Utilizing Semantic Equivalence Classes of Japanese Functional Expressions in Machine Translation

Akiko Sakamoto Takehito Utsuro
University of Tsukuba
Tsukuba, Ibaraki, 305-8573, JAPAN

Suguru Matsuyoshi
Nara Institute of Science and Technology
Ikoma, Nara, 630-0192, JAPAN

ABSTRACT

This paper applied "Sandglass" machine translation architecture to the task of translating Japanese functional expressions into English. We employ the semantic equivalence classes of a recently compiled large scale hierarchical lexicon of Japanese functional expressions. We examine each class whether it is monosemous or not. We realize this procedure by empirically studying whether functional expressions within a class can be translated into a single canonical English expression. Furthermore, in order to precisely identify the class of functional expressions to which our translation rule is directly applicable, we further introduce two types of ambiguities of functional expressions and identify monosemous functional expressions. We finally show that the proposed framework outperforms commercial machine translation software products.

Categories and Subject Descriptors

J.5.0 [**ARTS AND HUMANITIES**]: Language translation

General Terms

Languages

Keywords

machine translation, Japanese functional expressions, polysemy, sense disambiguation

1. INTRODUCTION

The Japanese language has various types of functional expressions, which are very important for understanding their semantic contents. Those functional expressions are also problematic in further applications such as MT of Japanese sentences into English. This problem can be partially recognized by the fact that the Japanese language has a large number of variants of functional expressions, where their total number is recently counted as over 10,000 in [5]. Based on those recent development in studies on lexicon for processing Japanese functional expressions [5], this paper studies issues on MT of Japanese functional expressions into English.

More specifically, in order to solve the problem of a large number of variants of Japanese functional expressions, in this paper, we employ the "Sandglass" MT architecture [14]. In the "Sandglass" MT architecture, variant expressions in the source language are first paraphrased into representative expressions, and then, a small number of translation rules are applied to the representative expressions. In this paper, we apply this architecture to the task of translating Japanese functional expressions into English, where we introduce a recently compiled large scale hierarchical lexicon of Japanese functional expressions [5]. We employ the semantic equivalence classes of the lexicon and examine each class whether it is monosemous or not. We realize this procedure by empirically studying whether functional expressions within a class can be translated into a single canonical English expression. Then, we create one translation rule for each of the monosemous semantic equivalence classes. Here, one of the most important assumption of applying those translation rules is that each functional expression to which those translation rules are applied must be monosemous. We introduce two types of ambiguities of functional expressions and further identify monosemous functional expressions. The experimental evaluation result shows that the proposed framework outperforms commercial MT software products.

2. JAPANESE FUNCTIONAL EXPRESSIONS

Even before [5] recently compiled the almost complete list of Japanese functional expressions, there had existed several collections which list Japanese functional expressions and examine their usages. For example, [6] examined 450 functional expressions and [1] also listed 965 expressions and their example sentences. Compared with those two collections, *Gendaigo Hukugouji Youreishu* [7] concentrated on 125 major functional expressions which have non-compositional usages, as well as their variants (337 expressions in total)[1], and collected example sentences of those expressions. For each of the 337 expressions, [12] developed an example database of, which is used for training/testing a chunker of Japanese (compound) functional expressions. The corpus from which they collected example sentences is 1995 Mainichi news-

[1]For each of those 125 major expressions, the differences between it and its variants are summarized as below: i) insertion/deletion/alternation of certain particles, ii) alternation of synonymous words, iii) normal/honorific/conversational forms, iv) base/adnominal/negative forms.

