The Network-based Multilingual ASR System Towards Multilingual Conversations in Medical Domain *

◎ Sakriani Sakti, Keigo Kubo, Sho Matsumiya, Graham Neubig,
Tomoki Toda, Satoshi Nakamura（NAIST), Fumihiro Adachi, Ryosuke Isotani (NEC)

1 Introduction

The language barrier has become the most notorious obstacle to free communication among people speaking different languages. In many parts of the world there are large recent immigrant populations that require medical care but are unable to communicate fluently in the local language. Spoken language translation is one of the innovative technologies that can help in situations where no common language between the diagnosing doctor and the patient exists. This paper outlines the recent development on multilingual medical data and multilingual speech recognition system for network-based speech-to-speech translation in the medical domain.

2 Medical Text and Speech Data Design and Construction

2.1 Text Materials

- Medical Phrasebooks

  It is designed based on sentences from Japanese-English bilingual phrasebooks designed for interpreters focusing on the medical domains. Chinese translations were obtained by translating each phrase from Japanese to Chinese. In total, we had 5130 sentences with good coverage of medical-domain terminology.

- Medical Conversation

  It consists of actual conversations between the patient and the receptionists, nurses or doctors recorded during a doctor's visit. Conversations were recorded in Japanese and all participants were native Japanese speakers. The conversations were then segmented by utterance and translated into English and Chinese. In total, we had 1007 sentences for each language.

For both text resources, we then allocate 67% as a training set, and 33% as development and test sets. More details can be found in [1].

2.2 Speech Materials

From the data described above, 200 sentences of the development and test set were selected and the recording was conducted in a sound proof room, at a 48 kHz sampling rate with 16-bit resolution. The sampling rate was later down-sampled to 16 kHz for our experiments. For Japanese, English, and Chinese, there were 27 speakers with a balance of gender and age. Each speaker uttered either 100 sentences from the development set or test set, resulting in a total of 27,000 utterances per language.

3 Speech Recognition System

For English, Japanese, Chinese acoustic model training, we utilize 157 hours of about 800 English TED speech talks (http://www.ted.com/talks), 518 hours of spontaneous Japanese speech (CSJ)[2], and about 250 hours of Chinese BTEC speech [3], respectively. For language model training, the 4,000 sentences of medical phrasebooks and conversation training set were also used. In addition to that, TED Talks transcripts and ATR BTEC text data were used for a total of 519k sentences.

For each utterance in the speech training data, 13 static MFCCs including zeroth order for each frame are extracted and normalized. Nine adjacent (center, 4 left, and 4 right) frames of MFCCs are stacked into one single feature vector (117=9x13 dimensions of super vectors). It is then reduced to an optimum 40 dimensions by applying LDA and MLLT [4]. In addition, we apply feature space MLLR for speaker adaptive training.

The context-dependent cross-word triphone HMMs units are derived from 39 phonemes for Japanese and English model, and 56 phonemes for Chinese model. Each phoneme is classified by its position in the word (4 classes: begin, end, internal and singleton). This model totally includes 80K Gaussians trained with both speaker adaptive training (SAT)[5] maximum likelihood (ML) estimation

*医療ドメインにおける多言語音声認識システム，Sakti Sakriani，久保慶伍，松宮翔，Graham Neubig，戸田智基，中村哲 (NAIST)，安達史博，磯谷亮輔 (NEC)

(denoted as SAT-ML) and boosted maximum mutual information (MMI)[6] discriminative training (denoted as SAT-bMMI).

We utilize the existing pronunciation dictionary: English CMU dictionary [7], Japanese CSJ dictionary, and Chinese BTEC dictionary, respectively. After that, we constructed a dictionary that would be used for medical domain. The out-of-vocabulary words were constructed based on Structured AROW G2P conversion [8]. The resulting pronunciation dictionary contains 50K, 40K and 33K vocabulary for Japanese, English, and Chinese, respectively.

Using the SRILM toolkit [9], we built n-gram language models with modified Kneser-Ney smoothing [10] from each of the text corpora (Medical, TED, and BTEC data). These were then combined using linear interpolation in which the weights were chosen to maximize the likelihood of a held-out medical development data set. The resulting language model contains 420K trigrams of Japanese, 300K trigrams of English, and 150K trigrams of Chinese.

Our decoding algorithms use weighted finite state transducers (WFSTs) [11] based on Kaldi [12]. Figure 1 shows the performance of our Japanese, English, and Chinese ASR system on the medical development and test sets. All systems could achieve a WER below 20%. SAT-bMMI provides a significant improvement achieving a WER of 14.38% for Japanese, a WER of 13.21% for English and a WER of 9.87% for Chinese on the medical test set.

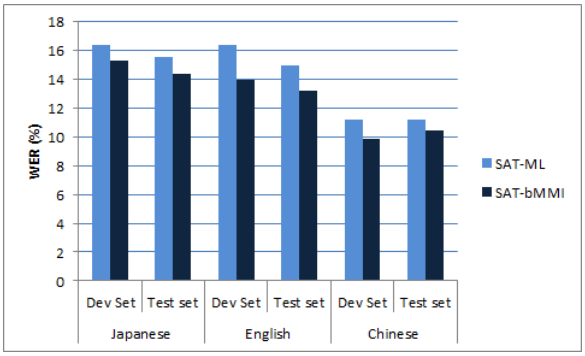

Fig. 1 The performance of our Japanese, English, and Chinese ASR system.

Currently, these Japanese, English and Chinese speech recognition systems have been implemented into Web servers. A user speaks an utterance on a client application. The speech signal is then send to the server and the server performs a speech-recognition operation and transfers the result back to the client by TCP/IP connection.

4 Conclusion

In this paper, we described our multilingual data collection and multilingual speech recognition system for a speech translation system that was designed to facilitate multilingual conversations in medical situations. Our final speech recognition system is based on a weighted finite-state transducers framework utilizing feature transformation, speaker adaptive training, the boosted maximum mutual information discriminative criterion and n-gram language models. Experimental results reveal that SAT-bMMI provide significant improvement achieving a WER of 14.38% for Japanese, a WER of 13.21% for English and a WER of 9.87% for Chinese on a medical test set.

5 Acknowledgements

Part of this work was supported by the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

参考文献

[1] G. Neubig, S. Sakti, T. Toda, S. Nakamura, Y. Matsumoto, R. Isotani, and Y. Ikeda, "Towards high-reliability speech translation in the medical domain," in *Proc. MedNLP*, Japan, 2013, pp. 22–29.

[2] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. LREC*, Greece, 2000, pp. 947–952.

[3] G. Kikui, , S. Yamamoto, T. Takezawa, and E. Sumita, "Comparative study on corpora for speech translation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1674–1682, 2006.

[4] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *CSL*, vol. 12, no. 2, pp. 75–98, 1998.

[5] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP*, USA, 1996, pp. 1137–1140.

[6] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. ICASSP*, USA, 2008, pp. 4057–4060.

[7] "The CMU pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[8] K. Kubo, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Grapheme-to-phoneme conversion based on adaptive regularization of weight vectors," in *Proc. INTERSPEECH*, France, 2013, pp. 1946–1950.

[9] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. ICSLP*, USA, 2002, pp. 901–904.

[10] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. ICASSP*, USA, 1995, pp. 181–184.

[11] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *CSL*, vol. 20, no. 1, pp. 69–88, 2002.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Moticek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, USA, 2011.