IEICE TRANS. INF. & SYST., VOL.E101–D, NO.3 MARCH 2018

PAPER

Classification of Utterances Based on Multiple BLEU Scores for Translation-Game-Type CALL Systems

Reiko KUWA[†], ***Nonmember***, **Tsuneo KATO**[†], ***Member***, ***and*** **Seiichi YAMAMOTO**[†a)], ***Fellow***

SUMMARY This paper proposes a classification method of second-language-learner utterances for interactive computer-assisted language learning systems. This classification method uses three types of bilingual evaluation understudy (BLEU) scores as features for a classifier. The three BLEU scores are calculated in accordance with three subsets of a learner corpus divided according to the quality of utterances. For the purpose of overcoming the data-sparseness problem, this classification method uses the BLEU scores calculated using a mixture of word and part-of-speech (POS)-tag sequences converted from word sequences based on a POS-replacement rule according to which words are replaced with POS tags in *n*-grams. Experiments of classifying English utterances by Japanese demonstrated that the proposed classification method achieved classification accuracy of 78.2% which was 12.3 points higher than a baseline with one BLEU score.

key words: *interactive CALL, classification, learner corpus, BLEU*

1. Introduction

Modern globalization has produced many more opportunities to communicate in a second language (L2) than ever. Learning an L2 is becoming important for an increasingly large number of people. As a convenient and economical self-learning method, computer-assisted language learning (CALL) has attracted great interest. Starting from a fixed and passive style, such as computer-assisted pronunciation-training (CAPT) systems [1]–[3], the development of CALL systems has expanded to be flexible and interactive, such as dialogue-game-type systems [4]–[8] and translation-game-type systems [9], [10], in accordance with advances in automatic speech recognition (ASR). Such interactive CALL systems have the advantages of training learners to produce utterances on their own then giving corrective feedback regarding errors in the utterances, as well as training in pronunciation.

Accurate ASR performance, appropriate corrective feedback regarding learner errors, and correct evaluation of proficiency are essential factors in providing effective tutoring through interaction. However, accurate recognition of L2 speech remains a challenging task, even for state-of-the-art ASR, because L2 speech has significantly different features, such as pronunciation, prosody, vocabulary, grammar, and disfluencies, from those of native tongue (L1) speakers. Interactive CALL systems adopt various methodologies to maintain high recognition performance for L2 speech. Dialogue-game-type CALL systems, which require learners to construct expressions by simulating real-life conversation exercises, maintain high prediction performance of learner utterances by severely restricting the domain and conversational flow of the topic. On the other hand, translation-game-type CALL systems guide the learner's utterances by presenting sentences in their L1 as their responses and maintain high recognition performance of learner utterances [9], [10].

Various methods of lexical- or grammatical-error detection have been proposed to give appropriate corrective feedback regarding errors that learners make when using interactive CALL systems, [11]–[14]. To accelerate advances of this technical area, the speech and language technologies in education (SLaTE) community recently set a shared task of binary content classification of learner responses into either a linguistically correct class or a reject class on a translation-game-type CALL system [15], and the results were presented at the SLaTE 2017 conference [16], [17]. In such tasks, no matter how carefully a method is designed, a proportion of learner utterances will greatly differ from expressions containing the expected errors, which makes it difficult not only to give corrective feedback but also to correctly recognize the words. The learner utterances containing the expected errors are *correctable* by giving adequate corrective feedback. In contrast, learner utterances, which contain erroneous expressions that greatly differ from the expected errors, cannot be corrected, and we call this type of utterance *uncorrectable*. To give appropriate corrective feedback regarding *correctable* utterances while not attempting to do this for *uncorrectable* utterances, an interactive CALL system should be able to classify such utterances.

This type of classification problem has been investigated as a rejection algorithm for an ASR result [18]. One basic approach to solve this problem is to compare outputs from two ASR systems: task-specific ASR, which exhibits high recognition performance regarding "in-grammar" utterances but misrecognizes "out-of-grammar" utterances as "in-grammar" utterances, and universal ASR, which usually exhibits moderate recognition performance for both classes of utterances.

As a method of achieving higher classification performance for translation-game-type CALL systems by extending the above method, Nagai et al. proposed a method of using such particular features that L2 speakers who have the

Manuscript received May 7, 2017.
Manuscript revised October 31, 2017.
Manuscript publicized December 4, 2017.