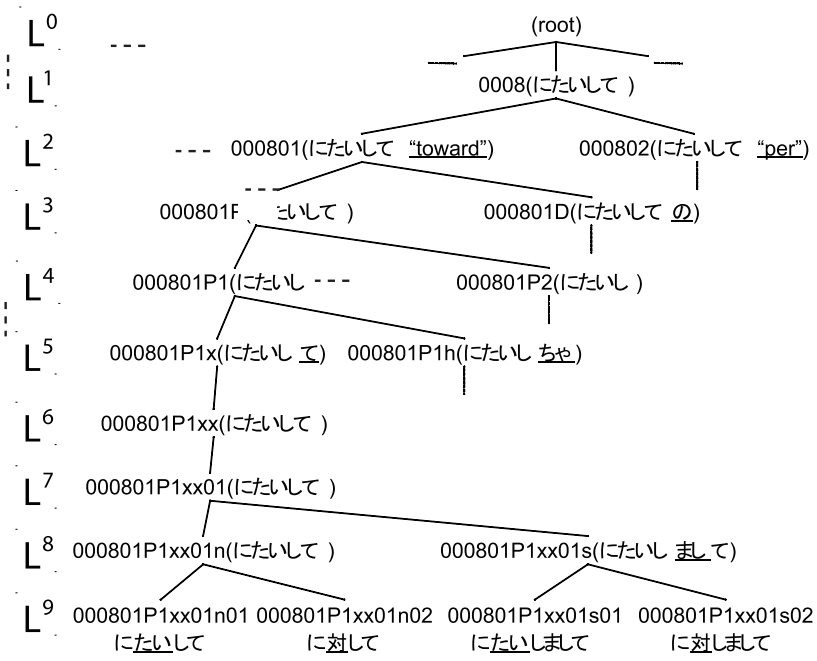

Figure 1: A Part of the Hierarchical Lexicon of Japanese Functional Expressions

paper text corpus (1,294,794 sentences, 473,553,300 bytes, 236,776,650 characters). For each of the 337 expressions, 50 sentences were collected and labels for chunking were annotated.

3. HIERARCHICAL LEXICON OF JAPANESE FUNCTIONAL EXPRESSIONS

3.1 Morphological Hierarchy

In order to organize Japanese functional expressions with various surface forms, [5] proposed a methodology for compiling a lexicon of Japanese functional expressions with hierarchical organization[2]. [5] compiled the lexicon with 341 headwords and 16,801 surface forms. The hierarchy of the lexicon has nine abstraction levels and Figure 1 shows a part of the hierarchy[3]. In this hierarchy, the root node (in L^0) is a dummy node that governs all the entries in the lexicon. A node in L^1 is an entry (headword) in the lexicon; the most generalized form of a functional expression. A leaf node (in L^9) corresponds to a surface form (completely-instantiated form) of a functional expression. An intermediate node corresponds to a partially-abstracted (partially-instantiated) form of a functional expression. The second level L^2 distinguishes senses of Japanese functional expressions. This level enables distinction of more than one senses of one functional expression. For example, "にたいして" (nitaishi-te)" has two different senses. The first sense is "toward"; e.g., "彼は私 にたいして 親切だ。" " (He is kind toward me). The second sense is "per"; e.g., "一人 にたいして5 つ。" " (five per one person). This level is introduced to distinguish such ambiguities. On the other hand, L^3 distinguishes grammatical functions, L^4 distinguishes alternations of function words, L^5 distinguishes phonetic variations, L^6 distinguishes optional focus particles, L^7 distinguishes conjugation forms, L^8 distinguishes normal/polite forms, and L^9 distinguishes spelling variations.

[2] http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

[3] In this lexicon, following [9], each functional expression is regarded as a fixed expression, rather than a semi-fixed expression or a syntactically-flexible expression.

3.2 Semantic Hierarchy

Along with the hierarchy of surface forms of functional expressions with nine abstraction levels, the lexicon compiled by [5] also has a hierarchy of semantic equivalence classes introduced from the viewpoint of paraphrasability. This semantic hierarchy has three abstraction levels, where 435 entries in L^2 (headwords with a unique sense) of the hierarchy of surface forms are organized into the top 45 semantic equivalence classes, the middle 128 classes, and the 199 bottom classes.

