THEMES IN SCIENCE AND TECHNOLOGY EDUCATION
Volume 3, Number 1-2, Pages 223-238

Second language writing classification system based on word-alignment distribution

Katsunori Kotani[1], Takehiko Yoshimi[2]
kkotani@kansaigaidai.ac.jp, yoshimi@nlp.i.ryukoku.ac.jp
[1]Kansai Gaidai University, Japan
[2]Ryukoku University, Japan

Abstract

The present paper introduces an automatic classification system for assisting second language (L2) writing evaluation. This system, which classifies sentences written by L2 learners as either native speaker-like or learner-like sentences, is constructed by machine learning algorithms using word-alignment distributions as classification features for detecting word-by-word translated expressions. The experimental results demonstrated that our classification system provided adequate classification results with respect to (i) high classification accuracy, (ii) classification results reflecting the properties of L2 learner sentences, and (iii) identification of learner-like expressions. These results suggest that our classification system can be used to evaluate L2 learner sentences.

Introduction

Teacher feedback is important in second language (L2) learning, and L2 learners who receive error feedback have been reported to show progress in L2 writing (Ferris, 2006). For this reason, language teachers must be able to identify expressions including errors in sentences written by students. As demonstrated by previous studies (Kotani et al., 2008; Lee et al., 2007; Baroni & Bernardini, 2006; Tomokiyo & Jones, 2001), an automatic classification system can reduce teacher burden in evaluating L2 learner sentences. These studies proposed classification systems that differentiate natural expressions from unnatural expressions.

A classification system performs binary-class classification, and the previous classification systems classify sentences written by L2 learners as either native speaker-like or learner-like sentences. Native speaker-like sentences are fluent and natural sentences, while learner-like sentences contain some unnatural expressions that might result from first language (L1) interaction. Thus, a learner writing more native speaker-like sentences will be judged as advanced learners, while a learner writing more learner-like sentences will be judged as immature learners.

As shown in Figure 1, the classification system (1) extracts classification features (linguistic properties that are clues to classification) from L2 learner sentences, and then (2) classifies the sentences as either native speaker-like or learner-like based on these classification features. Such classification systems are constructed using statistical methods or machine learning algorithms (Duda et al., 2001).

When constructing a classification system using machine learning algorithms, it is important to choose appropriate classification features for differentiating the various linguistic properties of native speaker-like and learner-like sentences. Kotani et al. (2008), for example, focused on the L1 effect on L2 learner sentences in determining appropriate classification features. Since the L1 effect is a critical problem in L2 writing, our classification system uses the classification features proposed by Kotani et al. (2008).

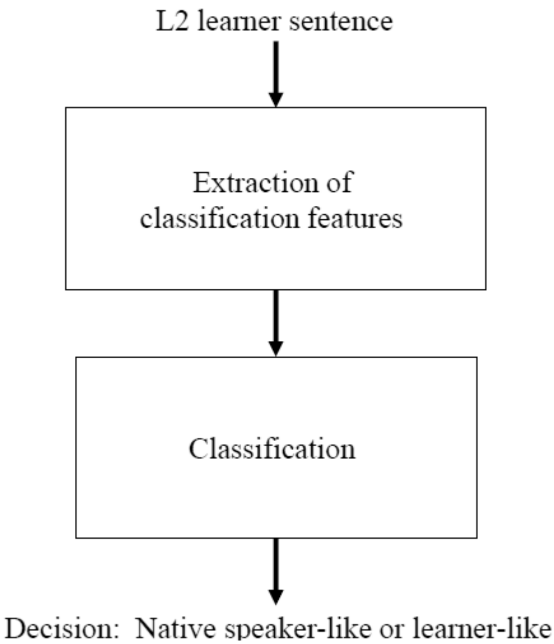

Figure 1. Automatic classification system

As shown in this paper, our classification system identifies native speaker-like sentences and learner-like sentences based on word-alignment distribution, which is used to determine the presence or absence of word-by-word translated expressions. The L1 effect can be detected by determining the existence of word-by-word translation in a sentence, as L2 learners, especially lower-level learners, must depend on translation when they cannot directly express an idea in the L2.

