



SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 27 (2011) 42 – 49

Procedia
Social and Behavioral Sciences

Pacific Association for Computational Linguistics (PACLING 2011)

A Scoring Method for Second Language Writing Based on Word Alignment

Katsunori Kotani^{a*}, Takehiko Yoshimi^b^aKansai Gaidai University, 16-1 Nakamiyahigashino-cho, Hirakata 573-1001, Japan^bRyukoku University, 1-5 Yokoya Seta Oe-cho, Otsu 520-2194, Japan

Abstract

Automatic evaluation reduces the burden on second language (L2) teachers when checking for errors made by L2 learners. One evaluation approach is to automatically classify L2 learner sentences into either natural or unnatural sentences. However, because the distinction between natural and unnatural sentences is complex, such a binary classification cannot accurately reflect the naturalness of L2 learner sentences. To solve this problem, we developed a method that assigns scores (continuous values) to L2 learner sentences by examining the degree to which L2 learners use erroneous expressions arising from unnatural literal translations of phrases from a L2 learner's first language. Experimental results suggest that our scoring method is effective for the automatic evaluation of L2 learner sentences.

© 2011 Published by Elsevier Ltd. Open access under [CC BY-NC-ND license](http://creativecommons.org/licenses/by-nc-nd/3.0/).

Selection and/or peer-review under responsibility of PACLING Organizing Committee.

Keywords: automatic scoring; second language writing evaluation; word alignment technique

1. Introduction

Second language (L2) learners often use unnatural expressions that must be identified and corrected by L2 teachers. Several authors [1, 2, 3, 4] have developed automatic methods that perform binary classification of sentences written by L2 learners (L2 learner sentences) as either unnatural (L2 learner-like) or natural (native speaker-like). However, as L2 learner sentences vary in their degree of naturalness, binary classification does not adequately reflect their level of naturalness. Therefore, instead of coarse

* Corresponding author.

E-mail address: kkotani@kansai-gaidai.ac.jp, yoshimi@rins.ryukoku.ac.jp

binary classification, an automatic method should perform a fine evaluation that is capable of multiple-level rating.

In response to this need, we developed a method of assigning a score (continuous value) to L2 learner sentences by adopting evaluation methods for machine translation sentences [5, 6] (described in more detail in Section 3). Our evaluation method differs from the previous evaluation methods [1, 2, 3, 4] in that it can perform multiple-level evaluations.

Our method uses the evaluation method [5] to determine a score by examining the degree to which a sentence involves erroneous expressions arising from unnatural literal translations from a learner's first language (L1) due to a lack of adequate L2 vocabulary. Unnatural expressions are identified through on corresponding lexical properties (word alignment) between L2 learner sentences/native speaker sentences and L1 sentences. Our scoring method is constructed by a machine learning algorithm, which requires training data. The training data consists of L2 learner sentences, sentences written by a native speaker (native speaker sentences), and sentences written by L2 learners in L1 (L1 sentences). We extracted L2 learner sentences and L1 sentences from an L2 learner corpus that was developed by National Institute for Japanese Language [7]. Native speaker sentences were obtained separately. The language data for all three types of sentences (L2 learner sentences, L1 sentences and native speaker sentences) were then compiled. Note that in this paper the term L1 refers to English, and L2 to Japanese.

2. Related Studies

Previous studies [1, 2, 3, 4] proposed classifying L2 learner sentences as either native speaker-like or L2 learner-like. These studies adopted binary classification methods developed for machine-translated sentences [6, 8, 9]. Here, we briefly review the linguistic properties (that is, machine learning features) used by these classification methods.