Figure 2 shows examples of the bottom 199 classes, where each of the leaf labels "B13", "B31", "B32", "C11", . . ., "d11", "d12", "d13", . . . represents a label of the bottom 199 classes. In [4], the bottom 199 semantic equivalence classes of Japanese functional expressions are designed so that functional expressions within a class are paraphrasable in most contexts of Japanese texts.

4. IDENTIFYING MONOSEMOUS SEMANTIC EQUIVALENCE CLASSES OF FUNCTIONAL EXPRESSIONS IN TRANSLATION

In this paper, in terms of translation in English, we identifying monosemous semantic equivalence classes of Japanese functional expressions. More specifically, we examine the effects of the bottom 199 semantic equivalence classes in MT, where we empirically study whether functional expressions within a class can be translated into a single canonical En-

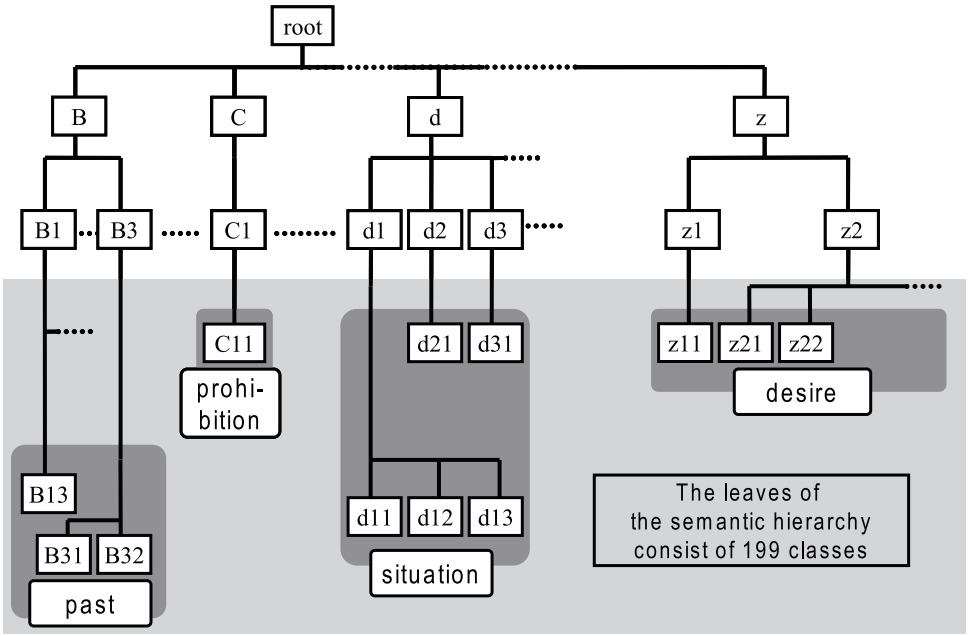

Figure 2: A Part of the Hierarchy of Semantic Equivalence Classes

glish expression. This section gives the description of the procedure.

First, we use a Japanese corpus of about 8,000 sentences for Japanese language grammar learners [1] as a repository for collecting example sentences of Japanese functional expressions. For each of the 199 semantic equivalence classes, we collect example sentences from this corpus. Here, for each of the 199 classes, we manually judge whether the sense of the functional expression in each sentence corresponds to that of the target class. Then, we keep 91 classes that are with at least five example sentences and we use the total 455 (5 sentences for each of the 91 classes) collected example sentences in further examination for translation into English.

The 455 example sentences are next manually translated into English. Here, if all of the five Japanese functional expressions can be translated into a single canonical English expression, we classify the class as "single translation", and otherwise, as "multiple translations". The "single translation" semantic equivalence classes are considered as monosemous. The result of the procedure is shown in Figure 3, where 49 out of the 91 classes are classified as "single translation", and the remaining 42 as "multiple translations". Furthermore, 11 classes out of the 49 "single translation" classes can be merged into 5 classes, each of which can be regarded as one "single translation" class. The 49 "single translation" classes cover more than 6,000 functional expressions.