In the fourth section of this paper, the validity of our classification system is investigated. First, we examined the classification accuracy of our classification system. We then determined whether our classification system could detect expressions unique to L2 learners. If our classification system can properly classify L2 learner sentences and detect expressions commonly used in learner-like sentences, the system may be able to alert teachers to sentences that should be evaluated more carefully.

Related studies

L2 writing

L2 learners use linguistic techniques different from those used by native speakers. For instance, L2 learners and native speakers choose different lexical items and/or syntactic constructions. In addition to these grammatical properties, L2 learner sentences and native speaker sentences also differ from a sociolinguistic perspective, including aspects such as politeness, as discussed by Birdsong (2005). Although sociolinguistic factors should be taken into account during classification, let us focus now on grammatical properties. As the primary goal of our classification system is to assist teachers in evaluating the progress of elementary and intermediate learners, there is an urgent need to identify unnatural expressions caused by grammatical problems such as incorrect lexical choice (henceforth, we refer to these simply as "learner-like expressions").

According to Tashiro (1995) and Asai (2002), the distinction between native speaker-like and learner-like sentences can be identified using lexical and syntactic properties such as sentence length, sentence patterns, and syntactic complexity. Lexical difference can be seen as a morphological choice in the following unnatural Japanese sentence (1) (Sakairi et al., 1999). Unlike the grammatical Japanese sentence (2), the verb in sentence (1) lacks the suffix-*gatteiru*. Here, the subject is third person, and therefore the verb must use the reportive form, which requires -*gatteiru*. Thus, sentence (1) is judged as unnatural. As English verbs have no such restriction, as shown in the English sentences (3a) and (3b), L2 learners whose L1 is English often make this type of error. As we will discuss below, this problem can be identified using word-alignment distribution.

(1) [L2 learner sentence]
***Tonari-no kazoku-wa uresii*
Next door-GEN family-TOP happy
"The neighbors are happy"

Cf. *Watasi-wa uresii*
I-TOP happy
"I am happy"

(2) [Native speaker sentence]
Tonari-no kazoku-wa uresi-gatteiru
Next-door-GEN family-TOP seems-happy
"The neighbors seem happy"

(3) a. I am happy.
b. The neighbor is happy.
(GEN, genitive case marker; TOP, topic marker)

Classification approach

Previous studies (Kotani et al., 2008; Lee et al., 2007; Baroni & Bernardini, 2006; Tomokiyo & Jones, 2001) constructed automatic classification systems for assisting the evaluation of L2 learner sentences using machine learning algorithms. These classification systems essentially classify L2 learner sentences as either native speaker-like or learner-like sentences. However, these systems use different machine learning algorithms and classification features.

Tomokiyo & Jones (2001) constructed an automatic classification system for utterances spoken by L2 learners using a naïve Bayes classifier, and Baroni & Bernardini (2006) constructed a system using Support Vector Machines (SVMs) (Vapnik , 1998). Their classification systems examine sequences of *N* words (word *N*-gram) and sequences of *N* parts of speech (POS *N*-gram). These classification systems examine L2 learner sentences at the text level rather than the sentence level, and have relatively high classification accuracy; the word *N*-gram-based classification system of Tomokiyo & Jones (2001) had a 94% classification accuracy, while their POS *N*-gram-based system had 100% classification accuracy; the classification system of Baroni & Bernardini (2006) had a 90% classification accuracy.

Lee et al. (2007) constructed an automatic classification system for L2 learner utterances using SVMs. Their classification system examines L2 learner sentences with respect to entropy from the trigram language model, parse score of a statistical parser, parse tree derivations, co-occurrence relation of a head noun and its determiner, and dependency relation such as subject-verb. This classification system had a relatively low classification accuracy (65.8%) compared with the classification systems of Tomokiyo & Jones (2001) and Baroni & Bernardini (2006).

