On Improving Speech Recognition and Keyword Spotting With Automatically Generated Morphological Units

Arseniy Gorin[1], Lori Lamel[1], Jean-Luc Gauvain[1], Thiago Fraga-Silva[2]

[1]LIMSI, CNRS, Université Paris-Saclay, 508 Campus Universitaire F-91405 Orsay
{gorin, lamel, gauvain}@limsi.fr

[2]Vocapia Research, 28 rue Jean Rostand, 91400 Orsay, France
thfraga@vocapia.com

Abstract

This paper presents an experimental study on using morphological units for both automatic speech recognition (ASR) and keyword spotting (KWS) for the Kazach language. Similar to other morpholoigically rich languages, the words in Kazach are composed from fixed morphemes, which specify the meaning and may change form depending on the context. This typically leads to a relatively large lexical variety, which in turn requires large amounts of textual resources for reliable language modeling. This paper demonstrates that automatically discovered sub-word units can be efficiently used for ASR and KWS with limited training resources. The experiments are conducted on conversational telephone speech data from the Babel project using word and three types of sub-word units. Using language models based on sub-word units, the KWS performance for OOV keywords is tripled. At the same time the sub-word-based systems perform as well or better than the word-based ones for the ASR task. Using a high order neural network language model also improves the ASR performance for all systems. When Web texts are used for language modeling, all sub-word systems outperform the word based one on the KWS task by finding more OOV keywords, without degrading ASR performance.

1. Introduction

Morphologically rich languages (the majority of Slavic languages, French, German, Turkish, Hungarian, Kazach, etc) pose specific problems for various tasks related to statistical language modeling. The main reason is that a given word can have a different morphological structure depending on grammatical, syntactic, or semantic context.

Language models for automatic speech recognition (ASR) systems typically rely on word sequence counts and do not take into account the language morphology. When the amount of the training texts is large enough, this standard approach usually works fine. However, in low-resource conditions, taking into account language morphology may significantly reduce the number of out-of-vocabulary words (OOV) and result in more reliable language models.

Several works have reported ASR performance gains using various types of word morphological decomposition (morphological units). For example, Kurimo et al. (2006) report improvements for Finnish, Estonian and Turkish large vocabulary speech recognition. Pellegrini and Lamel (2009) reported significant OOV word reduction for Amharic broadcast news system with only small improvements in word error rate. Tarjan et al. (2013) demonstrated large improvements for low-resource ASR in a Hungarian telephone speech transcription system.

Other types of word decomposition (so called cross-word lexical sub-word units, or character n-grams (Szöke et al., 2008)) have also been used for improving keyword spotting (KWS) of various languages (Hartmann et al., 2014). Character n-grams typically find more OOV keywords, but perform worse on in-vocabulary keywords, which may result in a degradation in ASR performance and KWS for in-vocabulary words.

This paper focuses on improving a joint ASR/KWS system designed for Kazach telephone speech, using resources provided in the context of IARPA Babel program[1]. Our study starts from the experiments in low-resource condition. At the end of the paper, additional experiments with language modeling using Web texts are described. Although the paper focuses on Kazach, the proposed approach can be extended to other morphologically rich languages.

The objective is to show that sub-word units (also called morphs in this paper) can be seen as universal units for both tasks in low-resource conditions for morphologically rich languages. Throughout the paper, we not only aim to compare words and morphs, but also try to understand, which style of morph decomposition leads to the best performance.

The remainder of the paper is organized as follows. Section 2 describes the Kazach language and the experimental data. Section 3 presents the three word-to-morph tagging schemes used in this work. Section 4 presents the experiments with low-resource language modeling. Section 5 analyzes the impact of adding Web data for language model training. Section 6 summarizes the results and draws conclusions.

2. Kazach language and data description

Kazach is the official language of Kazachstan, which is also spoken by a part of the population in China, Mongolia and Russia. It is a member of the Turkish language family. Kazach uses the Cyrillic alphabet for writing (all Russian alphabet plus 9 specific letters) since 1940. Older scripts were based on Arabic and Latin alphabets.

The pronunciation rules of modern Kazach are quite directly derived from the written form. It is a nomina-

[1]http://www.iarpa.gov/images/files/programs/babel/Babel_Overview_UNCLASSIFIED-2011-05-31.pdf

tive and agglutinative language. The nouns have 7 cases (noun endings change depending on the context). Verbs also have many forms determined by suffixes, depending on the grammatical category and tense.

All experiments in this work use the Kazach full language pack (iarpa-babel302b-v1.0a) from IARPA Babel program. About 40 hours of manually transcribed conversational telephone speech data are available for both acoustic and language model training. All results are reported on the supplied 10 hour development set.

