Structural and Semantic Indexing for Supporting Creation of Multilingual Web Pages

Hiroshi URAE, Taro TEZUKA, Fuminori KIMURA, and Akira MAEDA

***Abstract*—Translating webpages by machine translation is the easiest and fastest way a webmaster can multilingualize his/her webpages. Machine translation, however, often causes unnatural and mistranslated sentences with meanings that webmasters do not intend. Therefore, we propose a method that helps the webmaster to create multilingual web pages while avoiding mistranslations by adding metadata about analyzed sentence structures and word meanings. We have developed a prototype system that implements our proposed method. We evaluate our system and prove that it is able to translate sentences that machine translation mistranslates.**

***Index Terms*— Web, Multilingualization, Translation**

I. Introduction

The World Wide Web has enabled us to access information from across the whole world. However, differences in the languages in which webmasters write the contents of web pages are an obstacle to us accessing all the information on the Internet. To overcome this problem, web pages have been multilingualized in various ways. One of the most precise and natural ways to translate webpages is using a professional translation service. However, this is usually costly, so most webmasters of small businesses or personal web sites cannot use these services. In such cases, they translate webpages in one of the two ways. One is self-translation and the other is automatic translation. Self-translation refers to webmasters translating web pages manually and publishing them by themselves. The translated web pages may have relatively natural sounding sentences, depending on the webmaster's proficiency in the target language. This imposes, however, a burden on webmasters. Automatic translation refers to visitors using web translation services, such as Google Translate[1], Yahoo! Babel Fish[2], and Microsoft Translator[3]. This does not impose any burden on webmasters. However, it often produces unnatural and mistranslated sentences with meanings webmasters do not intend.

In this paper, we propose a new method for supporting translation of web pages that produces natural sentences by analyzing sentence structures and what each word means. This system lightens the burden on webmasters by doing this almost automatically. Webmasters are able to correct the system results if these results contain incorrect sentence structures or word meanings. Then the translated sentences become more precise and natural.

The rest of this paper is organized as follows. First, we describe related works in Section II. After that, we describe three steps of our method ("analysis of sentence structures", "analysis of what each word means", and "translating sentences") in Section III and system implementation in Section IV. Finally, we evaluate our system in Section V, and conclude the paper in Section VI.

II. Related Work

A. Describing grammar of several natural languages in the same way

Recently, it has become clear that only using statistical methods for analyzing natural language is not enough. Thus, there is a trend to combine statistical methods with linguistic theories to analyze natural language more deeply. There are various points of view about what the deep analysis of natural language means. Masuichi et al. [1] defined it as "not only analysis of relations of modification between the structural elements but also analysis of predicate argument structure." They developed a system that enables grammar of several natural languages to be described in the same way and the natural sentences to be restored from this grammar by deeply analyzing natural language. They use Lexical Functional Grammar (LFG)[2][3] to describe grammar of several natural languages in the same way. LFG produces two types of structure. One is the c-structure and the other is f-structure. C-structure describes sentence structures as trees. F-structure describes sentence structures as a matrix. The languages using c-structure differ greatly. In contrast, the languages using f-structure differ little.

In this paper, we resolve sentences into their elements like f-structure to describe grammar of several natural languages in the same way.

B. Translation repair using back translation

"Translation repair" is the method to repair incorrect translations by changing words of original text that may cause incorrect translations. It is, however, difficult to find words of original text that may cause incorrect translations. To solve this problem, Miyabe et al. [4] proposed a method

H. Urae is with the Graduate School of Science and Engineering, Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga, Japan, e-mail:(cm002067@ed.ritsumei.ac.jp).

T. Tezuka is with the Graduate School of Library, Information and Media Studies, University of Tsukuba, 1-2 Kasuga, Tsukuba City, Ibaraki, 305-8550, Japan, email: (tezuka@slis.tsukuba.ac.jp).

F. Kimura, and A. Maeda are with the Colleage of Information Science and Engineering, Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga, Japan, e-mail: (amaeda@media, fkimura@is.ritsumei.ac.jp).

