AN INVESTIGATION INTO LANGUAGE MODEL DATA AUGMENTATION FOR LOW-RESOURCED STT AND KWS

Guangpu Huang[⋆] *Thiago Fraga da Silva*[†] *Lori Lamel*[⋆] *Jean-Luc Gauvain*[⋆]
Arseniy Gorin[⋆] *Antoine Laurent*[†] *Rasa Lileikyte*[⋆] *Abdel Messouadi*[†]

[⋆] LIMSI, CNRS, Université Paris-Saclay, Rue John von Neumann, F-91405 Orsay Cedex, France
[†] Vocapia Research, 28 Rue Jean Rostand, 91400 Orsay Cedex, France
{huang,lamel,gauvain,gorin,lileikyte}@limsi.fr {thfraga,laurent,abdel}@vocapia.com

ABSTRACT

This paper reports on investigations using two techniques for language model text data augmentation for low-resourced automatic speech recognition and keyword search. Low-resourced languages are characterized by limited training materials, which typically results in high out-of-vocabulary (OOV) rates and poor language model estimates. One technique makes use of recurrent neural networks (RNNs) using word or subword units. Word-based RNNs keep the same system vocabulary, so they cannot reduce the OOV, whereas subword units can reduce the OOV but generate many false combinations. A complementary technique is based on automatic machine translation, which requires parallel texts and is able to add words to the vocabulary. These methods were assessed on 10 languages in the context of the Babel program and NIST OpenKWS evaluation. Although improvements vary across languages with both methods, small gains were generally observed in terms of word error rate reduction and improved keyword search performance.

***Index Terms*—** multilingual, low resourced languages, speech recognition, keyword search

1. INTRODUCTION

Language models (LMs), trained on large corpora are useful for many speech recognition tasks. However, large quantity of in-domain text data are not always readily available, especially for relatively low-resourced languages. Standard back-off n-gram LMs predict the following word based on the previous n-1 words, e.g., n = 3. Words are represented in a discrete space, i.e., the vocabulary. For languages with sparse training data, n-grams have poor generalization to low-frequency and unseen words. This problem becomes more severe when the vocabulary size increases. A weak LM, with limited size of the training vocabulary and high OOV rate, leads to poor speech-to-text (STT) and keyword spotting (KWS) performance.

In contrast to n-grams, LMs with continuous word representations using recurrent neural networks (RNNs) have become increasingly popular [1] [2]. RNNs capture syntactic and semantic regularities in the text data of a language [1]. Moreover, with advances in training algorithms and Graphics Processing Units, it has become much easier to train RNNs and to generate artificial text [3]. Efforts aiming at the OOV problem include augmenting the word-level LMs with LMs based on subword, character, and other linguistic units [4]. Character-level RNNLMs are able to generate text data that resemble the training data. They can also introduce new words, some of which are legitimate, e.g., proper nouns [3]. Subword LMs share the advantages of word- and character-level models [5]. Unlike n-grams, these LMs assign nonzero probability to OOV words. They can generalize to previously unseen word forms by recognizing them as sequences of shorter familiar word fragments.

With the growing availability of bilingual documents with aligned texts, it also becomes plausible to use a resource-rich language to improve the LM of a resource-deficient language [6] [7]. Previously, machine translation (MT) was employed to translate Mandarin transcripts to Cantonese [8], and English to Lithuanian [9] . A similar method was explored by Kim and Khudanpur [10], who used latent semantic analysis (LSA) for cross-lingual language modelling. Both methods require bilingual text resources, except that LSA uses document-aligned texts where we use sentence aligned texts. So the system performances are subject to the quantity and quality of these external resources.

In this paper, RNNLMs and MT models were used to generate text data for 10 IARPA Babel low-resourced languages. The generated texts were used to expand the lexicon, and an LM was estimated with them was interpolated with the baseline LM. The interpolation weights were chosen to minimize the perplexity on development data. We tested the quality of the augmented LMs on STT and KWS tasks in the context of the Babel program and NIST OpenKWS evaluation. While the performance gain differs across the 10 languages, the new LM reduces the perplexity by about 10% relative abd small gains in terms of word error rate reduction and keyword search performance are generally observed.

