Towards High-Reliability Speech Translation in the Medical Domain

Graham Neubig[1], **Sakriani Sakti**[1], **Tomoki Toda**[1], **Satoshi Nakamura**[1],
Yuji Matsumoto[1], **Ryosuke Isotani**[2], **Yukichi Ikeda**[2]

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

[2] NEC Corporation
5-7-1 Shiba, Minato-ku, Tokyo, Japan
{r-isotani@bp,y-ikeda@df}.jp.nec.com

Abstract

In this paper, we describe the overall design for a speech translation system that aims to reduce the problems caused by language barriers in medical situations. As first steps to building a system according to this design, we describe a collection of a medical corpus, and some translation experiments performed on this corpus. As a result of the experiments, we find that the best of three modern translation systems is able to translate 33%-81% of the sentences in a way such that the main content is understandable.

1 Introduction

One of the most important elements to provision of high-quality medical service is communication between medical practitioners and patients. However, in situations where practitioners and patients do not share a common language, the language barrier prevents effective communication, making proper diagnosis and treatment much more difficult. Language barriers occur in medical situations with immigrants who may speak the language of their country of residence to some extent, but not enough to effectively communicate medical symptoms. There is also the case of medical tourism, where tourists may visit another country to receive high-quality or affordable medical treatment that is not available in their home country.

One potential method for overcoming the communication barrier in medical situations is through the use of automatic speech translation technology (Nakamura, 2009). Automatic translation of speech in medical situations can be expected to be challenging for a number of reasons. The first reason is that communication of incomplete or incorrect information could lead to a mistaken diagnosis with severe consequences, and thus extremely high levels of *accuracy* and *reliability* are required. The second reason is that conversation in the medical domain has its own unique vocabulary and expressions, and thus it is natural to assume that we must *adapt* the system appropriately to the medical domain.

There has been some previous work attempting to adapt communication technology to meet these two challenges. Eck et al. (2004) focus on adapting a translation system to medical vocabulary, although the focus on text translation of medical documents instead of speech translation for communication. Miyabe et al. (2007) propose a system for reliable multilingual communication, but rely on a graphical interface that is something like a powerful bilingual phrasebook adapted to communication at a hospital reception desk.

In this paper, we describe our vision for full speech translation in medical situations, and some first steps to achieve this vision. First, in Section 2 we describe our overall design for the speech translation system. This system includes the common components of automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS), augmented to adapt each component to the task at hand. We also consider what is necessary to ensure the reliability of translation results, and consider the use of a system to allow the conversation to be forwarded to human medical interpreters when necessary.

In the first step towards achieving a translation system for the medical domain we have also collected a medical-domain corpus for Japanese-English and Japanese-Chinese translation, as described in Section 3. We share some insights gained in collecting this corpus, particularly comparing and contrasting text data from a medical domain bilingual phrasebook, and actual conversational data gathered during doctors' visits.

Based on this data, we then build several prototype translation systems for the four language pairs as described in Section 4. We perform au-

International Joint Conference on Natural Language Processing Workshop on Natural Language Processing for Medical and Healthcare Fields, pages 22–29,
Nagoya, Japan, 14-18 October 2013.

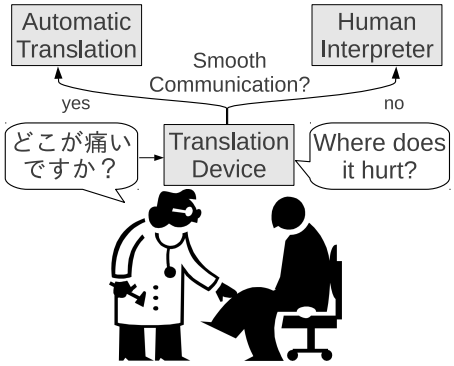

Figure 1: An overview of the use scenario for the medical translation system.

tomatic and manual evaluation of the results and evaluate how close we are to our goal of creating a system that can provide a first wave of assistance in medical situations. In particular, we find that over all four (relatively difficult) language pairs, we are on our way towards creating a practical medical machine translation system, with from 33%-81% of sentences over two tasks and four language pairs having all content understandable with some effort.