The official development keyword list is used for the KWS experiments reported in this work. It contains about 4k keywords, about half of which are composed of several words. If any one of the tokens in a compound keyword is OOV, then the whole sequence is considered as OOV.

The same principle is applied for KWS scoring: the keyword sequence is considered as detected only if all tokens are detected (case-insensitive match of the exact word form without normalization). About 20% keyword tokens are OOV with respect to the vocabulary of the training transcriptions.

3. Morphological decomposition

The morphs in this work are automatically extracted using Morfessor toolkit (Virpioja et al., 2013). The advantage of this tool is that no manual segmentation and even no language knowledge are required.

An important detail is how to represent the decomposed words in terms of morphs (morph sequence tagging). Ideally, we would like to keep some information about the word boundaries, but also allow additional flexibility when reconstructing word sequences from morph sequences after decoding.

One type of such word-to-morph mapping is called non-initial tagging (NI) (Arisoy et al., 2009). In this technique, all morphs of a word except for the first one are tagged with a special symbol ("@" symbol is used in this work). When reconstructing the ASR output, all non-initial morphs are merged with the associated left-context units.

Another form considered in this work is referred to as fully connected (FC) tagging. The approach is similar to the one used by Pellegrini et al. (2007). In this type of decomposition a special symbol is added to those morph boundaries that are located inside the original word. During the reconstruction, FC morphs are connected only if one of them has the special symbol on the side to be merged.

Finally, word boundary (WB) tagging is used, similar to Hartmann et al. (2014) work. In this case, a special symbol is added to the word boundaries. To reconstruct back the word sequence, all morphs are connected and the special symbols are replaced with spaces.

Figure 1 summarizes the three proposed decomposition techniques. In all tagging schemes the filler and silence units are treated separately, i.e. they are always considered as a word separator. For an isolated word (without a prefix/suffix), its original form is preserved for NI and FC tagging, while word boundary symbols are added in WB tagging scheme.

NI:	"prestemsuf"	⇒	"pre"	+ "@stem"	+ "@suf"
FC:	"prestemsuf"	⇒	"pre@"	+ "@stem@"	+ "@suf"
WB:	"prestemsuf"	⇒	"@pre "	+ "stem"	+ "suf@"

Figure 1: Three types of morph tagging: non-initial (NI), fully-connected (FC) and word boundary (WB)

The complexity of decomposition, decoding and word reconstruction is similar for all three types of tagging. NI tagging in our experiments typically leads to a smaller vocabulary, because there is no difference between stem and word ending composed of the same character sequences.

4. ASR and KWS experiments

This section presents the ASR and KWS experiments conducted under the Babel full-language pack condition, in which only the provided data could be used for acoustic and language model training. The goal is to compare the performance of full word and morph-based systems. The analysis is first done with conventional 3-gram language models, then with hybrid neural network language models.

4.1. ASR and KWS performance measures

ASR performance is traditionally reported in terms of word error rate (WER), which is similar to word-level Levenshtein distance.

The performance of KWS systems in Babel program is measured with maximum term-weighted value (MTWV) and actual term-weighted value (ATWV)[2]. ATWV for the keyword k at the specific threshold t is defined as

$$ATWV(k,t) = 1 - P_{FR}(k,t) - 999.9 \cdot P_{FA}(k,t) \quad (1)$$

where P_{FR} and P_{FA} are probabilities of false reject (miss) and false accept, respectively.

MTWV is computed as a maximal ATWV over all possible values of t. We report our results in terms of MTWV, as the currently used normalization techniques put ATWV and MTWV very close to each other.

4.2. Baseline ASR and KWS system

The ASR system is based on LIMSI STK toolkit. It is used for generating word (and morph) lattices, and 1-best hypotheses for scoring. The decoder uses a 2-gram language model to produce word lattices, which are then rescored with a 3-gram language model and then converted to consensus networks for KWS.

The dictionary is generated by grapheme-to-phoneme mappings extracted from a short language description file provided by the IARPA Babel program[3]. The resulting phone set consists of 38 units.

For acoustic modeling, we used multilingual (trained on 11 Babel languages) fine-tuned stacked bottleneck features provided by our partners from Speech@FIT group from Brno University of Technology (Grézl and Karafiát, 2014).

[2]http://www.nist.gov/itl/iad/mig/upload/KWS14-evalplan-v11.pdf

[3]http://www.nist.gov/itl/iad/mig/upload/IARPA_Babel_Performer-Specification-08262013.pdf

The HMM consists of roughly 10k tied states, which model word position-independent triphones. There are 150k Gaussian densities in our models. More details on the acoustic model training and KWS (although for different languages) can be found in (Lamel et al., 2011; Le et al., 2014).