2. DATA GENERATION WITH RNNLMS AND MT

Prior to training the RNN models, the word-based transcripts for all languages were normalized with the following rules:

- remove special tags (e.g., '<int>', '<ring>')
- split spelled words (e.g., 'w1_w2' to 'w1 w2')
- replace word fragments by a unique tag (e.g., '-w', and 'w-' to '{frg}')
- process compounded words (e.g., 'w1-w2')
  - if words 'w1' and 'w2' are already in the word list, keep 'w1-w2'
  - if one or the both are missing, then keep twice, once as 'w1-w2' and once as 'w1 w2'

The normalized transcripts are used to prepare subword, character, and morph based transcripts to train RNNLMs. For all languages, except Cantonese, the training transcripts are decomposed into subwords units (n-gram of characters), with maximum length ranging from 3 to 7-characters. Decomposition is cross-word and not unique. The subword units are obtained using an iterative procedure, which attempts to maximize the likelihood of the data while minimizing the total number of units, as described in [11]. They are then used to train RNNLM models which in turn generate the subword-based text data. Afterwards the generated subword transcripts are recomposed back into words to build LMs. The vocabularies are automatically discovered on the subword units, and the pronunciations are added to the lexicon. Other approaches to find pronunciations for low resource languages are in [12].

For Cantonese, we use CJK characters, unicode block: [\u4e00-\ufaff], as the subword unit to train RNNLM, as words are usually 2 to 5 characters long. We also segment Cantonese words into morpheme-like units using Morfessor [13]. Sub-strings occurring frequently enough in several different word forms are proposed as morphs, and the words in the corpus are then represented as a concatenation of morphs. We use an existing Jyutping dictionary to look up the pronunciations of the unknown words, Python CJKLIB[1]. If words are not found, the pronunciations are denoted as '&&' in the lexicon.

We randomly shuffle and split the train transcripts into 5 non-overlapping subsets. For each split, we train a RNNLM using 4 sets and the 5^{th} set for validation. In this way, the RNNLMs cover all the vocabulary in the train transcripts. For all languages, each word-based RNNLM generates about 20M tokens, and each subword-based RNNLM generates 50M tokens. Cantonese character-based RNNLM generates 100M tokens in total. The generated transcripts are used to train LMs, which are then interpolated with the baseline LMs for each Babel language.

We previously used the BUT RNNLM toolkit to generate the text data[2]. Training takes about 24 to 48 hours, and the trained model generates about 0.7M tokens/hour, averaged across all the languages in our experiments. Here an alternative implementation that speeds up RNN training and text data generation[3] was used. On average, training takes 2 to 12 hours, and the trained model generates about 3M tokens/hour. The size of the RNN hidden layer is 512, and the learning rate is 0.01. The implementation supports truncated back propagation through time (bptt). Gradients from hidden to input are back propagated on each time step. Gradients from hidden to previous hidden are propagated for 6 steps within each bptt-period block.

For MT based text data generation, the Moses default training scheme [14] was applied to 3 language pairs: Mandarin to Cantonese, English to Lithuanian, and English to Georgian. The Mandarin to Cantonese MT system is trained on a parallel corpus collected by [15]. The MT model translates Mandarin conversational telephone speech (CTS) text data into Cantonese. Since Mandarin and Cantonese share the same writing, simplified Chinese in this experiment, the Mandarin CTS transcripts are also added to LM training material. The English to Lithuanian/Georgian MT systems are trained on the OPUS corpus with parallel subtitles [16]. The MT models translate English Fisher text data into the target languages [17]. The translated texts are then added to LM training material.