Baroni & Bernardini [1] compared a sequence of N-words (word N-grams) found in L2 learner sentences and native speaker sentences. Tomokiyo & Jones [2] also examined a sequence of N-words, but used information on the parts of speech of N-words (POS N-gram). Lee et al. [3] examined the syntactic properties of L2 learner sentences and native speaker sentences, analyzing syntactic parsing results such as syntactic dependency relations between words, co-occurrence relations between verbs and subject/object nouns, and head words and their determiners in base noun phrases. Kotani et al. [4] examined the corresponding lexical properties (word alignment) between L2 learner sentences/native speaker sentences and L1 sentences in order to determine whether L2 learners use literally translated expressions.

Other studies [10, 11] developed fine rating evaluation methods. Lee et al. [10] developed a scoring method for L2 learner sentences that differs from our scoring method in the following respects. First, their method adopts a rule-based approach, not a machine learning approach: its rules are morphological and syntactic. Second, it deals with answers written in single sentences for questions in L1 (e.g., "Where is your hometown?"). Our scoring method does not require correct answer sentences when evaluating L2 learner sentences, but the scoring method of Lee et al. [10] requires correct answer sentences from L2 teachers in order to compare L2 learner sentences with correct answer sentences.

Tsubaki et al. [11] developed an automatic scoring method of L2 learner translation that does not require correct answer sentences when evaluating L2 learner sentences. This method uses measures commonly used in statistical machine translation: that is, translation probability and word N-gram probability. According to them, the word N-gram probability-based evaluation showed no correlation with subjective evaluation, but the translation probability-based evaluation showed a weak correlation with L2 learners' proficiency in terms of TOEIC scores ($r = 0.29$). They also found that their method seemed to depend on the sentence length, as the correlation varied with the sentence length. The

correlation coefficient of short sentences (less than 5 words) was 0.24, and that of long sentences (more than 21 words) was 0.47.

3. Scoring L2 Learner Sentences

Our scoring method determines scores for L2 learner sentences using support vector machines (SVMs). SVMs are well-known machine learning algorithms that have a high generalization performance [12]. They perform binary classification by producing a separating hyper-plane defined by a vector w , a constant b , and an inner product $\langle \cdot, \cdot \rangle$ in the feature space. Classification is determined by the side of the separating hyper-plane on which a test example x falls: $\text{sign}(\langle w, x \rangle - b)$.

An SVM classifier (a classification model) can function as a regression model. That is, an SVM classifier can also output an evaluation score (continuous value) for a test example [6]. An evaluation score can be obtained by computing the distance between a separating hyper-plane and the test example. The distance is computed by removing the operator sign in the expression above. Kulesza & Shieber [6] proposed a scoring method for machine-translated sentences using an SVM classifier, and Kotani & Yoshimi [5] later constructed a classifier based on machine learning features differing from those of Kulesza & Shieber [6]. We describe machine learning features of Kotani & Yoshimi [5] in detail in Section 4.

A novelty of the proposed method lies in the use of the scoring methods for machine translations [5, 6] as a scoring method for L2 learner sentences. We developed a method of assigning scores to L2 learner sentences by computing the distance between the separating hyper-plane and a test example. If an example is located deep within the L2 learner-like sentence half-space, this example is considered to contain unnatural expressions. In contrast, an example closer to the separating hyper-plane might share more qualities in common with native speaker-like sentences, even if it still falls into the L2 learner-like sentence class.

Regression models have an advantage over classification models (classifiers) because they can perform a fine evaluation rather than a coarse binary one. On the other hand, classification models have the advantage of low-cost evaluation since they do not require any manually labeled training examples. Our method is a classification model functioning as a regression model and therefore offers both advantages.

4. Properties of Word Alignment Results

L2 learner sentences often contain unnatural expressions derived from literal translations of phrases from L1 to L2. Understandably, L2 learners with limited vocabulary attempt to compensate for the lack of relevant vocabulary by incorporating literal translations from L1 using L2 vocabulary that they know. The resulting unnatural expressions can be identified by examining the word alignment between L2 learner sentences and L1 sentences, as reported by Kotani et al. [4]. Another novelty of the proposed method lies in the use of word alignment for the classification of L2 learner sentences [4] as machine learning features for scoring L2 learner sentences. Our scoring method identifies unnatural expressions by analyzing the word alignment between L2 learner sentences/native speaker sentences and L1 sentences.

