An Approach toward Automatic Error Detection in Learners' English Writing Based on the Source Language

Koichi Kawamura[1], Harumi Kashiwagi[2], Min Kang[1]
[1]Graduate School of Intercultural Studies
[2]Institute for Promotion of Higher Education
Kobe University
Kobe, Japan
email: kawamura51@stu.kobe-u.ac.jp, kasiwagi@kobe-u.ac.jp, kang@kobe-u.ac.jp

Abstract— **Automatic error detection systems for English writing have been improved since they were first introduced and are being applied to foreign language learning. However, these systems mainly focus on local errors, such as grammatical aspects in the target language and ignore the meaning intended in the source language. As a result, teachers must spend an inordinate amount of time to detect global errors. In this paper, we propose an approach to an automatic error detection system to solve this problem. In order to determine whether the structure of an English sentence is in error or not, criteria for error determination are first needed. Our approach is based on the idea that criteria for error determination are created by the correspondence relation between Japanese and English using sentence patterns. In order to evaluate our approach, by way of illustration, four sentence patterns were selected from the authors' original six sentence patterns. Automatic error detection using these four sentence patterns was carried out on 100 Japanese sentences with subjects and their corresponding English sentences. As a result, we concluded that, using the sentence patterns in the source language, automatic error detection is effective when based on our criteria for error determination.**

Keywords-Error Detection; Sentence Pattern; Global Error; Parser; Source Language; Criteria for Error Determination.

I. INTRODUCTION

For English learners, compared to speaking, reading and listening, writing is the most difficult skill to improve. "Writing abilities are not naturally acquired; they must be culturally (rather than biologically) transmitted in every generation, whether in school or in other assisting environments" [1]. Despite this linguistic feature, writing is not taught enough in schools relative to other skills [2]. It is especially difficult to teach as teachers must detect and correct learners' errors one by one which is very time consuming.

It is accepted that English essays written by learners with low proficiency contain a lot of errors. Of these errors, global errors negatively affect the structure of the whole sentence, and this limits the readers' comprehension. Therefore, it is necessary for teachers not to overlook such errors when proofreading the essay. However, in order to detect global errors teachers must devote an inordinate amount of attention discovering all the potential structural errors. Thus, teachers may have a tendency to overlook some structural errors due to time constraints. For this reason, a support tool for structural error detection is needed in order to reduce the burden on teachers.

On the other hand, a number of automatic error detection systems using natural language processing technology have been tried. They are applied to foreign language learning classes to reduce the burden on teachers and support students to acquire better writing skills. These systems perform excellently with single grammatical errors, such as spelling, article usage, prepositions and aspect. However, few error detection systems look at structural errors. Thus, current automatic error detection systems are limited in that they do not cover all types of learners' errors. In addition, most of these systems are designed to analyze the target language (English) only. This unilateral approach may cause a discrepancy between the system's automatic correction feedback and the learner's intention [3]. In order to overcome these problems, the source language (Japanese) should also be an object of analysis. For these reasons, these systems are of limited general use in classrooms.

In order to support teachers and improve error detection accuracy, an automatic error detection system which can easily identify sentence structure errors, and cope with various types of global errors, and recognize the learners' intention is needed. Therefore, this study proposes just such an automatic error detection system, one which can easily determine whether a sentence is correct or not by comparing the basic sentence elements (subject and predicate) of Japanese and English using parsers based on sentence patterns. This approach is based on the results of our previous study, which showed that "detecting English

 ISBN: 978-1-61208-619-4

errors using sentence patterns is more promising than detection that depends on full sentences" [4].

In Section 2, we propose an approach for an automatic error detection system that can determine whether an English sentence structure is in error or not. In Section 3, we detect structural errors according to criteria for error determination. Then we evaluate the accuracy of criteria for error determination created by the correspondence relation between Japanese and English, based on the four sentence patterns selected for illustration. In Section 4, we refer to the efficiency of automatic error detection using the sentence patterns in the source language.

II. APPROACH

To make the detection of global errors easy, we focus on basic sentence elements, by comparing them in the source language and the corresponding target language. To conduct the comparison, we classify the sentence patterns and create criteria for error determination: a rule created based on the correspondence relation between Japanese and English using sentence patterns. We compare the basic sentence elements (primary subject and predicate) of the source language (Japanese) and the corresponding target language (English) using parsers based on sentence patterns and criteria, this approach follows the procedure below.