5. IDENTIFYING MONOSEMOUS FUNCTIONAL EXPRESSIONS

One of the most important assumption of applying the translation rules invented in the previous section is that each functional expression to which those translation rules are applied must be monosemous. Unless each functional expression is monosemous, it is necessary to apply certain disambiguation techniques and then apply translation rules that are appropriate for the actual usage of the target functional expression.

This section first overviews two types of ambiguities of *functional* expressions (in a broad sense). Then, we give the procedure of identifying monosemous functional expressions which do not have either ambiguities of the two types.

5.1 Ambiguities of Functional/Content Usages

The first type of ambiguity is for the case that one compound expression may have both a literal (i.e. compositional) *content word* usage and a non-literal (i.e. non-compositional) *functional* usage. This type of ambiguity often happens when the surface form of a functional expression can be decomposed into a sequence of at least one content word and one or more function words. In such a case, the surface form of the compound expression may have both a literal (i.e. compositional) *content word* usage where each of its constituents has its own literal usage, and a non-literal (i.e. non-compositional) *functional* usage where its constituents have no longer their literal usages.

For example, Table 1 (b) shows two example sentences of a compound expression "と (to) は (ha) いえ (ie)", which consists of a post-positional particle "と (to)", a topic-marking particle "は (ha)", and a conjugated form "いえ (ie)" of a verb "いう (iu)". In the sentence (2), the compound expression functions as an adversative conjunctive particle and has a non-compositional functional meaning "*although*". On the other hand, in the sentence (3), the expression simply corresponds to a literal concatenation of the usages of the constituents: the post-positional particle "と (to)", the topic-marking particle "は (ha)", and the verb "いえ (ie)", and has a content word meaning "*can not say*". Compared to Table 1 (b), Table 1 (a) shows an example of a functional

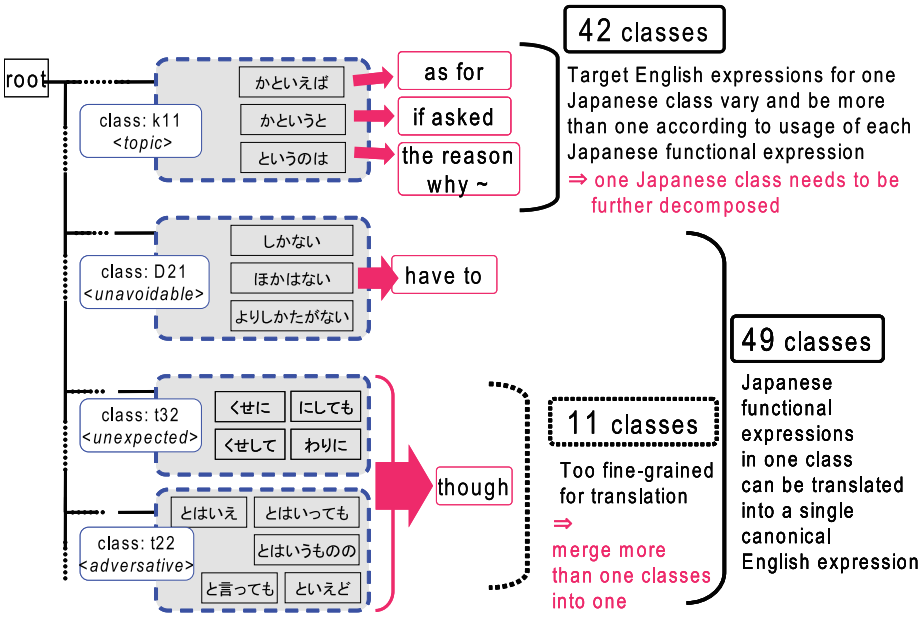

Figure 3: Translation of Japanese Functional Expressions through Semantic Equivalence Classes

expression without ambiguity of functional/content usages. In this case, the compound expression "こと (koto) が (ga) できる (dekiru)" consists of a formal noun "こと (koto)", a post-positional particle "が (ga)", and an auxiliary verb "できる (dekiru)". In almost all the occurrences in a newspaper corpus, the surface form of this compound expression functions as an auxiliary verb and has a non-compositional functional meaning "*can*".