Let us illustrate how word alignment assesses the naturalness of L2 learner sentences. In this study, the target of our scoring method is L2 learners of Japanese whose L1 is English. Thus, L2 learner sentences and native speaker sentences are in Japanese and L1 sentences are in English. Sentence (1) is an L2 learner sentence that contains unnatural expressions derived from literally translating the L1 sentence (2). In sentence (1), the adjectives *takakute* (high) and *tsuyokatta* (strong) make this sentence unnatural. If we compare the L2 learner sentence (1) with the L1 sentence (2), we find that these adjectives arose from a literal translation of the English adjectives “high” and “strong” in the L1 sentence (2). In contrast, the

native speaker sentence (3) does not contain the unnatural adjectives found in the L2 learner sentence (1); instead, the meaning of these adjectives is conveyed with the predicates *yaruki ga aru* (motivation-NOM exists) and *sainoo nimo megumarete ita* (talent is also endowed).

(1) [L2 learner sentence]

Kono kurasu no gakusei no dooki ga totemo

This class-GEN student-GEN motivation-NOM very

takakute nooryoku ga tuyokatta desu.

high and talent-NOM strong-PST-BE

(The students in these classes had very high motivation and their talents were very strong.)

(2) [L1 sentence]

The students in these classes had very high motivation and their talents were very strong.

(3) [native speaker sentence]

Kono kurasu no gakusei wa taihen yaruki ga ari

This class-GEN student-TOP very motivation-NOM exists

sainoo nimo megumarete ita.

talent also endowed-BE.

[NOM, nominative case marker; GEN, genitive case marker; PST, past tense marker; BE, copular; TOP: topic marker]

An examination of the word alignment shows the unnaturalness of the L2 learner sentence (1) and the naturalness of the native speaker sentence (3). The words in the L2 learner sentence are classified either as those matched with a corresponding expression in an L1 sentence (aligned word pairs) or as words without corresponding expressions in the L1 sentence (non-aligned words). Expressions derived from literal translation tend to be aligned word pairs, while non-aligned words usually do not include unnatural expressions resulting from literal translation.

Table 1 shows the results of a word alignment examination of the L2 learner sentence (1) and the L1 sentence (2), as well as the results of the native speaker sentence (3) and the L1 sentence (2). Here, “align (A, B)” indicates that an English word A and a Japanese word B compose an aligned word pair, while “non-align (C)” means that an English or Japanese word C remains unaligned. The adjectives “high” and “strong” are aligned with the corresponding expressions between the L2 learner sentence (1) and the L1 sentence (2). In contrast, the adjectives “high” and “strong” remain unaligned between the native speaker sentence (3) and the L1 sentence (2). We performed this word alignment examination with a word alignment tool [13]; we used this same tool in the experiment described in the following section.

Table 1. Word alignment results of the L1 sentence (2) with the L2 learner sentence (1) and with the native speaker sentence (3)

L2 Learner Sentence (1)	Native Speaker Sentence (3)
align (student, ‘gakusei’)	align (student, ‘gakusei’)
align (in, ‘no’)	align (in, ‘no’)
align (these, ‘kono’)	align (these, ‘kono’)
align (class, ‘kurasu’)	align (class, ‘kurasu’)
align (very, ‘taihen’)	align (very, ‘taihen’)
align (motivation, ‘yaruki’)	align (high, ‘takai’)
align (talent, ‘sainoo’)	align (motivation, ‘yaruki’)

align (be, ‘iru’)	align (talent, ‘nooryoku’)
align (., ‘.’)	align (be, ‘desu’)
non-align (the)	align (strong, ‘tuyoi’)
non-align (have)	align (., ‘.’)
non-align (high)	non-align (the)
non-align (and)	non-align (have)
non-align (their)	non-align (and)
non-align (very)	non-align (their)
non-align (strong)	non-align (very)
non-align (‘wa’)	non-align (‘tati’)
non-align (‘ga’)	non-align (‘no’)
non-align (‘aru’)	non-align (‘ga’)
non-align (‘ni’)	non-align (‘te’)
non-align (‘mo’)	non-align (‘ga’)
non-align (‘megumareru’)	non-align (‘ta’)
non-align (‘te’)	