[1] Google Translate http://translate.google.co.jp/
[2] Yahoo! Babel Fish http://babelfish.yahoo.com/
[3] Microsoft Translator http://www.microsofttranslator.com/

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

that use back translation. In this method, they estimate the words that make incorrect translations by finding words that differs between original text and the result of back translation. Thus, translators can easily find these words to repair the original text. Repaired text shows that the method is effective to decrease incorrect translations. It imposes, however, a burden on translators that they have to consider new words to replace the words that makes incorrect translations.

In contrast, we propose a method that repairs incorrect translations directly by using the results of analyzing original text. The translator does not have to change the original text, and only has to select what each word means.

III. Proposed Method

Our proposed method has three steps: "analysis of sentence structures", "analysis of what each word means" and "translating sentences". We show the system outline in Fig 1.

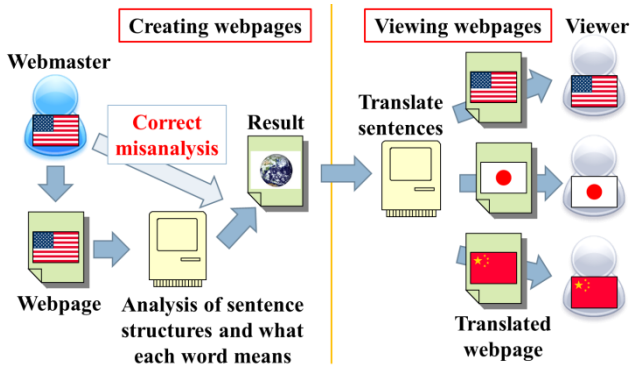

Fig. 1. System outline

A. Analysis of sentence structures

In this step, we resolve sentences into phrases that consist of principle elements such as subjects (S), verbs (V), complements (C), and objects (O) or modifiers (M). We name a sentence that consists of only principle elements as a "fundamental sentence". Each modifier has metadata by which a principle element is modified. We use the Apple Pie Parser (APP), which is a tool for analyzing English sentence structures automatically, and our system determines whether each word is a principle element or a modifier. If this determination fails, the webmaster can correct a misanalyzed word manually.

For example, suppose we resolve the sentence "The river overflowed its banks after a typhoon" into a fundamental sentence and modifiers. The results are shown in Fig 2.

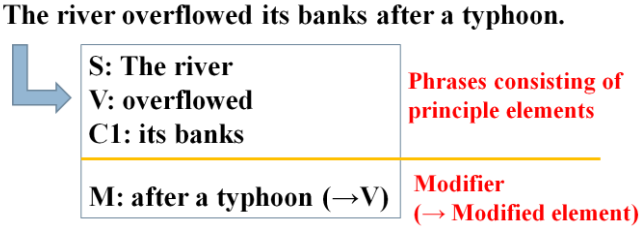

Fig. 2. Example of resolving the sentence "The river overflowed its banks after a typhoon" into a fundamental sentence structure and modifiers

B. Analysis of what each word means

In this step, we analyze what each word means by referring to the word ID database. The word ID database is a language resource created for the system. We made it from the Japanese-English dictionary "Eijiro[4]". We assigned Word IDs to each meaning. Therefore, we assign different IDs for homonyms (words with the same spelling but different meanings) as shown in table 1. Word IDs are assigned not only for words but also for phrases and idioms. Our system analyzes what each word means automatically. If this analysis fails, the webmaster can correct the meanings of each word. Then, our system converts each word, phrase, and idiom into Word ID.

For example, suppose we analyze what each word of the sentence "The river overflowed its banks after a typhoon" means. The results are shown in Fig 3.

TABLE I
EXAMPLE OF WORD ID DATABASE

Word ID	English	Japanese
157833	bank	(meaning: a geographic bank) 土手 岸
157844	bank	(meaning: pile up A) ～を積み上げる ～を山にする
157850	bank	(meaning: a financial institution) 銀行

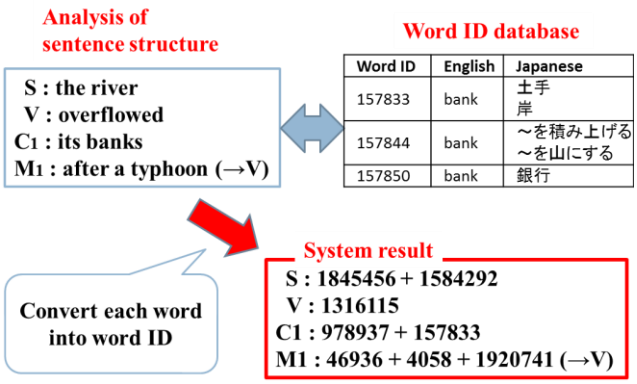

Fig. 3. Analyzing what each word of the sentence "The river overflowed its banks after a typhoon" means