This type of ambiguity has been well studied in [12], [11], and [10]. [12] reported that, out of about 180 compound expressions which are frequently observed in the newspaper text, one third (about 60 expressions) have this type of ambiguity. Next, [11] formalized the task of identifying Japanese compound functional expressions in a text as a machine learning based chunking problem. The proposed technique performed reasonably well, while its major drawback is in its scale. So far, the proposed technique has not yet been applied to the whole list of over 10,000 Japanese functional expressions. [10] also studied applying manually created rules to the task of resolving functional/content ambiguities, where their approach has limitation in that it requires human cost to create manually and to maintain those rules.

To summarize, any existing tool for resolving the ambiguity of functional/content usages for the whole list of over 10,000 Japanese functional expressions is not yet publicly available. Considering this situation, we conclude that we should avoid expressions which have this type of ambiguity when evaluating our translation rules.

5.2 Ambiguities of Functional Usages

The second type of ambiguity is for the case that the surface form of a functional expression has more than one *functional* usages. For example, Table 1 (c) shows two example sentences of a compound expression "ため (tame) に (ni)", which consists of a noun "ため (tame)" and a post-positional particle "に (ni)". In the sentence (4), the compound expression functions as a case-marking particle and has a non-compositional functional meaning "*for the purpose of*". Also in the sentence (5), the compound expression functions as a case-marking particle, but in this case, has another non-compositional functional meaning "*because of*". Compared to Table 1 (c), Table 1 (a) shows an example of a functional expression without ambiguity of functional usages. In this case, the functional expression "こと (koto) が (ga) できる (dekiru)" has only one non-compositional functional meaning "*can*".

This type of ambiguity includes issues on typical polysemies and homographs, where the issues on sense disambiguation of content words have been well studied in NLP community (e.g, in SENSEVAL tasks [2, 3]). However, in the areas of semantic analysis of Japanese sentences as well as machine translation of Japanese sentences, the issue of sense disambiguation of functional expressions has not been paid much attention so far, and any standard tool for sense disambiguation of Japanese functional expressions have not been publicly available. Considering the current situation on this type of ambiguity of functional usages, again we conclude that we should avoid expressions which have this type of ambiguity when evaluating our translation rules.

5.3 The Procedure of Identifying Monosemous Functional Expressions

Now, we present the procedure of identifying monosemous functional expressions which do not have either ambiguities of the two types. This procedure is applied to 166 L^2 entries as well as 6379 L^9 entries which belong to the 49 "single translation" semantic equivalence classes identified in section 4.

As shown in Table 2, first, 166 L^2 entries as well as 6379 L^9

Table 1: Example of Functional Expressions in 49 Monosemous Semantic Equivalence Classes

(a) *w/o* ambiguity of functional usages AND *w/o* ambiguity of functional/content usages

	Expression	Example sentence (English translation)	Usage
(1)	ことができる (koto-ga-dekiru)	彼は英語を話す[ことができる]。 (He *can* speak English.)	functional, semantic class = *possible* (ことができる (koto-ga-dekiru) = *can*)

(b) *w/o* ambiguity of functional usages AND *with* ambiguity of functional/content usages

	Expression	Example sentence (English translation)	Usage
(2)	とはいえ (to-ha-ie)	状況は改善している[とはいえ]、まだ安心できない。 (*Although* it has become better, we can not feel easy.)	functional, semantic class = *adversative* (∼ とはいえ (to-ha-ie) = *although* ∼
(3)	とはいえ (to-ha-ie)	状況が改善した[とはいえ]ない。 (We *can not say* that it has become better.)	content (∼ とはいえ (ない) (to-ha-ie(-nai))) = *cat not say* ∼)