A. Procedure

1. Prepare Japanese sentences and the corresponding English sentences as analytical data.
2. Set up a Japanese parser, Cabocha and an English parser, the Stanford Parser.
3. Automatically extract sets of sentence elements, a primary subject and a predicate (verb) by a parser based on extraction rules.
4. Automatically classify the sets of a primary subject and a predicate (verb) based on the Japanese sentence patterns.
5. Compare the defined sentence patterns with the extracted sentence patterns based on criteria for determination.
6. Obtain the results of error determination as feedback (ERROR, POSSIBLE, UNKNOWN).

*ERROR stands for "an outright error." POSSIBLE stands for "not an error, but may not be a correct answer." UNKNOWN stands for "indeterminable."

B. Sentence Elements

Although each Japanese and English sentence contains various elements, such as subjects, predicates (verbs), objects, complements, etc., this study examines the set of a primary subject and predicate (verb) only. This is because all major sentence patterns contain a subject and a predicate verb in academic writing [5]-[7]. Additionally, it is efficient for teachers to determine whether the learners' English is grammatically correct by checking sets of a primary subject and a predicate verb only. This will support teachers to detect errors easily since learners' errors are not always clear, and teachers have difficulty determining where the problems lie.

C. Parsers and Extraction Rules

To extract sets of primary subjects and predicates from Japanese sentences, the parser, Japanese Dependency Structure Analyzer, Cabocha [8] was utilized. To extract sets of primary subjects and predicate verbs from the corresponding English sentences, the Stanford Parser [9] was utilized. Table I indicates details of both parsers and extraction rules of subjects and predicates (verb).

TABLE I. EXTRACTION RULES OF CABOCHA AND THE STANFORD PARSER

Parser		**Cabocha 0.69**	**The Stanford Parser 3.6.0**
Target Language		***Japanese***	***English***
Extraction Rule	Subject	A clause including a case particle "が (GA)" or a binding particle "は (WA)" or "も (MO)" which has a dependency structure with the predicate	A nominal subject or a clausal subject
	Predicate (Verb)	The last clause	A verb (transitive or intransitive) or a "be" verb + copula which has a dependency structure with the subject

*が (GA), は (WA), も (MO) are particles in Japanese grammar that immediately follow a noun, a verb, an adjective, and indicate the subject of a sentence.

D. Sentence Patterns and Criteria for Determination

In order to clarify the determinate language behavior, we selected the following six sentence patterns, because they appear frequently in learners' writing. The patterns were classified into two categories (predicate based and subject based). First, the predicate based sentence pattern was sub-classified into four sentence patterns: A) Subject + (ARU / IRU), B) Subject + Noun + (DESU / DEARU / DA), C) Subject + Reporting Verb (OMOU / KANGAERU / KANJIRU / JIKKANSURU), D) Subject + Verb (excluding "be" verb *existence* and reporting verb). Second, the subject based sentence pattern was sub-classified into two sentence patterns: E) ~ (SURU) KOTO + (WA / GA / MO) + predicate (excluding a modal auxiliary verb), F) WATASHI + (WA / GA / MO) + predicate. Table II indicates the Japanese sentence patterns.

The following is a supplementary explanation of each sentence pattern: A) ARU and IRU represent a "be" verb *existence*, B) DESU, DEARU and DA represent an auxiliary verb *state*, C) ~ (TO) OMOU and KANGAERU represent a reporting verb *mental state*, KANJIRU and JIKKANSURU represent a verb *perception*, E) ~ (SURU) KOTO represents an inanimate subject, such as a formal subject, a gerund or

 ISBN: 978-1-61208-619-4

an infinitive in English, F) WATASHI represents the personal pronoun "I". Note on Japanese verbs, plain form is used.

TABLE II. JAPANESE SENTENCE PATTERNS

Type		Sentence Patterns with Subject
Predicate Based	A	主語 ＋(ある / いる) Subject + (ARU / IRU)
	B	主語 ＋ 名詞 ＋(です / である / だ) Subject + Noun + (DESU / DEARU / DA)
	C	主語 ＋ ～と(思う / 考える / 感じる / 実感する) Subject + Reporting Verb (OMOU / KANGAERU / KANJIRU / JIKKANSURU
	D	主語 ＋ 動詞(存在, 伝達除く) Subject + Verb (excluding "be" verb *existence* and reporting verb)
Subject Based	E	～(する)こと ＋(は / が / も)＋ 述語(法助動詞除く) ~ (SURU) KOTO + (WA / GA / MO) + predicate (excluding a modal auxiliary verb)
	F	私 ＋(は / が / も)＋ 述語 WATASHI + (WA / GA / MO) + predicate

