

Towards Multilingual Conversations in the Medical Domain: Development of Multilingual Medical Data and A Network-based ASR System

Sakriani Sakti¹, Keigo Kubo¹, Sho Matsumiya¹, Graham Neubig¹,
Tomoki Toda¹, Satoshi Nakamura¹, Fumihiko Adachi², Ryosuke Isotani²

¹Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, Japan
{ssakti,keigo-k,sho-m,neubig,tomoki,s-nakamura}@is.naist.jp

²NEC Corporation
5-7-1 Shiba, Minato-ku, Tokyo, Japan
{f-adachi@aj,r-isotani@bp}.jp.nec.com

Abstract

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. The overall speech-to-speech translation (S2ST) system was designed to translate spoken utterances from a given source language into a target language in order to facilitate multilingual conversations and reduce the problems caused by language barriers in medical situations. Our final system utilizes a weighted finite-state transducers with n-gram language models. Currently, the system successfully covers three languages: Japanese, English, and Chinese. The difficulties involved in connecting Japanese, English and Chinese speech recognition systems through Web servers will be discussed, and the experimental results in simulated medical conversation will also be presented.

Keywords: multilingual data, speech recognition, medical domain

1. Introduction

The ongoing proliferation of information technology is having an increasingly large impact on many aspects of our daily lives. As globalization rapidly expands, the nature and volume of world migration has also changed, and the interconnectedness of people across national boundaries has increased. On the other hand, language barriers has become the greatest obstacle to effective communication among people speaking different languages. The issue of communication barriers is particularly critical during intercultural health care, as miscommunication between patient and provider can lead to misdiagnosis and inadequate, even fatal, medical care.

In Japan, the total number of foreign nationals has been increasing year by year with the number of registered foreign residents at the end of 2008 reaching a record high of 2.22 million, and with the percentage standing at 1.74% of the total population of Japan (Japanese Ministry of Justice, 2010). Given the rapid growth in the foreign resident population, health care organizations can no longer focus solely on meeting the needs of the Japanese population, but also the needs of foreign nationals. Overcoming language barriers to health care is critical to the well-being of millions of immigrants. One way is to provide qualified medical interpretation services for a growing population of patients who don't understand local language. However, interpretation is a complex skill that takes many years to master. Consequently, professional interpretation services are expensive. The use of volunteer interpreters may be more cost-effective solution, but disadvantages include compromising the patient's right to privacy and relying on someone without training as an interpreter.

Spoken language translation is one of the innovative tech-

nologies that may help in situations where no common language between the diagnosing doctor and the patient exists. Yet developing such a system poses a challenging task because it integrates a set of complex technologies which attempt to simulate sophisticated human intellectual activities: automatic speech recognition (ASR), machine translation (MT), and text-to-speech synthesis (TTS). An ASR system should be robust enough to recognize speech in noisy environments and with different speaking styles, an MT system must be domain portable and achieve high translation quality for a wide variety of topics, and a TTS system must realize more natural and expressive speech quality. It should also be noted that the ultimate goal of the system is to surmount language barriers and facilitate free communication among those speaking different languages. Thus, the system should realize multilingual speech translation, which requires constructing ASR, MT, and TTS systems for all possible source and target language pairs.

In this paper, we focus on the development of our multilingual ASR system for network-based speech-to-speech translation in the medical domain. The overall speech translation system was designed to translate spoken utterances from a given source language into a target language in order to facilitate multilingual conversations and reduce the problems caused by language barriers in medical situations. Currently, the system successfully covers three languages: Japanese, English, and Chinese. Here, we also discuss the issues of multilingual speech data collection as well as the difficulties involved in connecting Japanese, English and Chinese speech recognition systems through Web servers. Furthermore, the experimental results in simulated medical conversation will also be presented.

In the following sections, an overview of speech translation

system and multilingual medical data construction are described. ASR systems development is described in Section 4, and the integration into network-based S2ST is detailed in Section 5. Experiment results are reported in Section 6, and conclusions are drawn in Section 7.