3. STT AND KWS SYSTEMS

The experiments were conducted in the context of the Babel program and NIST OpenKWS evaluation, including 10 IARPA-Babel languages: Cantonese (IARPA-babel101b-v0.4c), Lithuanian (IARPA-babel304b-v1.0b), Igbo (IARPA-babel306b-v1.0b), Pashto (IARPA-babel104b-v0.4aY), Javanese (IARPA-babel402b-v1.0b), Mongolian (IARPA-babel401b-v2.0b), Guarani (IARPA-babel305b-v1.0b), Amharic (IARPA-babel307b-v1.0b), Dholuo (IARPA-babel403b-v1.0b), and Georgian (IARPA-babel404b-v1.0a).

All STT systems use BUT 28L features [18][4], except Cantonese. The Cantonese STT system uses two bottle-neck MLPs, combining PLP and pitch features on one side, and TRAP-DCT features on the other side [19–21]. This results in a set of 88 features which are transformed using a speaker-based CMLLR transform estimated with a GMM-HMM.

For all systems, the acoustic models are sets of tied-state, word-position dependent triphones. The baseline LMs are the standard Kneser-Ney back-off 3-gram models. Word lattices from the STT system are converted to consensus networks (CN) for KWS [22]. Search on the CNs ignores word boundaries, which handles a portion of the OOVs even for a baseline system. In this work, the raw scores are first normalized with a linear fit model, after which keyword-specific thresholding and exponential normalization is applied [23]. This is

[1] https://code.google.com/archive/p/cjklib/

[2] http://rnnlm.org/

[3] https://github.com/yandex/faster-rnnlm

[4] Mult28Lv0.noisesv0.wpe1onFarF.1stage.cmllr

LM text	CER	ATWV (all / iv-iv / oov-iv / oov-oov)	ppx	oov(%)	vocab
trs	40.7	0.490 / 0.534 / - / 0.192	135	2.4	18.3K
s1: trs+rnn-w	40.4	0.495 / 0.539 / - / 0.199	106	2.4	18.3K
s2: trs+rnn-m	40.6	0.489 / 0.533 / - / 0.195	113	2.4	18.3K
s3: trs+rnn-c	40.3	0.497 / 0.541 / 0.321 / 0.197	110	2.1	112.2K
s4: trs+rnn-(c,m,w)	40.3	0.498 / 0.541 / 0.329 / 0.204	108	2.1	113.7K
s5: trs+mt	40.4	0.502 / 0.534 / 0.403 / 0.232	122	1.7	37.9K
comb. a: s1⊕s2⊕s3	40.2	0.504 / 0.544 / 0.250 / 0.237	-	-	-
comb. b: s1⊕s2⊕s3⊕s4	40.1	0.505 / 0.544 / 0.259 / 0.239	-	-	-
comb. c: s1⊕s2⊕s3⊕s5	40.1	**0.519 / 0.548 / 0.479 / 0.288**	-	-	-

Table 1: Cantonese STT and KWS performance on LMs with RNNLM and MT generated text data. trs: train transcripts; rnn-w/m/c: RNNLM generated transcripts using the word/morph/character unit; mt: MT generated transcripts from Mandarin CTS, Mandarin transcripts are also used in LM interpolation; ppx: LM perplexity on the development data; iv: in-vocabulary word; oov: out-of-vocabulary word; vocab: the number of unique words in the 3-gram LM; oov-iv: out-of-vocabulary word becomes in-vocabulary word.

to balance between true positives and false alarms. The official development keyword list distributed by NIST is used to evaluate the KWS performance.

The STT performance is measured with word error rate (WER) on all languages, except Cantonese, which uses character error rate (CER). KWS performance is measured with actual term-weighted value (ATWV)[5]. ATWV for the keyword k at the specific threshold t is defined as

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

where $C = 999.9$ is a constant, P_{FR} and P_{FA} are the missing probabilities and false accept, respectively.

