A Method for Evaluating Task-oriented Spoken Dialog Translation Systems Based on Communication Efficiency

Toshiyuki Takezawa †‡, *Masahide Mizushima* §, *Tohru Shimizu* †‡, *and Genichiro Kikui* §

† National Institute of Information and Communications Technology
‡ ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan
§ NTT Cyberspace Laboratories, Japan

```
toshiyuki.takezawa@{nict.go.jp, atr.jp}, mizushima.masahide@lab.ntt.co.jp,
    tohru.shimizu@{nict.go.jp, atr.jp}, kikui.genichiro@lab.ntt.co.jp
```

Abstract

We propose a method for measuring communication efficiency from the viewpoint of conveying essential information in task-oriented spoken dialog translation. We present the results of one dialog experiment using speech-to-speech translation systems and a similar experiment using the Wizard of Oz method, which was carried out using hidden interpreters instead of a speech-to-speech translation system. We also present the relative performance score of the speech-to-speech translation system, which was obtained by comparing the machine's performance with that of humans, i.e., hidden interpreters. Finally, we discuss the relationship between users' linguistic behavior and system performance. We found that users of the system tended to make shorter utterances without decreasing the number of essential items needed to achieve a task and to improve transmission efficiency by using strategy to control dialogs.

Index Terms: Speech communication, natural language interfaces, speech recognition, interactive systems, human factors.

1. Introduction

Objective metrics such as Bleu [1] have recently been developed to evaluate speech-to-speech translation (S2ST). Though experts are able to understand the results of such evaluation, it is usually very difficult for non-experts to understand them.

To make S2ST system easier for non-experts to use, their usability must be evaluated. The three ISO-recommended usability parameters, i.e., effectiveness (measured as dialog success rate), efficiency (measured as time taken to complete a task), and user satisfaction (as reported in a questionnaire) [2] are often used in evaluations of usability. However, it is difficult to conduct comprehensive evaluations because factors such as difficulty of tasks and personalities of prospective users can confound results.

If we limit the system to the covered domain, effectiveness mainly concerns the amount of essential information in task-oriented dialog, i.e., task difficulty. User satisfaction is the most subjective and highly dependent feature of user interface. To improve the validity of usability evaluations, researchers have developed quantitative metrics to measure user satisfaction. For example, the PARADISE framework enables designers to predict user satisfaction from a linear combination of objective metrics such as mean recognition score and task completion [3].

In this paper, we focus on efficiency. To evaluate a system for processing spoken dialog between a human and a machine, Glass et al. proposed a method based on measuring the efficiency of conveying essential information in task-oriented dialogs [4]. To measure spoken language processing technologies for speech communication as natural language interfaces, we propose a method of evaluating communication efficiency based on expanding Glass et al.'s method from one-way to two-way communication. We also present a new measure of speakers' linguistic behaviors.

In contrast to Glass et al.'s system, Paek proposed an empirical method for evaluating dialog systems [5]. The method measures the distance of the system's performance from the gold standard, i.e., human performance data obtained by a carefully controlled Wizard of Oz (WOZ) experiment. To improve usability for non-experts such as prospective users, it would be useful to measure the performance of the S2ST system against that of humans, i.e., hidden interpreters, as in the WOZ experiment. Thus, we present the results of one dialog experiment using S2ST systems and another using WOZ, which was conducted with hidden interpreters instead of an S2ST system. We then compare the performance of the machine with that of humans, i.e., hidden interpreters. The scoring can be considered an expansion of Paek's empirical method from dialog systems to S2ST systems.

2. Evaluation metrics

Certain utterances are essential for achieving tasks. For example, a customer must make utterances to convey information such as "iced coffee," "large size," "a hot dog," and "no mustard", and a shop assistant must make utterances to convey price information to the customer. Using the items conveyed in these utterances, we define three measures, *Provided Density (PD)*, *Transmission Density (TD)*, and *Transmission Efficiency (TE)*, as follows:

$$PD = \frac{\# \ of \ Items \ in \ Utterances}{\# \ of \ Informative \ Utterances} \tag{1}$$

$$TD = \frac{\# \ of \ Transmitted \ Items}{\# \ of \ Informative \ Utterances} \tag{2}$$

$$TE = \frac{\# \ of \ Transmitted \ Items}{\# \ of \ Items \ in \ Utterances} \tag{3}$$

