Quantitative Evaluation of Effects of Speech Recognition Errors on Speech Translation Quality

Kenko Ohta[†‡], *Keiji Yasuda*[†], *Genichiro Kikui*[†] *and* *Masuzo Yanagida*[‡]
[†]ATR Spoken Language Translation Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
[‡]Doshisha University
1-3, Tatara-Miyakodani, Kyo-Tanabe, 610-0394, Japan
{kenkoh.ohta, keiji.yasuda, genichiro.kikui}@atr.jp myanagid@mail.doshisha.ac.jp

Abstract

This paper investigates the relationship between the quality of speech translation outputs and the errors in a speech recognition subsystem. In this study, we assume that a speech translation system is a sequential combination of speech recognition and automatic translation subsystems. We conducted speech translation experiments while changing parameters in the speech recognition subsystem to get different speech recognition results, which were then fed into the translation subsystem. We applied regression analysis to the interrelationship between the speech recognition outputs and the final translation results. We found that particular kinds of speech recognition errors, including deletion of punctuation marks and substitutions of nouns, cause severe semantic errors in speech translation. We also found that the final translation quality degrades logarithmically with respect to the number of speech recognition errors.

1. Introduction

Speech translation integrates the processes of speech recognition, machine translation and speech synthesis. This paper concentrates on the relation between the first two functionalities, specifically how speech recognition errors affect the output of machine translation. This work assumes a single-best cascaded architecture, where a single output of the speech recognition subsystem is fed into the machine translation subsystem. This architecture is widely used [1], and it provides the foundation of a more complex architecture that employs an *N*-best list or a word lattice as intermediate data between two subsystems [2].

It is natural to think that speech recognition errors in a cascaded system deteriorate final translation outputs. There have been, however, few analyses on the quantitative relationship between these factors, much less analyses on what kinds of speech recognition errors (e.g., deletion of nouns, insertion of prepositions, etc.) affect the accuracy of translation and to what degree. These detailed analyses are crucial in optimizing system parameters such as noun insertion penalty and the threshold of rejecting speech recognition results.

In this work, we first conducted speech translation experiments using our speech translation system. Then we applied regression analyses between error types in each speech recognition result and the accuracy of the corresponding translation.

Section 2 briefly describes the experimental system. Section 3 overviews the corpus used for analyses and explains how to extract evaluation data from this corpus by subjective evaluation. Section 4 explains the analysis procedure used to clarify the relationship between speech recognition errors and the quality of automatic translation, and this section also gives analysis results. In Section 5, we discuss the verification of the analysis results. Then, in Section 6, we conclude this study and describe future works.

2. Outline of experimental system

2.1 Speech recognition subsystem

The speech recognition subsystem is a large-vocabulary continuous-speech recognizer that works as a 2-pass decoder [3]. In the first pass, it recognizes the input speech using HMM as acoustic models and a class-based 2-gram language model (LM), and then it outputs a word lattice. In the second pass, it re-scores the word lattice using each of the 3-gram LM, and then it outputs the 1-best hypothesized word sequence.

2.2 Automatic translation subsystem

The automatic translation subsystem of the experimental system is SAT [4], a statistical translation system. The decoding algorithm for finding the best translation result is based on a "greedy" search, where it retrieves multiple translation candidates from a bilingual corpus then applies modification operations to these candidates to improve translation scores. This is essentially the operation performed by the IBM model 4.

3. Corpus

3.1 Machine-Aided Dialogues (MAD) corpus

We used the Japanese part of a bilingual dialog corpus called MAD4 [5]. MAD4 data were collected by asking English and Japanese native speakers to talk through our experimental speech-to-speech translation system. MAD4 contains 1370 utterances by Japanese speakers and 1293 utterances by English speakers. The corpus in each language is divided into two groups: 502 test utterances and the rest for training language models.

3.2 Extracting evaluation data by subjective evaluation

From the 502 Japanese test utterances, we further eliminated utterances whose translations were very bad, even if we used correct transcriptions (i.e., not necessarily speech recognition

outputs) as inputs for translation. The reason is as follows: in these cases, the translation results cannot get worse even if speech recognition fails, in other words, the translation quality of such utterances cannot be degraded by speech recognition errors. To eliminate these utterances, we first evaluated the results of machine translation for transcriptions of 502 Japanese utterances. There are two ways of evaluating translation quality [6-9]: objective and subjective. Here, we employed subjective evaluation. Although subjective evaluation requires a lot of cost and time, it reflects human intuition. An evaluator is asked to classify the translation results into the following four ranks.