4. RESULTS ON CANTONESE STT AND KWS

Table 1 summarizes our most recent results on Cantonese STT and KWS systems using the RNNLM and MT generated text data in LMs. We observe improvements on both OOV and in-vocabulary (IV) words. The LM including all the RNNLM generated texts improve CER by 0.4% and ATWV by 0.8 point (s4 in Table 1). Word-based RNNLMs keep the same system vocabulary so do not impact the OOV. The morph-based RNNLMs do not impact the OOV for Cantonese, probably due to the fact that the generated morphs are actual words. Still the generated transcripts reduce the perplexity, and they give marginal gains in CER and ATWV. Character-RNNLMs generate useful word tokens by learning the text structure. The generated transcripts expand the 18.4K word vocabulary to 112.2K, and reduce the OOV rate by 0.3% absolute (s3 in Table 1). It is interesting to note that they gain 0.3 point on ATWV for OOV words that become IV words (oov-iv in Table 1).

The best performance was obtained with RNNLM generated transcripts, by combining the STT system outputs and the keyword hit lists (comb. b in Table 1). The combined system improves the CER by 0.6% and the ATWV by 1.5 points over the baseline system. The 1-best STT system outputs are combined via ROVER [24]. The keyword hits are combined using the maximum of the raw scores, with score normalization applied to the combined list. However, there is a limitation of the gain from RNNLMs. Beyond a certain point, any improvements must result from a deeper understanding of the text, or from using external data such as the MT generated transcripts. MT generated transcripts reduce OOV rate by 0.7% absolute without significantly increasing the lexicon size. They obtain 0.4 ATWV for OOV words that become IV words (oov-iv in Table 1). They prove more effective than the character-based RNNLM. Using both the Mandarin transcripts and the translated Cantonese transcripts yields an ATWV of 0.502 and an ATWV of 0.519 is obtained by combining the systems (comb. c in Table 1). The ATWV is improved by 2.9 points over the baseline system.

5. RESULTS ON 10 IARPA BABEL LANGUAGES

Table 2 compares the perplexity of the interpolated LMs with the baseline back-off 3-grams estimated only on the training transcripts for 10 IARPA Babel languages. The interpolated LM incorporating the RNNLM or MT generated text provides better vocabulary coverage and reduces the OOV rate. Combining the generated word and subword transcripts obtains about a 10% relative reduction in LM perplexity for most languages. The MT generated transcripts reduce the LM perplexity for both Cantonese and Lithuanian, but not for Georgian.

For Georgian, we compared using 57 manually selected affixes from scholar-seeded knowledge provided by IBM, with the automatically determined n-gram subword units, but these did not reduce the perplexity. We also trained Morfessor with the training transcript with and without BBN and IBM web text data, updated the Morfessor vocabulary with the 57 manual affixes, but still observed no gain with the generated morph-based transcripts for Georgian.

Table 3 summarizes the overall performance chart of STT and KWS systems using the RNNLM and MT generated text

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

LM text	vocab	trs	trs +rnn-w	trs +rnn-sw	trs +mt
Cantonese	18.3K	135	106	110	121
Lithuanian	28.3K	236	-	-	230
Igbo	15.3K	117	110	115	-
Pashto	12.5K	166	152	162	-
Javanese	13.2K	228	202	224	-
Mongolian	20.9K	128	119	121	-
Guarani	24.3K	182	169	177	-
Amharic	31.4K	301	285	300	-
Dholuo	15.9K	170	158	166	-
Georgian	30.4K	298	299	-	309

Table 2: LM perplexity on the development data of 10 limited resourced languages. trs: train transcript, rnn-w/sw: RNNLM generated transcripts using word/subword units; mt: MT generated transcripts. For Cantonese the rnn-sw transcripts are generated by character based RNNLMs. Since the subword-level RNNLMs and the MT generated trasncripts modify the vocabulary, a theoretical vocabulary of 100K words was used when calculating perplexity.