Keyword search is done on the consensus network without considering word boundaries. This allows to handle a part of the OOV keywords even on a baseline full word-based system. It is known that keyword score normalization is crucial for achieving the right balance between true positives and false alarms. In this work, the raw scores are first normalized with a linear fit model (Karakos and Schwartz, 2015), after which keyword-specific thresholding and exponential normalization (KST) is applied (Karakos et al., 2013).

4.3. Comparing word and morph-based systems

In order to compare performance of different lexical units, the training data are decomposed into morphs, which are encoded with additional symbols for word reconstruction, as discussed in Section 3. Then, the pronunciations are generated for the resulting morphs in the same way as for words. Table 1 summarizes the main results and shows the lexicon size for word- and morph-based systems with three types of tagging.

Units	# Units	WER	MTWV (All / IV / OOV)
Word	20257	50.62	0.4116 / 0.4628 / 0.0668
NI	14471	50.65	**0.4186 / 0.4541 / 0.1829**
FC	17158	50.50	0.4133 / 0.4496 / 0.1747
WB	17139	50.46	0.4123 / 0.4494 / 0.1654

Table 1: Number of units in the lexicon (for word-based systems equivalent to the vocabulary size), ASR (WER) and KWS (MTWV) performances for word-based and three morph-based systems (NI, FC and WB)

Morphs with not-initial (NI) tagging scheme result in the smallest vocabulary size and triple the detection of OOV keywords with the least degradation on the in-vocabulary (IV) words. This means that some portion of OOV words can be found as a sequence of morphs in consensus network.

As for ASR, only a small improvement is observed for morphs with the WB and FC tagging schemes. Looking at a small drop of IV keyword spotting with morph-based units, we could conclude that full words are still more robust for recognizing in-vocabulary words.

4.4. Lattice re-scoring with a neural network LM

While the comparison of words and morphs in the previous section shows the advantage of the latter in the context of KWS task with no degradation of ASR performance, the language models of morph-based systems can be improved by using a longer context. Due to the fact that a word can be decomposed into several morphs, the 3-gram probability estimates of the last morph of a word would frequently not use any information outside of this word. However, preliminary experiments demonstrated that conventional back-off 4-gram LMs cannot be reliably trained with limited data even using morphological decomposition.

This section presents the experiments with re-scoring word lattices using a 4-gram hybrid neural network language model (NNLM) (Schwenk, 2004; Schwenk, 2013). The NNLM projects the word indices onto a continuous space and uses a probability estimator operating on this space. These models have been shown to be particularly helpful when the training resources are limited (Oparin et al., 2012).

For each system, four NNLMs with a varying number of parameters were trained. The resulting language model is achieved by interpolating back-off 4-gram language model and all NNLMs. The interpolation weights are estimated with EM algorithm on development set. The parameters and the associated perplexities are summarized in Table 2.

A known problem of NNLM is that the computational cost grows significantly with the size of the output layer, i.e., with the vocabulary size. To cope with this problem, NNLM is frequently used for computing probabilities of a smaller subset of words - shortlist. See (Schwenk, 2004) for further details. The size of the shortlist is 12k words for all NNLMs in these experiments.

model	P	H	ppx (dev)	weight
back-off	–	–	188.9	0.366
NNLM 1	300	500	183.9	0.151
NNLM 2	250	450	184.0	0.160
NNLM 3	200	500	184.3	0.159
NNLM 4	220	430	183.6	0.164
Interpolation	–	–	165.1	–

Table 2: 4-gram back-off language model and neural network language models trained using transcriptions of 40 hours train set. Projection layer size (P), hidden layer size (H), perplexities on development data and the interpolation weights are reported (12k shortlist)

Overall, the combined model improves the perplexity from 188.9 to 165.1. This model is used for re-scoring the lattices from the baseline decoder described in the previous section. The resulting ASR and KWS performances are summarized in Table 3.

Units	WER	MTWV (All / IV / OOV)
Word	50.15	0.4116 / 0.4644 / 0.0574
NI	**49.86**	**0.4219 / 0.4578 / 0.1820**
FC	50.07	0.4184 / 0.4536 / 0.1826
WB	**49.84**	0.4177 / 0.4544 / 0.1733

Table 3: ASR and KWS performances for word-based and three morph-based systems after lattice re-scoring with a 4-gram hybrid neural network language model

Comparing the results of Table 1 and Table 3, the following conclusion can be drawn. First, the absolute WER improvement from using NNLM ranges from 0.5 to 0.8. Second, the WER improvement for the morph-based systems with NI and WB tagging are larger than for the word-based system, which supports the intuition that even with small amounts of training texts, morphs can benefit from the language models with larger contextual dependencies.

