Development of a Simultaneous Interpretation System for Face-to-Face Services and Its Evaluation Experiment in Real Situation

Akiko Sakamoto, Nayuko Watanabe, Satoshi Kamatani and Kazuo Sumita
Knowledge Media Laboratory,
Corporate Research & Development Center,
Toshiba Corporation
akiko7.sakamoto@toshiba.co.jp

Abstract

We developed a simultaneous interpretation system for face-to-face services at shops, front desks, and so forth. The system supports interpretation between Japanese and English, or Japanese and Chinese speakers. It incrementally processes user's continuous and spontaneous speech, and then incrementally produces interpretation results. We conducted a field test of the system to evaluate the "solved task ratio" for tasks including buying souvenirs, asking a bus route. As a result, we achieved the solved task ratio of 81%.

1 Introduction

Automatic interpretation system has been studied from the earliest days of the computer science. Recent progress of automatic speech recognition (ASR) technology and machine translation (MT) technology achieves high interpretation accuracy enough to use for sentence-by-sentence translation in travel situation or simple daily conversation.

However, it is not widely introduced into practical use in tourist information, retail stores, receptions at a government office and that kind of situation despite of the great needs for communication in foreign language. It is because that existent interpretation systems force user to speak only one sentence at once, and show each interpretation result one by one. Such intermittent process does not meet with face-to-face business conversation which requires smooth communication.

To give a solution for business use of machine simultaneous speech interpretation, we developed a system which recognizes speakers' spontaneous and continuous speech, and automatically divides into semantically reasonable units, and then consecutively interprets each unit. This system allows users to speak freely without paying any attention to speech length for one time.

Following chapters describe experiment details. Chapter 2 presents a comparison with relevant studies. Chapter 3 describes our simultaneous interpretation system. Chapter 4 mentions setup of our evaluation experiments in Chiba-city Japan. Chapter 5 reports the experiment result. Chapter 6 concludes this paper.

2 Related work

Many studies such as (Waibel et al., 1991), (Metze et al., 2002) and (Wahlster, 1993) have been held for speech-to-speech translation (S2ST). In early stage of S2ST technology studies, those systems restricted to accept certain topic and/or speech style. Recently, systems which can incrementally interpret utterances have emerged (Matsubara et al., 1997), and some of them are commercially available (NTT docomo, 2012). Some complex applications can be a target of the S2ST system, like lectures interpretation (Fügen et al., 2007).

Most previous works on S2ST systems have been evaluated from the viewpoint of recognition and translation accuracy. Therefore, it is not enough argued that what kind of support the current systems can provide and what level of user satisfaction it attains.

As S2ST technology, the systems need to be evaluated in the practical use, and our study contributes in this aspect. We developed our own simultaneous interpretation system and evaluated it in terms of conversation goal achievement.

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 85–92.

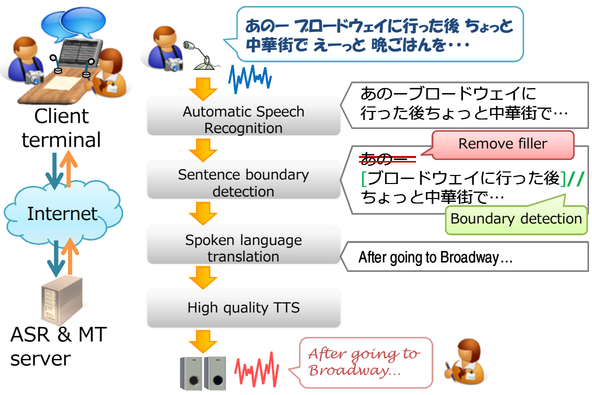

Figure 1: Process flow of our system

Knowledge and information obtained by this study helps to improve the S2ST performance from a viewpoint of practical use.

3 Simultaneous interpretation system

Figure 1 illustrates a sketch of our simultaneous interpretation system. The left hand side of the figure describes the system configuration of ASR and MT server, and a client terminal. The server and the client terminal communicate over the internet.

The right hand side of the figure describes the brief flow of our interpretation process. First, the system recognizes user's spontaneous speech segmented by certain length of silence and continuously outputs a transcribed text. Secondly, it detects a sentence boundary to split the text into segments suitable for translation, and then examine each segment is necessary to translate or not. Each segment is translated in time order. This procedure enables the system to start ASR and MT process without waiting the end of the whole speech by a speaker, and to interpret users' utterances in a short delay for original user's utterance. In addition, if needed, text-to-speech engine synthesizes a voice sound for the translation result.