[†]The authors are with the Graduate School of Science and Engineering, Doshisha University, Kyotanabe-shi, 610–0394 Japan.
a) E-mail: seyamamo@mail.doshisha.ac.jp

DOI: 10.1587/transinf.2017EDP7151

same L1 tend to make similar lexical or grammatical mistakes affected by the characteristics of their L1. They used two task-specific ASRs: one designed to receive *correct* and *correctable* utterances and the other for *uncorrectable* utterances, in addition to a universal ASR to classify learner utterances [19]. They showed that these particular features are effective in classifying utterances for CALL systems. However, their method uses a finite state automaton (FSA) as the language model, which makes it difficult to create a language model that accepts various expressions.

Inspired by the result that classification performance can be improved using these particular features, we set out to develop a classification method that is effective, even when the variety of collected expressions is relatively small, by using the partial-word sequences of utterances as features for classification.

Most *uncorrectable* utterances are not completely different from *correctable* or correct utterances lexically. They contain partial word sequences with perfectly correct ones, but they have serious lexical or syntactic errors. Therefore, to correctly classify learner utterances having such features into three classes, a classification method using such features of partial-word sequences that learner utterances have is needed.

The "bilingual evaluation understudy (BLEU) score" was proposed as an evaluation method of translation quality based on the partial-word sequences of utterances in the field of machine translation [20]. BLEU is formulated as a geometric mean of n-gram precisions between machine-translated sentences and their human reference translation to evaluate the quality of machine-translated sentences, and it was shown to be highly correlated with subjective evaluation from a statistical perspective. Recently, the BLEU score has been used as an additional feature of a support-vector-machine based content-classification system for the shared task at SLaTE 2017 [17]. However, BLEU is not as highly correlated in a bunch of sentences if it is used to evaluate the quality of a single sentence.

We propose a classification method for "translation-game-type" CALL systems, which uses multiple BLEU scores as features for the classifier. The purpose with our method is to classify utterances into three classes (*correct*, *correctable*, and *uncorrectable*) in accordance with a learner corpus. Based on the observation that L2 speakers of the same L1 make similar types of lexical or grammatical mistakes, we assumed that utterances in the three classes have similar partial-word sequences as other utterances in the same class, and BLEU scores calculated using utterances in the same class are higher than those calculated using utterances in the other classes. The proposed method applies three types of BLEU scores at the same time as features of a classifier of CALL systems.

Considering that an expression in a source language has multiple translations in the target language, creation of a high-quality bi-text corpus requires a variety of corresponding expressions in the target language. It is in principle very difficult to collect a sufficiently large number of *uncorrectable* utterances, though there is a tendency of L2 speakers of the same L1 making similar types of lexical or grammatical mistakes. When the number of reference translations is not sufficient, the performance of BLEU degrades because there are few matches between the n-grams of a learner utterance and those of the reference translations. To solve this data-sparseness problem, the proposed method introduces a rule with which words are to be replaced with their corresponding part-of-speech (POS) tags in each n-gram of both learner utterances and reference translations when their occurrence frequencies in a learner corpus are lower than a threshold.

It should be noted that our proposed method does not request FSA as the language model and can use stochastic language models that have in principle larger coverage than FSA.

The rest of this paper is structured as follows. In Sect. 2, we give an overview of our previously proposed translation-game-type CALL system. In Sect. 3, we describe our proposed classification method using multiple BLEU scores. In Sect. 4, we present a rule of replacing words with POS tags in calculating the BLEU scores to solve the data-sparseness problem. We explain our experiments and the results in Sect. 5 and conclude this paper and discuss future work in Sect. 6.

2. Overview of the Interactive CALL System

We have developed a translation-game-type CALL system with which Japanese learners can practice English conversation while enjoying a role-playing game in situations such as shopping and making hotel reservation [21]. Figure 1 shows a screenshot of our translation-game-type CALL system and Fig. 2 shows a schematic of its flow.

After scene selection, the system asks a question to a learner in audio and displays its transcription at the top of the screen. A Japanese sentence of an expected response a learner should utter in English is also displayed at the top line of the bottom part of the screen. There are two purposes

Fig. 1 Screenshot of translation-game-type CALL system.

