A Prototype Automatic Simultaneous Interpretation System

Xiaolin Wang **Andrew Finch** **Masao Utiyama** **Eiichiro Sumita**
Advanced Translation Research and Development Promotion Center
National Institute of Information and Communications Technology, Japan
{xiaolin.wang, andrew.finch, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

Simultaneous interpretation allows people to communicate spontaneously across language boundaries, but such services are prohibitively expensive for the general public. This paper presents a fully automatic simultaneous interpretation system to address this problem. Though the development is still at an early stage, the system is capable of keeping up with the fastest of the TED speakers while at the same time delivering high-quality translations. We believe that the system will become an effective tool for facilitating cross-lingual communication in the future.

1 Introduction

Interpretation is the oral translation of speech from one language to another. Simultaneous interpretation is one type of real-time interpretation where the interpreter performs the translation within the time permitted by the pace of source speech. Compared to another type of interpretation – consecutive interpretation – where the speaker pauses after completing one or two sentences, simultaneous interpretation has the advantages of saving time, and also not interupting the natural flow of the speaker [1].

Simultaneous interpretation is an effective way to bridge language gaps. A good example of events where simultaneous interpretation is used are the United Nations and European Union conferences. The interpreter sits in a soundproofed booth and speaks into a microphone, while clearly seeing and hearing the speaker. The delegates in the meeting room select the relevant channel to hear to interpretation in the his or her native language [2]

Simultaneous interpretation is an expensive service due to the cost of interpreters. First, the number of simultaneous interpreters is small, because the job requires many years of experience and subject matter expertise. Second, for a real-world event, employing one interpreter is normally insufficient, because the task demands so much concentration that any individual can only hope to be effective for periods of 20 minutes or less. Several interpreters are required for continuous service of more than two hours [3].

Inspired by both the merits and the demands of simultaneous interpretation, we have developed a fully automatic simultaneous interpretation system, as presented in this paper. Recently some other simultaneous interpretation systems such as (Müller et al., 2016) have also been presented. Unfortunately, cross-comparison is currently not possible without access to these systems, and will hopefully become interesting future work. This paper first explains how the system works (Section 2), then describes how to use the system (Section 3), then shows how well the system works (Section 4), then presents an example of the system's performance on a TED talk (Section 5), and finally concludes with a description of future work (Section 6).

2 The System in a Nutshell

The simultaneous interpretation system is a fully automatic speech-to-speech system that is currently capable of English-Japanese bidirectional interpretation. The method is general, and can be applied to other language pairs directly.

[1] https://en.wikipedia.org/wiki/Language_interpretation
[2] http://ec.europa.eu/dgs/scic/what-is-conference-interpreting/simultaneous/index_en.htm
[3] http://www.londontranslations.co.uk/our-services/simultaneous-interpreters/

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 30–34, Osaka, Japan, December 11-17 2016.