This chapter will explain each module.

3.1 Automatic speech recognition

We customized and used our own technology of large vocabulary continuous speech recognition (LVCSR) engine ((Nakamura et al., 2012), (Ding et al., 2008)). It is trained with large scale text corpus collected from the web and originally developed bilingual corpus in travel domain.

200 thousand Japanese words are registered as entries of recognition dictionary. These entries are selected according to frequency of appearance in the corpus. In addition , words related to Chiba-city (eg. sightseeing spots, transport facilities, etc.) where we conducted an evaluation experiment described in Chapter 4, are registered to the dictionary. In the same way, we developed Chinese and English recognition dictionaries which contain 30 thousand word entries respectively.

ASR module outputs a recognition result for every speech section separated by 300 ms pause.

3.2 Sentence boundary detection

The speech segment processed by ASR is not always appropriate to translation. ASR runs and converts voice sound into text while it detects voice sound, in other words, it ends when it detects some pauses. When a speaker makes pauses between sentences, every ASR result contains just one sentence. However, when a speaker puts several pauses in a sentence, the ASR result for one sentence is divided into several ASR results. When a speaker makes no pause between sentences, one ASR result includes more than one sentence.

A human interpreter often interprets at the end of every sentence or clause, because a sentence or clause is a basic unit to express an events. It is clear that sentence boundary detection module for ASR result is required to make input text for MT.

3.2.1 Detection model

Sentence boundaries were detected in 2 steps.

In the first step, the system performs morphological analysis on the result of ASR and obtain word segmentation on Japanese and Chinese and also POS tags on Japanese, Chinese and English. Then, we removed fillers and other redundant parts using simple pattern matching to POS.

In the second step, we used machine learning based classifier to detect sentence boundaries. We treated sentence boundary detection task as labeling task to each word (Liu et al., 2005). We prepare spontaneous speech corpus in which words at the beginning of a sentence has "B" labels and other words has "I" labels. We used CRF++ (Kudo, 2005), a machine learning tool with conditional random field, and created a discrimination model for the labeling. As for learning features, we used surface form of two morphemes (two words for Chinese) before and after each morpheme.

3.2.2 Training corpus

To create Japanese, English and Chinese sentence boundary detector, we used three different corpora: 140,000 sentences from "Corpus of Spoken Japanese (CSJ) (Maekawa et al., 2000)" for Japanese, 110,000 sentences form WIT[3] (Cettolo et al., 2012) data including transcription of TED talk for English, and 400,000 sentences from our original travel domain corpus for Chinese.

These corpora do not contain any tags denoting a suitable unit for translation. We regarded a punctuation mark as a boundary marker in English and Chinese. As for Japanese, we regarded a clause to be suitable unit for translation (Takanashi et al., 2003), and prepared simple rules to give clause boundary on the training corpus.

3.2.3 Detection performance

We evaluated precision and recall of boundary detection on test sets. The test sets had been ideally segmented into 244 Japanese sentences, 1664 English sentences, and 3648 Chinese sentences. We regarded punctuations as reference. Table 1 shows detection accuracy. In this table, we calculate precision and recall value as follows:

$$\text{Precision}=\frac{\text{No. of correctly estimated sentence boundaries}}{\text{No. of estimated sentence boundaries}}$$

$$\text{Recall}=\frac{\text{No. of correctly estimated sentence boundaries}}{\text{No. of period in original corpus}}$$

□

3.3 Machine Translation

3.3.1 Forest Driven Rule-based MT

Rule-based machine translation (RBMT) technology has achieved the considerable market in a written language domain including instruction manual and patent translation. RBMT has been tested for a long time, and answered users' request. Therefore, it is advantageous for spoken language translation (SLT) system to utilize fine translation rules of the existent RBMT system.

However, these rules are designed for grammatically written language, and it sometimes fails to process ungrammatically spoken language. One spoken utterance often consists of a sequence of fragmental phrases. When some fragments gather and make a clause, an internal structure of the clause is similar to that of written language.