Finally, in Section 5 we conclude the paper with a discussion of future work.

2 Medical Translation System Design

We show the overall use scenario in Figure 1. In a typical doctor's visit, patient first enters the doctor's office, speaks with receptionists, and fills out forms. The patient will then enter the doctor's office and communicate with doctors/nurses. In order to introduce a speech translation system for use in this scenario, we will provide a device that will then translate between the language of the patient and that of the medical practitioners.

The device provides two possible methods of communication. The first is through the use of automatic speech translation technology, where the speaker's voice will be recorded, recognized, translated, and synthesized entirely automatically. In addition, as speech translation technology is still far from perfect, the device will also have the ability to connect to an actual human medical interpreter located in a call center. However, as the cost of hiring and maintaining medical interpreters is quite high, we would also like to reduce our reliance on human effort as much as possible. Thus, each device will use automatic translation by default, but also have functionality to connect to human interpreters, either based on a manual request of one of the users, or through automatic detection of when the dialogue is going poorly, such as the method described by Walker et al. (2000).

Even with this fall-back to human interpreters, it is still desirable that the automatic translation system is effective as possible. In order to ensure this, we must be certain that the ASR, MT, and TTS models are all tuned to work as well as possible in medical situations. Some potential problems that we have identified so far based on our analysis of data are as follows:

Specialized Vocabulary: Perhaps the most obvious problem is that the ASR, MT, and TTS systems must all be able to handle the specialized vocabulary and usage that occurs in the medical domain. For example, most we found that unadapted systems had trouble handling specialized terms such as "cardiogram," medicine names such as "Sudafed," and disease names such as "chicken pox." This will require the creation of domain specific corpora/dictionaries, and domain adaptation for each of the components (Leggetter and Woodland, 1995; Bellegarda, 2004).

Conversational Speech: The speech during doctor's visits will generally be somewhat informal and conversational when compared to that of speeches, news, or other more formal locations. As a result, we can expect ASR to be more difficult due to fillers, disfluencies and other factors (Goldwater et al., 2010).

Translation/Synthesis of Erroneous Input: As we can expect ASR not to be perfect, it will be necessary to be able to translate input that contains errors. This problem can potentially be ameliorated by passing multiple speech recognition hypotheses to translation (Ney, 1999), and jointly optimizing the parameters of ASR and MT (Zhang et al., 2004; Ohgushi et al., 2013). In addition, it will also be necessary to resolve difficulties in TTS due to grammatical errors, lack of punctuation, and unknown words (Parlikar et al., 2010).

While all of these problems need to be solved to provide high-reliability speech translation systems, in this paper as a first step we focus mainly on the MT system, and relegate the last problem of integration with ASR to future work.

3 Medical Translation Corpus Construction and Analysis

In this section, we describe our collection of a tri-lingual (Japanese, English, Chinese) corpus to serve as an initial testbed for our medical translation experiments, and an analysis of the corpus.

3.1 Corpus Construction

In general, when creating a corpus for training/testing a machine translation system, it is important to collect content that is as close as possible to that which we will encounter when the system is actually used. In our medical translation situation, this is true for both vocabulary (the corpus must cover special medical terms) and for speaking style (the corpus must have a similar style to that used by actual doctors and patients speaking through the system). There is also the practical concern that the cost of corpus collection is high, so we would like to perform collection in efficient a manner as possible.

Based on these principles, we designed and collected the following two corpora:

Medical Phrasebooks: The first corpus consists of sentences 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. This corpus has the advantages of relatively efficient construction, and good coverage of medical-domain terminology, but the conversations are not necessarily exactly representative of the conversations that actually occur at a doctor's office.

Medical Conversation: The second corpus we gathered consists of actual conversations between the patient and the receptionists or doctors recorded during a doctor's visit. The doctors and receptionists were all actual practicioners, 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, but requires a large amount of time and effort for the creation of scenarios, gathering the participants, execution of the actual dialog, and transcription/translation of the results.