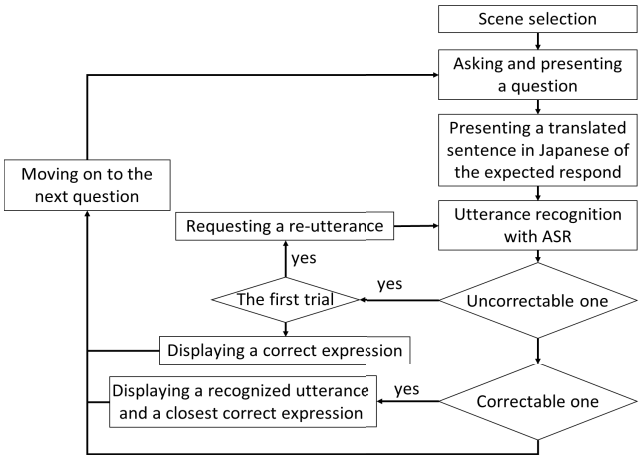

Fig. 2 Flow of translation-game-type CALL system.

of displaying the Japanese sentences. One is to help learners respond more easily, and the other is to maintain high recognition accuracy by restricting variations in learner utterances.

After a learner's utterance is recognized using ASR, the subsequent process branches into the following three patterns depending on the decision of classifying the utterance; *correct*, *correctable*, or *uncorrectable*.

If the learner's utterance is a *correct* response, the system must be able to notify the learner that his/her utterance is perfect and has no need of modification. Therefore, the system moves on to the next question.

If the learner's utterance is a *correctable* utterance, the system displays the transcription of the recognized utterance and displays, as feedback, one correct expression selected as the closest in the corpus to enable the learner to compare his/her utterance with it.

If the learner's utterance is an *uncorrectable* utterance, the system requests a re-utterance. If the re-utterance is again an *uncorrectable* utterance, the system gives a correct expression and moves on to the next question.

3. Utterance Classification Based on Multiple BLEU Scores

The classification flow of the proposed method is shown in Fig. 3. To recognize not only correct utterances accurately but also *correctable* and *uncorrectable* utterances with sufficient accuracy, a learner utterance is recognized using ASR with a task-specific n-gram language model which was trained with various utterances from correct to both erroneous classes. To classify learner utterances accurately, the proposed classification method uses three BLEU scores as features of a classifier.

The three BLEU scores are defined with the following formulation;

$$BLEU = BP * \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log P_n\right), \tag{1}$$

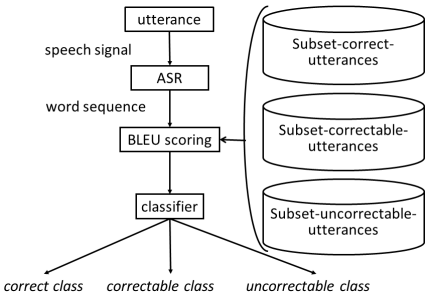

Fig. 3 Schematic classification flow of recognized learner utterances.

where BP denotes the brevity penalty, which prevents very short candidates from receiving too high a score, N denotes the maximum length of word series n of the n-grams; N generally equals 4, and P_n is the modified n-gram precision in the reference translations in each class. A BLEU score ranges from 0 (low quality) to 1 (high quality).

The higher the n of n-gram, the fewer matches there are between n-grams of a learner utterance and those of the reference translations. If there is no match, the BLEU score becomes zero, which often occurs when evaluated at sentence level. To compensate for this phenomenon by smoothing, a modified BLEU ("BLEU+1") was proposed [22]. The proposed method adopts the modified BLEU, instead of the standard BLEU.

For utterance classification, a linear classifier (LC) is used as a classifier because the feature vector of the proposed classification method is low-dimensional. The LC outputs y, where $y \in$ *correct*, *correctable*, and *uncorrectable*, using three BLEU scores as feature parameters;

$$\begin{aligned} y = LC(&BLEU_{\text{Subset-of-correct-utterances}}, \\ &BLEU_{\text{Subset-of-correctable-utterances}}, \\ &BLEU_{\text{Subset-of-uncorrectable-utterances}}) \end{aligned} \tag{2}$$

4. POS-Replacement Rule for BLEU Scoring

BLEU generally evaluates translation quality based on word n-gram precisions (n = 1, 2, 3, 4). Considering that there is often more than one possible English translation for a single Japanese sentence, it is preferable to collect as many references as possible in calculating BLEU scores.

As the number of reference translations becomes smaller, the relation between the BLEU score and subjective evaluation becomes less correlated. This is because the fewer the reference translations, the fewer matches there are between n-grams of a leaner utterance and those of the reference translations. To control sparseness of n-gram space, the proposed method converts word sequences to a mixture of word and POS-tag sequences based on a POS-replacement rule and calculates BLEU scores using the mixture of word and POS-tag sequences.