Table 1: Segment detection accuracy

	Precision	Recall	F-value
Japanese	0.739	0.672	0.705
English	0.720	0.809	0.763
Chinese	0.679	0.685	0.681

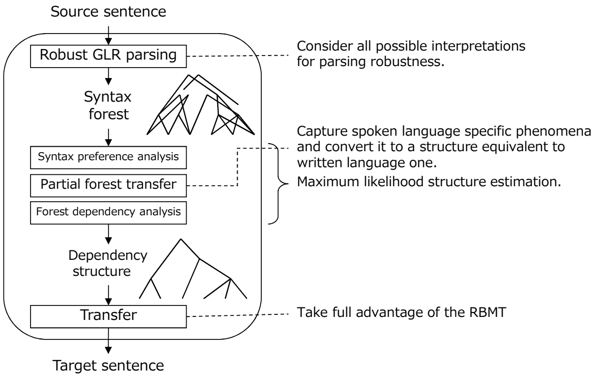

Figure 2: Process flow of forest driven RBMT

Forest driven RBMT (Kamatani et al., 2009) parses input sentence by generalized LR parsing algorithm based method which can accept an ungrammatical part by using an original CFG based grammar to capture the clause structure and deal with various ambiguities. Then it generates possible syntax structures as one forest structure and transfer one best structure to target language structure, according to syntactic and semantic preferences.

This procedure allows us to effectively acquire totally preferable structures from a syntax forest. Therefore, we can utilize rich translation rules used in conventional RBMT, while handling difficulty of spoken language.

3.3.2 Hybrid MT

Statistical MT (SMT) can generate natural translation result for restricted and specific domain, like tourism information, department stores, and public office window. But, in some cases, well developed RBMT engine outputs more suitable translation and covers larger domain.

RBMT method translates input text by referring to many translation rules: parsing, transfer, generation rules and so on. Some rules are described generally to handle various fundamental linguistic phenomena, on the other hand, some rules are elaborated concretely to translate practically. That gives robustness to a system, but sometimes causes lack of fluency.

We considered that these strong and weak points are complementary, then we used SMT and RBMT engine as one hybrid MT engine. Specifically, when SMT result has lower translation probability than threshold, RBMT result is selected as final result of hybrid MT engine (Kamatani et al., 2009). This engine selection runs for each segment detected by the sentence boundary detection.

We utilized phrase-based statistical machine translation (SMT) method (Wang et al., 2008). For Japanese-English and English-Japanese SMT, we trained the engine on our originally developed 220 thousand sentence pairs corpus and 20 thousand sentence pairs corpus (Akiba et al., 2004) distributed by ALAGIN (Advanced Language Information Forum). Both corpora contain example sentences in a travel domain. Japanese-Chinese and Chinese-Japanese SMT is trained by our own developed 210 thousand sentence pairs corpus.

3.3.3 Translation quality

We conducted two kinds of translation quality evaluation. First we conducted manual evaluation on English-to-Japanese (EJ), Japanese-to-English (JE), Chinese-to-Japanese (CJ) and Japanese-to-English (JE) MT engines, and second, we conducted detailed evaluation on JE and EJ engines.

First, we manually evaluated translation result of EJ, JE, CJ and JE engines. We used originally developed 100 sentences in travel conversation The evaluation metrics is following; “4. Impeccable”, “3. Grammatical, but not fluent”, “2. Ungrammatical, but understandable”, “1. Incomprehensible and/or misleading” and “0. No interpretation given”. Figure 3 shows the evaluation result. In the figure, “H” denotes a hybrid MT engine, “R” rule-based MT, “S” statistical MT.

Improvement of the hybrid method strongly depends on RBMT performance, because RBMT works as a fail guard for SMT. In this evaluation, because JE RBMT had better improvement on translation quality than JC RBMT, hybrid MT quality of JE/EJ was superior to that of JC/CJ.

Assuming that 2nd and higher grade translation quality is necessary to establish a conversation, translation accuracy for JE/EJ achieved about 90%, and CJ/JC about 80%.

Second, we conducted detailed evaluation of JE and JE SMT, RBMT and Hybrid MT engines with automatic and manual evaluation. We used English and Japanese sentences from IWSLT 2004 test set (Akiba et al., 2004). Full 500 sentence pairs were used for automatic evaluation and the first 100 sentence pairs were used for manual evaluation.