A: perfect, B: fair, C: acceptable and D: nonsense

We discarded utterances evaluated as rank C or D. So, the number of analysis targets was reduced to 209 from the 502 utterances.

4. Relation between Speech Recognition Quality and Speech Translation Quality

4.1 Analysis procedure

First, 209 utterances were fed to ATRASR, where five variations of beam width are used for finding a word sequence. Three types of acoustic models (male & female, male-only, and female-only) were used for each value of beam width, where beam width for the male-only acoustic model had only four variations. So, the total number of analysis data is 2926 (=209×5×2+209×4).

Second, speech recognition errors are examined by carrying out DP matching between speech recognition results and prescribed transcriptions.

Third, 2926 recognition results are translated into English.

Fourth, 2926 translation results were evaluated by both subjective evaluation and objective evaluation. The objective evaluation used here is Word Error Rate (WER), defined as follows:

$$WER = \frac{insertion + substitution + deletion}{total} \quad (1)$$

Here, *insertion*, *substitution* and *deletion* denote the number of speech recognition or translation errors of the corresponding types detected by DP matching and *total* denotes the number of words given in the transcription.

Finally, regression analysis was carried out based on the regression model between WER and speech recognition errors. From these results, the relationship between speech recognition errors and the WER of automatic translation results (WER_{trans}) was quantitatively evaluated.

4.2 Statistics on speech recognition errors

Two types of statistics on the results of speech recognition are described in this section. First, classification of speech recognition errors is shown in Table 1.

Table 2 shows the percentage of detected speech recognition errors for each word category.

Table 1. Percentage of insertion, deletion and substitution errors (%)

Deletion	Insertion	Substitution
12	21	67

Table 2. Percentage of each word category (%)

	Deletion	Insertion	Substitution	
			from	to
noun (n.)	5.19	7.46	25.75	24.79
verb (v.)	0.81	1.58	8.20	8.00
auxiliary verb (aux.)	0.82	1.21	7.81	6.59
particle (pa.)	2.28	6.79	15.82	20.95
adjective (adj.)	0.12	0.23	1.04	1.36
adverb (adv.)	0.20	0.46	0.84	1.32
pre-noun adjectival (pre-n.)	0.07	0.20	0.34	0.69
conjunction (conj.)	0.11	0.23	0.81	0.59
interjection (kantoushi) (int.1)	0.23	0.38	1.36	0.52
interjection (kandoushi)(int.2)	0.36	0.32	0.70	0.49
prefix (pref.)	0.47	0.94	1.77	0.99
suffix (suf.)	0.04	0.22	0.62	0.56
punctuation mark (pu.)	1.38	0.58	2.24	0.45

4.3 Analysis based on regression models

4.3.1 Using a simple regression model

A) Simple regression using all data was carried out. The independent variable is taken as WER in speech recognition results (WER_{recog}) and WER_{trans} is regarded as a dependent variable. The regression equation is expressed as follows:

$$WER_{trans} = 1.07 \times WER_{recog} + 0.074 \quad (2)$$

The contribution ratio is 0.65. Figure 1 shows a scatter plot and a regression line expressed by eq. (2).

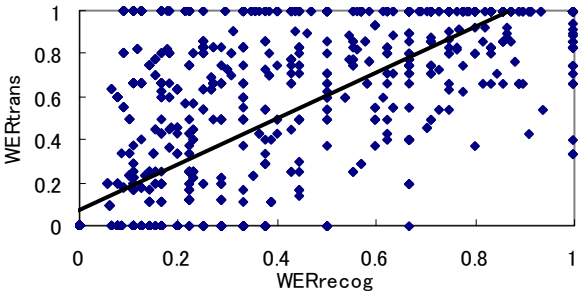

Fig. 1. Scatter plot with all data and regression line

Figure 1 implies that line regression cannot fully describe the relationship when we use all of the data, including those that have a low subjective evaluation.