LM text	trs+rnn-w		trs +rnn-sw	trs+mt	
	WER	ATWV	ATWV	WER	ATWV
Cantonese	-0.3	+1.0	+0.7	-0.4	**+1.2**
Lithuanian	+0.1	-0.1	-	-3.0	**+5.9**
Igbo	-0.4	+0.8	+0.1	-	-
Pashto	-0.9	**+2.2**	-	-	-
Javanese	-0.3	+0.4	**+2.3**	-	-
Mongolian	-0.2	+0.5	+0.6	-	-
Amharic	-0.4	+0.2	**+3.2**	-	-
Georgian	0.0	-0.1	**+1.5**	0.0	0.0

Table 3: Overall STT and KWS performance gains on 8 IARPA Babel limited resourced languages using LMs with RNNLM and MT generated text data. WER: word error rate; Cantonese uses character error rate (CER).

data in LMs. Improvements over the baseline systems are language specific. We observe small gain from MT and RNNLM generated transcripts on most languages, except for Georgian. For Georgian, subword based RNNLM generated transcripts improve ATWV by 1.5 points, though MT and word-based RNNLM generated transcripts give no gain.

RNNLM generated word transcripts get the most gain for Pashto STT and KWS results: WER is reduced by 0.9% absolute, ATWV is improved by 2.2 points. RNNLM generated subword transcripts get the most gain for Amharic KWS result: ATWV is improved by 3.2 points. They also get 2.3 points ATWV gain on Javanese, and 1.5 points on Georgian. We also observe small gain from MT generated transcripts on several languages. MT generated transcripts get the most gain on Lithuanian STT and KWS results: WER is reduced by 3.0% absolute, ATWV is improved by 5.9 points. They also get 3.2 points ATWV gain on Mongolian, and 1.2 points on Cantonese. The detailed results are given in Table 4.

LM text	trs+rnn-w		trs +rnn-sw	trs+mt	
	WER	ATWV	ATWV	WER	ATWV
Cantonese	40.4	0.495	0.497	40.3	0.502
Lithuanian	43.1	0.568	-	40.0	0.628
Igbo	58.0	0.289	0.282	-	-
Pashto	48.2	0.366	-	-	-
Javanese	50.3	0.395	0.414	-	-
Mongolian	47.9	0.458	0.459	-	-
Amharic	39.5	0.607	0.637	-	-
Georgian	41.8	0.574	0.590	41.8	0.575

Table 4: STT and KWS performance on 8 IARPA Babel limited resourced languages using LMs with RNNLM and MT generated text data. Cantonese uses CER.

6. CONCLUSIONS

Large amounts of in-domain text are required in order to train accurate and robust n-gram language models. In this paper, we investigated two techniques, recurrent neural networks and machine translation models, for text data augmentation to improve language models of 10 IARPA Babel languages with low resources. We combine the advantages of word, subword, morph, and character based RNNLMs to generate additional text data. Word-based RNNLMs do not impact the OOV rate, whereas the subword and character based RNNLMs can reduce the OOV rates. Machine translation method is also used to translate English Fisher and Mandarin CTS text data to the target languages. The two techniques are complementary to each other, where we obtain the best results on Cantonese via system combination on the generated transcripts. We observe small gains on STT and KWS system performance for the other languages. The improvements vary across languages with both techniques. Word-based RNNLMs generated transcripts get the most gain for Pashto STT and KWS results. Subword-based RNNLMs generated transcripts get the most gain for Amharic KWS result. MT generated transcripts get the most gain on Lithuanian STT and KWS results.

ACKNOWLEDGMENTS

This work was in part supported by the French National Agency for Research as part of the SALSA (Speech And Language technologies for Secur ity Applications) project under grant ANR-14-CE28-0021 and by Intelligence Advanced Research Projects Activity (IARPA-babel-babel) 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

[1] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig, "Linguistic regularities in continuous space word representations," in *HLT-NAACL*, 2013, pp. 746–751.

[2] Ilya Oparin, Martin Sundermeyer, Hermann Ney, and Jean-Luc Gauvain, "Performance analysis of neural networks in combination with n-gram language models," in *ICASSP*, 2012, pp. 5005–5008.

[3] Ilya Sutskever, James Martens, and Geoffrey E Hinton, "Generating text with recurrent neural networks," in *ICML*, 2011, pp. 1017–1024.