C. Translating sentences

In this step, we translate sentences into the user's native language. The accuracy of translation, however, heavily depends on each language. Thus, we propose a translation method that combines the result of the previous step (analysis of what each word means) with a machine translation API. We use Google Translate as a machine translation API in our system. First, our system utilizes the machine translation API to translate the fundamental sentence into the user's native language. After that, the system obtains the translated fundamental sentence from the machine translation API, and our system converts the Word IDs of principle elements into the words' meanings in the user's native language by referring to the word ID database. We call this process

[4] Eijiro http://www.alc.co.jp/

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

"converted meanings". Then our system checks that the translated fundamental sentence contains each converted meaning. If the translated fundamental sentence contains all of the converted meanings, our system assumes that machine translation API has translated the sentence successfully. If, however, the translated fundamental sentence does not contain all of the converted meanings, our system assumes that machine translation API has mistranslated the sentence. In such cases, our system corrects mistranslations of the translated fundamental sentence. Our system restores the Word ID into the original word that the webmaster spelled. Then we search all of the converted meanings into which the original word may be translated. We call them "possible mistranslated meanings". Then our system rechecks whether the translated fundamental sentence contains any possible mistranslated meanings. If the translated fundamental sentence contains any possible mistranslated meanings, our system assumes that these possible mistranslated meanings are mistranslations of the converted meanings and replaces these possible mistranslated meanings with the converted meanings.

After correcting the translated fundamental sentence, our system adds modifiers to the translated fundamental sentence. In the same way as for principle elements, our system obtains the converted meanings of modifiers. By referring to the results of the first step (analysis of sentence structures), our system obtains metadata the principle element of which is modified by each modifier and uses them to add each modifier to the translated fundamental sentence.

For example, suppose we translate the sentence "The river overflowed its banks after a typhoon." First, our system obtains the sentence "川は、その銀行をオーバーフローしました" as the translated fundamental sentence. In this translated fundamental sentence, there are two mistranslations. One is "銀行", which means a financial institution. This is mistranslation of "bank". The other is "オーバーフローしました", which means arithmetic overflow. This is mistranslation of "overflowed". By checking that the translated fundamental sentence contains converted meanings as shown in Fig. 4, correcting mistranslation as shown in Fig. 5, and adding modifiers to the translated fundamental sentence as shown in Fig. 6, finally, our system produces the sentence "川は、その土手を台風後のあふれるしました". This result is not a fluent sentence but is a much more accurate translation in terms of the selection of translated words.

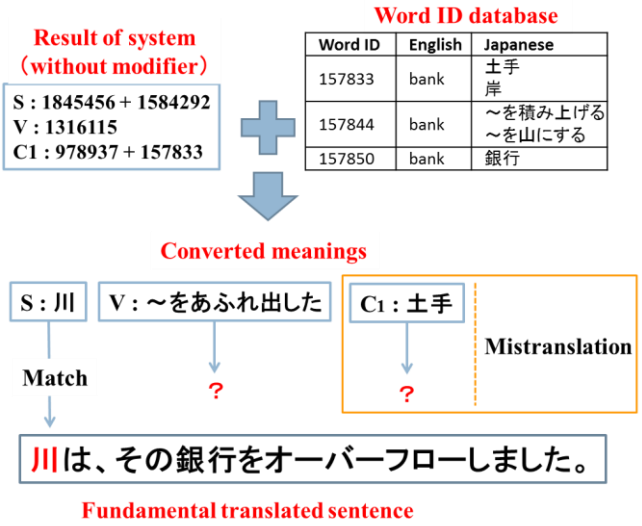

Fig. 4. Checking that the translated fundamental sentence contains converted meanings