2. Overview of Medical Speech-to-Speech Translation System

Figure 1 illustrates an example of the overall architecture of our medical speech translation system, in which a spoken Japanese utterance of the medical practitioners is translated into English or Chinese target speech of the patient, and vice versa. This mechanism can be performed within two scenarios: (1) when the patient first enters the hospital and communicate with receptionists; (2) when the patient enter the doctor’s office and communicate with doctors/nurses during medical diagnosis.

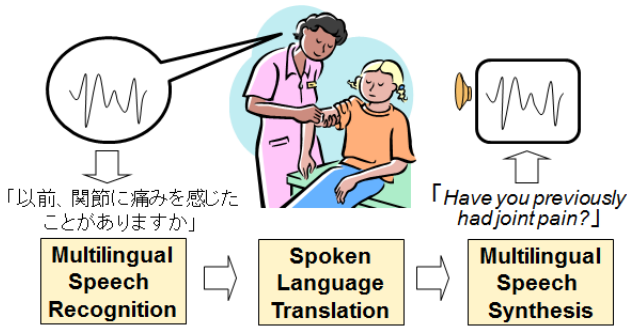


Figure 1: Overall architecture of speech translation system in medical situations.

The speech recognition module recognizes the Japanese input speech and converts it into a string of Japanese words, so that the probability of the string of words is maximized. These words are then translated by the machine translation module, replacing each Japanese phrase in the string with the corresponding English or Chinese phrase. These phrases are then appropriately rearranged based on a re-ordering model trained from word alignments in a bilingual corpus. Finally, the text-to-speech synthesis module produces the speech waveform of the resulting English or Chinese sentence.

3. Multilingual Medical Text and Speech Data Design and Construction

To provide an initial testbed for our medical translation experiments, we design and collect the following Japanese, English, and Chinese text and speech materials.

3.1. Text Materials

To build speech translation system in medical domain, we design and collect the following two corpora:

- Medical Phrasebooks

It is designed based on sentences from Japanese-English bilingual phrasebooks designed for interpreters focusing on the medical domains. Thus, it has

good coverage of medical-domain terminology. However, the conversations may not exactly representative of the conversations that actually occur during doctor’s visit. In total, we had 5130 Japanese-English sentences. Chinese translations were obtained by translating each phrase from Japanese to Chinese. We then divide it into training, development and test set as described in Table 1.

Data Set	Sent	Word		
		Ja	En	Zh
Train	3420	68k	43k	38k
Dev	855	17k	12k	9.6k
Test	855	17k	12k	9.6k

Table 1: Size in sentences and words of Japanese (Ja), English (En), Chinese (Zh) medical phrasebook.

- Medical Conversation

It consists of actual conversations between the patient and the receptionists, nurses or doctors recorded during a doctor’s visit. The doctors and receptionists were all actual practitioners, but for privacy reasons the person acting as a patient was actually healthy, but given a scenario to act out. 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. This corpus has the advantage of being highly natural and covering medical domain terminology. In total, we had 1007 sentences, which then divided into training, development and test set as described in Table 2.

Data Set	Sent	Word		
		Ja	En	Zh
Train	671	5.6k	4.7k	3.4k
Dev	168	1.4k	1.3k	900
Test	168	1.5k	1.2k	880

Table 2: Size in sentences and words of Japanese (Ja), English (En), Chinese (Zh) medical conversation.

More details of both text materials can be found in (Neubig et al., 2013).

3.2. Speech Materials

From the development and test set of medical phrasebooks and conversation text data described above, 200 sentences (100 sentences development set, 100 sentences of test set) of medical dialog conversation were selected and the recordings was conducted in a sound proof room, at a 48 kHz sampling rate with 16 bits resolution. The sampling rate was later down-sampled to 16 kHz for our experiments. For each Japanese, English, and Chinese language, there were 27 speakers with balance of gender (males, females) and age (twenties, thirties, forties) as described in Table 3. Each speaker uttered either 100 sentences from development set or test set, resulting in a total of 27,000 utterances per language.