This rule replaces each word with its corresponding POS tag in each n-gram of both learner utterances and reference translations when its occurrence frequencies in a

Table 1 n-gram expressions based on POS replacement rule.

n-gram	Examples of n-gram (e.g. I want to return this item.)
uni-gram	i, want, to, return, this, item
bi-gram	i-want, want-to, **TO**-return, **VB**-this, **DT**-item
tri-gram	i-want-**TO**, want-to-**VB**, **TO**-return-**DT**, **VB**-this-**NN**
four-gram	i-want-to-**VB**, want-**TO**-return-**DT**, **TO**-**VB**-this-**NN**

learner corpus are lower than a threshold. Furthermore, this POS-replacement rule is provided to prevent a case in which all the words in an n-gram are replaced with the POS tags. If such a case occurs, the evaluation is conducted without taking into account any semantic aspect. This POS-replacement rule is described below.

- POS replacement rule
  If the occurrence frequency of every word that precedes or follows the middle word in each n-gram is lower than a threshold α, the word is replaced with its corresponding POS tag, which is defined as follows;

$$\begin{cases} freq(w) < \alpha & \text{(POS replacement)} \\ freq(w) \geq \alpha & \text{(non-POS replacement),} \end{cases} \quad (3)$$

  Table 1 gives examples in which this POS-replacement rule was applied. The bold uppercase letters denote POS tags. The method of tagging follows the Brown corpus tagset [23]. Hence, "**TO**" is a tag of an infinitive marker, "**VB**" is a tag of a verb (base form), "**DT**" is a tag of a singular determiner, and "**NN**" is a tag of a singular or mass noun. For uni-gram, there is no word to which the POS-replacement rule is applied. For bi-gram, the second word is regarded as the middle word; therefore, the first word is the target word of the POS-replacement rule. For tri-gram, the second word is the middle word; therefore, the first and third words are the target words of the POS-replacement rule. For four-gram, the third word is regarded as the middle word; therefore, the first, second, and fourth words are the target words of the POS-replacement rule.

5. Experiments

5.1 Learner Corpus

We used a learner corpus as reference material, which consists of 2,803 utterances of 41 Japanese-speaking students and one bilingual English/Japanese speaker [21], [24]. Each Japanese student orally produced 66 utterances in English responding to 66 questions in situations such as shopping and making hotel reservations. The bilingual English/Japanese speaker made multiple expressions per response for model answers, considering that a single Japanese sentence has more than one possible translation in English. The quality of each transcribed utterance by the Japanese-speaking students was evaluated on a scale of 1 to 5 in terms of English fluency/adequacy by the bilingual English/Japanese speaker. Table 2 shows details of the scale,

Table 2 Subjective evaluation metric and example sentences.

Level	Criterion	Example
5	Perfect	Where is the cashier?
4	Good	Where do I pay?
3	Non-native	Where pay?
2	Disfluent	Where is I can pay?
1	Incomprehensible	Where can I cash?

Table 3 Acoustic analysis conditions.

Sampling rate	16 kHz, 16 bit
Frame length	20 ms
Frame shift	10 ms
Window type	Hamming
Acoustic features	MFCC 12-order
	log power
	Δ MFCC 12-order
	Δ log power
	ΔΔ MFCC 12-order
	ΔΔ log power

which is a subjective evaluation metric used at the International Workshop of Spoken Language Translation [25], and example sentences that belong to each level.

5.2 Creation of Reference Materials

The learner corpus is divided into three subsets according to the quality of transcribed utterances evaluated by the bilingual English/Japanese speaker: Subset-of-*correct* consisting of the bilingual speaker's utterances and students' utterances graded as 5, Subset-of-*correctable* consisting of students' utterance graded as 3 or 4, and Subset-of-*uncorrectable* consisting of students' utterances graded as 1 or 2. The proposed method uses these three subsets of the learner corpus for calculating three types of BLEU scores. The numbers of utterances in Subset-of-*correct*, Subset-of-*correctable* and Subset-of-*uncorrectable* are 585, 1,285, and 933, respectively.

The utterances in the learner corpus were tagged using a transformation-based tagging system called Brill tagger [26], which has the characteristics of both rule-based taggers and stochastic taggers, and was trained with the Brown Corpus [23]. The word sequences of each utterance were converted to a mixture of word and POS sequences and stored as reference materials.