		Sent.	Word ja	Word en	Word zh
Phrase	Train	3420	68k	43k	38k
	Dev	855	17k	12k	9.6k
	Test	855	17k	12k	9.6k
Conv.	Train	671	5.6k	4.7k	3.4k
	Dev	168	1.4k	1.3k	900
	Test	168	1.5k	1.2k	880

Table 1: Size in sentences and words of each language for each split for the phrasebook and conversation corpora.

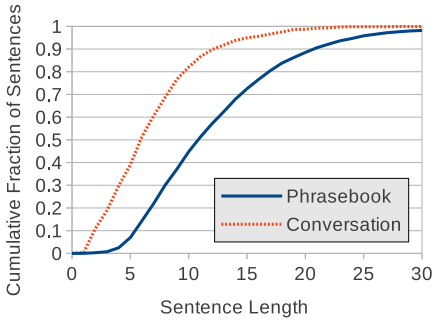

Figure 2: The cumulative length distribution of sentences in each corpus.

At the end of the collection, we had 5130 and 1007 sentences for the phrasebook and conversation corpus respectively. In addition, we create three splits of the corpus for use in the training, tuning, testing of our machine translation system with a ratio of 4:1:1. The final size of the data in all of these corpora is shown in Table 1.

3.2 Corpus Analysis

In this section, we describe some insights gained from the analysis of both corpora, with some examples illustrating the features of each corpus in Table 2.

One feature of the data with major implications is that there were large differences in speaking style between the phrasebook and conversation corpora. The data in the phrasebook corpus generally consisted of longer sentences, while the

Phrase	1)	I was sewing my jeans using a sewing machine and the needle broke and stabbed my left cheek.
	2)	I have been told that I have early indications of liver cirrhosis.
Conv.	1)	No more than two vials of blood. Possibly three, if for a blood sugar test.
	2)	Go straight, and on your left there is a green chair. / Here? Which way should I face?

Table 2: Examples of sentences (or several sentences separated by slashes) from the phrasebook and conversational corpora.

majority of the utterances in the conversation corpus contained short questions, requests, responses, and commands. This trend of longer sentences in the phrasebook corpus is shown clearly in Figure 2, which shows the cumulative length distribution of English sentences under a certain length in both corpora. Focusing on sentences under length 15, we can see that this covers a total of 95% of the conversation corpus, but only 72% of the phrasebook corpus.

In addition, the language in the conversation corpus is significantly less formal, particularly in Japanese where spoken language includes features such as dropped subjects or particles and abbreviations, which rarely occur in written language (Neubig et al., 2012). We hypothesize that in a cross-lingual medical conversation situation, the content of the utterances will fall somewhere between these two situations, as the content will be conversational, but the kind of natural and informal interaction seen two native speakers will be difficult to achieve through an automatic translation system.

The second enlightening feature of the two corpora that we noticed was that medical terminology was significantly less prevalent in the conversation corpus. This is also natural, as actual patients to a doctors office will likely be unfamiliar with difficult medical terms, and thus the doctors will tend to explain in language that is understandable for their audience. This observation will likely carry over to computer-mediated medical communication as well. As a result, it is likely that adapting to medical terminology of the domain is somewhat less important than adapting to the conversational speaking style of the speech.

4 Preliminary Evaluation of Medical Machine Translation

In this section we describe a preliminary evaluation of the effectiveness of automatic translation on the medical domain data described in the previous section. In particular, we focus on the MT component, leaving evaluation of ASR, TTS, and the system as a whole for future work.

4.1 Experimental Setup

For the tuning and test data for our translation system, we use the data described in the previous section. For training, 4,000 sentences is not enough to build an accurate MT system, so we add several additional corpora for each language pair. For Japanese-English parallel training data, we add the Eijiro dictionary[1] and its accompanying sample sentences, the BTEC corpus(Takezawa et al., 2002), and Wikipedia data from the Kyoto Free Translation Task (Neubig, 2011), for a total of 1.33M parallel sentences and 1.97M dictionary entries. For Japanese-Chinese parallel training data, we add a dictionary extracted from Wikipedia's language links[2], the BTEC corpus, and TED talks (Cettolo et al., 2012) for a total of 519k sentences and 184k dictionary entries. In addition, we add monolingual from English GigaWord with 22.5M sentences and Chinese Wikipedia with 841k sentences.