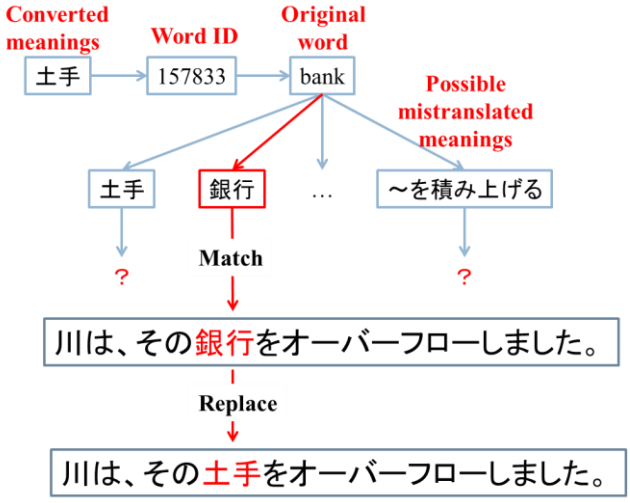

Fig. 5. Correcting mistranslation

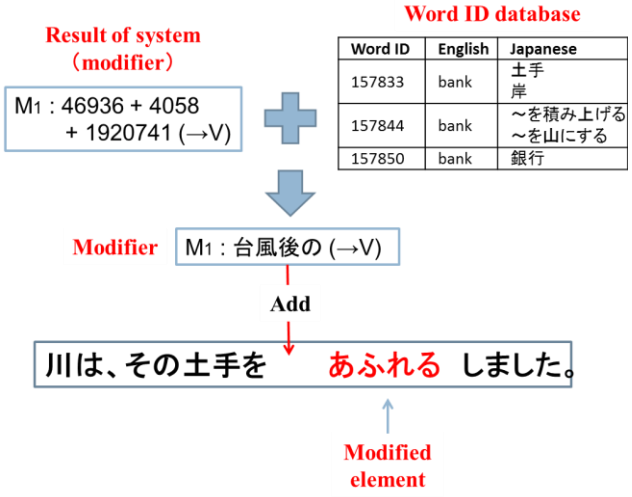

Fig. 6. Adding modifiers to the translated fundamental sentences

IV. System Implementation

In this section, we describe the functions and features of the user interface of our system.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

A. *Part of sentence and Sub-element*

A sentence does not always have only one fundamental sentence. If a sentence contains conjunctions, it consists of more than one fundamental sentence. To handle these complex sentences, we divide such a sentence into parts. Each part contains either a fundamental sentence and modifiers or other elements such as conjunctions and periods.

Sometimes, an element contains a description of a parallel relationship between the words. If these words are modified, the element becomes long and complex. Therefore, we divide the element into sub-elements. Modifiers can have metadata by which a sub-element of an element is modified.

Finally, the elements and Word IDs that are the results of analyzing sentence structures are described like "part – element – sub-element : word ID".

B. *Combining the meanings of several words*

If several words consist of a phrase or an idiom, the webmaster is able to select their meanings as a phrase or an idiom. If the webmaster wants to combine words into a phrase, he/she can click the combine/separate button as shown in Fig 7.

Fig. 7. Combining the meanings of several words

V. Evaluation

We experimented to evaluate the effectiveness of our system. In this experiment, we translated sentences from English into Japanese by using our system. We prepared 25 sentences for this evaluation. We obtained these sentences from a Japanese-English corpus and a dictionary. In the same way, we used Google Translate as a comparative translation system for our system. We evaluate the results of experiment in two ways. One is an automatic evaluation using BLEU [5], and the other is a subjective evaluation.

A. *Evaluating the systems by BLEU*

BLEU is a method for automatically evaluating the quality of machine-translated text by comparing the reference human-translated text and the system result. BLEU is calculated as follows:

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

$$P_n = \frac{\sum_{C \in \{\mathrm{Candidates}\}} \sum_{\text{n-gram} \in C} \mathrm{Count}_{\mathrm{clip}}(\text{n-gram})}{\sum_{C' \in \{\mathrm{Candidates}\}} \sum_{\text{n-gram}' \in C'} \mathrm{Count}(\text{n-gram}')}$$

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

where N is n-gram length, C is references, $\mathrm{Count}_{\mathrm{clip}}(\text{n-gram})$ is the number of n-grams in which reference and system result match, C' is system results, $\mathrm{Count}(\text{n-gram}')$ is the number of system result n-grams, c is the total length of system results, and r is the total length of references.