5.3 Experimental Setup

The hidden Markov model (HMM) toolkit [27] was used as the ASR engine. A tri-phone HMM acoustic model was trained with an L2 English-speech database created from 200 Japanese students (100 males and 100 females) [28]. Table 3 lists the conditions of the acoustic analysis. The pronunciation lexicon consisted of about 35,000 vocabulary words. We developed a *bi*-gram language model trained with 4,472 English transcriptions of utterances made by 65 Japanese students and a native English speaker to reliably estimate stochastic values considering the relatively small

Table 5 Classification results of both proposed methods. Proposed method-1 is classification method using three BLEU scores, and Proposed method-2 is that using three BLEU scores and POS-replacement rule.

	correct			*correctable*			*uncorrectable*		
Method	recall	prec.	F-m.	recall	prec.	F-m.	recall	prec.	F-m.
Baseline method	0.80	0.81	0.80	0.47	0.64	0.54	0.78	0.59	0.68
Proposed method-1	0.86	0.75	0.80	0.64	0.80	0.71	0.82	0.73	0.77
Proposed method-2	0.86	0.82	0.84	0.71	0.81	0.76	0.82	0.73	0.77

Table 4 Word accuracies of utterances in each class.

	correct	*correctable*	*uncorrectable*	all
word accuracy	95.3%	92.6%	88.8%	91.7%

value of transcribed speech data. It covered various expressions from *correct*, *correctable*, and *uncorrectable* utterances.

An LC was trained with three types of BLEU scores calculated using Subset-of-*correct*, Subset-of-*correctable*, and Subset-of-*uncorrectable* utterances as reference materials. The BLEU scores of each utterance were computed based on only its corresponding reference materials for the target question. The parameter vector of hyperplanes dividing each class was calculated as closed-solutions minimizing the total square error for the training data. A classification experiment was conducted with 10-fold cross validation. The trained LC was evaluated based on classification accuracy. As no hyper-parameter was included to the LC, we did not use a development set.

For comparison, another experiment was conducted in which only the BLEU score calculated on Subset-of-*correct* utterances was used as a feature for a classifier, and each utterance was classified based on two thresholds, 0.4 and 0.9, which gave the minimal total error for the training data.

5.4 Experimental Results

The test dataset consisted of 924 utterances by 14 university students (7 males and 7 females) who were recruited for the test set. Each student made 66 utterances responding to 66 target questions. The quality of each transcribed utterance in the evaluation dataset was rated by the same bilingual speaker. Table 4 lists the word accuracies of the ASR for the utterances in each class and for all utterances.

Figure 4 compares the classification accuracies with the baseline method using a single BLEU score calculated using only utterances in Subset-of-*correct* with the proposed method using three BLEU scores (proposed method-1) and proposed method using three BLEU scores with POS-replacement rule (proposed method-2).

As shown in Fig. 4, the overall accuracy of proposed method-1 was 9.9 points higher than that of the baseline method. Also, proposed method-2 was 2.4 points higher than that of proposed method-1. Question-based paired t-test revealed a significant difference between the baseline method and proposed method-1 ($p < 0.001$) and marginally significant difference between proposed method-1 and method-2 ($p < 0.06$).

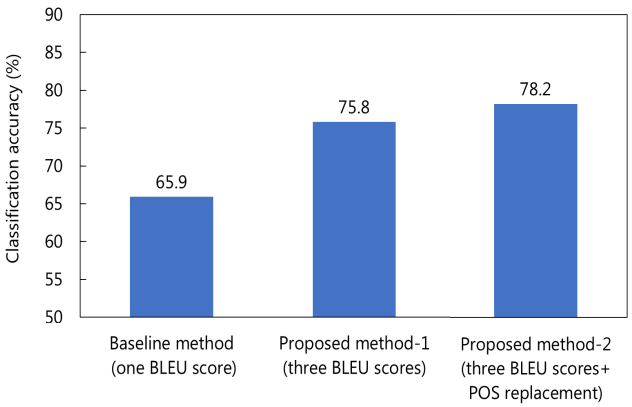

Fig. 4 Accuracies of baseline with one BLEU score and proposed method with three BLEU scores without/with POS-replacement rule.

Table 5 lists the measures on classification accuracy of three classes of the baseline method and the two versions of the proposed method. Both versions of the proposed method using three BLEU scores had better or equal F-measures for all three classes compared with the baseline method. Regarding *correctable* and *uncorrectable*, both versions of the proposed method had better F-measures than the baseline method. These results suggest that using three types of BLEU scores as classification features contribute to the improvement in overall classification accuracy, especially in the classification accuracy of the *correctable* class.