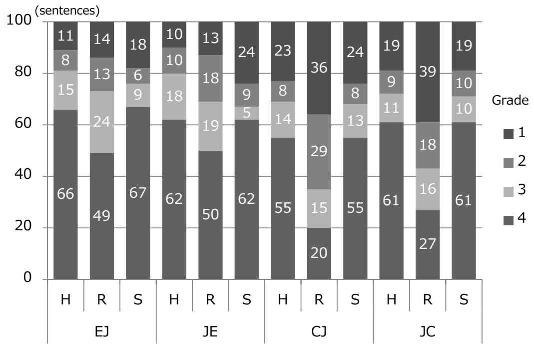

Figure 3: Translation quality

Table 2: Translation quality (IWSLT)

		BLEU	RIBES	Adeq.	Flue.
JE	RBMT	20.64	0.575	3.93	3.69
	SMT	33.97	0.650	3.90	4.12
	Hybrid	28.54	0.631	4.01	3.89
EJ	RBMT	22.21	0.755	4.15	3.94
	SMT	34.28	0.807	4.25	4.29
	Hybrid	32.27	0.790	4.30	4.25

Adeq. = Adequacy, Flue. = Fluency

We used BLEU (Papineni et al., 2002) and RIBES (Isozaki et al., 2010) for automatic evaluation. We also manually evaluated with fluency and adequacy metrics (Koehn, 2006). Table 2 shows the evaluation result.

We assumed that adequacy of manual translation reflects correctness of meaning, and we chose the hybrid engine for our simultaneous interpretation system. This performance of our hybrid MT engine suit for our idea that adequacy is most important in a communication.

3.4 Application user interface

A host and a guest share a display of a terminal and communicate with each other through our system. We developed a client application for an Android tablet. Figure 4 shows the user interface.

A speaker start her/his speech after pressing down to “speak” button. While she/he continues to speak, it is not necessary to press the button. When she/he re-presses down the button, system processes it as an explicit utterance end.

Until recognition result is fixed, a recognition

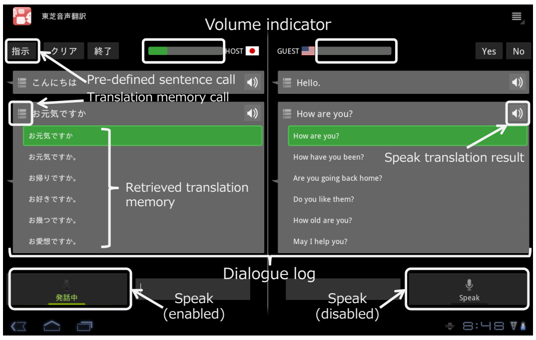

Figure 4: User interface of Client Application

Figure 5: Displayed interpretation result

candidate is shown in gray color text. When translation result is fixed, system shows ASR and MT text in visible. In addition, we developed following functions to help host side subjects:

- Translation memory retrieval
- Short cut button to input typical expression
  - Yes / No response.
  - Direction ("speak more slowly", etc.)
  - Frequently used utterances
- Interpretation result removal by swipe action

These functions are difficult to use for a guest user who sees the system for the first time, but we can expect that she/he learns the usage by observing a host user's usage. Figure 5 shows an application display of Japanese-English interpretation. Translation unnecessary part removed by the sentence boundary detection is not displayed.

4 Evaluation experiment

We gave American and Chinese guest side subjects a conversation task sheet. They communicate with a Japanese host side subject through our system, and solve each task.

4.1 Experimental environment

Chiba City Tourist Information Center and Chiba City International Communication Plaza provided experiment place for us. We evaluated tourism conversation at the information center, and daily conversation at the communication plaza.

We used Toshiba REGZA Tablet AT700 as a terminal (Figure 6). We used two monaural microphones. These microphones are connected together to one monaural audio input of the tablet using audio splitter. In this experiment, interpretation result outputs in synthesized voice by Text-to-Speech, only when users pressed down a button. The terminal communicates with the server via public LTE[1] line.

4.2 Subjects

We recruited 6 American and 6 Chinese native speakers as guest side subjects. Table 3 shows their brief information including place of birth and length of living experience in Japan. Before the experiment, we ask them not to use Japanese. We regard the simultaneous interpretation system as a support tool for motivated two participants, so, we rather expect users to make use of any information that is caught from the opponent user's utterance.

As for Japanese subjects, two staffs of the tourist information center took part in the experiment, and we also recruited 6 Japanese students of Chiba University for the experiment in the International Communication Plaza. Their brief information is shown in table 4.