Finally, a consistent gain on KWS performance is observed for all morph-based systems, while no improvement is seen for the word-based system. Overall the best ASR and KWS performance is achieved with NI tagged morphs.

5. Language modeling with Web data

This set of experiments aims to evaluate the ASR and KWS performance when additional texts retrieved from the Web are used for language model training. Intuitively, using morphs rather than words makes more sense with limited text resources. In somewhat similar way to sub-word decomposition, enlarging training texts for language modeling results in a smaller number of OOV words. The goal of this evaluation is to understand if these techniques are complementary for the ASR and KWS tasks.

The Web data used in these experiments were filtered, normalized and provided to the Babelon team by our partner BBN (Zhang et al., 2015). The texts collected by BBN and IBM are comprised of various Web and Wikipedia documents. In total, there are about 15M words (587k unique) available for the Kazach language.

To understand how many words are actually useful to select, several 3-gram back-off language models were trained using the selected vocabularies of various sizes, and a fast decoding (small beam) was done on the development set. The resulting WER and OOV are shown in Figure 2.

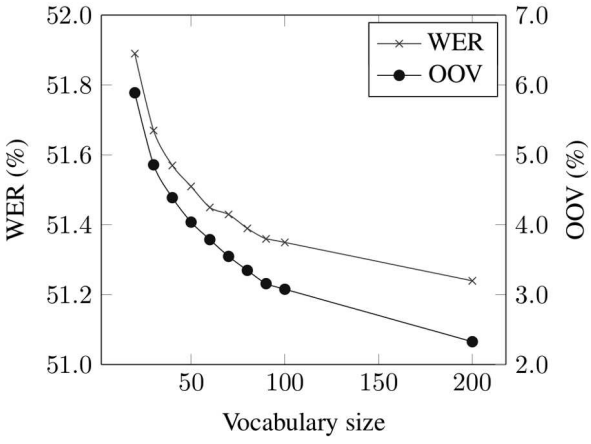

Figure 2: Word error rate (WER) and out-of-vocabulary range (OOV) on development data as a function of the vocabulary size

5.1. Morph and word-based LMs with Web data

For the final set of experiments with words, morphs, back-off and neural network language models, a 100k vocabulary has been selected. The same vocabulary and the amount of Web data are used to train the word-based system and for the Morfessor-based decomposition.

The architecture of the hybrid NNLM is exactly the same as in the previous experiments (interpolation of 4-gram back-off LM and four NNLMs with varying number of parameters). The only difference is that when training NNLMs, the weights of Web sources (BBN web, Wikipedia and IBM web) were slightly perturbed to introduce more variance across NNLMs.

The resulting ASR and KWS performances for three morph-based systems are summarized in Table 4, which also provides the corresponding lexicon sizes.

Units	# Units	LM	WER	MTWV (All / IV / OOV)
Word	100k	3-gram	49.20	0.4403 / 0.4581 / 0.1124
		NNLM	48.53	0.4461 / 0.4652 / 0.0930
NI	83k	3-gram	49.50	**0.4431 / 0.4586 / 0.1580**
		NNLM	**48.50**	**0.4491 / 0.4656 / 0.1434**
FC	95k	3-gram	49.50	0.4429 / 0.4586 / 0.1509
		NNLM	48.63	0.4471 / 0.4636 / 0.1423
WB	95k	3-gram	49.45	0.4429 / 0.4586 / 0.1509
		NNLM	48.54	0.4461 / 0.4627 / 0.1394

Table 4: Summary of the experiments with Web data used for language modeling: number of units in the lexicon, ASR and KWS performances for word-based and three morph-based systems with 3-gram back-off and 4-gram neural network language models

These experiments lead to slightly different conclusions than the earlier ones. First, applying NNLM significantly reduces the WER for all models, making them almost identical across word and morph-based systems. Remarkably, a slight drop of MTWV for in-vocabulary words is no longer observed in morph-based systems. The improvement in KWS is consistent for all three morph-based systems, but the gain is less since there are fewer OOV words with the larger vocabulary (2.3% compared to 5.9%).

6. Conclusion

This paper reports on an experimental analysis of several types of morphological units for Kazach conversational speech recognition and keyword spotting. In contrast to the conventional system combination based approaches, we show that sub-word units can be efficient for both ASR and KWS.