Concerning the effect of POS-replacement rule on the data-sparseness problem, the F-measures of both *correct* and *correctable* classes obtained with proposed method-2 were better than those with proposed method-1. The precision of the *correct* class with proposed method-2 increased compared with that with proposed method-1, and the recall of the *correctable* class with proposed method-2 improved compared to that with proposed method-1. These results imply that the classifier trained with the BLEU scores calculated based on the POS-replacement rule reduced the proportion of the utterances of the *correctable* class that were mistakenly classified as the *correct* class.

From these results, it can be said that replacing every word that precedes or follows the middle word in an n-gram with its corresponding POS tag when its occurrence frequency is lower than the threshold can contribute to solving the data-sparseness problem in the classification for CALL systems.

Table 6 lists the word accuracies of the ASR of the classified utterances of three classes: *correct*, *correctable*, and *uncorrectable*. For the *correctable* utterances, the word ac-

Table 6 Word accuracies of utterances in each class classified with proposed method-2.

	correct	*correctable*	*uncorrectable*
word accuracy	96.1%	96.3%	85.7%

Table 7 Matching rates and Cohen's kappa coefficients between scores by two raters.

	Matching rate [%]	Kappa coefficient
Rater 1 - Rater 2	50.2	0.27
Rater 1 - Rater 3	60.7	0.42
Rater 2 - Rater 3	71.6	0.55
average	60.9	-

curacy of utterances was 96.3%. This result indicates that the utterances classified as *correctable* have almost the same recognition accuracy as those of *correct*.

5.5 Effect of Difference among Human Raters on Classifier Performance

Two bilingual English/Japanese speakers re-rated the collected utterances from 41 Japanese-speaking students for the learner corpus, and the evaluation data of 14 Japanese university students on a scale of 1 to 5 to examine the effect of the difference among human raters on classifier performance.

Table 7 lists the matching rates of rated utterances (924 utterances) between both raters and Cohen's kappa coefficients, which were computed to assess agreement between the two raters. The average of three matching rates was 60.9%. The results of the kappa coefficients indicate that there were no sufficient agreements among all three patterns. These results show that the rated scores of utterances largely depend on the subjective view of each human rater.

To verify whether the proposed method can adapt to differences in human raters, two additional experiments were conducted using the data by two additional raters (Raters 2 and 3).

One experiment involved applying the proposed method to the data rated by Raters 2 and 3, respectively, and the classification accuracies were compared among all three raters. Table 8 lists the classification accuracies from "the classifier by each single rater".

The other experiment involved the learner corpus, which was re-classified into three subsets based on the median of the subjectively evaluated scores by the three raters. The LC was trained with the class labels determined based on the median. Hereafter, this LC is called "The classifier with the median". To evaluate the LC, three types of test datasets were used; dataset rated by Rater 1 ("the data rated by Rater 1"), dataset rated by Rater 2 ("the data rated by Rater 2"), and dataset rated by Rater 3 ("the data rated by Rater 3"). Table 8 lists the classification accuracies of "the classifier with the median", which were calculated for each of the three test datasets mentioned above.

The classification accuracies using the evaluation data

Table 8 Classification results using data with three raters' rated scores. Figures in parentheses show performance when baseline method was used.

Test set data	Classifiers by each single rater [%]	Classifier with median [%]
Data rated by Rater 1	78.2 (65.9)	55.0
Data rated by Rater 2	69.6 (60.4)	65.0
Data rated by Rater 3	71.0 (67.0)	68.4
average	72.9 (64.4)	62.8

by each rater ranged from 70 to 78%, and the difference among the raters affected the accuracy of the proposed classification method, as shown in Table 8. However, the accuracy of the proposed method was higher than that of the baseline method using a single BLEU score (values shown in parentheses).

The performance of "the classifier with the median" was 62.8%, which was 10.1% worse than the average classification accuracies of "the classifier by each single rater", which can be easily predicted from difference of the rated scores of utterances between the human raters. However, it was 1.9 points higher than the average of the matching rates among the three raters (shown in Table 7). This suggests that the proposed method can exhibit almost the same evaluation performance as the average values among human raters and perform better than the baseline method.