In this paper, PD is specifically defined to measure speakers' linguistic behaviors. PD indicates the average number of
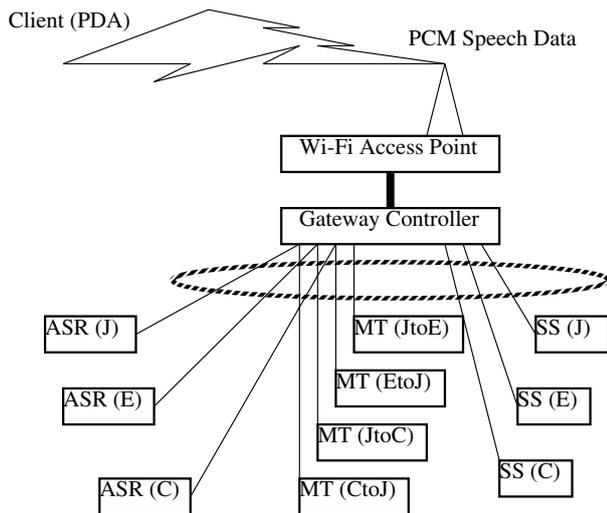
Figure 1: Configuration of experimental S2ST system

Table 1: Training data for speech recognition

		Japanese	English	Chinese
AM	No. of speakers	400	384	540
	Total recording time [h]	38	150	257
LM	No. of sentences	852k	710k	510k
	No. of word tokens	8.7M	6.1M	3.5M
	Vocabulary size	66k	44k	38k

essential items in informative utterances, which are utterances that convey new essential items to others. TD is a measure based on the listener's viewpoint, and TE is a measure of the performance of communication mediators such as S2ST systems or hidden interpreters.

Glass et al. proposed two metrics for evaluating the efficiency of spoken dialog systems: *query density (QD)* and *concept efficiency (CE)* [4]. A 'query' corresponds to an informative utterance in this paper and a 'concept' corresponds to an item essential for achieving a task. Query density (QD) is the mean number of concepts understood by the system per user query. Concept efficiency (CE) is the ratio of concepts understood by the system to the total number of concepts uttered by the user. That is, $0 \leq CE \leq 1$.

In this paper, we considered all the utterances made by each speaker participating in the dialogs. TD and TE can be considered expansions of QD and CE, respectively, from one-way to two-way communication; Thus, $0 \leq TE \leq 1$.

3. Experimental system

Figure 1 shows the overall configuration of the server-client S2ST system. It consists of several terminals (clients) and a speech translation server. The terminals and server are connected via a wireless data network.

The speech translation server consists of a "gateway" and component engines. The functions of the gateway include controlling information comprised of speech data, text data, and system messages between the terminals and component engines. The three major functions of the S2ST are automatic speech recognition (ASR), which includes a noise-suppressing function, machine translation (MT), and speech synthesis (SS) for every language or language pair. These run on the server.

An utterance spoken to one terminal is sent to the gateway. The gateway server then calls the ASR, MT, and SS engines, in that order, to obtain the translated text and speech. Finally, the gateway sends the resulting text and compressed synthesized speech back to the terminal.

3.1. Automatic speech recognition (ASR)

The speech recognition engine, which is called ATRASR, uses a two-pass decoding scheme. In the first pass, a frame-synchronous decoder creates a word graph using HMM acoustic models and class-based bigram language models. The second pass re-scores the word graph using word-based trigrams.

3.1.1. Acoustic model (AM)

The MDL-SSS algorithm [6] is used to train acoustic models. This algorithm automatically determines both the total number of states and maximum number of states per triphone based on the Minimum Description Length (MDL) criterion. Japanese, English, and Chinese acoustic models were trained using approximately 500 speakers, as shown in Table 1. The feature extraction parameters were 25 dimensional vectors (12 MFCC + 12 Δ MFCC + Δ log power) extracted from 20-ms-long windows with a 10-ms shift.

3.1.2. Language model (LM)

The statistical language model was trained using a large corpora from the travel domain [7], as shown in Table 1. To obtain a robust language model, the multi-class composite N-gram was used [8].

3.2. Machine translation (MT)

The translation modules were automatically constructed from a large corpus from the travel domain. The translation component consists of two corpus-based machine translation modules called SAT [9] and HPATR2 [10], and a selection module [11]. First, these two translation modules separately translate a given sentence, and then the selection module chooses the better result.

3.3. Speech synthesis (SS)

The speech synthesis module, which is called XIMERA, consists of four major modules: a text processing module, a prosodic parameter generation module, a segment selection module, and a waveform generation module [12]. The languages XIMERA is trained in are Japanese, Chinese, and English. We used XIMERA for Japanese and Chinese and AT&T Natural Voice[TM] for English because English XIMERA was not available when the dialog experiments were conducted.

