtain mixed-units. We choose a full-word vocabulary size of 300k. To create a conventional morpheme/syllable HLM, the top N frequent words are preserved as full-words, and the remainder is decomposed into sub-words. As described in Section 2.1, we use the trained G2P model, to generate the pronunciations for the full-words (except for the the top N frequent words). Then, we generate the sub-word (morphemes/syllable) graphones by aligning the full-word pronunciation to its graphemic sub-word sequence. For alignment, we use Dynamic Programming and Expectation Maximization algorithm as described in [16]. To use mixed-units, we implement a threefold approach, i.e., the value of N is optimized for each type of sub-word (morphemes: N=70k ; syllables: N=130k) over the development corpus to obtain the best WER. We also use M graphones (M=100k), where the graphemic component of the graphone is a morpheme/syllable. To compare this threefold approach to the conventional sub-word HLM, we replace graphones with normal morphemes/syllables.

2.3. MaxEnt HLMs

Words can be represented by either full-words, or grapheme or graphone based sub-words. Elements from the combined set U of all full-words and sub-words will be denoted as *unified words*. We use unified words as N-gram features to create MaxEnt HLMs. If u is an unified word, $f(.)$ is the feature function, λ is the optimal weight, h is the context, $Z(h)$ is the normalization factor for all observed contexts then the MaxEnt model is given by Equation (4). After estimating the optimal weights λ_i, the MaxEnt model is smoothed using Gaussian priors.

$$p_{me}(u|h) = \frac{e^{\sum_i \lambda_i f_i(u,h)}}{Z(h)} \tag{4}$$

$$Z(h) = \sum_{u_i \epsilon U} e^{\sum_j \lambda_j f_j(u_i,h)} \tag{5}$$

2.4. Adaptation

We perform *Maximum a-posteriori* (MAP) adaptation using Gaussian priors over trained MaxEnt models. The MaxEnt model is trained on background data including the features of in-domain data. The prior parameters computed from background data are used to learn parameters from in-domain data. During MaxEnt training as described in Section 2.3, the prior has zero mean during Gaussian prior smoothing. But during adaptation, the prior distribution is centered at the background data parameters. The regularized log-likelihood of the adaptation training data is maximized during adaptation. In our experiments as an in-domain data, we investigate the use of development data as a type of a supervised adaptation. Alternatively, we also use the transcriptions from a first recognition pass as a type of unsupervised adaptation. We created MaxEnt and adapted models using open-source MaxEnt SRILM-extension [6].

2.5. Interpolation

In general, N-gram LMs are known to perform better in capturing short range context dependencies. If u is a unified word, then we effectively preserve the advantages of N-gram backoff HLM $p_{bo}(u)$ and MaxEnt HLM $p_{me}(u)$ by linear interpolation as shown in Equation (6). λ is optimized over development corpora.

$$p(u) = \lambda p_{bo}(u) + (1 - \lambda) p_{me}(u) \tag{6}$$

3. Experimental Setup

For German (2-pass) and Polish LVCSR (3-pass), we use the acoustic models, LMs and recognition setup using Constrained Maximum Likelihood Linear Regression (CMLLR), and Maximum Likelihood Linear Regression

(MLLR) speaker adaptation as described in [13] and [12] respectively. The German LM training corpora (Broadcast News (BN)) consist of around 188 Million running full-words. The Polish LM training corpora (BN, web, podcasts) consist of around 658 Million running full-words. For German and Polish LVCSR, we construct a 4-gram and 5-gram LM (full-word/hybrid) respectively using modified Kneser-Ney smoothing. In the recognition setup, after the final pass the N-best (N=500) lists are generated for MaxEnt HLM rescoring and adaptation experiments. Due to huge training corpora, we created multiple partitions to create MaxEnt models followed by linear interpolation to merge LMs. In our experiments we compute *effective OOV rate* by considering a word is an OOV if it is not found in the vocabulary and it is not possible to compose it using in-vocabulary sub-words. For our experiments, we use the development and evaluation corpora from German:Quaero-2009 (dev09: 7.5h; eval09: 3.8h) and Polish:Quaero-2010 (dev10: 3.2h; eval10: 3.5h) system. Both corpora consist of audio from BN, web and podcast sources.