Japanese							
	Dev			Test			Total
	20s	30s	40s	20s	30s	40s	
Males	1	4	2	2	3	3	15
Females	1	3	2	1	3	2	12
Total	2	7	4	3	6	5	27
English							
	Dev			Test			Total
	20s	30s	40s	20s	30s	40s	
Males	1	3	3	1	4	2	14
Females	2	3	2	1	3	2	13
Total	3	6	5	2	7	4	27
Chinese							
	Dev			Test			Total
	20s	30s	40s	20s	30s	40s	
Males	1	3	2	2	4	2	14
Females	1	3	2	1	3	3	13
Total	2	6	4	3	7	5	27

Table 3: Speaker distribution of Japanese (Ja), English (En), Chinese (Zh) medical speech data.

4. Multilingual Speech Recognition System

4.1. Training and Test Resources

To evaluate the performance of our speech recognition system, development and test set of Japanese, English, Chinese medical speech data described in previous section were used. For English, Japanese, Chinese acoustic model training, we utilize the following speech corpora:

- Japanese
Corpus of Spontaneous Japanese (CSJ)(Maekawa et al., 2000) which is a richly annotated speech and language database of spontaneous speech. It contains 658 hours of speech consisting of approximately 7.5 million words. The speech materials were provided by more than 1,400 speakers of ages ranging from twenties to eighties. About 95% of the CSJ is devoted to spontaneous monologues (academic presentations, public speaking), and the remaining consists of spontaneous dialogues and reading aloud. The overall corpus is divided into training and test set. In this project, we used only the training part with about 518 hours of speech.
- English
Open-domain spontaneous speech of TED Talks¹ downloaded from the TED websites with the corresponding subtitles. TED talks bring together the world’s most fascinating thinkers and doers, who are challenged to give the talk of their lives in about 5-25 minutes covering topics related to technology, entertainment and design (TED). Spanning everything, from internet trends to solving the world’s water supply problems, today TED is a global movement ”riveting talks by remarkable people free to the world”. Here, we have successfully collected 157 hours of speech from about 800 TED talks.

¹<http://www.ted.com/talks>

- Chinese

The ATR basic travel expression corpus (BTEC) has served as the primary source for developing broad coverage speech translation systems (Kikui et al., 2006). The sentences were collected by bilingual travel experts from Japanese/English sentence pairs in travel domain phrasebooks. ATR-BTEC has also been translated into 18 different languages including French, German, Italian, Chinese, Korean, and Indonesian. Each language consists of 160,000 sentences (with about 20,000 unique words) of training set and 510 sentences of test set with 16 references per sentence. Here, we use the Chinese data covering 1600 Chinese speakers with Putonghua, Beijing, Shanghainese, Cantonese, and Taiwan accents, resulting in a total of 250 hours of speech utterances.

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 565k sentences of Japanese, 519k sentences of English, and 260k of Chinese.

4.2. Training Procedure

4.2.1. Front-End Processing

We trained the systems with a front-end based on the widely used mel-frequency cepstral coefficients (MFCC). The frontend provides features every 10ms with 25ms width. For each utterances in the speech training data, 13 static MFCCs including zeroth order for each frame are extracted and normalized with cepstrum mean normalization. To incorporate the temporal structures and dependencies, 9 adjacent (center, 4 left, and 4 right) frames of MFCCs are stacked into one single feature vector leading to 117 dimensional super vectors (9x13 dimensions). These are then projected down to an optimum 40 dimensions by applying a linear discriminant analysis (LDA). After that, the resulting features are further de-correlated using maximum likelihood linear transformation (MLLT)(Gopinath, 1998), which is also known as global semi-tied covariance (STC)(Gales, 1998a) transform. Moreover, speaker adaptive training (SAT)(T. Anastasakos and Makhoul, 1996) is performed using a single feature-space maximum likelihood linear regression (fMLLR)(Gales, 1998b) transform estimated per speaker.