4. Dialog experiments

4.1. Conditions

Even when a large corpus from a specific domain (i.e. travel) is used to train S2ST systems, the sentence coverage of the

Table 2: Basic characteristics of dialog data and system performance

	S2ST				WOZ	
Language direction	JtoE	JtoC	EtoJ	CtoJ	JtoC	CtoJ
Number of tasks	40	46	40	46	67	67
Number of utterances	326	419	392	572	363	366
Average number of words in one utterance	6.9	6.7	6.0	5.4	10.5	9.2
Perplexity of language model	32	28	38	92	25	139
Word accuracy in ASR	96%	97%	88%	84%	–	–
Correct utterances in ASR	82%	87%	64%	55%	–	–
Canceled utterances	23%	22%	29%	45%	–	–
Correctly translated utterances	81%	80%	76%	64%	–	–

domain is generally inadequate. In particular, lack of certain proper nouns causes some recognition and translation errors. It is difficult to include in advance enough of the proper nouns generally used in the domain. To focus on evaluating communication efficiency, we decided to use tasks that users would expect to complete with a normal amount of effort. The guidelines were as follows:

Task Tasks should be selected from categories with a large amount of training data available in the corpora. The three categories of shopping, hotel reservations, and minor trouble-shooting in a hotel or restaurant were selected.

Proper noun All the proper nouns essential to achieving the given tasks should be included in the lexicon.

4.2. Dialog experiment using S2ST

We conducted a dialog experiment using the S2ST. The experimental conditions were as follows:

Subjects Japanese-English (6 pairs); Japanese-Chinese (6 pairs). There were no duplications.

Instructions When speaking to the system, subjects were instructed to speak clearly and concisely. To ensure that they concentrated on completing the task, subjects were asked to utter only essential information and to avoid uttering non-task-related sentences.

Training Each pair conducted over 20 dialogs in the experiments, with about half of the dialogs being completed during the morning session. The evaluation was done using the data from the afternoon session. This meant that each subject had completed 10 dialogs in the morning session and was very familiar with the system.

Rejection of recognized text The subject was allowed to cancel his/her own recognition results and to repeat the utterance when the quality of the recognition result was poor.

4.3. Dialog experiment using hidden interpreters

We also conducted a dialog experiment using the WOZ method. The experiment involved using hidden interpreters instead of the S2ST. The experimental conditions were as follows:

Subjects Japanese-Chinese (6 pairs). There were no duplications.

Interpreters There were two interpreters in a different room from the subjects. One interpreted from Japanese to Chinese and the other from Chinese to Japanese. They heard the subjects' voices and typed their translations on PCs. The translated text and synthesized speech were transmitted to the subjects, who were unaware that hidden interpreters were translating the dialog.

Tasks and instructions The same tasks and instructions as for the S2ST system were given to the subjects.

Time restriction for each utterance Each utterance had to be made within eight seconds.

Rejection of erroneous utterances The hidden interpreters sent a message asking subjects to try again if they made over-long utterances, i.e., more than eight seconds, or if errors occurred.

5. Experimental results

5.1. Basic characteristics and system performance

Table 2 shows the basic characteristics of the collected dialogs for the S2ST and WOZ experiments, as well as the speech recognition performance and translation performance for the S2ST experiment (the latter was evaluated subjectively). Word accuracy in ASR, correct utterances in ASR, and correctly translated utterances in Table 2 were calculated using valid utterances. Valid utterances were those processed by the S2ST system, i.e., those the subject did not cancel. Note that correctly translated utterances included erroneous sentences that were able to convey correct information with insignificant, correctable errors in spite of which the subjects were easily able to infer meaning.

Table 2 shows that the system in the S2ST experiment achieved more than 95% word accuracy for ASR and approximately 80% correct translations for JtoE and JtoC, although about 20% of sentences were canceled by the subject as too erroneous. The system correctly translated 76% of the EtoJ sentences even though about 30% of the sentences were canceled. The system correctly translated 64% of the CtoJ sentences with 45% of the sentences being canceled.

5.2. Communication efficiency and relative score

Table 3 shows the experimental results for Provided Density (PD), Transmission Density (TD), and Transmission Efficiency (TS), which were calculated using valid informative utterances.