(c) *with* ambiguity of functional usages

	Expression	Example sentence (English translation)	Usage
(4)	ために (tame-ni)	世界平和の[ために]国際会議が開かれる。 (An international conference is held *for the purpose of* world peace.)	functional, semantic class = *purpose* (ために (tame-ni) = *for the purpose of*
(5)	ために (tame-ni)	雨の[ために]彼の到着が遅れた。 (He arrived late *because of* rain.)	functional, semantic class = *reason* (ために (tame-ni) = *because of*

entries in the 49 "single translation" classes are divided into those *with* the ambiguity of functional usages and *without* the ambiguity of functional usages. Here, if the surface form of a functional expression of an entry X (i.e., ID) in the lexicon is identical to that of a functional expression of another entry Y (i.e., ID) in the lexicon, then we regard both of the entries X and Y as *with* the ambiguity of functional usages. Next, for each of the surface forms of functional expressions *without* the ambiguity of functional usages, we collect example sentences from 1995 Mainichi newspaper text corpus and blog text (about 130,000,000 characters, which includes colloquial forms of functional expressions more often than in the newspaper text). Then, we keep surface forms with more than or equal to 20 occurrences in either of the newspaper text or the blog text. Finally, for each of the surface forms of the remaining functional expressions, we observe the collected example sentences and judge whether their usages have the ambiguity of functional/content usages. The distribution of the numbers of functional expressions in terms of that of entries (i.e., ID) in the lexicon is shown in Table 2. As shown in the table, 42 L^2 entries as well as 2756 L^9 entries are identified as monosemous functional expressions.

6. EVALUATION OF TRANSLATION RULES

For each of the 49 "single translation" classes identified in section 4, we evaluate the rule of translation into a single canonical English expression with 272 held-out example sentences collected from the 8,000 sentences of [1] as well as newspaper text and blog text. We evaluate the English translation of Japanese functional expression into the three levels: "correct", "partially correct", and "error". Here, we achieved 96.3% "correct" rate.

Next, in order to compare this correct rate with commercial MT software products[4], we divide the 272 sentences for evaluation into 121 sentences which include monosemous functional expressions identified in section 5.3 and the remaining 151 sentences. To the monosemous functional expressions in the 121 sentences, our translation rule can be directly applied without any disambiguation techniques. As we show in Figure 4, in the evaluation against the monosemous functional expressions in the 121 sentences, we outperformed the commercial MT product, although the scale of the evaluation is small. This result partially supports the effects of the proposed approach.

Also in the evaluation against the ambiguous functional expressions in the remaining 151 sentences, we outperformed the commercial MT product. This is simply because in the 151 sentences we selected, functional expressions have exactly the same usage/sense as assumed in the translation rule we use in this comparison. In actual situation of translating those ambiguous functional expressions, however, it

[4]We compared 7 commercial J/E MT softwares and selected one of them with the best correct rate in translation of Japanese functional expressions.

Table 2: # of Functional Expressions in 49 Monosemous Semantic Equivalence Classes (L^2 entries / L^9 entries, both in # of IDs in the hierarchical lexicon)

w/o ambiguity of functional usages			*with* ambiguity of functional usages
w/o ambiguity of functional/content usages	*with* ambiguity of functional/content usages	less than 20 occurrences in newspaper/blog corpora	
42 / 2752	22 / 749	33 / 2188	
97 / 5689			69 / 690
166 / 6379			

is of course necessary to apply certain disambiguation techniques before applying any translation rule. Since the comparison through this 151 sentences is too advantageous to our rule, we show the result of evaluating the commercial MT product against the ambiguous 151 sentences, not for comparing to our rule, but just for comparing to that against the 121 sentences with monosemous functional expressions.

Table 3 shows example sentences used for creating the translation rules and those used for evaluation. With the sentences for evaluating the translation rules, we also show English translation returned by both the proposed translation rule and the commercial MT product. The first semantic equivalence class "M11" includes functional expressions which express the sense of "unnecessary" such as "ないでもよい (nai-de-mo-yoi)" and "までもない (made-mo-nai)". For this class, a translation rule with an English translation *"not have to"* is created. As shown in the table, in the results of evaluating the translation rule for this class as well as that for the second class "o21", the proposed method outperformed the commercial MT product. However, in the result of evaluating the translation rule for the third class "b11", the judgements both for the proposed method and for the commercial MT product were "error".

