

MACHINE TRANSLATION BASED DATA AUGMENTATION FOR CANTONESE KEYWORD SPOTTING

Guangpu Huang, Arseniy Gorin, Jean-Luc Gauvain, Lori Lamel

LIMSI CNRS, Spoken Language Processing Group, 91405 Orsay Cedex France
{huang, gorin, gauvain, lamel}@limsi.fr

ABSTRACT

This paper presents a method to improve a language model for a limited-resourced language using statistical machine translation from a related language to generate data for the target language. In this work, the machine translation model is trained on a corpus of parallel Mandarin-Cantonese subtitles and used to translate a large set of Mandarin conversational telephone transcripts to Cantonese, which has limited resources. The translated transcripts are used to train a more robust language model for speech recognition and for keyword search in Cantonese conversational telephone speech. This method enables the keyword search system to detect 1.5 times more out-of-vocabulary words, and achieve 1.7% absolute improvement on actual term-weighted value.

Index Terms— keyword spotting, data augmentation, language modelling, neural networks, low-resourced languages

1. INTRODUCTION

Training robust language models (LMs) on sparse data is a major challenge in automatic speech recognition (ASR). Several data augmentation approaches have been proposed to cope with this problem [1] [2]. For well-resourced languages, e.g., Mandarin and English, additional resources such as meeting and Web data have been successfully used to improve LM in broadcast news (BN) and conversational telephone speech (CTS) recognition through text normalization [3] and topic adaptation [4]. For low-resourced languages, it is also achievable to harvest Web data to improve LM in BN recognition, e.g., on Luxembourgish [5] and Latvian [6]. However, data augmentation remains difficult for CTS recognition of low-resourced languages such as Cantonese. More recently, Mendels et al. collected Web data to improve LM for CTS recognition in several low resources languages: Kurmanji, Tok Pisin, Kazakh, Telugu, and Lithuanian [7]. Yet this is challenging for Cantonese. The reason is that traditionally Cantonese is a spoken dialect without any universally recognized standard written form. Though currently both Cantonese and Mandarin speakers write in standard Chinese, Cantonese also contains a number of words and expressions

that are unique to the dialect [8]. These factors make it difficult to collect Web data for Cantonese CTS LM training.

In this paper, we propose a framework to generate CTS transcripts for low-resourced Cantonese. Using statistical machine translation (MT) models trained on a small corpus of parallel Mandarin-Cantonese subtitles, we convert a large set of Mandarin CTS transcripts to Cantonese. Our method makes use of the abundant resources available in Mandarin Chinese to train a more robust Cantonese CTS LM. In addition, it decreases the amount of out-of-vocabulary (OOV) words, which pose a serious problem for keyword search (KWS). Previous works on Cantonese ASR and KWS within BABEL project are reported in [9–13].

We show that the simple translation-based method can improve the ASR and KWS performance with significant gains in OOV detection. We report results with and without using a recurrent neural network (RNN) LM [14].

2. DATA AUGMENTATION USING MT

The amount of transcription data for conversational speech in Cantonese is substantially less than that for Mandarin. This poses a serious problem for LM training. Cantonese and Mandarin are both Chinese dialects and their written forms share many similarities in vocabulary, syntactic, and lexical compositions. They also share many unique words and characters [15]. However, there are notable differences in morphology, e.g., suffixes for plurals used in Mandarin are optional in Cantonese. Moreover, conversational speech exhibits some noticeable differences, e.g., different word order in predicative adjectives, comparison of quantities, double objects, omission of numerals, etc. To capture and generalize these regular differences, a statistical MT model was trained using the Moses toolkit [16] on a small corpus of parallel Cantonese and Mandarin TV subtitles [8], and used to convert a corpus (3.2M word tokens) of Mandarin CTS transcripts to Cantonese.

The words in the subtitle corpus are pre-segmented and separated with a space. The subtitle corpus consists of 4,135 pairs of aligned sentences, with a total of 36K characters in Mandarin, and 39K in Cantonese. In order to be consistent

with Mandarin and Cantonese CTS transcripts, the parallel corpus was converted to simplified Chinese.