TABLE III. SENTENCE PATTERN AND ITS CRITERIA FOR ERROR DETERMINATION

S.P.	Criteria for Error Determination
A	If predicate verb is not "be" verb, it should be ERROR.
B	If predicate verb is not "be" verb, it should be ERROR.
C	If predicate verb is not "reporting" verb, it should be ERROR.
D	If predicate verb does not meet semantic agreements, it should be ERROR
E	If subject is not "it," "to verb" or "verb-ing," it should be ERROR.
F	If subject is not "I," it should be ERROR.

S.P. is an acronym of "sentence pattern." The above highlighted sentence patterns are dealt with in this study as an illustration.

The predicate based sentence pattern A) "Subject + (ARU / IRU)" and B) "Subject + Noun + (DESU / DEARU / DA)" always correspond with a 'be' verb in English, without the 'be' verb in the English sentences there would be errors. Sentence pattern C) "Subject + reporting Verb (OMOU / KANGAERU / KANJIRU / JIKKANSURU)" always responds with a reporting verb "think" or "feel" in English, without reporting verb "think" or "feel" in the English sentences there would be errors. Sentence pattern D) "Subject + Verb" is the most common, if semantic agreement in terms of predicate (verb) is missing, an error would occur.

The subject based sentence pattern E) "~ (SURU) KOTO + (WA / GA / MO) + predicate" always corresponds with an inanimate subject, such as a formal subject, a gerund or an infinitive in English, without the inanimate subject in the English sentences there would be an error. Sentence pattern F) "WATASHI + (WA / GA / MO) + predicate" is the most basic, without the subject "I" in the English sentences there would be an error.

Table above shows six sentence patterns and their original criteria for determination whether a sentence is correct or not.

III. RESULTS AND DISCUSSION

A. Results

In order to evaluate our approach, by way of illustration, automatic error detection using four sentence patterns (A, B, C and F) was carried out on Japanese sentences with subjects and their corresponding English sentences.

This study utilized 1499 sentences for analysis from essay data written by 110 Japanese EFL (English as a foreign language) college students. The proficiency level of all the learners was equivalent to the A1 level of the Common European Framework of Reference (CEFR). All the participants were required to write an essay in Japanese with the following prompts: "It is important for college students to have a part time job" and "Smoking should be completely banned at all the restaurants in the country." They then had to translate the Japanese essay into English. The essay had to be 200 - 300 words, written in under 1 hour, with no use of a dictionary or internet enabled devices.

For parsing, 100 Japanese sentences with subjects and the corresponding English sentences were randomly selected from essay data. As a result of parsing, 75 sentences were analyzed since they had the required one subject only. These 75 sentences were classified into six sentence patterns. In order to obtain feedback, comparisons between Japanese primary subjects and predicates and the corresponding English primary subjects and predicate verbs were conducted based on the extraction by parser and sorted based on sentence patterns.

TABLE IV. SAMPLE RESULTS OF EXTRACTION AND ERROR DETERMINATION

	Results of Extraction				Results of Error Determination			
	JPN		*ENG*		*Type of S.P.*		*S.S.*	
	Sub.	*Pre.*	*Sub.*	*Pre.*	*Sub-based*	*Pre-based*	*Sub-based*	*Pre-based*
1	ことは	大切です	It	is important	E	B	-	POSSIBLE
2	人が	困ります	smoking	is difficult	UNKNOWN	D	UNKNOWN	-
3	可能性も	ある	we	have	UNKNOWN	A	UNKNOWN	ERROR
•	•	•	•	•	•	•	•	•
75	私も	思います	I	think	F	C	POSSIBLE	POSSIBLE

*Sub. is an abbreviation of "subject." Pre. is an abbreviation of "predicate." S.P. is an acronym of "sentence pattern." S.S. is an acronym of "sentence structure." (-) is "unanalyzed" on behalf of N/A in this paper.

 ISBN: 978-1-61208-619-4

Table IV shows sample results of extraction and determination.