5. Properties of Word Alignment Results

We tested the proposed scoring method by Spearman rank correlation between the evaluation scores of our scoring method and the learning experience of L2 learners and the mean evaluation scores of three human evaluators. The L2 learner sentences were collected from L2 learners with study experience ranging from one to four years, so we divided the sentences into four groups. We assumed that sentences written by L2 learners in the 1-year group were the least natural, while those written by L2 learners in the 4-year group were the most natural. In this experiment, three native Japanese speakers evaluated the naturalness of the L2 learner sentences on a 100-point scale.

Our scoring method was trained with the results of the word alignment between L2 learner sentences/native speaker sentences and L1 sentences. The L2 learner sentences and the L1 sentences were extracted from an L2 learner corpus [7] that contains essays written by L2 learners; 689 sentences (49 essays) were extracted from this corpus. Native speaker sentences were written by a native Japanese speaker based on the L2 learner sentences and the L1 sentences. We performed a five-fold cross-validation test using the 689 L2 learner sentences, the L1 sentences, and the native speaker sentences.

In this experiment, we used the following natural language processing tools. We performed machine learning with TinySVM [14]; we chose the d -th polynomial kernel function ($d = 1, 2, 3, 4$), varying the soft-margin parameter ($C = 1, 0.1, 0.01, 0.001, 0.0001, 0.00001$), and defaults were used for the other settings. We also used a word alignment tool [13] that segments Japanese sentences into word units and aligns English and Japanese words using a bilingual dictionary/thesaurus and dependency analysis.

We also compared the results obtained using two other scoring methods to those obtained using the proposed scoring method. These methods were constructed with SVMs using machine learning features different from those used in our scoring method; the features were similar to those used in the classification methods reported in previous studies [1, 2, 3]. One of the two scoring methods was trained with machine learning features of lexical properties of L2 learner sentences and native speaker sentences. This scoring method was based on a sequence of the parts of speech of N-words (POS N-gram) in a sentence. This type of machine learning feature is essentially the same as the one used by Baroni & Bernardini [1] and Tomokiyo & Jones [2]. The other scoring method was trained with machine learning features of syntactic properties of L2 learner sentences and native speaker sentences. This scoring method used syntactic information on the phrase dependency relation between a modifier and modifée that was derived using the syntactic parser Cabocha [15]. The phrase dependency relation was described in terms

of POS information such as adjective-noun. This type of machine learning feature is essentially the same as the one used by Lee et al. [3].

The results of the Spearman correlation are shown in Fig. 1. The values shown are the mean Spearman rank correlation coefficients for 689 test sentences. Each value is the highest mean value derived from the optimal combination of a d -th polynomial parameter kernel function ($d = 1, 2, 3, 4$) and a soft margin parameter ($C = 1, 0.1, 0.01, 0.001, 0.0001, 0.00001$). Our scoring method was trained with three types of word alignment results, and thus there are three scoring methods based on word alignment: one trained only with aligned word pairs (aligned-based scoring method), one trained only with non-aligned words (non-aligned-based scoring method), and one trained with both aligned word pairs and non-aligned words (both-based scoring method). The scoring method referred to as “N-gram-based” in Fig. 1 was trained with POS N-gram features; the scoring method referred to as “dependency-based” was trained with the dependency relation.