Classification features

It is well known that L2 learner sentences contain linguistic errors that arise from the interference of the L1 in the construction of L2 expressions (Ellis, 1994). As we will see below, this type of error can be identified in sentences containing word-by-word translated expressions. Note that our classification system requires three types of linguistic data: L2 learner sentences, sentences in the L2 learner's first language parallel with the L2 learner sentences, and native speaker sentences parallel with the L2 learner sentences. The following L2 learner sentence (4) contains unnatural word-by-word translated expressions such as the Japanese verbal expression *kagayai* for the English expression “shine” and the Japanese nominal expression *taiyo* for the English nominal expression “the sun.” In contrast, the native speaker sentence (5), which is completely natural, contains no word-by-word translated expressions. We consider that L2 learners use word-by-word translation from the L1 when faced with a lack of relevant L2 knowledge.

(4) [L2 learner sentence]
Kyoo taiyoo-wa kagayai-teiru
Today the-sun-TOP shine-BE-ING
“Today the sun is shining”

(5) [Native speaker sentence]
Kyoo-wa seiten-da
Today-TOP nice-weather-BE
“The weather is nice today”

(6) [Intended meaning of sentence (4)]
The weather is nice today.

(BE, copular verb; ING, gerundive verb form)

The distinction between sentences (4) and (5) can be identified using the word-alignment distribution, which reveals the presence or absence of word-by-word translated expressions. Table 1 shows the word-alignment distribution obtained using word-alignment software (Whitelock & Poznanski, 2006) that automatically segments Japanese sentences into word units and aligns English and Japanese words using a bilingual dictionary/thesaurus and dependency analysis. The symbol “align(A, B)” refers to the formation of an aligned pair composed of the English word “A” and the Japanese word “B”. The symbols “non-align_eng(C)” and “non-align_jpn(D)” refer to instances of non-aligned words; the former shows that the English word “C” remains unaligned, while the latter shows that the Japanese word “D” remains unaligned. Based on this word-alignment distribution, the learner-like sentence (4) contains more aligned words than the native speaker-like sentence (5).
Given this result, we used word-alignment distribution as a classification feature in our classification system. Our classification system was constructed using SVMs, well-known machine learning algorithms with high generalization performance. SVMs are relatively insensitive to the number of classification features.

Table 1. Word-alignment distribution of sentences (4) and (5)

Learner-like Japanese sentence (4)	Native speaker-like Japanese sentence (5)
align(today, *kyoo* [today])	align(today, *kyoo-wa* [today TOP])
align(is, *teiru* [BE/ING])	align(is, *da* [BE])
align(sun, *taiyoo* [sun])	nonalign_jpn(*seiten* [nice weather])
align(shining, *kagayai* [shine])	nonalign_eng(the)
nonalign_jpn(*wa* [TOP])	nonalign_eng(sun)
nonalign_eng(the)	nonalign_eng(shining)

Experiments

Design

Our classification system was constructed and attested using the following linguistic data: (i) Japanese sentences written by L2 learners; (ii) English sentences parallel to the L2 Japanese sentences; and (iii) Japanese sentences written by a native speaker of Japanese. The L2 learner Japanese sentences and the English sentences were extracted from a corpus of L2 learners of Japanese (The National Language Research Institute 2001). This corpus contains L2 learner Japanese sentences and English sentences that are parallel with the L2 learner Japanese sentences. The L2 learners in this corpus are native/native-like speakers of English. Six hundred and eighty-nine sentences were taken from the corpus. Native speaker Japanese sentences that have the same meaning with L2 learner sentences were prepared by a native speaker of Japanese. These sentences were written based on L2 learner Japanese sentences and the parallel English sentences (if necessary).

The fluency of Japanese sentences of L2 learners were evaluated on a 100-point scale by a native speaker of Japanese. Based on the results of this evaluation, the Japanese L2 learner sentences were divided into sentences with more than 50 points (higher-level sentences) and those with less than 50 points (lower-level sentences). The former group consisted of 478 sentences (69.4%), and the latter group consisted of 211 sentences (30.6%).

Word-alignment distribution with the English sentences was analyzed in the Japanese sentences of both L2 learners and native speakers using word-alignment software (Whitelock and Poznanski 2006).[1] This software automatically carries out word-alignment using a bilingual dictionary/thesaurus and dependency analysis. Word-alignment distribution consists of aligned pairs and non-aligned words. Each instance of word alignment was taken as a classification feature.