[4] Moonyoung Kang, Tim Ng, and Long Nguyen, "Mandarin word-character hybrid-input neural network language model," in *INTERSPEECH*, 2011, pp. 625–628.

[5] Tomas Mikolov, Ilya Sutskever, Anoop Deoras, Hai-Son Le, and Stefan Kombrink, "Subword language modeling with neural networks," 20012.

[6] Mark J. F. Gales, Kate M. Knill, Anton Ragni, and Shakti P. Rath, "Speech recognition and keyword spotting for low resource languages: Babel project research at CUED," *SLTU Keynote*, 2014.

[7] Gideon Mendels, Erica Cooper, Victor Soto, Julia Hirschberg, Mark Gales, Kate Knill, Anton Ragni, and Haipeng Wang, "Improving speech recognition and keyword search for low resource languages using Web data," in *INTERSPEECH*, 2015.

[8] Guangpu Huang, Arseniy Gorin, Jean-Luc Gauvain, and Lori Lamel, "Machine translation based data augmentation for Cantonese keyword spotting," in *ICASSP*, 2016, pp. 6020–6024.

[9] Arseniy Gorin, Rasa Lileikyte, Guangpu Huang, Lori Lamel, Jean-Luc Gauvain, and Antoine Laurent, "Language model data augmentation for keyword spotting in low-resourced training conditions," in *INTERSPEECH*, 2016.

[10] Woosung Kim and Sanjeev Khudanpur, "Cross-lingual latent semantic analysis for language modeling," in *ICASSP*. IEEE, 2004, vol. 1, pp. I–257.

[11] William Hartmann, Lori Lamel, and Jean-luc Gauvain, "Cross-word subword units for low-resource keyword spotting," in *SLTU*, 2004.

[12] Marelie Davel, D Karakos, E Barnard, C van Heerden, R Schwartz, S Tsakalidis, et al., "Exploring minimal pronunciation modeling for low resource languages," *INTERSPEECH*, pp. 538–542, 2015.

[13] Sami Virpioja, Peter Smit, Stig-Arne Grnroos, and Mikko Kurimo., "Morfessor 2.0: Python implementation and extensions for morfessor baseline," Tech. Rep., Aalto University, Helsinki, 2013.

[14] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al., "Moses: Open source toolkit for statistical machine translation," in *ACL*, 2007, pp. 177–180.

[15] John Lee, "Toward a parallel corpus of spoken Cantonese and written Chinese," in *IJCNLP*, 2011, pp. 1462–1466.

[16] Jörg Tiedemann, "Parallel data, tools and interfaces in opus," in *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, 2012.

[17] Christopher Cieri David, David Miller, and Kevin Walker, "The fisher corpus: a resource for the next generations of speech-to-text," in *Proceedings of 4th International Conference on Language Resources and Evaluation*, 2004, pp. 69–71.

[18] František Grézl and Martin Karafiát, "Bottle-neck feature extraction structures for multilingual training and porting," *Procedia Computer Science*, vol. 81, pp. 144–151, 2016.

[19] Petr Fousek, Lori Lamel, and Jean-Luc Gauvain, "On the use of MLP features for broadcast news transcription," in *Text, Speech and Dialogue*, 2008, p. 303.

[20] Petr Fousek, Lori Lamel, and Jean-Luc Gauvain, "Transcribing broadcast data using MLP features.," in *INTERSPEECH*, 2008, pp. 1433–1436.

[21] František Grézl and Petr Fousek, "Optimizing bottle-neck features for LVCSR," in *ICASSP*, 2008, pp. 4729–4732.

[22] Viet-Bac Le, Lori Lamel, Abdel Messaoudi, William Hartmann, Jean-Luc Gauvain, Cécile Woehrling, Julien Despres, and Anindya Roy, "Developing STT and KWS systems using limited language resources," in *INTERSPEECH*, 2014.

[23] Damianos Karakos and Richard Schwartz, "Combination of search techniques for improved spotting of oov keywords," in *ICASSP*, 2015.

[24] Jonathan G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *ICASSP*, 1997, pp. 347–354.