4.3 Conversation tasks

We made conversation tasks of various difficulties. They included such tasks that can be solved with

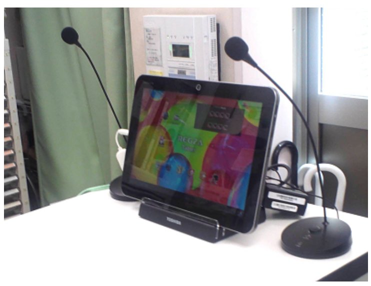

Figure 6: Experiment at the information center.

[1] downlink 75Mbps, uplink 25Mbps (best effort)

Table 3: American and Chinese Subjects' Profile

	Sex	Age	Place of birth	Years in Japan
China	f	31	Shenyang	1
	m	42	Chengdu	3
	m	28	Beijing	1
	f	31	Chengdu	2
	f	25	Shanghai	1.5
	f	40	Beijing	2
USA	m	22	New York	1
	m	30	California	4
	f	25	Utah	3
	m	32	Pennsylvania	4
	m	40	Claifornia	4
	f	48	Florida	3

Table 4: Japanese Subjects' Profile

Affiliation	Sex	Age	Place of birth
Tourist Info. Center	f	-	Wakayama
	f	-	Gunma
Chiba Univ.	m	23	Nagasaki
	m	23	Tochigi
	m	23	Saitama
	f	23	Chiba
	f	21	Chiba
	f	23	Fukushima

very simple communication, and such that requires complex conversation.

We prepared 8 conversation tasks related to tourism information to ask following items; 1) tour reservation, 2) train route to a theme park, 3) train fare, 4) find an exchange, 5) bus route map and time table in Chiba-city, 6) best souvenir from Chiba-prefecture, 7) sightseeing spots and 8) activity spots.

As for daily conversation, we made 10 tasks to ask items below; 1) day of the week, 2) birth place, 3) way to downtown, 4) find a lost bag, 5) experience to visit your home country, 6) experience to visit foreign countries, 7) hobby, 8) food recommendation, 9) souvenir recommendation and 10) common interest.

4.4 Instruction for subjects

We assume the situation that this device is used between a host who is already get used to using the device and a guest who uses the device for the first time. In this experiment, we asked Japanese subjects to have a role of host and practice using the device before the experiment. On the other hand, we asked American and Chinese subject to take a role of guest, and we only prepared a simple leaflet of instruction about the device and did not explain

課題文	回答欄
Ask the way to get to a money exchange shop near here.	Place [ ] Did you complete this task? ☐Yes/☐No
Now you would like to know a bus routemap and its schedule in Chiba city. Ask how you can get these information.	Did you complete this task? ☐Yes/☐No
Ask the best souvenir of Chiba. Ask its features and how to get to a store where you can buy it.	Souvenir [ ] Did you complete this task? ☐Yes/☐No

Figure 7: Task sheet for American subjects.

how to use the device.

To American and Chinese subject, we briefly explained the experiment procedure , and gave them a task sheet. As shown in figure 7, a task sheet includes several tasks and some spaces to fill in the answers. Before the experiment, we asked guest side subjects to talk to the host using the device and write the information that they obtained from the host side subjects. In addition, we asked guest side subjects to mark a checkbox if they think they successes to obtain the right answer from the host side subject. After the experiment, a Japanese native speaker analyzed the answer and the conversation log data, then evaluated whether the answer is correctly transferred from a host side subject to a guest side subject or not.

The order of tasks on the task sheet is shuffled for each subject. As for the time to solve the task, we allowed 30 minutes for all the tasks at the tourist information center, and 60 minutes at international communication plaza. If some tasks were not solved within the time, we regarded these tasks as failed. During the experiment, both host side and guest side speaker were assumed to be able to understand only their native languages. We asked them not to get information from the conversation opponent's language which is displayed on the device. At the tourist information center, both of the host side and guest side subjects were allowed to use city maps, route maps of bus or train, and other materials for tourist guide.

5 Evaluation result

5.1 Solved task ratio

We defined Solved Task Ratio (STR) as a degree of interpretation success. It indicates the number of achieved tasks out of all the tasks. Figure 8 shows the evaluation result.

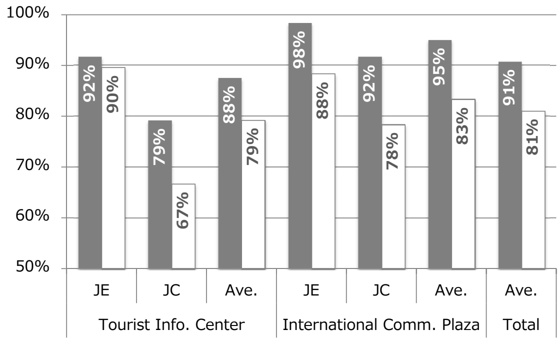

Figure 8: Solved task ratio.