6. Conclusion

We proposed a classification method of learner utterances based on BLEU scores for translation-game-type CALL systems. The proposed method classifies learner utterances into three classes using three types of BLEU scores computed in accordance with three subsets of a corpus (Subset-of-*correct*, Subset-of-*correctable*, and Subset-of-*uncorrectable*) as feature parameters of a linear classifier. The purpose of calculating the three types of BLEU scores in accordance with the three subsets of a corpus is that there are similarities in errors made by learners who have the same L1. Furthermore, for overcoming the data-sparseness problem, our classification method uses BLEU scores calculated using a mixture of word and POS-tag sequences converted from word sequences based on a POS-replacement rule with which the number of matches does not become too few, even if the number of reference translations is not sufficient. This rule replaces each word with its corresponding POS tag in each n-gram of both learner utterances and reference translations when its occurrence frequencies in a learner corpus are lower than a threshold.

To evaluate the proposed method, we conducted experiments to examine the classification accuracy of the learner utterances of our previously developed translation-game-type dialogue-based CALL system for Japanese learners of English. The results indicate that the proposed classification method achieved a classification accuracy of 78.2%, which was 12.3 points higher than the baseline with one BLEU score

Thus, we found that the proposed classification method

can classify learner utterances with high accuracy. The recognition accuracy of *correctable* utterances is almost the same as that of *correct* ones, and the method is considered useful for providing effective instructive feedback regarding various *correctable* utterances.

For future work, we plan to develop a method that can specify inappropriate parts of utterances classified as *correctable*.

Acknowledgments

The authors would like to thank Dr. Xiaoyun Wang of Doshisha University (currently at Nanyang Technological University) for her suggestions and various discussions. This research was supported by a contract with MEXT number 15K02738.

References

[1] A. Neri, C. Cucchiarini, and H. Strik, "Effective feedback on L2 pronunciation in ASR-based CALL," Proc. Workshop on Computer Assisted Language Learning, Artificial Intelligence in Education Conference, AIED, San Antonio, Texas USA, pp.40–48, 2001.

[2] B. Mak, M. Siu, M. Ng, Y.-C. Tam, Y.-C. Chan, K.-W. Chan, K.-Y. Leung, S. Ho, F.-H. Chong, J. Wong, and J. Lo, "PLASER: Pronunciation learning via automatic speech recognition," Proc. HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing, vol.2, pp.23–29, Association for Computational Linguistics, 2003.

[3] M. Eskenazi, A. Kennedy, C. Ketchum, R. Olszewski, and G. Pelton, "The nativeaccent™ pronunciation tutor: Measuring success in the real world," SLaTE, pp.124–127, 2007.

[4] A. Raux and M. Eskenazi, "Using task-oriented spoken dialogue systems for language learning: Potential, practical applications and challenges," InSTIL/ICALL Symposium 2004, pp.35–38, 2004.

[5] H. Morton and M.A. Jack, "Scenario-based spoken interaction with virtual agents," Computer Assisted Language Learning, vol.18, no.3, pp.171–191, 2005.

[6] J. Brusk, P. Wik, and A. Hjalmarsson, "DEAL: A serious game for CALL practicing conversational skills in the trade domain," SLaTE 2007, pp.88–91, 2007.

[7] P. Wik and A. Hjalmarsson, "Embodied conversational agents in computer assisted language learning," Speech Communication, vol.51, no.10, pp.1024–1037, 2009.

[8] K. Lee, S.-O. Kweon, S. Lee, H. Noh, G.G. Lee, "POSTECH immersive English study (POMY): Dialog-based language learning game," IEICE Trans. Inf. & Syst., vol.E97-D, no.7, pp.1830–1841, July 2014.

[9] E. Rayner, P. Bouillon, N. Tsourakis, J. Gerlach, Y. Nakao, and C. Baur, "A multilingual CALL game based on speech translation," Proc. LREC, 2010.

[10] S. Seneff and C. Wang, "Web-based dialogue and translation games for spoken language learning," SLaTE, pp.9–16, 2007.

[11] N.-R. Han, M. Chodorow, and C. Leacock, "Detecting errors in English article usage by non-native speakers," Natural Language Engineering, vol.12, no.2, pp.115–129, 2006.

[12] T. Anzai, A. Hahm, A. Ito, and S. Makino, "Grammatical error detection from Englsih utterances spoken by Japanese," Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2010 Asia-Pacific, pp.482–485, IEEE, 2010.

[13] S. Lee, H. Noh, K. Lee, and G.G. Lee, "Grammatical error detection for corrective feedback provision in oral conversations," AAAI, pp.797–802, 2011.

[14] T. Anzai and A. Ito, "Recognition of utterances with grammatical mistakes based on optimization of language model towards interactive CALL systems," Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, pp.1–4, IEEE, 2012.