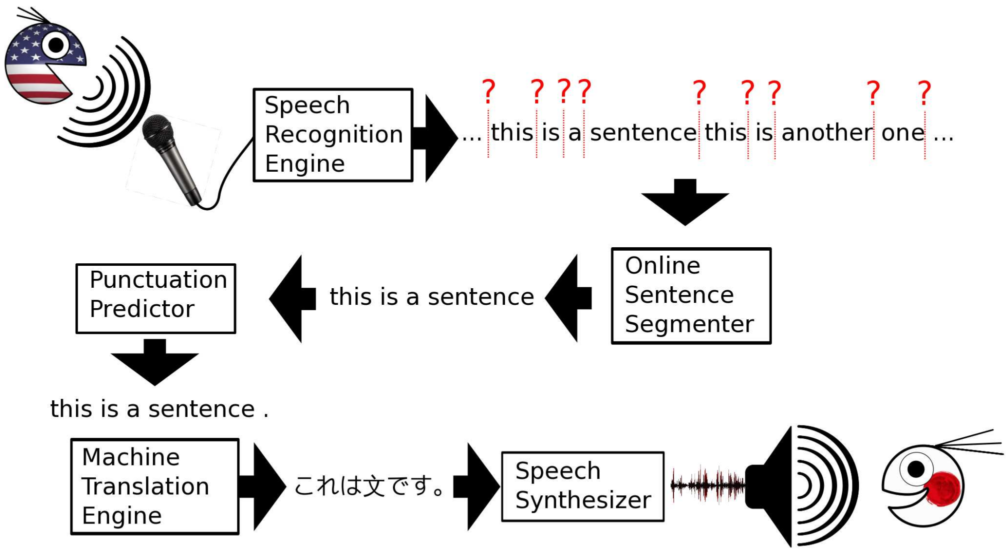

Figure 1: Architecture of the Simultaneous Interpretation System

Figure 1 illustrates the architecture of the system. The key element in the design is an online sentence segmenter that bridges the speech recognition engine and the machine translation engine. The whole system is a pipeline of six components: a speech recognition engine, a sentence segmenter, a punctuation predictor, a machine translation engine, and a speech synthesizer.

The **Speech Recognition Engine** converts audio signals into a stream of words. The current implementation is an online decoder based on the Kaldi open source toolkit (Povey et al., 2011)[4]. We plan to integrate our own in-house speech recognition engine – SprinTra (Shen et al., 2014) in the future.

Our system is able to perform speech detection. That is to say the system is always listening, and will respond to any speech it hears (see Section 3 for details). Speech detection is done by applying a threshold to the energy of the input audio signals. We determined empirically that this heuristic works well in actual use. In the case when loud noises exceed the threshold and trigger the system, the speech recognition engine normally outputs no words, thus little damage is caused.

The **Online Sentence Segmenter** converts the stream of words into sentences. The implementation is based on the method proposed in (Wang et al., 2016a). The implementation uses a linear combination of a language model, a length model and a prosodic model to calculate the confidence of segmentation boundaries, and uses a threshold-latency-based heuristic to make decisions.

The **Punctuation Predictor** converts an un-punctuated sentence into a punctuated sentence. The implementation is based on the findings in (Wang et al., 2016b). It uses a hidden N-gram model (Stolcke et al., 1998; Matusov et al., 2006), which is available in the toolkit of SRILM (Stolcke, 2002), to insert punctuation.

The **Machine Translation Engine** translates a source-language sentence into a target-language sentence. The implementation is our in-house pre-ordering translation system, called the General Purpose Machine Translation (GPMT) engine. The system is publicly accessible through a Web API [5]

The **Speech Synthesizer** converts sentences into speech. The implementation is based on the HTS open-source toolkit (Tokuda et al., 2013)[6]

[4]`https://github.com/kaldi-asr/kaldi`
[5]`https://mt-auto-minhon-mlt.ucri.jgn-x.jp/`
[6]`http://hts.sp.nitech.ac.jp/`