B) The analysis described in A) was carried out using the data having subjective evaluation rank A or B. Here, we carried out an analysis to see what happens if we use only the data having subjective evaluation rank A or only that having rank B. Results are shown in Fig. 2. From the two figures, we can see that the gradient of the regression line for rank A in Fig. 2 is smaller than that in Fig. 1 and that the gradient of the regression line for rank B in Fig. 2 is larger than that in Fig. 1. This means that an utterance with a low subjective score is easily degraded by speech recognition errors.

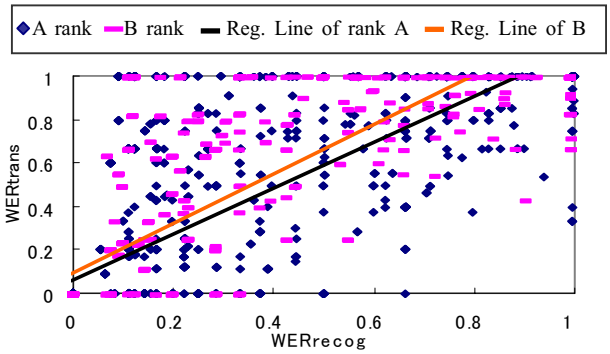

Fig. 2. Scatter plots and regression lines of the data of A and B ranks

C) The analysis in A) is based on a simple linear regression model. However, looking at Fig. 1, WER_{trans} tends to increase logarithmically with respect to WER_{recog}. So, we introduce $LOGWER_{recog}$ as follows:

$$LOGWER_{recog} = \log\{(WER_{recog})^{0.8} + 1\} \quad (3)$$

$LOGWER_{recog}$ was used to carry out linear regression, and the contribution ratio improved to 0.71. Figure 3 shows a scatter plot between $LOGWER_{recog}$ and WER_{tran} together with the regression line.

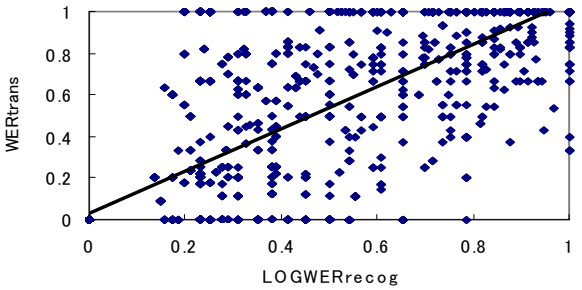

Fig. 3. Linear regression using $LOGWER_{recog}$

4.3.2 Using a multiple regression model

A) A multiple regression analysis using the number of substitution, deletion and insertion errors as separate independent variables was carried out, where three independent variables were transformed by eq. (3). The contribution ratio for this case was 0.67. Each normalized regression coefficient was compared with the others to investigate its influence on speech translation quality. As a normalized regression coefficient becomes larger, an independent variable corresponding to the larger coefficient has more influence on speech recognition quality. Table 3 shows values of normalized regression coefficients.

Table 3. Normalized regression coefficients of three types of recognition errors

Deletion	Insertion	Substitution
0.22	0.25	0.61

According to Table 3, it turns out that a substitution error has the most severe influence on speech translation quality among the three types of recognition errors.

Figure 4 shows a scatter plot, where the abscissa denotes WER estimated by the regression model (WER_{model}) and the ordinate represents WER_{trans}.

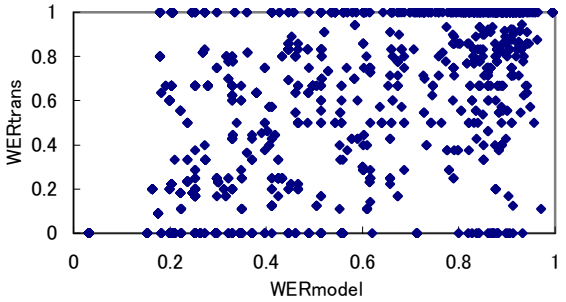

Fig. 4. Multiple regression using substitution, deletion and insertion errors

B) Multiple regression analysis was carried out by using three types of speech recognition errors (deletion, insertion and substitution) of word categories (noun, verb, adjective, adverb, etc) as independent variables, and WER_{trans} was regarded as the dependent variable. The number of coefficients was 195, and the contribution ratio of this analysis was 0.69. To compare the influence on speech translation quality among independent variables, normalized regression coefficients are shown in Table 4.