TABLE V. EVALUATION RESULTS OF THE PREDICATE BASED SENTENCE PATTERNS AND THE SUBJECT BASED SENTENCE PATTERNS

			Results of Determination by Error Detection System			Results of Manual Determination
Type	***S.P.***		***ER.***	***PO.***	***UN.***	***ER.***
Predicate Based	A	8	4	4	0	5
	B	28	9	19	0	13
	C	9	2	7	0	1
	D	30	-	-	-	-
	Total (A+B+C)	45	23	22	0	17
Subject Based	E	4	-	-	-	-
	F	10	4	0	6	3
	Others	61	-	-	-	-
	Total (F)	10	4	0	6	3

*ER. stands for "ERROR." PO. stands for "POSSIBLE." UN. stands for "UNKNOWN." (-) is "unanalyzed" due to being non applicable.

Table V shows the evaluation results on the accuracy to criteria for error detection of the predicate based sentence patterns (A, B, C) and the subject based sentence pattern (F). Manual determination follows the determination method in our previous works. The numbers in Results of Manual Determination are errors identified by criteria for determination (Table III) based on the meaning.

B. Discussion

Comparing the results of determination by error detection system with the results of manual determination in Table V, we obtained the following information.

Sentence Pattern A: In the results of manual determination, nine errors were the same as in the results of determination by error detection system. The coverage of determination was 80%. To increase the accuracy of error determination, distinction between *existence* "ARU" and *probability* "SURUKOTOMO-ARU" is needed.

Sentence Pattern B: In the results of manual determination, seven errors were the same as in the results of determination by error detection system. The coverage of determination was 69.2%. To increase the accuracy of error determination, the problem arising from conjugation must be solved. Cabocha determines conjugation as a noun, although conjugation is a part of verb. Therefore, adding detailed conditions regarding conjugation to sentence pattern is needed.

Sentence Pattern C: In the results of manual determination, two errors were the same as in the results of determination by error detection system. The coverage of determination was 100%. To maintain the accuracy of error determination, adding more reporting verbs in the list is needed.

Sentence Pattern F: In the results of manual determination, two errors were the same as in the results of determination by error detection system. The coverage of determination was 100%. However, the error detection system over-detects errors. To reduce over-detection it is necessary to create detailed sentence patterns and criteria for error determination to deal with sentences which contain multiple subjects, such as a compound sentence or a complex sentence. This will be the subject of further study.

IV. CONCLUSION

In this study, we proposed an approach toward an automatic error detection system. Developing sentence patterns, which differentiates the system from other language error detection systems, is the key point of our approach. We concluded that using the sentence patterns in the source language, automatic error detection is effective when based on our criteria for error determination. The remaining issue is to expand the number of sentence patterns in order to respond to as wide a range of English essays as possible. Furthermore, developing sentence patterns enables our system to be applied to other languages.

ACKNOWLEDGEMENT

This work was partially supported by JSPS KAKENHI Grant Number JP17K01081.

REFERENCES

[1] W. Grabe and R. B. Kaplan, "Theory and practice of writing: An applied linguistics perspective," Harlow: Pearson Education, p.6, 1996.

[2] Y. Takada, "Saishuuhoukokusho," Eigokyouiku Ni Kansuru Kenkyuu, Monbukagakushou itakukenkyuu, ["Final report," Research on English Education, Contract Research Commissioned by the Ministry of Education, Culture, Sports, Science and Technology], p.12, 2004.

[3] K. Kawamura, H. Kashiwagi, and Min Kang, "On the Potential of Error Analysis Using Source Language in Learner's Written English," Research report of JET conference, JSET, 17(2), pp.225-232, 2017.

[4] K. Kawamura, H. Kashiwagi, and M. Kang, "An Approach of Semi-Automatic Correction Feedback for Learners' English Writing Using the Source Language," Proceedings of Symposium on Language and Sustainability in Asia 2017, SELSA2017, pp.11-16, 2017.

[5] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, "A comprehensive grammar of the English language," London, UK: Longman, 1985.

[6] D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, "Longman grammar of spoken and written English," Essex, UK: Pearson Education Limited, 1999.

[7] S. Ando, "Eigo no ronri, nihongo no ronri [The logic of English, the logic of Japanese]," Tokyo: Taishukan, 1986.

[8] T. Kudo and Y. Matsumoto, "Fast Methods for Kernel-Based Text Analysis," ACL 2003 in Sapporo, Japan, 2003.

[9] D. Klein and C. D. Manning, "Accurate Unlexicalized Parsing," Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp.423-430, 2003.

 ISBN: 978-1-61208-619-4