7. RELATED WORKS

[14] proposed the "Sandglass" machine translation architecture in which variant expressions in the source language are first paraphrased into representative expressions, and then, a small number of translation rules are applied to the representative expressions. In this paper, we apply the "Sandglass" architecture to the task of translating Japanese functional expressions into English, where we introduce a recently compiled large scale hierarchical lexicon of Japanese functional expressions [5, 4].

[13] and [8] studied syntactic analysis of functional expressions in sentences. [13] studied how to incorporate the process of analyzing compound non-compositional functional expressions into the framework of Japanese statistical dependency parsing. [8] also reported improvement of Swedish parsing when multi word units are manually annotated.

8. CONCLUDING REMARKS

This paper applied "Sandglass" MT architecture [14] to the task of translating Japanese functional expressions into English. We employed the semantic equivalence classes of a recently compiled large scale hierarchical lexicon of Japanese functional expressions. We examined each class whether it is monosemous or not. by empirically studying whether functional expressions within a class can be translated into a single canonical English expression. Furthermore, in order to precisely identify the class of functional expressions to which our translation rule is directly applicable, we further introduced two types of ambiguities of functional expressions and identified monosemous functional expressions. We then showed that the proposed framework outperformed commercial MT software products. Future work includes scaling up the procedure of empirical examination on discovering "single translation" semantic equivalence classes into the whole 199 classes.

9. REFERENCES

[1] Group Jamashii, editor. *Nihongo Bunkei Jiten*. Kuroshio Publisher, 1998. (in Japanese).

[2] A. Kilgarriff and M. Palmer. Introduction to the special issue on SENSEVAL. *Computers and Humanities*, 34:1–13, 2000.

[3] S. Kurohashi and K. Uchimoto. SENSEVAL-2 Japanese Translation Task. *Journal of Natural Language Processing*, 10(3):25–37, 2003.

[4] S. Matsuyoshi and S. Sato. Automatic paraphrasing of Japanese functional expressions using a hierarchically organized dictionary. In *Proc. 3rd IJCNLP*, pages 691–696, 2008.

[5] S. Matsuyoshi, S. Sato, and T. Utsuro. Compilation of a dictionary of Japanese functional expressions with hierarchical organization. In *Proc. ICCPOL*, LNAI: Vol. 4285, pages 395–402. Springer, 2006.

[6] Y. Morita and M. Matsuki. *Nihongo Hyougen Bunkei*, volume 5 of *NAFL Sensho*. ALC, 1989. (in Japanese).

[7] National Language Research Institute. *Gendaigo Hukugouji Youreishu*. 2001. (in Japanese).

[8] J. Nivre and J. Nilsson. Multiword units in syntactic parsing. In *Proc. LREC Workshop, Methodologies and Evaluation of Multiword Units in Real-World Applications*, pages 39–46, 2004.

[9] I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proc. 3rd CICLING*, pages 1–15, 2002.

[10] K. Shudo et al. MWEs as non-propositional content indicators. In *Proc. 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pages 32–39, 2004.