To construct the classification system, machine learning was carried out using TinySVM software (http://chasen.org/~taku/software/TinySVM/). We chose the d-th polynominal kernel function (d=1, 2, 3, 4), varying the soft-margin parameter (C=1, 0.1, 0.01, 0.001, 0.0001, 0.00001). The other settings were default.

Classification features and accuracy

We examined to what extent different classification systems could correctly identify L2 learner sentences. Three different classification systems were used: aligned pair-based; non-aligned word-based; and aligned pair- and non-aligned word-based. These classification systems were constructed with the fourth polynomial kernel

[1] For explanation of the alignment software see http://www.slc.atr.jp/IWSLT2006/proceedings/EC_15_SLE_slides.pdf

function and the soft-margin parameter C set to 0.00001. The classification accuracy of our classification systems was examined in five-fold cross-validation tests.

The classification accuracy of each classification system is shown in Table 2. Classification accuracy was measured with the mean values of the five-fold cross-validation tests. The classification system using both aligned pair- and non-aligned word-based classification had the highest classification accuracy. The non-aligned word-based classification system had higher classification accuracy than the aligned pair-based classification system. Hence, we concluded that non-aligned word-classification features are more effective for the classification of L2 sentences.

Table 2. Classification accuracy of classification systems

Classification system	Classification accuracy (%)
Aligned pair-based	78.7
Non-aligned word-based	85.3
Aligned pair- and non-aligned word-based	92.7

Comparison with systems based on classification features of previous studies

We compared our classification systems with those described in previous studies (Lee et al., 2007; Baroni & Bernardini, 2006; Tomokiyo & Jones, 2001). Two classification systems were then constructed based on the classification features shared among these previous studies.

Lee et al. (2007) used classification features consisting of parsing results such as (i) context-free grammar rules used for parsing sentences, (ii) co-occurrence relations between a verb and its subject/object noun, and (iii) parsing scores. We constructed a classification system using dependency relations analogous to the first two classification features used by Lee et al. (2007). Although this syntactic dependency relation does not precisely reproduce the experimental conditions used by Lee et al. (2007), we consider this comparison to indicate the validity of our classification system. Dependency relation of POS between modifier and modifee was obtained by syntactic analysis using the dependency parser Cabocha (http://chasen.org/~taku/software/cabocha/).

The dependency relation-based classification system marked 80.0% classification accuracy in the five-fold cross-validation tests. The classification accuracy was highest among the various kernel functions and the soft-margin parameters. Although the classification accuracy for the dependency relation-based classification system was higher than that for the aligned words-based system (Table 2), our classification systems using non-aligned word classifications outperformed the dependency relation-based classification system.

Next, we compared our classification systems with the system using POS N-gram-classification features (N=1, 2, 3) (Baroni & Bernardini, 2006; Tokokiyo, 2001). The N-gram-based classification system had 71.2% classification accuracy in the five-fold cross-validation tests. The classification accuracy of the N-gram-based classification system was highest among the various kernel functions and the soft-margin parameters. All of our classification systems outperformed this classification system.

Based on these results, we concluded that our word alignment-based classification systems should provide valid evaluation results.

Lower rates of learner-like sentences for higher-level L2 learner sentences

We further examined whether our classification systems could properly distinguish between native speaker-like and learner-like sentences. An adequate classification system should exhibit a lower rate of learner-like sentences for L2 learner sentences written by advanced L2 learners and a higher rate for sentences written by non-advanced L2 learners. Hence, our classification systems were examined with respect to the relation between the levels of L2 learner sentences and native speaker- or learner-like classification results. We examined the rate of learner-like sentences and native speaker-like sentences in the classification results of the three classification systems using different classification features. These classification systems were constructed with the fourth polynomial kernel function and the soft-margin parameter C set to 0.0001.