In using BLEU, we set the parameter N to 4. Before evaluation, we manually changed the references to the best match of each result because Japanese natural sentences have the following features.

Synonym and difference of characters

Japanese, like other languages, has many synonyms. Moreover, Japanese has three types of script: *Hiragana*, *Katakana* and *Kanji*. Japanese natural sentences are mainly written in a mix of these three scripts. For example, *tempura*, a Japanese dish, is usually written "天ぷら". "天" is *Kanji*, while "ぷ" and "ら" are *Hiragana*. Sometimes, tempura is written "てんぷら" (only using *Hiragana*), "テンプラ" (only using *Katakana*), or "天麩羅" or "天婦羅" (only using *Kanji*). These mean the same thing, the only difference being the characters. Therefore, we changed these words of references into other words or other characters to best match each result.

Position of modifier in the sentence

Modifiers in Japanese natural sentences do not always set a position related to modified words. It, of course, cannot be placed anywhere. Thus, we changed some positions of modifiers as long as these changes were not unnatural.

Differences in these auxiliary verbs

In Japanese, auxiliary verbs, for example "です"(-desu) / "ます"(-masu), are sometimes used at the end of a sentence instead of auxiliary verbs "だ"(-da) / "である"(-dearu) in order to represent more polite or softer nuances. However, main meanings do not differ even if any of these auxiliary verbs are used. Therefore, we ignore the difference of these auxiliary verbs in the corpus sentences.

The results of evaluation by BLEU are shown in Table II.

TABLE II
Evaluation Results of the Systems by BLEU

Translation System	Score
Our System	0.0544
Google Translate	0.0881

B. *Subjective evaluation*

In subjective evaluation, we use two viewpoints (adequacy and fluency) and evaluate the results of our translation system by comparing it with Google Translate. The results of subjective evaluation are shown in Table III and Table IV.

TABLE III
Evaluation Results of the Systems by Subjective Evaluation in Terms of Adequacy

	Improved	Worsened
Total	7	13

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE IV
EVALUATION RESULTS OF THE SYSTEMS BY SUBJECTIVE EVALUATION IN TERMS OF FLUENCY

	Improved	Worsened
Total	1	12

C. Discussion

The "improved" column in Table III indicates that our system is able to correct mistranslations of Google Translate. Our system, however, sometimes fails to add modifiers into the translated sentence as shown in the "worsened" column in Table III. In addition, our system just replaces words with other words without considering the connection between them. Thus, from the viewpoint of fluency, our system is worse than Google Translate.

D. Re-evaluation experiment

In the previous experiment, most of the sentences we chose were relatively simple and clear. Thus, Google Translate did not mistranslate these sentences. In this re-experiment, we prepared another 25 sentences that Google Translate mistranslated and re-evaluated our system as the same way as the previous evaluation.

E. Re-evaluate systems by BLEU

The results of re-evaluation by BLEU are shown in Table V.

TABLE V
RE-EVALUATION RESULTS OF THE SYSTEMS BY BLEU

Translation System	Score
Our System	0.1518
Google Translate	0.0248

F. Subjective evaluation

The results of re-evaluation by subjective evaluation are shown in Tables VI and VII, and the correction rate of the system in terms of adequacy is shown in Table VIII. In Table VIII, the "Sentences" column shows the results from a viewpoint of sentences, the "Correct points" column shows the results from the viewpoint of mistranslated points of words, phrases, and idioms that sentences have (one sentence may have two or more correct points), the "Total" row shows the total number of sentences or correct points, "System correction" row shows how many sentences or correct points the system corrected, and the "Correction rate" row shows the correction rate of sentences or correct points that is calculated by the value of the "System correction" row divided by the value of the "Total" row. "Full correction", which appears in "System Correction" row and "Correction rate" row, shows the number of sentences or correction rate of sentences in which the system corrected all the mistranslated points. We investigated the reasons the system mistranslated in each case, and the results are shown in Table IX.

