Identifying and Utilizing the Class of Monosemous Japanese Functional Expressions in Machine Translation

Akiko Sakamoto[a], Taiji Nagasaka[a], Takehito Utsuro[a], and Suguru Matsuyoshi[b]

[a] Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, JAPAN
[b] Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, 630-0192, JAPAN

Abstract. In the "Sandglass" machine translation architecture, we identify the class of monosemous Japanese functional expressions and utilize it in 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 then study whether functional expressions within a class can be translated into a single canonical English expression. Next, we introduce two types of ambiguities of functional expressions and identify monosemous functional expressions. In the evaluation of our translation rules for Japanese functional expressions, we directly apply those rules to monosemous functional expressions, and show that the proposed framework outperforms commercial machine translation software products. We further study how to extract rules for translating functional expressions in Japanese patent documents into English. In the result of this study, we show that translation rules manually developed based on the corpus for Japanese language grammar learners is reliable also in the patent domain.

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 Matsuyoshi et al. (2006). Based on those recent development in studies on lexicon for processing Japanese functional expressions (Matsuyoshi et al., 2006), 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 (Yamamoto, 2002) [1]. 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 (Matsuyoshi et al., 2006). We employ the semantic equivalence classes of the lexicon and examine each class whether it is monosemous or not. We then study whether functional expressions within a class can be translated into a single

[1] A similar idea was proposed also in Shirai et al. (1993).

canonical English expression. Next, we introduce two types of ambiguities of functional expressions and identify monosemous functional expressions. In the evaluation of our translation rules for Japanese functional expressions, we directly apply those rules to monosemous functional expressions, and show that the proposed framework outperforms commercial machine translation software products. We further study how to extract rules for translating functional expressions in Japanese patent documents into English. In the result of this study, we show that translation rules manually developed based on the corpus for Japanese language grammar learners is reliable also in the patent domain.

2 Japanese Functional Expressions

Even before Matsuyoshi et al. (2006) 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, Morita and Matsuki (1989) examined 450 functional expressions and Group Jamashii (1998) also listed 965 expressions and their example sentences. Compared with those two collections, *Gendaigo Hukugouji Youreishu* (National Language Research Institute, 2001) concentrated on 125 major functional expressions which have non-compositional usages, as well as their variants (337 expressions in total), and collected example sentences of those expressions. For each of the 337 expressions, Tsuchiya et al. (2005) 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 newspaper text corpus. For each of the 337 expressions, 50 sentences were collected and labels for chunking were annotated.

3 Hierarchical Lexicon of Japanese Functional Expressions

In order to organize Japanese functional expressions with various surface forms, Matsuyoshi et al. (2006) proposed a methodology for compiling a lexicon of Japanese functional expressions with hierarchical organization[2]. Matsuyoshi et al. (2006) compiled the lexicon with 341 headwords and 16,801 surface forms. The hierarchy of the lexicon has nine abstraction levels. 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. 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.

Along with the hierarchy of surface forms of functional expressions with nine abstraction levels, the lexicon compiled by Matsuyoshi et al. (2006) 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 1 shows examples of the bottom 199 classes, where each of “k11”, “D21”, “t32”, and “t22” represents a label of the bottom 199 classes. In Matsuyoshi and Sato (2008), 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.

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

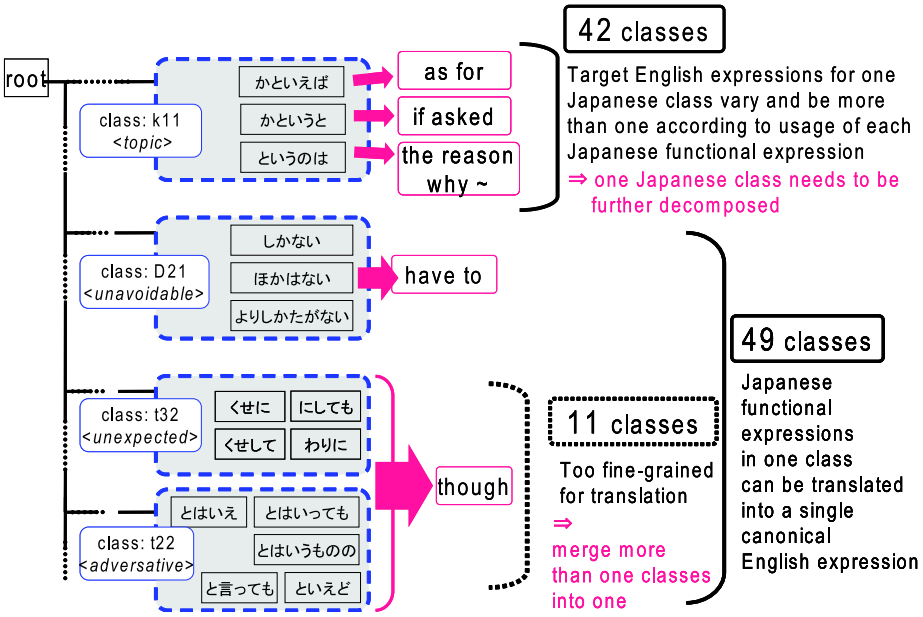

Figure 1: Translation of Japanese Functional Expressions through Semantic Equivalence Classes

4 Ambiguities of Functional/Content Usages

One of the most important assumption of applying the translation rules invented in this paper 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 and the next section overview two types of ambiguities of *functional* expressions (in a broad sense).