Tables 3, 4 and 5 show the classification results of these classification systems. The levels of L2 learner sentences were determined based on the manual evaluation results stated in section 4.1. We statistically analyzed the classification results using the Chi-square test. Let us first examine Table 3. The analysis showed no significant difference for the aligned pair-based classification system ($\chi^2(1)=1.47$, $p=0.224$).

Table 3. Rate of learner-like sentences by aligned pair-based classification

Classification Level	Learner-like sentences	Native speaker-like sentences	Rate of learner-like sentences (%)
Lower-level sentences	172	39	81.5
Higher-level sentences	370	108	77.4

Table 4. Rate of learner-like sentences by non-aligned word-based classification

Classification Level	Learner-like sentences	Native speaker-like sentences	Rate of learner-like sentences (%)
Lower-level sentences	194	17	91.9
Higher-level sentences	394	84	82.4

In contrast, a statistically significant difference was observed in the classification results for the non-aligned word-based classification system ($\chi^2(1)=10.60$, $p<0.01$) in Table 4, and the aligned pair- and non-aligned word-based classification system ($\chi^2(1)=5.43$, $p<0.05$) in Table 5. These results suggest that classification systems using non-aligned word-based classification features can accurately distinguish between native speaker-like and learner-like sentences.

Table 5. Rate of learner-like sentences by aligned pair- and non-aligned word-based classification

Classification Level	Learner-like sentences	Native speaker-like sentences	Rate of learner-like sentences (%)
Lower-level sentences	203	8	96.2
Higher-level sentences	436	42	91.2

Analysis of linguistic properties of learner-like sentences based on weighted classification features

If our classification system can properly detect learner-like expressions in L2 learner sentences and alert teachers to these sentences that should be evaluated more carefully, the system should be useful in assisting teacher evaluations of L2 writing. In this section, we examine the validity of our classification system by analyzing whether the system can detect learner-like expressions.

Our classification system assigns weight to classification features derived from SVM learning. Thus, our classification system can quantify classification features in terms of the degree of learner-likeness or native speaker-likeness. A classification feature is weighted by either a negative value or a positive value. Classification features with negative weight indicate learner-like expressions and those with positive weight indicate native speaker-like expressions. The absolute value of a weight indicates the degree of learner-likeness/native speaker-likeness. Given this property of weighted classification features, we examined whether a highly weighted classification feature actually identifies learner-like expressions. If this is true, we can conclude that our classification system can detect learner-like expressions for alerting teachers to sentences that include these expressions. Then, the teacher can focus on the sentences for evaluation, and hence our classification system can assist the teacher's evaluation.

We analyzed whether both aligned pair- and non-aligned word-based classification features reveal linguistic properties of learner-like sentences. Among the 4,013 aligned pair- and non-aligned word-based classification features, 2,098 had negative weights. It is highly likely that sentences including these features are learner-like sentences.

Figure 2 shows the normalized negative weights of classification features characterizing learner-like sentences. The weights of classification features were normalized by dividing each weight by the highest absolute value among the weights. Most features have weights with small absolute, near-zero, values, but a few features bear weights with relatively large absolute values. The latter classification features may more strongly characterize linguistic properties of learner-like sentences. Below, we examine in detail L2 learner sentences containing these classification features.