The analysis was carried out for several conditions: with text resources limited to the audio transcriptions as well with using additional texts from the Web for language modeling. The word and sub-word units were also used in combination with a 4-gram hybrid neural network language model.

Under low-resource conditions, sub-word units where shown to triple the MTWV score for OOV keywords without degrading the ASR performance. In addition, with NNLM re-scoring in the low-resource setup, NI and WB morphs slightly outperform words in ASR task. When additional Web corpora are used for LM training, morphs again result in the same ASR performance, but consistently outperform words in terms of MTWV.

Considering the fact that using NNLMs significantly improves the performance of the morph-based systems, it would be interesting to continue the experiments with recurrent neural network LMs, which take into account the whole context of the utterances. Evaluation on other morphologically rich languages (such as French and Arabic) can also be considered.

Acknowledgements

We would like to acknowledge the help and contribution of other colleagues from Vocapia, including Abdel Messaoudi, Viet Bac Le and Antoine Laurent. We would also like to thank other partners of the Babelon team on the IARPA Babel project for exchanging the resources (BUT for the bottle-neck features, BBN and IBM for collecting and preparing Web texts).

This work was partially supported by the French National Agency for Research as part of the SALSA (Speech And Language technologies for Security Applications) project under grant ANR-14-CE28-0021 and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12- C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

7. References

Arisoy, Ebru, Doğan Can, Siddika Parlak, Haşim Sak, and Murat Saraçlar, 2009. Turkish broadcast news transcription and retrieval. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(5):874–883.

Grézl, František and Martin Karafiát, 2014. Combination of multilingual and semi-supervised training for under-resourced languages. In *Proceedings of Interspeech*.

Hartmann, William, Lori Lamel, and Jean-Luc Gauvain, 2014. Cross-word subword units for low-resource keyword spotting. In *Proceedings of SLTU*.

Karakos, Damianos and Richard Schwartz, 2015. Combination of search techniques for improved spotting of oov keywords. In *Proceedings of ICASSP*.

Karakos, Damianos, Richard Schwartz, Stavros Tsakalidis, Le Zhang, Shivesh Ranjan, T Tim Ng, Roger Hsiao, Guruprasad Saikumar, Ivan Bulyko, Long Nguyen, et al., 2013. Score normalization and system combination for improved keyword spotting. In *Proceedings of ASRU*.

Kurimo, Mikko, Antti Puurula, Ebru Arisoy, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Tanel Alumäe, and Murat Saraclar, 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*.

Lamel, Lori, Sandrine Courcinous, Julien Despres, Jean-Luc Gauvain, Yvan Josse, Kevin Kilgour, Florian Kraft, Viet Bac Le, Hermann Ney, Markus Nußbaum-Thom, et al., 2011. Speech recognition for machine translation in Quaero. In *Proceedings of IWSLT*.

Le, Viet-Bac, Lori Lamel, Abdel Messaoudi, William Hartmann, Jean-Luc Gauvain, Cécile Woehrling, Julien Despres, and Anindya Roy, 2014. Developing STT and KWS systems using limited language resources. In *Proceedings of Interspeech*.

Oparin, Ilya, Martin Sundermeyer, Hermann Ney, and Jean-Luc Gauvain, 2012. Performance analysis of Neural Networks in combination with n-gram language models. In *Processing of ICASSP*.

Pellegrini, Thomas and Lori Lamel, 2007. Using phonetic features in unsupervised word decompounding for asr with application to a less-represented language. In *Proceedings of Interspeech*.

Pellegrini, Thomas and Lori Lamel, 2009. Automatic word decompounding for ASR in a morphologically rich language: Application to Amharic. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(5):863–873.

Schwenk, Holger, 2004. Efficient training of large neural networks for language modeling. In *Proceedings of International Joint Conference on Neural Networks*, volume 4.

Schwenk, Holger, 2013. CSLM-a modular open-source continuous space language modeling toolkit. In *Proceedings of Interspeech*.

Szöke, Igor, Lukaš Burget, Jan Černocky, and Michal Fapšo, 2008. Sub-word modeling of out of vocabulary words in spoken term detection. In *Proceedings of SLT*.

Tarjan, Balazs, Gellért Sárosi, Tibor Fegyo, and Péter Mihajlik, 2013. Improved recognition of hungarian call center conversations. In *Proceedings of Speech Technology and Human-Computer Dialogue (SpeD)*.

Virpioja, Sami, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al., 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

Zhang, Le, Damianos Karakos, William Hartmann, Roger Hsiao, Richard Schwartz, and Stavros Tsakalidis, 2015. Enhancing low resource keyword spotting with automatically retrieved web documents. In *Proceedings of Interspeech*.