The parallel corpus consists of pre-planned speech, free from false starts, repairs, repetitions, and other errors. We use the Moses toolkit in connection with GIZA++ for word alignment and IRSTLM [17] for target language modelling. 80% of the sentence pairs are randomly selected for training, and 20% for tuning and testing. Prior to tokenisation in Moses, we removed the Chinese punctuation marks in the parallel corpus. After tuning, the MT model is used to translate a corpus of Mandarin transcript to Cantonese.

The raw Mandarin CTS transcript contains 418.1K sentences and 3.2M tokens. The MT translated Cantonese transcript contains 4.7M tokens. Unknown character(s) were filtered out from translated transcripts. For new words the pronunciations were generated using GIZA++ and Moses trained on the initial pronunciation lexicon [18]. We keep only 1-best pronunciation for the new words. Furthermore, we collected 665 Cantonese words and short-phrases commonly used in conversations with their Mandarin translations from an online Baidu archive¹. If these words and phrases are found in the raw Mandarin transcript, they are directly mapped to Cantonese via table look-up. Thus the augmented Cantonese transcripts include the raw Mandarin CTS, the MT translated, and the table mapped transcripts.

Figure 1 illustrates the system architecture of translation-based data augmentation to improve LM. We trained separate LMs for the translation augmented transcripts, which are then interpolated with baseline LM. For simplicity, we refer to the LMs on the augmented transcripts as MT based LM. The mixture weights are calculated through Expectation Maximisation using a held out set. The resulting trigram is used for decoding and rescoring the word lattices.

3. EXPERIMENTAL SETUP

3.1. ASR and KWS Data

The experiments are conducted using BABEL Cantonese full language pack (babel-101b-v0.4c)². The training set contains 138 hours of manually transcribed spontaneous telephone conversations. The results are reported on the 20 hour development set. In the baseline experiments we use BABEL reference pronunciation dictionary, which contains 25.7K word types and a total of 29.1K pronunciations variants. The training transcripts contain 78K sentences and 804K word tokens. Words in the training data are pre-segmented and separated with a space. The official development keyword list is used in this evaluation. It contains 1050 in-vocabulary (IV) and 258 OOV keyword phrases. In total, keyword list has 2.1K words. The average length of keyword phrase is 3

¹Online: <http://wenku.baidu.com/view/5525fbc24028915f804dc225.html>

²Online: www.iarpa.gov/images/files/programs/babel/Babel_Overview_UNCLASSIFIED-2011-05-31.pdf

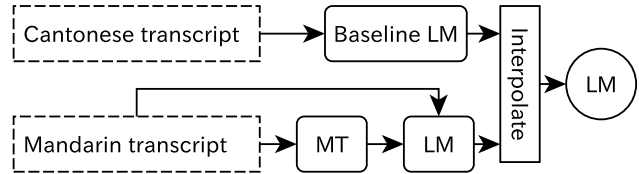


Fig. 1. System architecture for MT-based data augmentation and LM Interpolation.

characters, and the longest keyword has 10 characters and 5 compound words.

3.2. ASR System

The speech recognizer uses n -gram statistics estimated on speech transcripts for language modelling and HMMs with MLP posteriors for acoustic modelling. The acoustic features are obtained using 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 (42+46) which are then transformed using a speaker-based CMLLR transform estimated with a GMM-HMM.

The acoustic models are sets of tied-state, word-position dependent triphones. Each phone model is a left-to-right, 3-state triphone HMM. These triphones are word-position dependent in the sense that different models are used for word internal phones and word boundary phones. The decision tree state clustering is based on a set of about 800 questions automatically generated from the GMM-HMM triphones with a set of 66 phones. Clustering results in a set of 10k tied states. The MLP used to estimate the tied state posteriors has 5 hidden layers and a total of 10M weights.

The baseline language model is a standard Kneser-Ney backoff 3-gram model with a test perplexity of 181 measured on the official development data. The word decoder generates a word lattice for each speech segment. Each word lattice is then converted to a word confusion network and the 1-best word consensus hypothesis is obtained by taking the word with the highest confidence score in each confusion network slot.

3.3. KWS System

Word lattices from ASR system are converted to consensus network (CN) for KWS [22]. The search on CN ignores word boundaries, which handles a portion of the OOVs even on a baseline system. Score normalization is crucial for the right balance between true positives and false alarms. In this work, the raw scores are first normalized with a linear fit model, after which keyword-specific thresholding and exponential normalization (KST) is applied [23].