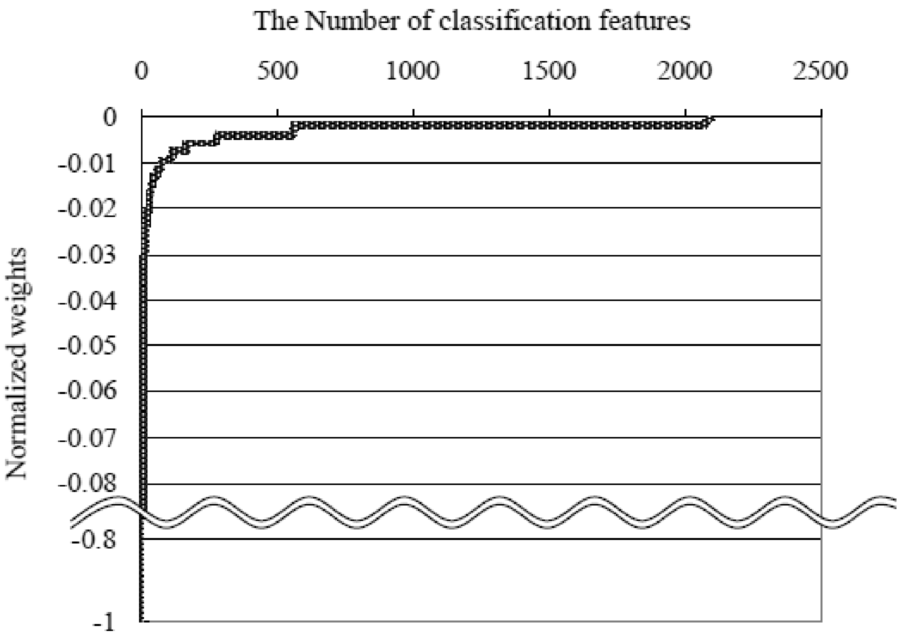

Figure 2. Distribution of normalized weights of learner-like classification features

In principle, L2 learners, especially beginners, use expressions different from those used by advanced L2 learners or native speakers. Suppose that the English word *x* can be translated into Japanese as *a*, *b*, or *c*. A beginner may use only the Japanese word *a*, while an advanced L2 learner may use any of the three words *a*, *b*, and *c* depending on the context or style. Thus, it is highly possible that only the aligned pair "align(x, a)" will be found in learner-like sentences, and all the aligned pairs "align(x, a)", "align(x, b)" and "align(x, c)" will be found in native speaker-like sentences. Given this situation, our classification system should identify the linguistic properties of learner-like sentences in terms of "align(x, a)" and those of native speaker-like sentences in terms of "align(x, a)", "align(x, b)" and "align(x, c)."

Table 6 shows typical pairs of the learner-like and native speaker-like classification features. A learner-like classification feature is paired with a native speaker-like classification feature if both features have a common English expression. In the rest of this section, we will examine whether the classification features in Table 6 actually identify the linguistic properties that differ between L2 learner sentences and native speaker sentences.

Table 6. Typical learner-like classification features and corresponding native speaker-like classification features

Learner-like classification feature			Native speaker-like classification feature
Feature instance	**Normalized weight**	**Rank**	**Feature instance**
align (class, *kurasu* [class])	-0.02064	26	align (class, *jugyoo* [class])
align (question, *situmon* [question])	-0.01126	74	align (question, *mondai* [question])
align (American, *amerika-no* [American])	-0.00871	113	align (American, *amerika-jin* [American])
align (reception, *resepusyon* [reception])	-0.00750	165	align (reception, hirooen [wedding reception]
align (small, *tiisai* [small])	-0.00563	276	nonalign_eng (small)

The learner-like classification features include "align(class, *kurasu* [class])," suggesting that the Japanese word *kurasu* is often used in place of the English word "class" in learner-like sentences. This indicates that use of the word *kurasu* differs between L2 learners and native speakers. This usage sometimes makes a sentence incorrect, as shown in the L2 learner sentence (7). In this sentence, the word *jugyoo* should be used instead of *kurasu*. As expected, the English word "class" is aligned with *jugyoo* in the native speaker sentence (8). The word *kurasu* seems to be easier to use for L2 learners because this word is phonetically similar to the English word "class" and is written using simplified Chinese character, *katakana* (クラス). In contrast, the word *jugyoo* is phonetically dissimilar to the English word "class" and is written using Chinese characters (授業), which are more morphologically complex than *katakana*. Among 689 L2 learner sentences, we found 48 instances of *kurasu* aligned with "class(es)" in 42 L2 learner sentences. Thirty-six of these instances were correct, while 12 were incorrect; the rate of erroneous usage was thus 25%.

(7) [L2 learner sentence]

**Jiyuu-ni kurasu-o uketari...*

Freely-DAT class-ACC take

[Intended meaning of sentence (7)]

"One can take any classes..."

(8) [Native speaker sentence]

Jiyuu-ni jugyoo-o uketari...

Freely-DAT class-ACC take

"One can take any classes..."