We compare three different statistical translation methodologies: phrase-based MT (PBMT, (Koehn et al., 2003)), hierarchical phrase-based MT (Hiero, (Chiang, 2007)), and forest-to-string MT (F2S, (Mi et al., 2008)). The reason why we test these three methodologies is because the former two methodologies do not rely on syntactic analysis, and thus may be more robust to conversational input that is ill-formed and/or informal. On the other hand, using syntactic information has been shown to improve translation, particularly between language pairs with different syntactic structures such as those we are handling in our experiments. Thus it will be interesting to see which methodology can produce better results, and also if any difference in the effectiveness of

[1] http://eijiro.jp
[2] http://wikipedia.org

the methodologies will be seen between the two corpora.

For training the translation models and decoding, we use the Moses toolkit (Koehn et al., 2007) for PBMT and Hiero, and the Travatar (Neubig, 2013) toolkit for F2S with the default settings. For training language models, we use SRILM (Stolcke, 2002), training Kneser-Ney smoothed 5-gram models for each individual language model training corpus, and linearly interpolating these models to maximize likelihood on the tuning corpus.

For tokenization we use the Stanford Tokenizer/Segmenter for English and Chinese (Tseng et al., 2005), and the KyTea segmenter for Japanese (Neubig et al., 2011). For syntactic parsing in English and Chinese we use a modified version of the Egret parser,[3] and for Japanese we use the Eda parser (Flannery et al., 2011) and the dependency-to-CFG conversion rules in the Travatar toolkit. Alignment is performed using the unsupervised aligner GIZA++ (Och and Ney, 2003) for Japanese-Chinese, and the supervised aligner Nile for Japanese-English (Riesa and Marcu, 2010), with the alignment models being trained on the alignments distributed with the Kyoto Free Translation Task.[4]

To measure translation accuracy, we use the automatic evaluation measures of BLEU (Papineni et al., 2002) and RIBES (Isozaki et al., 2010) measured over all sentences in the test corpus. We also perform a manual evaluation on 120 sentences from the phrasebook corpus and 80 sentences from the conversation corpus. These were randomly selected from all sentences of length 1-30, and graded using 1-5 adequacy (Goto et al., 2011) as our evaluation measure. We also report the percentage of sentences that received a rating of greater than or equal to 2, indicating that the main points of the sentence can be understood, possibly with some difficulty.

4.2 Experimental Results

The results of the experimental evaluation are shown in Figure 3. This graph shows many results, but we first focus on the furthestmost right graph, which shows the percentage of sentences understandable to some extent for each of the systems. From this graph, we can see that the scores range from 81% understandable sentences for Japanese-Chinese phrasebook sentences, to only 33% understandable sentences on Japanese-English phrasebook sentences. On the other hand, for conversational sentences, most language pairs hovered at around 55% understandable, with Japanese-English being significantly worse.

An in-depth analysis of the mistaken sentences identified several issues for improvement that were generally shared by all three systems.

Omitted pronouns: Japanese is a pro-drop language, which means that pronouns, usually the subject of the sentence can be omitted and inferred from the context. This phenomenon is particularly prevalent in the types of dialogue contained in the conversation corpus, with the majority of sentences having their subject omitted. Given that it is difficult for the translation systems used in the experiments to accurately reproduce these omitted subjects in a non-pro-drop target language such as English, it is likely that replacing these subjects in a preprocessing step would lead to gains in accuracy (Taira et al., 2012)

Dropped words: There were many cases where words central to the sentence were missing from the translation output by the system. This problem is rooted in a number of problems, such as words being mistakenly unaligned in the training data.

Word segmentation: Both Chinese and Japanese require the segmentation of raw text into words, but occasionally word segmentation errors occurred either due to conversational speech or specialized medical terms. Thus, using domain adaptation techniques (Neubig et al., 2011) to fix the word segmentations in the medical domain could potentially improve down-stream accuracy of translation as well.

Medical domain terms: As expected, there were a few medical domain terms not covered by corpora from more general domains, such as "Benadryl." However, the number was also relatively small, with only 5 untranslatable words occurring in a 200 sentence corpus.