LM texts	Perplexity	OOV (%)	CER (%)	ATWV (all / IV / OOV)	MTWV (all / IV / OOV)
trn (baseline)	180.8	2.7	40.4	0.488 / 0.532 / 0.190	0.490 / 0.534 / 0.194
trn+MT	161.4	1.9	40.2	0.505 / 0.539 / 0.277	0.508 / 0.542 / 0.285
trn+RNN	157.4	2.7	40.0	0.492 / 0.536 / 0.198	0.493 / 0.537 / 0.200
trn+RNN+MT	149.2	1.9	39.7	0.508 / 0.542 / 0.274	0.511 / 0.544 / 0.292
combine	–	–	39.7	0.517 / 0.548 / 0.305	0.518 / 0.549 / 0.309

Table 1. System performance on the development data: perplexity, OOV (%), character error rate (%), actual and maximum term-weighted values (measured on development keyword list).

3.4. Performance Measures

ASR performance on Cantonese is measured with character error rate (CER), which is a conventional way of scoring Chinese speech recognition systems. KWS in BABEL program is measured with actual term-weighted value (ATWV) and maximum term-weighted value (MTWV)³. 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 probabilities of miss and false accept, respectively. MTWV is computed as a maximal ATWV over all possible values of t .

4. RESULTS

Table 1 summarizes the ASR and KWS performance with the baseline system and the improvements obtained by adding the MT transformed Mandarin-to-Cantonese transcripts in the LM. Interpolating the baseline LM with the MT LM (Section 2), reduces the dev set perplexity from 181 to 161. The interpolation weight is 0.96 for the baseline LM and 0.04 for the MT LM. The original 19.5K word lexicon obtained from the training transcripts was extended with 17.5K words, reducing the OOV rate by 30% relative.

The CER of the baseline system with the LM trained only on the audio transcripts is 40.4% and the overall ATWV is 0.488. The interpolated LM (trn+MT transcripts) gives a small CER reduction and improves the overall KWS performance by almost 2%, with a larger gain for the OOV keywords (0.190 to 0.277).

We also investigated using an MT-based data augmentation technique proposed by Mikolov et al. [14]. An RNN trained on 80% of the transcripts is used to generate 100 million words of texts. As was done for the MT texts, these pseudo transcripts are used to train a component LM, which was then interpolated with the baseline LM. These two approaches are complementary as the RNN finds long contextual regularities in Cantonese transcripts, but does not address the OOV problem. This can be seen in 3rd entry (trn+RNN) of Table 1, where although the CER is reduced by 0.4%, the RNN generated transcripts are less useful for KWS than the

MT-LM. The result of interpolating the 3 LMs are given in the 4th entry (trn+RNN+MT), where the dev set perplexity is reduced to 149, and the system obtains a CER of 39.7% CER and an ATWV of 0.508.

An additional gain is obtained by combining the outputs of all systems. For the ASR system outputs, a ROVER combination of 1-best hypotheses [24] was used. For KWS the keyword hits are combined based on the maximum of the raw scores, with score normalization applied to the combined list.

5. CONCLUSION

We proposed a novel approach to generate new transcripts for Cantonese from Mandarin transcriptions. An MT model trained on a small corpus of parallel Cantonese-Mandarin subtitles was used to convert a large corpus of Mandarin transcriptions to pseudo transcripts in the low-resourced Cantonese dialect. The produced transcripts contained 17K new words with respect to the original lexicon. N -gram language models were trained on the new texts and the resulting LM was interpolated with the baseline LM. With the interpolated LM, the dev data OOV rate and perplexity were substantially reduced, and the ASR and KWS performance improved. Using an RNN to generate additional pseudo transcripts further improves performance. The best results are obtained combining the three systems, achieving a CER of 39.7% and ATWV of 0.517. The proposed method could be applied to other language pairs with parallel corpus to improve the speech recognition and keyword search performance of low-resourced languages.

ACKNOWLEDGMENTS

This research was in part supported 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.