4.2.2. Acoustic Model Training

Acoustic models are trained on the LDA+STC+fMLLR features describe above. All models are context-dependent cross-word triphone with a standard three-state left-to-right HMM topology without skip states. The HMM units are derived from 39 phonemes of Japanese and English, and 56 phonemes of Chinese. Each phoneme is classified by its position in word (4 classes: begin, end, internal and singleton). Additionally, we added 9 special phoneme of non-speech sounds derived from TED speech sources. These include *SIL*, *SENTSTART* and *SENTEND* for silence, and *APPLAUSE*, *BEEP*, *LAUGHTER*, *MUSIC*, *NOISE*, and *VOICENOISE* for noise sounds that may appeared in speech sources.

The context-dependent cross-word triphone HMMs was first trained with GMM output probability. This model totally include 80k Gaussians trained with both speaker adaptive training (SAT)(T. Anastasakos and Makhoul, 1996) maximum likelihood (ML) estimation (denoted as SAT-ML) and boosted maximum mutual information (MMI)(Povey et al., 2008) criterion of discriminative training (denoted as SAT-bMMI).

4.2.3. Dictionary Construction

We normalized the training data sources of Medical, TED and BTEC, in a case-insensitive fashion. For tokenization we use the Chasen Tokenizer/Segmenter for Japanese (Matsumoto et al., 2000), and the KyTea segmenter for Chinese (Neubig et al., 2011). We utilize the existing pronunciation dictionary: (1) the CMU pronunciation dictionary of English (Carnegie Mellon University, 2007); (2) the CSJ pronunciation dictionary of Japanese; and (2) the BTEC pronunciation dictionary of Chinese. After that, we constructed a dictionary that would be used for medical domain. The pronunciations of out-of-vocabulary words were constructed based on Structured AROW G2P conversion (Kubo et al., 2013) which is an online discriminative training that extends AROW (Crammer et al., 2009) to structured learning (*SAROW*). The resulting pronunciation dictionary contains 50k, 40k and 33k vocabulary for Japanese, English, and Chinese, respectively.

4.2.4. Language Model Training

Using the SRILM toolkit (Stolcke, 2002), we built n-gram language models with modified Kneser-Ney smoothing (Kneser and Ney, 1995) from each of the text corpora (Medical, TED, and BTEC data). These were the 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 550k bigrams and 420k trigrams (with a perplexity of 25.11) for Japanese, 550k bigrams and 300k trigrams (with a perplexity of 39.71) for English, and 380k bigrams and 150k trigrams (with a perplexity of 79.97) for Chinese.

5. Integration on Network-based Speech Translation System

Figure 2 shows an example of the server-client speech-recognition scheme and Table 4 summarizes the configuration of the client and server development platform, including the library, operating environment, and hardware. 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. The input speech format is 16-kHz, 16-bit, mono-channel, and the output format is text string.

6. Decoding System and Evaluation Results

6.1. Decoding System

Our decoding algorithms use weighted finite state transducers (WFSTs) (Mohri et al., 2002) based on Kaldi (Povey et al., 2011), a free, open-source toolkit for speech recognition

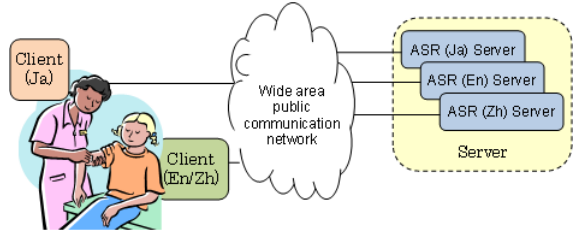


Figure 2: Example of the server-client speech-recognition scheme.

Table 4: Client and server configuration.

	Client	Server (ASR, MT, TTS)
Development Library	C platform	C++ platform
Operating Environment	Linux	Linux
Hardware	PC terminal	PC terminal (Minimum RAM 8GB)

research. The decoding-graph construction process is basically based on the conventional recipe described in (Mohri et al., 2002) with slightly modification to allow different phones to share the same context-dependent states.