Overall, an interesting shared point between the majority of members of the list is that they are not

[3] http://github.com/neubig/egret

[4] This preprocessing pipeline is available as part of the Travatar toolkit: http://phontron.com/travatar/preprocessing.html

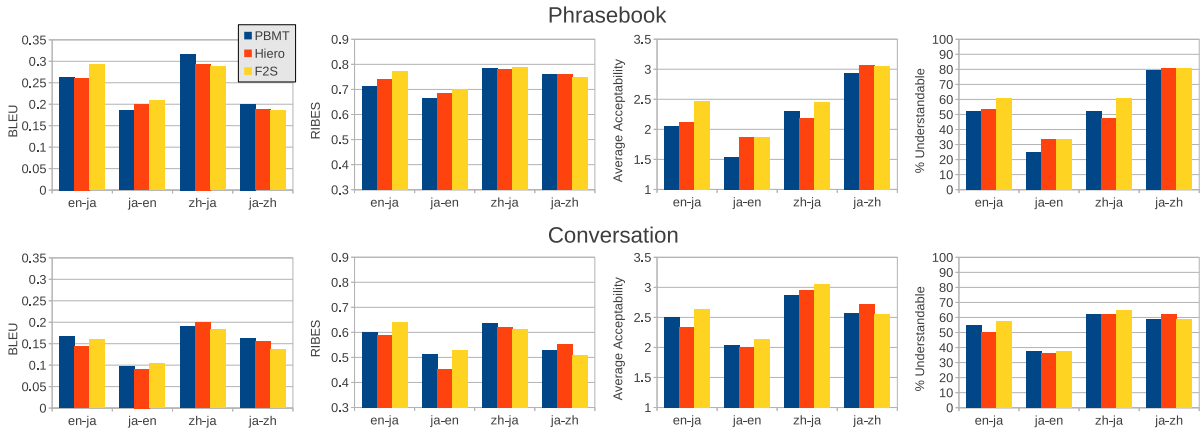

Figure 3: Results in BLEU, RIBES, Average Acceptability, and % Understandable for phrase-based (PBMT), hierarchical phrase based (Hiero), and forest-to-string (F2S) systems over translation of medical phrasebooks and conversations.

1	Input	糖尿病でインスリンを使用しています。
	Ref	I use insulin because I have diabetes.
	PBMT	I have diabetes using insulin.
	Hiero	Diabetes has been using insulin?
	F2S	I have been using insulin for diabetes.
2	Input	反対の目をお願いします。
	Ref	Let's test the other eye.
	PBMT	Other eye, please.
	Hiero	Please check your other eye.
	F2S	I'd like other eye.
3	Input	もう一回眼をぱちぱちしてください。
	Ref	Once again, open and close your eyes.
	PBMT	Their eyes again, please.
	Hiero	Their eyes again, please.
	F2S	Please ぱちぱち eye again.

Table 3: Examples of translations generated by each system for Japanese-English.

specific to medical translation, but more related to the style of the text. Thus while raising the level of medical MT will certainly involve covering medical terminology, it is also equally, if not more, important to overcome obstacles facing the more general speech translation task as well.

Finally, in Table 3, we show concrete examples for each of the three translation methods in Japanese-English translation. The first example is from the phrasebook data, uses some medical terms, and has a very typical syntactic structure for a written Japanese sentence. As a result F2S is able to translate almost perfectly, but PBMT and Hiero have reordering problems garbling the meaning of the sentence. The second example literally means "other eye, please," and PBMT is able to generate this very literal translation. Hiero, on the other hand, mistakenly makes the listener the subject of "check," and F2S mistakenly translates "please" as "I'd like," which doesn't make sense in this context. In the third example, all three systems have trouble translating the colloquial word for "blink one's eyes," with PBMT and Hiero dropping the word altogether, and F2S leaving it untranslated.

5 Conclusion and Future Work

In this paper, we described an overall design for a speech translation system that aims to reduce the problems caused by language barriers in medical situations. We describe a collection of a medical corpus, and some translation experiments performed on this corpus. As a result of the experiments, we find that the best of three modern translation systems is able to translate 33%-81% of the sentences in a way such that the main content is understandable.