(DAT, dative case marker; ACC, accusative case marker)

A similar morphological effect can also be seen in the classification feature "align(reception, *resepusyon* [reception])." As seen in the L2 learner sentence (9), L2 learners use the Japanese word *resepusyon* for the English word "reception," and this word can be written using *katakana*. However, another word, *hirooen*, written using Chinese characters, is used in the native speaker sentence (10). Sentence (11) is unnatural as a result of the incorrect lexical choice of *resepusyon* over the more natural *hirooen*. Among 689 L2 learner sentences, *resepusyon* aligned with "reception" in five cases; all of these usages were incorrect.

(9) [L2 learner sentence]

**Resutoran-toka hoteru-nado resepusyon-joo-o erabi-masu*

Restaurant-and hotel-and reception-place-ACC choose-PRES

[Intended meaning of sentence (9)]

"They choose a place to hold the reception, such as a restaurant or hotel"

(10) [Native speaker sentence]

Resutoran-toka hoteru-toitta hirooen-kaijyoo-o erabimasu

Restaurant-and hotel-and reception-place-ACC choose-PRES

"They choose a place to hold the reception, such as a restaurant or hotel"

(PRES, present tense marker)

The learner-like classification feature "align(question, *situmon* [question])" identifies a learner-like expression in an L2 learner sentence. In the L2 learner sentence (11), the Japanese word *situmon* is used, while another Japanese word, *mondai*, is used in the native speaker sentence (12). Although the English word "question" can be translated as either *situmon* or *mondai*, these Japanese words are not completely equivalent. In sentence (11), the L2 learner uses the Japanese word *situmon* for "exam question," causing the sentence to become unnatural. Among 689 L2 learner sentences, *situmon* aligned with "question" in six cases; two of these were correct and four were incorrect. The rate of erroneous usage was thus 67%.

(11) [L2 learner sentence]

Sore-kara gakusei-wa nyuugaku-siken-no situmon-o syuucyuusuru-koto-bakari-de manabe-nai

Then since student-TOP entrance-exam-GEN question-ACC concentrate-that-only-and study-not

[Intended meaning of sentence (11)]

"Therefore, they will concentrate solely on the questions that will be on the exam and nothing else."

(12) [Native speaker sentence]

Sonotame siken-mondai- dake-ni isiki-o mukeru

Therefore exam-question-only-DAT concentrate

"Therefore, they will concentrate solely on the questions that will be on the exam and nothing else."

A similar learner-like expression was identified with the classification feature "align(American, *amerika-no* [American])." L2 learners used the Japanese words *amerika-no* for the English word "American," as in the L2 learner sentence (13). In the native speaker sentence (14), the meaning of this English word is conveyed with the Japanese *amerika-jin no*. Among 689 L2 learner sentences, *amerika-no* aligned with "American" in 25 sentences; 14 of these were correct and 12 were incorrect. The rate of erroneous usage was therefore 46%.

(13) [L2 learner sentence]

Sara-ni sono-hutuu-no amerika-no hassoo-ni-yotte…

Moreover the typical-GEN American-GEN mindset according-to

[Intended meaning of sentence (13)]

"Moreover, according to this typical American mindset…"

(14) [Native speaker sentence]

Sara-ni tenkeiteki-na amerika-jin-no kangae-dewa…

Moreover typical American-GEN mindset according-to

"Moreover, according to this typical American mindset…"

Another learner-like expression could be identified with the classification feature "align(small, *tiisai* [small])." L2 learners used the Japanese word *tiisai* for the English word "small," as in the L2 learner sentence (15). In the native speaker sentence (16), the meaning of this English word is conveyed with the Japanese word *semai*. Among 689 L2 learner sentences, *tiisai*(*-na*) aligned with "small" in seven sentences; four of these instances were correct and three instances were incorrect. The rate of erroneous usage was thus 43%.

(15) [L2 learner sentence]

**Yooroppa-dewa kotonatta-gengo-oyobi bunka-no ooku-ga tiisai kuiki-ni…*

Europe-LOC-TOP different-language-and culture-GEN lots-of-NOM small area-LOC

[Intended meaning of sentence (15)]

"In Europe…many different languages and cultures…in a small area."