[15] C. Baur, J. Gerlach, M. Rayner, M. Russell, and H. Strik, "A shared task for spoken call?," Proc. LREC, pp.237–244, 2016.

[16] C. Baur, C. Chua, J. Gerlach, M. Rayner, M. Russell, H. Strik, and X. Wei, "Overview of the 2017 spoken CALL shared task," Proc. SLate 2017, pp.71–78, 2017.

[17] K. Evanini, M. Mulholland, E. Tsuprun, and Y, Qian, "Using an automated content scoring system for spoken CALL responses: The ETS submission for the spoken CALL challenge," Proc. SLate 2017, pp.97–102, 2017.

[18] T. Watanabe and S. Tsukada, "Unknown utterance rejection using likelihood normalization based on syllable recognition," IEICE Trans. Inf. & Syst., vol.E75-D, no.12, pp.2002–2009, Dec. 1992.

[19] Y. Nagai, T. Senzai, S. Yamamoto, and M. Nishida, "Sentence classification with grammatical errors and those out of scope of grammar assumption for dialogue-based CALL systems," International Conference on Text, Speech and Dialogue, Lecture Notes in Computer Science, vol.7499, pp.616–623, Springer, Berlin, Heidelberg, 2012.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," Proc. 40th Annual Meeting on Association for Computational Linguistics, pp.311–318, Association for Computational Linguistics, 2002.

[21] X. Wang, J. Zhang, M. Nishida, and S. Yamamoto, "Phoneme set design for speech recognition of English by Japanese," IEICE Trans. Inf. & Syst., vol.E98-D, no.1, pp.148–156, Jan. 2015.

[22] C.-Y. Lin and F.J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," Proc. 42nd Annual Meeting on Association for Computational Linguistics, Article No. 605, Association for Computational Linguistics, 2004.

[23] H. Kučera, W.N. Francis, et al., Computational analysis of present-day American English, Dartmouth Publishing Group, 1967.

[24] X. Wang, T. Kato, and S. Yamamoto, "Phoneme set design based on integrated acoustic and linguistic features for second language speech recognition," IEICE Trans. Inf. & Syst., vol.E100-D, no.4, pp.857–864, April 2017.

[25] E. Sumita, Y. Sasaki, and S. Yamamoto, "Frontier of evaluation method for MT systems," IPSJ Magazine, vol.46, no.5, pp.552–557, 2005 (in Japanese).

[26] E. Brill, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging," Computational Linguistics, vol.21, no.4, pp.543–565, 1995.

[27] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, J. Moore, D. Odell, and D. Povey, "HTK speech recognition toolkit version 3.4," Cambridge University Engineering Department, pp.67–75, 2006.

[28] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database read by Japanese to support CALL research," Proc. International Congresses on Acoustics, pp.557–560, 2004.

Reiko Kuwa received her B.S. and M.S. degrees from the graduate school of Science and Engineering, Doshisha University in 2015 and 2017. She joined KDDI Co., Ltd. She is a member of ASJ.

Tsuneo Kato received his B.E., M.E., and Ph.D. degrees from The University of Tokyo in 1994, 1996, and 2011. He joined Doshisha University in 2015, where he is currently an associate professor in the Department of Intelligent Information Engineering. Before his current position, he had worked at KDDI R&D Laboratories Inc. since 1996. He has been engaged in research and development of automatic speech recognition and intelligent user interfaces. He received an IPSJ Kiyasu Special Industrial Achievement Award in 2011. He is a member of IPSJ, ASJ, IEICE, and IEEE.

Seiichi Yamamoto received his B.S., M.S., and Ph.D. degrees from Osaka University in 1972, 1974, and 1983. He joined Kokusai Denshin Denwa Co. Ltd. in April 1974 and ATR Interpreting Telecommunications Research Laboratories in May 1997. He was appointed president of ATR-ITL in 1997. He is currently a professor in the Department of Information Systems Design, Faculty of Science and Engineering, Doshisha University, Kyoto, Japan. His research interests include digital signal processing, speech recognition, speech synthesis, natural language processing, spoken language processing, spoken language translation, and multi-modal dialogue processing. He received Technology Development Awards from the Acoustical Society of Japan in 1995 and 1997, a best paper award from the Information and Systems Society of IEICE in 2006, and a telecom-system technology award from the Telecommunications Advancement Foundation in 2007. Dr. Yamamoto is a member of ASJ, IPSJ, IEEE (Fellow), and IEICE Japan (Fellow).