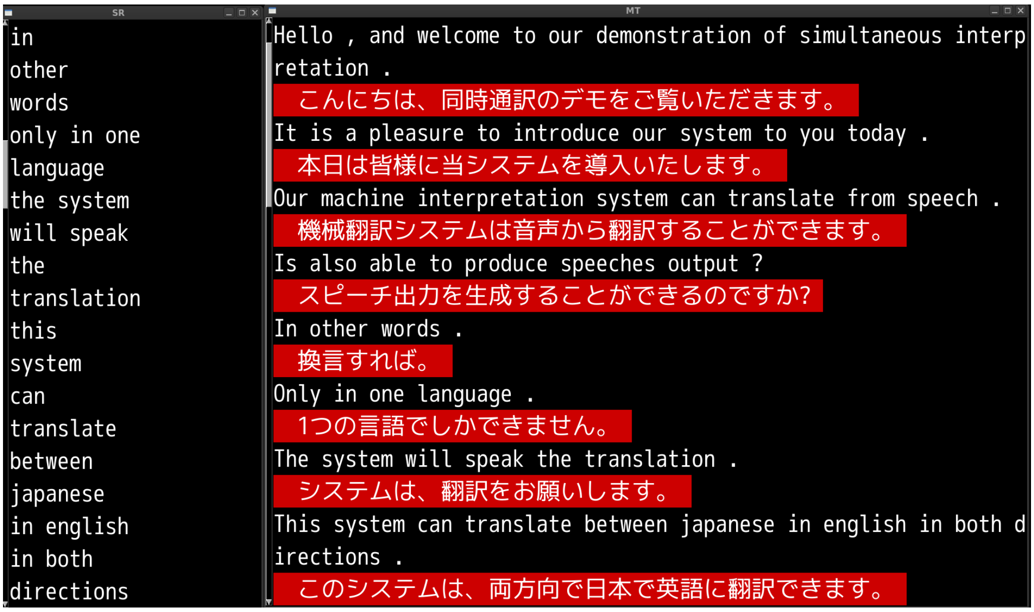

Figure 2: Logs of Simultaneous Interpretation System: Speech Recognition (left) and Machine Translation (right).

3 Usage

The system is designed to work in exactly the same manner as a human interpreter working in multilingual conferences. Once launched, the system can work continuously for hours or days without intervention. In operation, it receives audio signals constantly from its microphone. If no one is speaking, the system will produce no output. If someone is speaking, the system will speak out the translation, normally, in only a few seconds.

In addition to the speech output, two logs can be used to monitor the running of the system: the speech recognition log and the machine translation log (Figure 2). The speech recognition log shows the recognized words from the speakers. The machine translation log shows the recognized sentences and their translations. The content of both logs is updated in realtime.

4 Performance

The performance of our method was measured in (Wang et al., 2016a). Experiments were performed on translation between Japanese and English in both directions. The time efficiency was measured by average latency per source word using the definition given in (Finch et al., 2014). The translation quality was measured by the BLEU of end-to-end translation. Because the segmented source sentences did not necessarily agree with the oracle, translations were aligned to reference sentences through edit distance in order to calculate BLEU (Matusov et al., 2005).

The results of the measurement are presented in table 1. Different sentence segmentation methods were compared. Our system adopted the threshold-latency method which generally outperformed the other methods on both time efficiency and translation quality.

5 Example Analysis

Here is an example of interpreting a TED talk from English to Japanese by the system. The talk is ”Your elusive creative genius “ given by Elizabeth Gilbert in 2009[7]. The oracle transcript is,

> I am a writer. Writing books is my profession but it’s more than that, of course. It is also my great lifelong love and fascination. And I don’t expect that that’s ever going to change. But, that said, something kind of peculiar has happened recently in my life and in my career ...

[7] `https://www.ted.com/talks/elizabeth_gilbert_on_genius?language=en`

Sentence Segmenter	Dev. Set BLEU	Dev. Set Latency	Test Set BLEU	Test Set Latency
Japanese-to-English				
Oracle	13.82	NA	13.67	NA
Hidden N-gram †	13.30	NA‡	12.97	NA‡
Fixed-length	11.71	16.66	11.55	16.63
Threshold-based	**13.38**	14.20	13.16	13.68
Latency-based	13.21	18.04	13.20	18.03
Threshold-latency (our System)	**13.38**	**12.98**	**13.28**	**12.89**
English-to-Japanese				
Oracle	13.84	NA	14.15	NA
Hidden N-gram†	12.85	NA‡	13.10	NA‡
Fixed-length	11.86	8.19	12.15	8.20
Threshold-based	12.93	**7.13**	13.19	**7.18**
Latency-based	**13.18**	12.25	13.38	12.26
Threshold-latency (our System)	**13.18**	10.01	**13.42**	10.11

Table 1: Performance of interpretation systems that use different sentence segmenters. The confidence scores in threshold-based, latency-based and threshold-latency-based segmenters were calculated using Equation 4 in (Wang et al., 2016a). † Employed the segment tool from the SRILM toolkit (Stolcke, 2002). ‡ The method is not online since it operates on a whole sequence of words, thus the measurement of latency is not applicable.