	元発声	認識結果	翻訳結果
1ej	I'd like to buy some souvenirs.	i'd like to buy some souvenirs	お土産を買いたいのですが。
2ej	Can you recommend anything special?	can you recommend anything special	お勧めの特別の何かはありますか？
3je	最近東京スカイツリーができたので	最近東京スカイツリーができたので	The Tokyo sky tree was made recently.
4je	スカイツリーのグッズがおすすめです	スカイツリーのグッズがおすすめです。	The goods of a sky tree are recommended.
5ej	I plan to go see Tokyo Sky Tree,	i plan to go see this guy tree	この人木を見に行くことを計画します。
6ej	tommorow. That sounds like a great idea.	tommorow that sounds like a great idea.	明日です それは素晴らしい考えのように聞こえます
7je	夜のライトアップがとてもきれいなのでぜひ見て下さい	夜のライトアップがとてもきれいなのでぜひ見て下さい	Lighting of night is very beautiful. Please see.

Figure 9: English-Japanese conversation

	元発声	認識結果	翻訳結果
1cj	在哪里可以把外汇兑换成	在那里可以把外汇兑换成	あそこで 外貨にすることができます
2cj	应该怎么走	应该怎么走	どう行くべきである
3jc	（雑音）	えーええーえーえーので	是。
4cj	在哪里可以把外汇兑换成日元应该怎么走	在那里可以把外汇兑换成日元应该怎么走	あそこで 外貨にすることができます どう行くべきである
5jc	外貨両替ですね。	外貨両替ですね。	兑换外币。
6cj	二	二	二
7cj	是的	是的	はい。
8jc	ワールドカレンシーショップか	ワールドカレンシーショップか	是世界货币店铺吗？
9jc	ヨドバシカメラのビルの三階にあります。	ヨドバシカメラのビルの三階にあります。	友都八喜照相机的大楼的三楼有。

Figure 10: Chinese-Japanese conversation

Figure 8 consists of two parts. The black bar shows STR calculated based on subjects' self-report about task achievement. Next, we observed conversation log data and checked whether the subjects really achieved the task. The white bar shows the ratio calculated based on our check. The difference between black and white bar denotes amount of misinterpretation between subjects.

The system needs to interpret particular terms in tourism domain. This gets STR in tourist information center worse than in international plaza. From the viewpoint of a language pair, translation quality evidently reflects to the performance.

Further analysis on conversation log requires transcription of actual utterances and comparison with ASR and MT result. We consider this analysis as future work.

5.2 Time of completing the tasks

Average time for completing each task is shown in table 5. Measured time includes delay time caused by system and network failure.

All tasks are completed in less than 3 minutes. Despite of the high STR, Japanese-English subjects pair used longer time than Japanese-Chinese. According to the conversation log, this difference comes from conversation content, not from difficulty of the task or interpretation accuracy. From other point of view, if subjects can keep a lively conversation, conversation time gets longer.

5.3 Conversation logs

Figure 9 shows conversation between Japanese and American subjects for a task "Souvenir recommendation," and Figure 10 shows conversation between Japanese and Chinese subjects for a task "Find an exchange." In these figures, split columns (eg. 6ej and 4cj) indicate utterances segmented by sentence boundary detection.

Subjects reached the goal of conversation task by estimating partner's intention from context, even there are recognition errors (eg. *5ej* and *1cj*) and machine translation errors (eg. *7je* and *8cj*). Our system displays all interpretation results including non-informative part like *6cj*, but subjects browsed significant context and progressed their conversation.

When users noticed that an interpretation result contained some errors, they sometimes began to speak word by word. This speaking style brings worse results for ASR and MT, because of lack of context information. This is a difficult problem related to pause length, and can be a key to

Table 5: Conversation time.

(mm:ss)

		EJ	CJ
Tourist Info. Center	All	3:11	2:35
	Solved only	2:56	2:30
International Comm. Plaza	All	2:22	1:36
	Solved only	2:20	1:32

tackle spontaneous speech. An appropriate length of pause is different from each person, and it dynamically changes.