Learning experience was moderately correlated with the aligned-based scoring method ($r = 0.47$). The second highest correlation coefficient was obtained using the dependency-based scoring method ($r = 0.35$). The aligned-based method also showed a moderate correlation coefficient with human evaluation ($r = 0.48$). The other scoring methods showed correlation coefficients of approximately 0.20 with human evaluation. The aligned-based scoring method outperformed not only the scoring methods based on word alignment (non-aligned-based and both-based) but also the other scoring methods (N-gram-based and dependency-based).

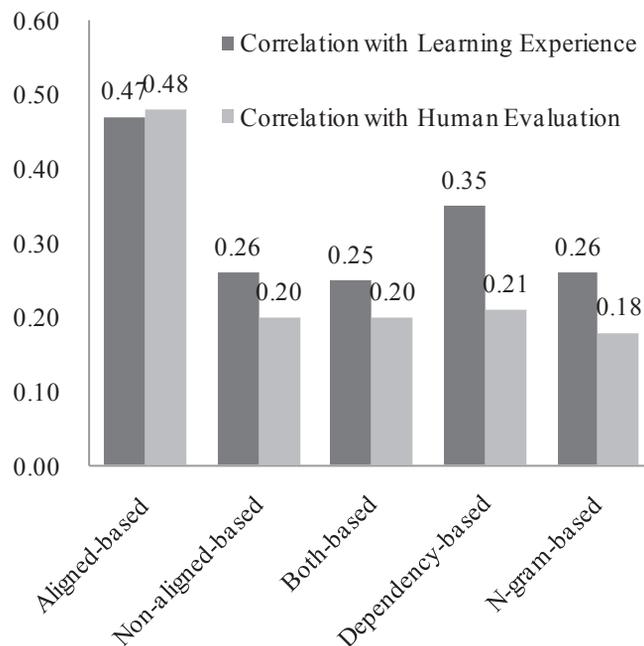


Fig. 1. Correlation of scoring methods with learning experience and human evaluation

Before the experiment, we assumed that of our three types of scoring methods (aligned-based, non-aligned-based, and both-based), the both-based scoring method would outperform the others. In actuality, the both-based scoring method underperformed compared to the aligned-based one. Underperformance of the proposed scoring method was observed when the methods were trained with machine learning features involving non-aligned words. We assume that the failure of non-aligned words can arise from unnatural expressions that are unrelated to literal translations from L1 to L2; for example, Japanese has a subjective case marker and an objective case marker, while English lacks either. Hence, errors clustering around case markers do not arise from literal translation, and case markers always appear as non-aligned words, as seen in Table 1. If an L2 learner confuses the subjective case marker *ga* with the objective case marker *wo*, the resulting unnaturalness cannot be identified by examining non-aligned words in L2 learner sentences.

We also determined the correlation between the three human evaluators (HE1, HE2, and HE3) in order to determine the target correlation coefficient. The inter-evaluator correlation was as follows: 0.91 (HE1-HE2), 0.83 (HE2-HE3), and 0.79 (HE1-HE3). The lowest inter-evaluator correlation (0.79) serves as the target correlation coefficient for the scoring methods. As seen in Fig. 1, none of the scoring methods reached the target. This low performance may have been caused by the failure of non-aligned words, as discussed above.

6. Conclusion

We developed a scoring method for L2 learner sentences using word alignment results as machine learning features. This proposed method can assign a score to each sentence based on the presence or absence of literally translated expressions caused by the interference of L1. We assessed the validity of our scoring method based on correlation between the evaluation scores of our scoring method and the learning experience of L2 learners and the mean evaluation scores of three human evaluators. We also compared our scoring method with other scoring methods that used machine learning features other than word alignment. The results showed that aligned-based scoring method (one of the proposed methods) had higher correlations with learning experience and human evaluation scores than the N-gram-based and dependency-based methods. Given these findings, we conclude that aligned-based scoring method (one of the proposed methods) is promising for the automatic evaluation of L2 learner sentences. This study also demonstrated the applicability of a machine translation evaluation method into an L2 writing evaluation method.