4. Experimental Results

4.1. Mixed-unit HLM

In Table 1, we record the Polish LVCSR recognition WERs using morphemic/syllabic HLMs and mixed-unit HLMs. The initial part of the vocabulary consists of full-words (70k in the case of morphemes, and 130k in the case of syllables). This is obtained after a series of optimization experiments not shown in this paper. Systems m1, m2, and s1, s2 are two-fold morpheme or syllable HLMs respectively containing only full-words and single type sub-words (morphemes or syllables or graphones). Systems m3 and s3 are mixed-unit type morpheme or syllable HLMs respectively as described in Section 2.2. The reason behind choosing 100k graphone entries is to achieve comparable OOV rate to the 500k baseline. Systems b1 and b2 are full-word baselines.

Table 1: *Polish LVCSR WERs based on full-word/mixed-unit 5-gram LMs (sbws: sub-words, wrds: words, grfs: graphones, mrf: morpheme, slb: syllable).*

sys	sbw type	#full wrds	# sbws	# grfs	Dev WER [%]	Eval WER [%]
b1	-	300k	-	-	22.7	26.8
b2	-	500k	-	-	22.1	25.6
m1	mrf	70k	230k	0k	22.7	26.2
m2			0k	277k	22.2	25.9
m3			130k	100k	22.4	26.0
s1	slb	130k	170k	0k	22.3	26.1
s2			0k	173k	22.7	26.6
s3			70k	100k	**22.1**	**25.8**

As shown in Table 1, the syllabic system (s1) performs better than the morphemic system (m1) in terms of WER. Where as the morphemic graphone system (m2) performs better than the syllabic graphone system (s2). The syllabic mixed-unit HLM (s3) outperformed all the conventional HLMs. This gives WER reductions of [dev10: 2.6% relative (0.6% absolute); eval10: 3.7% relative (1.0% absolute)] compared to the 300k baseline (b1). Moreover, the WERs of the best system (s3) are comparable to the 500k baseline system (b2).

4.2. MaxEnt HLM and Adaptation

We record the German and Polish LVCSR recognition results using MaxEnt HLMs and the adaptation experiments as shown in Table 2. We perform the experiments only using mixed-unit vocabulary, which gave the best results using conventional LM training. For German, we choose the mixed-unit vocabulary comprising of 5k full-words, 95k morphemes and 200k graphones as described in our previous work [13] for MaxEnt experiments. For all the German MaxEnt and adaptation experiments, we use 4-gram features to create HLMs. For Polish, we choose the best system (s3) mixed-unit vocabulary from Table 1 for MaxEnt experiments. For Polish MaxEnt experiment, we use 5-gram features. To create adapted models, due to resource constrains we use 4-gram features. In our experiments, we show results of adapted/MaxEnt HLMs interpolated with N-gram HLMs.

As shown in Table 2, for German LVCSR, the interpolated MaxEnt HLM with N-gram HLM performed better than the N-gram HLM. Adaptation did not show additional improvements. One of the reasons is the use of limited LM training data which is already in-domain. It should be noted that unsupervised adaptation resulted in relatively few mis-recognized in-vocabulary errors and high OOV recognition compared to the other systems. For all MaxEnt systems, we achieve relative WER reductions of 1.8%, compared to 300k full-word system on eval09 corpora.

For Polish LVCSR, the interpolated MaxEnt HLM with N-gram HLM performed better than the standard mixed-unit N-gram HLM. We report that the adapted MaxEnt system using the first recognition pass transcriptions outperformed all other systems. For the unsupervised adapted system we achieve relative WER reductions of 6.0% compared to 300k full-word system on eval10 corpora.