Here, for each Japanese, English and Chinese language, we employ 3-gram LM WFST-based decoder with the Kaldi toolkit. Two types of AMs described in Section 4.2.2., (a) SAT-ML and (b) SAT-bMMI, are employed individually, with n-gram LMs described in Section 4.2.4.. This step produces two lattices corresponding to the two AMs.

6.2. Evaluation Results

Figure 3 shows the performance of our Japanese, English, and Chinese ASR system on the medical development and test sets. All systems could achieves a WER below 20%. 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 medical test set.

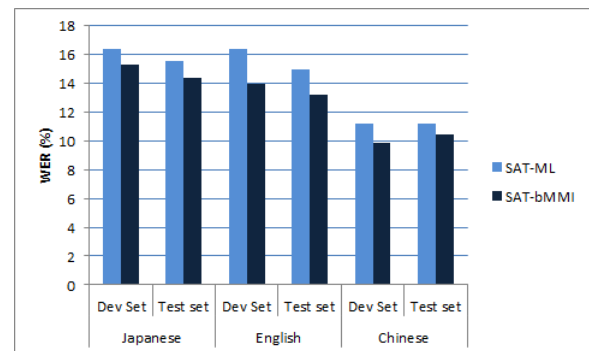


Figure 3: The performance of our Japanese, English, and Chinese ASR system on the medical development and test sets.

7. Conclusion

In this paper, we described our multilingual data collection and multilingual speech recognition system for speech translation system that was designed to translate spoken utterances from a given source language into a target language in order 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, 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 medical test set. In the future, we will improve to extend the vocabulary of the system leading to a large vocabulary in medical domain.

8. Acknowledgements

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

9. References

- Carnegie Mellon University. (2007). The carnegie mellon university pronouncing dictionary. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- Crammer, K., Kulesza, A., and Dredze, M. (2009). Adaptive regularization of weight vectors. *Advances In Neural Information Processing Systems*, pages 414–422.
- Gales, M. (1998a). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12(2):75–98.
- Gales, M. (1998b). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12(2):75–98.
- Gopinath, R. (1998). Maximum likelihood modeling with gaussian distributions for classification. In *Proc. of ICASSP*, pages 661–664.
- Japanese Ministry of Justice. (2010). Basic plan for immigration control. http://www.immimoj.go.jp/seisaku/keikaku_101006_english.pdf.
- Kikui, G., Yamamoto, S., Takezawa, T., and Sumita, E. (2006). Comparative study on corpora for speech translation. *IEEE Trans. on Audio, Speech and Language Processing*, 14(5):1674–1682.
- Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proc. ICASSP*, pages 181–184.
- Kubo, K., Sakti, S., Neubig, G., Toda, T., and Nakamura, S. (2013). Grapheme-to-phoneme conversion based on adaptive regularization of weight vectors. In *Proc. INTERSPEECH*, pages 1946–1950.
- Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous speech corpus of Japanese. In *Proc. LREC*, pages 947–952.
- Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., and Asahara, M. (2000). Morphological analysis system ChaSen version 2.2.1 manual. Technical Report NAIST-IS-TR99009, Nara Institute of Science and Technology.
- Mohri, M., Pereira, F., and Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 20(1):69–88.
- Neubig, G., Nakata, Y., and Mori, S. (2011). Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proc. ACL*, pages 529–533, Portland, USA.
- Neubig, G., Sakti, S., Toda, T., Nakamura, S., Matsumoto, Y., Isotani, R., and Ikeda, Y. (2013). Towards high-reliability speech translation in the medical domain. In *Proc. of Workshop on Natural Language Processing for Medical and Healthcare Fields (MedNLP)*, pages 22–29, Nagoya, Japan.
- Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., and Visweswariah, K. (2008). Boosted MMI for model and feature-space discriminative training. In *Proc. of ICASSP*, pages 4057–4060, Las Vegas, USA.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Moticek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proc. ASRU*, Hawaii, USA.
- Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proc. of ICSLP*, pages 901–904, Denver, USA.
- T. Anastasakos, J. McDonough, R. S. and Makhoul, J. (1996). A compact model for speaker adaptive training. In *Proc. ICSLP*, pages 1137–1140.