In spoken communication dialogs, participants often utter greeting utterances such as "May I help you?" and confirmation utterances. We considered only those utterances that were necessary to convey new essential items to other participants to be informative utterances. Along with informative utterances,

Table 3: Experimental results

	S2ST		WOZ
	JE/EJ	JC/CJ	JC/CJ
Provided density (PD)	1.45	1.43	1.59
Transmission density (TD)	0.97	0.88	1.46
Transmission efficiency (TE)	67%	62%	91%

Table 4: Relative performance score

	S2ST		WOZ
	JE/EJ	JC/CJ	JC/CJ
Average number of transmitted items per valid utterance	0.45	0.34	0.74
Relative performance score	60%	46%	100%

greetings and confirmations were also considered valid utterances. The number of confirmation utterances usually increased as the efficiency of the S2ST systems decreased. Table 4 shows total communication efficiency, which was calculated on the basis of all valid utterances, including greeting and confirmation utterances. Table 4 also shows the relative performance score of the S2ST system obtained by comparing the performance of the machine with that of humans, i.e., hidden interpreters.

6. Discussion

6.1. Users' linguistic behavior

As can be seen in Table 2, the average number of words in one utterance in the S2ST experiment was less than that in the WOZ experiment. As can be seen in Table 3, however, the PD of the S2ST experiment for both JE/EJ and JC/CJ was almost the same as that in the WOZ experiment. This means that users of the S2ST system made shorter utterances without decreasing the number of essential items required to achieve a task, i.e., they tended to make fewer redundant utterances.

6.2. System performance and dialog strategy

As can be seen in Table 4, the total communication efficiency for task achievement by the JE/EJ S2ST system was 60% when it was compared with the WOZ, but that of the JC/CJ S2ST system was 46%. As can be seen in Table 2, the ASR and MT performance for JtoE was almost the same as that for JtoC. However, the ASR and MT performance for EtoJ was better than that for CtoJ. The current performance for CtoJ lowered the overall communication efficiency for task achievement by the JC/CJ S2ST system.

However, as can be seen in Table 3, the TE for JC/CJ was almost the same as that for JE/EJ. Based on analysis of the content, subjects speaking Japanese for translation to Chinese often changed the dialog strategy from general questions to a yes/no style based on meaningful chunks in the translation results and contextual information. The accuracy of yes/no style speech was sufficient to continue the dialog even for Chinese. In such cooperative dialogs, experienced people seem to tend to adapt their dialog strategy to the system.

7. Conclusion

We proposed a method for measuring communication efficiency from the viewpoint of conveying essential information in task-oriented spoken dialog translation. We showed the results of one dialog experiment using S2ST systems and another using the WOZ method, which was conducted using hidden interpreters. We also presented a relative performance score for the S2ST system by measuring the performance of the machine against that of humans, i.e., hidden interpreters. We found that users of the system tended to make shorter utterances without decreasing the number of essential items required to achieve tasks and also to improve transmission efficiency by strategically controlling dialogs. In the future, we plan to conduct similar research on the usability of natural spoken language interfaces.

8. References

[1] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[2] ISO (International Standardization Organization), "ISO 9241: Ergonomic requirements for office work with visual display terminals (VDTs), Part 11: Guidance on usability," 1998, http://www.iso.org.

[3] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "PARADISE: A framework for evaluating spoken dialogue agent," in *ACL*, 1997, pp. 271–280.

[4] J. Glass, J. Polifroni, S. Seneff, and V. Zue, "Data collection and performance evaluation of dialogue system: The MIT experience," in *ICSLP*, 2000, vol. IV, pp. 1–4.

[5] T. Paek, "Empirical methods for evaluating dialog systems," in *ACL 2001 Workshop on Evaluation Methodologies for Language and Dialogue Systems*, 2001, pp. 3–10.

[6] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic generation of non-uniform context-dependent HMM topologies based on the MDL criterion," in *Eurospeech*, 2003, pp. 2721–2724.

[7] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Eurospeech*, 2003, pp. 381–384.

[8] H. Yamamoto, S. Isogai, and Y. Sagisaka, "Multi-class composite N-gram language model," *Speech Communication*, vol. 41, pp. 369–379, 2003.

[9] T. Watanabe and E. Sumita, "Example-based decoding for statistical machine translation," in *MT Summit IX*, 2002, pp. 410–417.

[10] K. Imamura, T. Watanabe, and E. Sumita, "Practical approach to syntax-based statistical machine translation," in *MT Summit X*, 2005, pp. 267–274.

[11] Y. Akiba, T. Watanabe, and E. Sumita, "Using language and translation models to select the best among outputs from multiple MT systems," in *COLING*, 2002, pp. 8–14.

[12] H. Kawai, T. Toda, J. Ni, and M. Tsuzaki, "XIMERA: A new TTS from ATR based on corpus-based technologies," in *5th ISCA Speech Synthesis Workshop*, 2004, pp. 179–184.