While these preliminary results are encouraging, this is just the first step towards a full medical speech translation system. As described, there are still a number of challenges related to the MT module itself, including the handling of informal speech. These will further be compounded when combined with the need for robust ASR and TTS. However, given the potential for speech translation technology to be useful in medical situations, we believe that meeting these research challenges is a worthy target for research and development in the near future.

Acknowledgments

Part of this research was executed under the Commissioned Research for "Research and Development on Medical Communication Support System for Asian Languages based on Knowledge and Language Grid" of National Institute of Information and Communications Technology (NICT), Japan.

References

Jerome R Bellegarda. 2004. Statistical language model adaptation: review and perspectives. *Speech communication*, 42(1):93–108.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: web inventory of transcribed and translated talks. pages 261–268.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2).

Matthias Eck, Stephan Vogel, and Alex Waibel. 2004. Improving statistical machine translation in the medical domain using the unified medical language system. In *Proc. COLING*, pages 792–798.

Daniel Flannery, Yusuke Miyao, Graham Neubig, and Shinsuke Mori. 2011. Training dependency parsers from partially annotated corpora. In *Proc. IJCNLP*, pages 776–784, Chiang Mai, Thailand, November.

Sharon Goldwater, Daniel Jurafsky, and Christopher D. Manning. 2010. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. 2011. Overview of the patent machine translation task at the NTCIR-9 workshop. In *Proceedings of NTCIR*, volume 9, pages 559–578.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proc. EMNLP*, pages 944–952.

Phillip Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. HLT*, pages 48–54, Edmonton, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL*, pages 177–180, Prague, Czech Republic.

Christopher J Leggetter and Philip C Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171–185.

Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. In *Proc. ACL*, pages 192–199.

Mai Miyabe, Kunikazu Fujii, Tomohiro Shigenobu, and Takashi Yoshino. 2007. Parallel-text based support system for intercultural communication at medical receptions. In *Intercultural Collaboration*, pages 182–192. Springer.

Satoshi Nakamura. 2009. Overcoming the language barrier with speech translation technology. *NISTEP Quarterly Review*, (31).

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proc. ACL*, pages 529–533, Portland, USA, June.

Graham Neubig, Yuya Akita, Shinsuke Mori, and Tatsuya Kawahara. 2012. A monotonic statistical machine translation approach to speaking style transformation. *Computer Speech and Language*, 26(5):349–370, October.

Graham Neubig. 2011. The Kyoto free translation task. `http://www.phontron.com/kftt`.

Graham Neubig. 2013. Travatar: A forest-to-string machine translation engine based on tree transducers. In *Proc. ACL Demo Track*, Sofia, Bulgaria, August.

Hermann Ney. 1999. Speech translation: Coupling of recognition and translation. In *Proc. ICASSP*, pages 517–520.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Masaya Ohgushi, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. An empirical comparison of joint optimization techniques for speech translation. In *Proc. 14th InterSpeech*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318, Philadelphia, USA.

Alok Parlikar, Alan W Black, and Stephan Vogel. 2010. Improving speech synthesis of machine translation output. In *Proc. 11th InterSpeech*, pages 194–197.

Jason Riesa and Daniel Marcu. 2010. Hierarchical search for word alignment. In *Proc. ACL*, pages 157–166.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. 7th International Conference on Speech and Language Processing (ICSLP)*.

Hirotoshi Taira, Katsuhito Sudoh, and Masaaki Nagata. 2012. Zero pronoun resolution can improve the quality of J-E translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 111–118, Jeju, Republic of Korea, July.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. LREC*, pages 147–152.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*. Jeju Island, Korea.

Marilyn Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman. 2000. Learning to predict problematic situations in a spoken dialogue system: experiments with how may I help you? In *Proc. 6th Conference on Applied Natural Language Processing*, pages 210–217.

Ruiqiang Zhang, Genichiro Kikui, Hirofumi Yamamoto, Taro Watanabe, Frank Soong, and Wai Kit Lo. 2004. A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation. In *Proc. COLING*, pages 1168–1174.