³Online: www.nist.gov/itl/iad/mig/upload/KWS14-evalplan-v11.pdf

6. REFERENCES

- [1] Ping Xu and P. Fung, “Cross-lingual language modeling for low-resource speech recognition,” *IEEE Trans. ASLP*, vol. 21, no. 6, pp. 1134–1144, 2013.
- [2] 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.
- [3] Ivan Bulyko, Mari Ostendorf, and Andreas Stolcke, “Getting more mileage from Web text sources for conversational speech language modeling using class-dependent mixtures,” in *HLT-NAACL’03*, 2003, vol. 2, pp. 7–9.
- [4] Tim Ng, Mari Ostendorf, Mei-Yuh Hwang, Man-Hung Siu, Ivan Bulyko, and Xin Lei, “Web-data augmented language models for mandarin conversational speech recognition,” in *ICASSP’05*, 2005, pp. 589–592.
- [5] Martine Adda-Decker, Lori Lamel, Gilles Adda, and Thomas Lavergne, “A first LVCSR system for luxembourgish, a low-resourced european language,” in *LTC’11*, 2011, pp. 479–490.
- [6] Ilya Oparin, Lori Lamel, and Jean-Luc Gauvain, “Rapid development of a latvian speech-to-text system,” in *ICASSP’13*, 2013.
- [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’15*, 2015.
- [8] John Lee, “Toward a parallel corpus of spoken Cantonese and written Chinese,” in *IJCNLP’11*, 2011, pp. 1462–1466.
- [9] Martin Karafiát, František Grézl, Mirko Hannemann, Karel Veselý, and Jan Cernocký, “BUT BABEL system for spontaneous Cantonese,” in *INTERSPEECH’13*, 2013, pp. 2589–2593.
- [10] Jia Cui, Xiaodong Cui, Bhuvana Ramabhadran, Jung-Ho Kim, Brian Kingsbury, Jonathan Mamou, Lidia Mangu, Michael Picheny, Tara N Sainath, and Abhinav Sethy, “Developing speech recognition systems for corpus indexing under the IARPA Babel program,” in *ICASSP’13*, 2013, pp. 6753–6757.
- [11] KM Knill, Mark JF Gales, Satish Prasad Rath, Philip C Woodland, Chenghui Zhang, and S-X Zhang, “Investigation of multilingual deep neural networks for spoken term detection,” in *ASRU’13*, 2013, pp. 138–143.
- [12] Jonathan Mamou, Jia Cui, Xiaodong Cui, Mark JF Gales, Brian Kingsbury, Kate Knill, Lidia Mangu, David Nolden, Michael Picheny, Bhuvana Ramabhadran, et al., “Developing keyword search under the IARPA Babel program,” in *ASPC’13*, 2013.
- [13] Brian Kingsbury, Jia Cui, Xiaodong Cui, Mark JF Gales, Kate Knill, Jonathan Mamou, Lidia Mangu, David Nolden, Michael Picheny, Bhuvana Ramabhadran, et al., “A high-performance Cantonese keyword search system,” in *ICASSP’13*, 2013, pp. 8277–8281.
- [14] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig, “Linguistic regularities in continuous space word representations,” in *HLT-NAACL’13*, 2013, pp. 746–751.
- [15] Stephen Matthews and Virginia Yip, *Cantonese: A Comprehensive Grammar*, Routledge, New York, 1994.
- [16] 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’07*. Association for Computational Linguistics, 2007, pp. 177–180.
- [17] Marcello Federico, Nicola Bertoldi, and Mauro Cettolo, “IRSTLM: an open source toolkit for handling large scale language models,” 2008.
- [18] Panagiota Karanasou and Lori Lamel, “Pronunciation variants generation using smt-inspired approaches,” in *ICASSP’11*, 2011, pp. 4908–4911.
- [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’08*, 2008, pp. 1433–1436.
- [21] František Grézl and Petr Fousek, “Optimizing bottleneck features for LVCSR,” in *ICASSP’08*, 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’14*, 2014.
- [23] Damianos Karakos and Richard Schwartz, “Combination of search techniques for improved spotting of oov keywords,” in *ICASSP’15*, 2015.
- [24] J.G. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” in *ICASSP’97*, Dec 1997, pp. 347–354.