Table 4. Normalized regression coefficients of each independent variable (Variables having normalized regression coefficient larger than 0.1 are shown in Italic bold font)

Type of error	Normalized regression coefficient	Type of error	Normalized regression coefficient
del. (n.)	0.077	sub. (n. and suf.)	0.046
del. (v.)	0.072	sub. (v. and n.)	0.058
del. (pu.)	***0.195***	***sub. (v. and v.)***	***0.113***
ins. (n.)	0.071	sub. (v. and pa.)	0.057
ins. (v.)	0.065	sub. (aux. and n.)	0.093
ins. (par.)	0.096	sub. (aux. and v.)	0.07
ins. (conj.)	0.042	sub. (aux. and aux.)	0.073
ins. (pan.)	0.079	sub. (pa. and n.)	0.056
sub. (n. and n.)	***0.319***	***sub. (pa. and pa.)***	***0.144***
sub. (n. and v.)	0.089	sub. (adv. and n.)	0.046
sub. (n. and pa.)	***0.14***	sub. (conj. and conj.)	0.076
sub. (n. and ad.)	0.08	sub. (int.2 and v.)	0.039
sub. (n. and pre-n.)	0.061	sub. (pref. and n.)	0.059
sub. (n. and conj.)	***0.1***	sub. (suf. and suf.)	0.077
sub. (n. and int.2)	0.075		

From Table 4., it can be seen that substitution errors between nouns have the most severe influence on speech translation quality.

Figure 5 shows a scatter plot, where the abscissa denotes WER_{model} and the ordinate shows WER_{trans}.

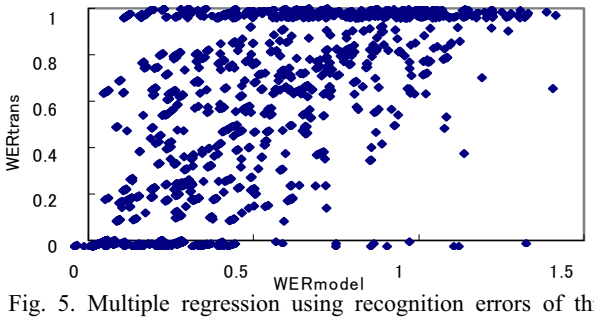

Fig. 5. Multiple regression using recognition errors of three types (substitution, deletion and insertion) for each word category as independent variables.

5. Discussion

Multiple regression analysis was done using speech data having low WER_{recog} values spanning between 0 and 0.2. The reason why we used such data is that for utterances of high WER_{recog}, WER_{trans} is also high, but even when WER_{recog} is low, WER_{trans} may not always be low. The number of data of low WER_{recog} value between 0 and 0.2 is 1939. As for the case mentioned above, the results of multiple regression analysis are compared to the results of multiple regression analysis of B). Table 5 shows normalized regression coefficients.

Table 5. Normalized regression coefficients of multiple regression using the data of low WER_{recog} between 0 and 0.2

Type of error	Normalized regression coefficient	Type of error	Normalized regression coefficient
del. (n.)	0.069	***sub. (v. and v.)***	***0.111***
del. (v.)	0.05	sub. (v. and pa.)	0.043
del. (pu.)	***0.242***	sub. (aux. and n.)	0.085
ins. (n.)	0.049	sub. (aux. and v.)	0.069
ins. (v.)	0.053	sub. (aux. and aux.)	0.076
ins. (pa.)	***0.129***	sub. (pa. and n.)	0.056
ins. (conj.)	0.037	***sub. (pa. and pa.)***	***0.135***
ins. (int.2)	0.078	sub. (pa. and int.1)	0.034
ins. (pu.)	0.065	sub. (pa. and suf.)	0.035
sub. (n. and n.)	***0.31***	sub. (adv. and n.)	0.042
sub. (n. and v.)	0.075	sub. (adv. and pre-n.)	0.031
sub. (n. and pa.)	***0.125***	sub. (pre-n. and pa.)	0.041
sub. (n. and ad.)	0.068	sub. (conj. and conj.)	0.097
sub. (n. and pre-n.)	0.069	sub. (int.2 and v.)	0.034
sub. (n. and conj.)	***0.125***	sub. (int.2 and int.2)	0.046
sub. (n. and int.2)	0.066	sub. (pref. and n.)	0.051
sub. (n. and suf.)	0.041	sub. (pref. and pa.)	0.057
sub. (v. and n.)	0.046	sub. (suf. and suf.)	0.094