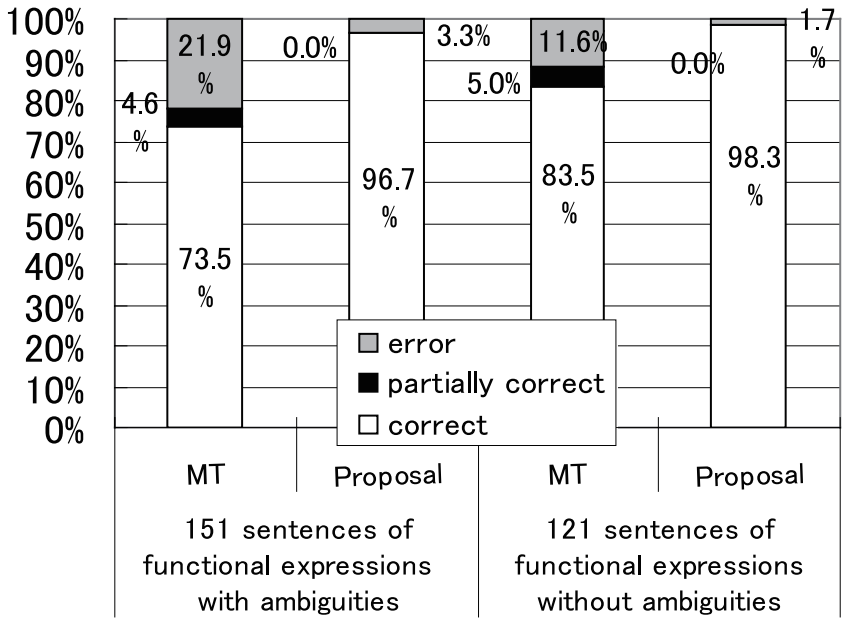

Figure 4: Evaluation Results

[11] M. Tsuchiya, T. Shime, T. Takagi, T. Utsuro, K. Uchimoto, S. Matsuyoshi, S. Sato, and S. Nakagawa. Chunking Japanese compound functional expressions by machine learning. In *Proc. Workshop on Multi-Word-Expressions in a Multilingual Context*, pages 25–32, 2006.

[12] M. Tsuchiya, T. Utsuro, S. Matsuyoshi, S. Sato, and S. Nakagawa. A corpus for classifying usages of Japanese compound functional expressions. In *Proc. PACLING*, pages 345–350, 2005.

[13] T. Utsuro, T. Shime, M. Tsuchiya, S. Matsuyoshi, and S. Sato. Learning dependency relations of Japanese compound functional expressions. In *Proc. Workshop on A Broader Perspective on Multiword Expressions (ACL-2007 Workshop)*, pages 65–72, 2007.

[14] K. Yamamoto. Machine translation by interaction between paraphraser. In *Proc. 19th COLING*, pages 1107–1113, 2002.

Table 3: Example Sentences for Creating Translation Rules and their Evaluation

Semantic Equivalence Class	Example Sentences	
M11 (unnecessary)	for creating the translation rule	(6) この欄には何も書か ないでもよい 。 (7) その程度の用事ならわざわざ出向く までもない 。電話で十分だ。
	for evaluation	(8) 被災直後の救援活動がうまくいかなかったことは、いまさら論じる までもない ことです。
	proposed method: translation = *"not have to"*, judgement = "correct"	(9) Now, we do *not have to* criticize emergent rescue immediately after the strike did not work well.
	translation by MT, judgement = "error"	(10) The emergency relief assistance immediately after suffering a calamity not having worked is now not discussing.
o21 (simultaneous)	for creating the translation rule	(11) 山田さんたら、来 たと思ったら すぐ帰っちゃった。 (12) 夜があける とまもなく 小鳥たちが鳴き始める。
	for evaluation	(13) つめたい雨が降ってき たと思う間もなく 、それは雪にかわった。
	proposed method: translation = *"just after"*, judgement = "correct"	(14) *Just after* cold rain started to fall, it changed into snow.
	translation by MT, judgement = "error"	(15) Also while thinking that it has rained coldly, there is nothing, and it changed to snow.
b11 (about)	for creating the translation rule	(16) その事件 に関して 学校から報告があった。 (17) 農村の生活様式 について 調べている。
	for evaluation	(18) 入力ラベルが得られる毎に、各単語 に関する スコアが更新される。
	proposed method: translation = *"about"*, judgement = "error"	(19) For every label input, the score *about* each word is updated.
	translation by MT, judgement = "error"	(20) an input label – obtaining – ら – れる – every – the score about each word is updated.