(16) [Native speaker sentence]

Yooroppa-dewa semai tiiki no naka-de kotonaru gengo-ya bunka…

Europe-LOC-TOP small area-GEN inside-LOC different languages-and cultures

"In Europe…many different languages and cultures…in a small area."

(LOC, location marker)

The present analysis demonstrated that our classification system could identify L2 learner sentences with a high probability of containing learner-like expressions based on the weighting of particular classification features. Therefore, our classification system is judged to be effective in assisting teachers to identify L2 learner sentences containing learner-like expressions.

Conclusion

We introduced an automatic classification system that assists in the evaluation of L2 learner sentences by classifying L2 learner sentences as either native speaker-like or learner-like sentences. Our classification system can detect expressions commonly found in learner-like sentences based on word-alignment distribution, showing the effect of L1 on L2 sentences, and can alert teachers to these sentences, allowing more careful and intensive evaluation.

We also examined the validity of our classification systems. First, the validity of our classification systems was confirmed based on the classification accuracy. In addition, our classification systems were judged to be valid because our systems outperformed the other classification systems not using word-alignment distortion as classification features. Statistical analysis in section 4.4 revealed that our classification systems could distinguish between native speaker-like and learner-like sentences, and qualitative analysis in section 4.5 showed that our classification systems could identify learner-likeness based on weights assigned to classification features by SVM learning.

This paper leaves several problems unresolved. First, we must examine classification properties based on word alignment in more detail. As shown in Table 2, classification accuracy varies depending on the type of word-alignment information provided. Second, our classification system might be improved by adding other classification features, such as lexical difficulties for L2 learners. Third, we must examine whether the number of L2 learner sentences and parallel native speaker sentences affects the weighting of classification features, which can be used to discriminate native speaker-likeness from learner-likeness. Last, we should examine the applicability of our classification systems. In the present study, we used our classification system for sentences written by L2 learners of Japanese whose L1 was English. In the future, we should examine whether our classification system can be used for learners with L1 other than English.

Acknowledgement

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

References

Asai, M. (2002). Analysis of sentence structure in Japanese composition: a comparison between compositions written by Chinese advanced Japanese learners and by Japanese native speakers. *Journal of Japanese Teaching*, 115, 51-60.

Baroni, M., & Bernardini, S. (2006). A new approach to the study of translations: machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3), 259-274.

Birdsong, D. (2005). Nativelikeness and non-nativelikeness in L2A research. *International Review of Applied Linguistics in Language Teaching*, 43(4), 319-328.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York: Wiley-Interscience.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Ferris, D. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in Second Language Writing: Contexts and Issues* (pp. 81-104). Cambridge: Cambridge University Press.

Kotani, K., Yoshimi, T., Kutsumi, T., & Sata, I. (2008). 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 (IALP08)* (pp. 210-216). Chiang Mai, Thailand.

Lee, J., Zhou, M., & Liu, X. (2007). Detection of non-native sentences using machine-translated training data. *Proceedings of the 2007 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 93-96). Rochester, New York, USA. The Association for Computational Linguistics.

The National Institute for Japanese Language (2001). *Contrastive linguistic database for Japanese language learners' written language*. Tokyo, Japan.

Sakairi, I., Sato, Y., Sakuragi, N., Nakamura, K., Nakamura, H., & Yamada, A. (1991). *Gaikokujin ga Nihongokyoosi ni Yokusuru 100 no Shitumon*. Tokyo: Babel Press.

Tashiro, H. (1995). A study of compositions by intermediate and advanced level learners: focusing on unnaturalness and incomprehensibility. *Journal of Japanese Teaching*, 85, 25-37.

Tomokiyo, L. M., & Jones, R. (2001). 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* (pp. 1-8). Pittsburgh, PA., USA.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley-Interscience.

Whitelock, P., & Poznanski, V. (2006). The SLE example-based translation system. *Proceedings of the International Workshop on Spoken Language Translation* (pp. 111-115). Kyoto, Japan.