In comparing Table 5 with Table 4, it can be seen that the speech recognition errors having a severe influence on speech translation quality in Table 5 are almost the same as those in Table 4, except for insertion errors in particles. The results predict that speech recognition errors common to both Table 4 and Table 5 have a severe influence on speech recognition quality. Errors showing a particularly strong effect include substitution between nouns, substitution between noun and particle, substitution between noun and conjunction, substitution between verbs, substitution between particles and deletion of punctuation marks.

Next, we discuss the validity of logarithmic transformation of independent variables. Accordingly, Table 6 shows the contribution ratios obtained for analyses using logarithmic transformation and those with no transformation.

Table 6. Comparison of contribution ratios

	No transformation	Logarithmic transformation
Simple regression	0.65	0.71
A) multiple regression	0.60	0.67
B) multiple regression	0.65	0.69

The result of a t-test shows significant difference at a 5% hazard rate. Therefore, it can be said that degradation of translation quality can be roughly approximated with a logarithmic function of speech recognition quality, rather than with a linear function of speech recognition quality.

6. Conclusion

Regression analyses were carried out to clarify the relationship between speech recognition errors and automatic translation quality in speech translation systems. Experimental results show that

1) According to simple regression analysis, it can be seen that utterances of low translation quality are easily degraded by speech recognition errors.
2) According to multiple regression analysis, it can be said that substitution errors have the most severe influence on speech translation quality among deletion, insertion and substitution.
3) It also can be said that substitution between nouns has the most severe influence on the speech translation quality among variations of insertion errors.
4) It can be said that degradation of automatic translation quality roughly approximated with a logarithmic function of speech recognition quality, rather than a linear function of speech recognition quality.

Future works should extend the results of the current study to research on the actual effects of optimizing a speech translation system.

Acknowledgement

The research reported here has been supported in part by a contract with the National Institute of Information and Communications Technology entitled, “A study of speech dialogue translation technology based on a large corpus.”

References

[1] Waibel, A., “Speech Translation: Past, Present and Future”, ICSLP2004
[2] Zhang, R., Kikui, G., Yamamoto, H., Soong, F., Lo, W., Watanabe, T. and Sumita, E., “Improved Spoken Language Translation Using N-best Speech Recognition Hypotheses”, ICSLP2004
[3] Nakamura, S., Markov, K., Jitsuhiro, T., Zhang, J, Yamamoto, H. and Kikui, G., “Multi-Lingual Speech Recognition System for Speech-To-Speech Translation”, Proc. of IWSLT2004, pp. 147-154
[4] Watanabe, T.et al., “Example-based Decoding for Statistical Machine Translation”, *MT Summit IX*, pp. 410-417, 2003.
[5] Takezawa, T. and Kikui, G., “Collecting Machine-Translation-Aided Bilingual Dialogues for Corpus Based Speech Translation”, *Proc. EUROSPEECH2003*, pp. 2757-2760, 2003
[6] Papineni, K. et al., “Bleu: a Method for Automatic Evaluation of Machine Translation”, *Proc. of ACL*, pp311-318, 2002.
[7] NIST, “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics”, http://www.nist.gov/, 2002.
[8] Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S. and Yanagida, M., “Automatic Evaluation Method of Translation Quality Using Parallel Corpus”, *IPSJ*, Vol. 43, No. 7, pp. 2108-2116, 2002.
[9] Sugaya, F., Takezawa, T., Yokoo, A. and Yamamoto, S., ”Proposal of an Evaluation Method for Speech Translation Capability by Comparing a Speech Translation System with Humans and Experiments Using the Method”, *EICJ*, D-II, Vol. J84-D-II, No. 11, pp. 2362-2370, 2001.