4.3. OOV Word Recognition Accuracy

In Table 2, we show the number of correctly recognized OOVs w.r.t. to the 300k full-words. Here, a word is considered an OOV if it is not found in the 300k full-words. For German, the unsupervised adapted system recognizes around 37% OOVs. For Polish, the mixed-unit HLM system recognizes around 36% OOVs with the cost of high in-vocabulary mis-recognitions. However, unsupervised adapted system is the best system in terms of both WER

Table 2: *Recognition results (ln.: Language, DE: German, PL: Polish, ME.: MaxEnt experiment, adap.: adaptation, sp: supervised adaptation, usp: unsupervised adaptation, wrd mod.: word modeling, tot. voc.: total vocabulary, fw: full-words, mix.: mixed-unit types, OV: effective OOV Rate [%], MIV: Fraction of mis-recognized in-vocabulary words w.r.t. 300k fw vocabulary [%], PPL: perplexity, COV: OOVs recognized w.r.t. 300k fw vocabulary [%])*

ln.	ME.	**adap.**	**wrd**	**tot.**	**dev**					**eval**				
	(yes/no)	(sp/usp)	**mod.**	**voc.**	**OV**	**PPL**	**WER**	**MIV**	**COV**	**OV**	**PPL**	**WER**	**MIV**	**COV**
DE	no	-	fw	300k	2.9	403	31.2	–	–	2.6	430	27.3	–	–
				500k	2.4	434	30.9	–	–	2.1	464	27.1	–	–
			mix.	300k	2.6	342	31.0	23.7	29.8	2.3	364	27.0	22.3	36.2
	yes	-				338	31.0	23.5	30.0		359	26.8	22.2	36.4
		sp				291	30.4	23.0	30.4		338	26.8	22.1	36.1
		usp				–	–	–	–		278	**26.8**	**22.0**	**36.8**
PL	no	-	fw	300k	1.7	580	22.7	–	–	1.9	536	26.8	–	–
				500k	1.1	620	22.1	–	–	1.2	600	25.6	–	–
			mix.	300k	0.8	576	22.1	21.8	28.6	1.0	613	25.8	24.8	**35.9**
	yes	-				572	22.1	21.9	28.8		609	25.7	24.6	34.4
		sp				451	20.9	20.7	28.3		529	25.4	24.0	33.0
		usp				–	–	–	–		444	**25.2**	**24.2**	**34.4**

and MIV, where 34% of OOVs are recognized.

5. Conclusions

We investigated the use of mixed type of sub-word units for building an open vocabulary Polish LVCSR system. We showed that using mixed-unit types during language modeling is helpful for morphologically rich German and Polish languages in order to model the OOVs and reduce the WERs. We applied MaxEnt techniques along with adaptation to hybrid LMs. For German and Polish LVCSRs mixed-units as N-gram features helped to reduce the WERs and also in-vocabulary error rates using MaxEnt hybrid language modeling. Furthermore, we recognized around 37% and 34% of OOVs using unsupervised adapted systems for German and Polish LVCSRs respectively compared to the 300k full-word baselines on evaluation corpora. The obtained WERs for the best systems outperformed the 500k full-word baselines.

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

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

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

[4] R. Sarikaya, M. Afify, Y. Deng, H. Erdogan, and Y. Gao, "Joint morphological-lexical language modeling for processing morphologically rich languages with application to dialectal Arabic," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 7, pp. 1330–1339, 2008.

[5] S. F. Chen, "Performance prediction for exponential language models," in *Proc. Human Language Technology Conf. of the North American Chapter of the ACL*, Boulder, Colorado, USA, May 2009, pp. 450 – 458.

[6] T. Alumäe and M. Kurimo, "Efficient estimation of maximum entropy language models with N-gram features: an SRILM extension," in *Interspeech*, Chiba, Japan, September 2010.

[7] C. Chelba and A. Acero, "Adaptation of maximum entropy capitalizer: Little data can help a lot," *Computer Speech and Language*, vol. 20, no. 4, pp. 382 – 399, Oct. 2006.

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

[9] 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, Sep. 2009, pp. 2679 – 2682.

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

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

[12] M. Shaik, A. El-Desoky, R. Schlüter, and H. Ney, "Using Morpheme and Syllable Based Sub-words for Polish LVCSR," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 4680–4683.

[13] M. Shaik, A. El-Desoky, R. Schlüter, and H. Ney, "Hybrid Language Models Using Mixed Types of Sub-lexical Units for Open Vocabulary German LVCSR," in *Interspeech*, Florence, Italy, Aug. 2011.

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

[15] KombiKor-v.8.0.website: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, Jun. 2004, pp. 209 – 214.