Recognized Sentence	Translation	Post Edited	Lat.(s)
I am a writer .	私は作家です。	私は作家です。	1.5
writing books is my profession .	書く仕事です。	本を書くのが私の仕事です。	3.3
but , it's more than that of course it is also my great lifelong love and fascination .	しかし、それはまた、私がたいへん好きや魅力のものより多い。	ですが、それは仕事以上のもので、私がずっと大好きで魅了されていることなのです。	2.5
and I don't expect that that's ever going to change .	そして私はそれが変わるのです、とは思っていません。	そして、今後もそれは変わらないと思っています。	2.1
but that said , something kind of peculiar has happened recently in my life , and in my career .	しかしそうは言っても、最近変わった体験をし私の人生において、ました。	ですが、最近、公私に渡り変わった体験をしました。	1.8

Table 2: Example of Simultaneous Interpretation System Working on an TED Talk

The result of the system is shown in Table 2. The system works rapidly, and can easily keep with up the speaker, with a latency ranging from 1.5 to 3.3 seconds for these sentences.

For analysis, the output was corrected by a professional translator ('Post Edited' in Table 2). The first sentence was translated perfectly; the second was good but omitted the translation for the word *books*. The third sentence's translation resolved the pronouns incorrectly, and this was subjectively the worst translation. The fourth sentence was semantically correct, but it is more natural to say: 'I expect not *X*' rather then 'I didn't expect *X*' in Japanese. The fifth sentence was also quite good but the word *career* was not translated. Overall, the translation quality is impressive, given the difficulty of translation between English and Japanese.

Note that although speech recognition errors rarely happen on this speech. Recognition error rate is speaker dependent and proved to be one of the main sources of errors in our tests. Therefore we believe that further improvements in speech recognition are vital for the future development of simultaneous interpretation systems.

6 Conclusion

This paper presents a prototype automatic simultaneous interpretation system. The system adopts a robust and effective pipeline framework. It is designed to behave like a human interpreter, and is very easy to use. In real-world use it is capable of producing useful translations while keeping up with the fastest of speakers.

Our system is still in early-stage, and we hope that by demonstrating this system we can encourage both academic research and industrial development in this field. In the future, we will constantly improve the system with an emphasis on the quality of final output. Future efforts may include handling disfluencies, applying neural networks to the task of sentence segmentation, integration with our in-house speech recognition engine of SprinTra, and improving our GPMT in-house machine translation engine.

References

Andrew Finch, Xiaolin Wang, and Eiichiro Sumita. 2014. An Exploration of Segmentation Strategies in Stream Decoding. In *IWSLT*.

Evgeny Matusov, Gregor Leusch, Oliver Bender, Hermann Ney, et al. 2005. Evaluating machine translation output with automatic sentence segmentation. In *IWSLT*, pages 138–144. Citeseer.

Evgeny Matusov, Arne Mauser, and Hermann Ney. 2006. Automatic sentence segmentation and punctuation prediction for spoken language translation. In *Proceedings of 3rd International Workshop on Spoken Language Translation*, pages 158–165.

Markus Müller, Thai Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger, Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani, Sebastian Stüker, and Alex Waibel. 2016. Lecture translator - speech translation framework for simultaneous lecture translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 82–86, San Diego, California, June. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, December.

Peng Shen, Xugang Lu, X Hu, N Kanda, M Saiko, and C Hori. 2014. The NICT ASR system for IWSLT 2014. In *Proceedings of the 11th International Workshop on Spoken Language Translation*.

Andreas Stolcke, Elizabeth Shriberg, Rebecca A Bates, Mari Ostendorf, Dilek Hakkani, Madelaine Plauche, Gökhan Tür, and Yu Lu. 1998. Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proceedings of 5th International Conference on Spoken Language Processing*, pages 2247–2250.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*.

Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. 2013. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252.

Xiaolin Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2016a. An efficient and effective online sentence segmenter for simultaneous interpretation. In *(to appear)*.

Xiaolin Wang, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016b. A study of punctuation handling for speech-to-speech translation. In *Proceedings of 22nd Annual Meeting on Natural Language Processing*, pages 525–528.