TABLE VI
RE-EVALUATION OF THE SYSTEMS BY SUBJECTIVE EVALUATION IN TERMS OF ADEQUACY

	Improved	Worsened
Total	24	5

TABLE VII
RE-EVALUATION OF THE SYSTEMS BY SUBJECTIVE EVALUATION IN TERMS OF FLUENCY

	Improved	Worsened
Total	0	7

TABLE VIII
CORRECTION RATE OF THE SYSTEM IN TERMS OF ADEQUACY

	Sentences	Correct points
Total	25	39
System Correction	19	24
(Full correction)	(12)	(-)
Correction rate	0.76	0.61
(Full correction)	(0.48)	(-)

TABLE IX
REASONS WHY THE SYSTEM IS INCORRECT

Reason	Incorrect points
Using unnatural nuance	3
Mismatching text because of differences of tense or part of speech	4
Mismatching text because of synonyms	1
Mismatching text because of mistranslated phrases or idioms	4
Mismatching text because of using non-existent meanings that machine translation made	1
Could not select correct meanings when the webmaster created the webpage	2

G. Discussion

The "improved" column in Table VI indicates that our system is able to correct many mistranslations of Google Translation. As shown in Table VI, improvements surpass degradations. Thus, our system is able to translate sentences more accurately especially when machine translation mistranslated. As shown in Table VIII, machine translation mistranslated 39 points of words, phrases, or idioms in this experiment. Of these, our system successfully corrected 24 points (correction rate: 0.61). Out of the total 25 sentences, our system successfully corrected 19 sentences (correction rate: 0.76), and fully corrected 12 sentences (correction rate: 0.48).

As shown in Table IX, most of the reasons the system was incorrect are mismatching of text. This means that the system failed to find mistranslated words, phrase, or idioms in various causes. This indicates that the accuracy can be improved if we can match texts more correctly.

VI. CONCLUSION

In this paper, we proposed a new method for supporting creation of multilingual web pages that creates natural sentences by analyzing sentence structures and what each word means, and we developed a system to create translated sentences using this method. Experimental results showed that our system is able to translate sentences more accurately when machine translation mistranslates. Our system, however, has some points that need improvement:

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

A. Matching translated meanings more exactly with translated fundamental sentences

When adding modifiers into fundamental translated sentences, our system searches for the positions of modified elements by the modifier to decide the position into which the system adds modifies. Our system, however, uses simple text matching to search for these positions, and these matchings often fail. We will try to solve this problem by using a part-of-speech analyzer.

B. Tense of words

The meanings contained in the word ID database are the original forms. Thus, the results of our system can only describe present forms. After matching the translated meanings more exactly with the translated fundamental sentences, we will try to correct the tense of each sentence by referring to a dictionary of tense or by some other method.

ACKNOWLEDGMENT

This work was supported in part by the Grant-in-Aid for the Global COE Program “Digital Humanities Center for Japanese Arts and Cultures (DH-JAC)” from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, MEXT Grant-in-Aid for Strategic Formation of Research Infrastructure for Private University “Sharing of Research Resources by Digitization and Utilization of Art and Cultural Materials” (Grant Number: S0991041), and MEXT Grant-in-Aid for Young Scientists (B) “Research on Information Access across Languages, Periods, and Cultures” (Leader: Akira Maeda, Grant Number: 21700271).

REFERENCES

[1] M. Butt, H. Dyvik, T. H. King, H. Masuichi, and C. Rohrer, The Parallel Grammar Project, Proc. the 19th International Conference on Computational Linguistics Workshop Grammar Engineering and Evaluation, 2002.

[2] R. M. Kaplan, and J. Bresnan, Lexical-Functional Grammar: A formal system for grammatical representation, in The Mental Representation of Grammatical Relations, pp.173 – 281, The MIT press, 1982.

[3] R. M. Kaplan, and J. Wedekind, LFG generation produces context-free languages, Proc. the 18th International Conference on Computational Linguistics, pp.425 - 431, 2000.

[4] M. Miyabe, T. Yoshino, T. Shigenobu, Effects of Undertaking Translation Repair using Back Translation, Proc. the 2009 ACM International Workshop on Intercultural Collaboration (IWIC'09), pp.33-40, 2009.

[5] A. Papineni, S. Roukos, T.Ward, and W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, Technical Report RC22176 (W0109-022), IBM Research Division, 2001.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)