6 Summary

In this paper, we described our simultaneous interpretation system designed for face-to-face services. It processes users' continuous speech and incrementally interprets it. This processing style enables users to speak without paying attention of a sentence boundary, and also reduces time required for a conversation.

We conducted the evaluation experiment under the real situation. In the result, we achieved the solved task ratio of 81%. This shows that subjects can manage to achieve a goal of dialogue by communication through our system and other information obtaining from face-to-face environment.

References

Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, and J. Tsujii, "Overview of the IWSLT04 evaluation campaign", In *Proc. of the International Workshop on Spoken Language Translation*, pp.1-12, (2004).

S. Bangalore, V. K. R. Sridhar, P. Kolan, L. Golipour and A. Jimenez, "Real-time Incremental Speech-to-Speech Translation of Dialogs", In *Proc. of NAACL-HLT*, pp.437-445, (2012).

M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web inventory of transcribed and translated talks", In *Proc. of EAMT*, pp.261-268, (2012).

Hongfei Ding, Koichi Yamamoto and Masami Akamine, "Comparative evaluation of different methods for voice activity detection", In *Proc. Interspeech 2008*, pp.107-110, (2008).

C. Fügen, A. Waibel, M. Kolss, "Simultaneous translation of lectures and speeches", Machine Translation, 21, pp.209-252, (2007),

H. Isozaki, T. Hirao, K. Duh, K. Sudoh and H. Tsukada, "Automatic Evaluation of Translation Quality for Distant Language Pairs", In *Conference on Empirical Methods on Natural Language Processing*, pp.944-952, (2010).

S. Kamatani, T. Chino and K. Sumita, "Hybrid Spoken Language Translation Using Sentence Splitting Based on Syntax Structure", In *Proc. of Machine Translation Summit XII*, (2009).

P. Koehn and C. Monz, "Manual and automatic evaluation of machine translation between European languages", In *Proc. of the Workshop on Statistical Machine Translation*, pp.102-121, (2006).

T. Kudo 2005. "CRF++: Yet Another CRF toolkit", Available at https://code.google.com/p/crfpp/

Y. Liu, A. Stolcke, E. Shriberg and M. Harper, "Using Conditional Random Fields For Sentence Boundary Detection In Speech", In *Proc. of the 43rd Annual Meeting of the ACL*, pp.451-458, (2005).

K. Maekawa, H. Koiso, S. Furui and H. Isahara, "Spontaneous Speech Corpus of Japanese", In *Proc. of LREC2000*, pp.947-952, (2000).

S. Matsubara and Y. Inagaki, "Incremental Transfer in English-Japanese Machine Translation", IEICE TRANSACTIONS on Information and Systems, Vol.E80-D No.11, pp.1122-1130, (1997).

F. Metze, J. McDonough, H. Soltau, A. Waibel, A. Lavie, S. Burger, C. Langley, L. Levin, T. Schultz, F. Pianesi, R. Cattoni, G. Lazzari, N. Mana, and E. Pianta, "The NESPOLE! speech-to-speech translation system", In *Proc. of HLT 2002*, (2002).

M. Nakamura, H. Fujimura, Y. Shinohara, T. Masuko and A. Kawamura, "Evaluation of Group Delay-based Features in Noisy Environments", In *Proc. of Spring Meeting of the Acoustic Society of Japan*, pp.947-952, (2000).

NTT docomo, "NTT DOCOMO to Introduce Mobile Translation of Conversations and Signage", http://www.nttdocomo.co.jp/english/info/media_center/pr/2012/001611.html, Press Release, (2012).

K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation", In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pp.311-318, (2002).

K. Takanashi, T. Maruyama, K. Uchimoto, and H. Isahara, "Identification of "Sentence" in Spontaneous Japanese – Detection and modification of clause boundaries –", In *Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp.183-186, (2003).

W. Wahlster, "Verbmobil: translation of face-to-face dialogs", In *Proc. of European Conf. on Speech Communication and Technology*, pp.29-38, (1993).

A. Waibel, A. Badran, A. W. Black, H. Saito, A. G. Hauptmann, and J.Tebelskis, "JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies", In *Proc. of the ICASSP*, pp.793-796, (1991).

H. Wang, H. Wu, X. Hu, Z. Liu, J. Li and D. Ren and Zhengyu Niu, "The TCH Machine Translation System for IWSLT 2008", In *Proc. of IWSLT 2008*, pp.124–131, (2008).