Several points remain to be clarified in future studies. First, as we noted above, we must investigate why non-aligned words fail to contribute to the scoring in more detail. This investigation would help improve the performance of our scoring method. Second, we plan to carry out a qualitative analysis of the scoring results (in this study, we statistically confirmed the validity of our scoring method). Third, we have to compare our method with other scoring methods such as the scoring method [11]. Finally, we will investigate the effectiveness of our scoring method for L2 learner sentences in languages other than Japanese.

Acknowledgements

This work was supported in part by Grant-in-Aid for Scientific Research (B) (22300299).

References

- [1] Baroni, M. and S. Bernardini, "A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text," *Literary and Linguistic Computing*, Vol.21-36, 2006, pp. 259-274.
- [2] Tomokiyo, L. M. and R. Jones, "You're Not from Round Here, Are You? Naive Bayes Detection of Non-native Utterances," *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001, pp. 1-8.
- [3] Lee, J., Zhou, M. and X. Liu, "Detection of Non-native Sentences Using Machine-translated Training Data," *Proceedings of the 2007 Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2007, pp. 93-96.
- [4] Kotani, K., Yoshimi, T., Kutsumi, T. and I. Sata, "An Automatic Evaluation of Writing in Japanese as a Second Language Using a Word-alignment-based Classifier," *Proceedings of the International Conference on Asian Language Processing*, 2008, pp. 210-216.
- [5] Kotani, K. and T. Yoshimi, "A Machine Learning-based Evaluation Method for Machine Translation," *Artificial Intelligence: Theories, Models and Applications*, Vol. 6040 of Lecture Notes in Computer Science, Springer Berlin, Heidelberg, 2010, pp. 351-356.
- [6] Kulesza, A. and S. M. Shieber, "A Learning Approach to Improving Sentence-level MT Evaluation," *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, 2004, pp. 75-84.
- [7] The National Institute for Japanese Language. *Contrastive Linguistic Database for Japanese Language Learners' Written Language*, 2001.
- [8] Corston-Oliver, S., Gamon, M. and C. Brockett, "A Machine Learning Approach to the Automatic Evaluation of Machine Translation," *Proceedings of the 39th Meeting of Association for Computational Linguistics*, 2001, pp. 148-155.
- [9] Kotani, K., Yoshimi, T., Kutsumi, T. and I. Sata, "Validity of an Automatic Evaluation of Machine Translation Using a Word-alignment-based Classifier," *Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*, Vol. 5459 of Lecture Notes in Artificial Intelligence, Springer Berlin, Heidelberg, 2009, pp. 91-102.
- [10] Lee, K.J., Choi, Y-S., and J. E. Kim, "Building an Automated English Sentence Evaluation System for Students Learning English as a Second Language," *Computer Speech & Language*, Vol.25-2, 2011, pp. 246-260.
- [11] Tsubaki, H., Yasuda, K., Yamamoto, H. and Sagisaka, Y. "Objective evaluation of second language learner's translation proficiency using statistical translation measures," *Proceedings of International Speech Communication Association Tutorial and Research Workshop on Experimental Linguistics*, 2008, pp. 217-220.
- [12] Vapnik, V. *Statistical Learning Theory*, Wiley-Interscience, New York, 1998.
- [13] Whitelock, P. and V. Poznanski, "The SLE Example-based Translation System," *Proceedings of the International Workshop on Spoken Language Translation*, 2006, pp. 111-115.
- [14] Kudo, T. *TinySVM*, retrieved from <http://www.chasen.org/~taku/software/TinySVM/>
- [15] Kudo, T. and Y. Matsumoto, "Japanese Dependency Analysis Using Cascaded Chunking," *Proceedings of The Sixth Workshop on Natural Language Learning*, 2002, pp. 63-69.