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 “to ha ie”, which consists of a post-positional particle “to”, a topic-marking particle “ha“, and a conjugated form “ie” of a verb “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 “to”, the topic-marking particle “ha“, and the verb “ie”, and has a content word meaning “*can not say*”. Compared to Table 1 (b), Table 1 (a) shows an example of a functional expression without ambiguity of functional/content usages. In this case, the compound expression “koto ga dekiru” consists of a formal noun “koto”, a post-positional particle “ga”, and an auxiliary verb “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 Tsuchiya et al. (2005) and Tsuchiya et al. (2006). Tsuchiya et al. (2005) 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, Tsuchiya et al. (2006) 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. Considering this situation, we conclude that we should avoid expressions which have this type of ambiguity when evaluating our translation rules.

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)	koto ga dekiru	Kare ha eigo wo hanasu koto ga dekiru. (He *can* speak English.)	functional, semantic class = *possible* (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)	to-ha-ie	Jokyo ha kaizen shite iru to ha ie , mada anshin deki nai. (*Although* it has become better, we can not feel easy.)	functional, semantic class = *adversative* ($\sim$ to ha ie = *although* $\sim$)
(3)	to ha ie	Jyokyo ga kaizen shita to ha ie nai. (We *can not say* that it has become better.)	content ($\sim$ to ha ie (nai) = *can not say* $\sim$)

(c) *with* ambiguity of functional usages

	Expression	Example sentence (English translation)	Usage
(4)	tame ni	Sekai heiwa no tame ni kokusai kaigi ga hiraka reru. (An international conference is held *for the purpose of* world peace.)	functional, semantic class = *purpose* (tame ni = *for the purpose of*)
(5)	tame ni	Ame no tame ni kare no touchaku ga okureta. (He arrived late *because of* rain.)	functional, semantic class = *reason* (tame ni = *because of*)

5 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 "tame ni", which consists of a noun "tame" and a post-positional particle "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 "koto ga dekiru" has only one non-compositional functional meaning "*can*". 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, we conclude that we should avoid expressions which have this type of ambiguity when evaluating our translation rules.

6 Monosemous Semantic Equivalence Classes of Functional Expressions in Translation

Next, in terms of translation in English, we identify monosemous semantic equivalence classes of Japanese functional expressions. We examine the effects of the bottom 199 semantic equivalence classes in MT. We empirically study whether functional expressions within a class can be translated into a single canonical English 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 (Group Jamashii, 1998) 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. Then, for each of the 91 classes, English translation of the Japanese functional expressions in the collected five sentences are compared. 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 1, 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.

7 Identifying Monosemous 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			

This section presents how we identified monosemous functional expressions which do not have either ambiguities of the two types introduced in sections 4 and 5. 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 6.

As shown in Table 2, first, 166 L^2 entries as well as 6379 L^9 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, 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 2752 L^9 entries are identified as monosemous functional expressions.

8 Evaluation of Translation Rules

For each of the 49 "single translation" classes identified in section 6, 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 Group Jamashii (1998). We evaluate the English translation of Japanese functional expression into 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[3], we divide the 272 sentences for evaluation into 121 sentences which include monosemous functional expressions identified in section 7 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 Table 3, 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.

Table 3: Evaluation Results for 121 Sentences of Functional Expressions *without* Usage/Sense Ambiguities (correct / partially correct / error (%))

MT	proposed
83.5 / 5.0 / 11.6	**98.3** / 0.0 / 1.7

9 Extracting Translation Rules from Parallel Patent Sentences

In this paper, we further study how to extract rules for translating functional expressions in Japanese patent documents into English. In this study, we use about 1.8M Japanese-English parallel sentences automatically extracted from Japanese-English patent families, which are distributed through the Patent Translation Task at the NTCIR-7 Workshop (Fujii et al., 2008). Then, as a toolkit of a phrase-based SMT (Statistical Machine Translation) model, Moses (Koehn et al., 2007) is applied and Japanese-English translation pairs are obtained in the form of a phrase translation table. Finally, we extract translation pairs of Japanese functional expressions from the phrase translation table.

Out of the 49 "single translation" classes, with the lower bound of the phrase translation probability as 0.05 and that of the phrase translation frequency as 10, we extract translation rules for 29 semantic equivalence classes. Within this 29 semantic equivalence classes, we actually extract translation pairs for 72 Japanese functional expressions, where the number of extracted translation pairs for those 72 expressions is 133. Here, it is quite important to note that, in the parallel patent sentences, three semantic equivalence classes out of the 29 are not actually "single translation"

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

classes. To put it another way, 26 classes out of the 29 are actually "single translation" classes in the parallel patent sentences. This means that the result of the procedure in section 6 based on the corpus for Japanese language grammar learners (Group Jamashii, 1998) is reliable also in the patent domain to the extent that 26 out of the 29 "single translation" classes are actually with *single translation into English*. For each of the three "multiple translations" classes, the following lists its sense description as well as multiple translations into English.

- In the class with a label "n12" with the sense of "addition":
  - A Japanese functional expression "ue" is translated into an English preposition / conjunction "after".
  - Another Japanese functional expression "dake-de-naku" is translated into an English conjunctive phrase "not only".
- In the class with a label "m21" with the sense of "restriction":
  - A Japanese functional expression "hoka" is translated into an English prepositional phrase "in addition to".
  - Another Japanese functional expression "igai" is translated into an English preposition "except".
- In the class with a label "P21" with the sense of "exemplification - extreme case":
  - A Japanese functional expression "sae" is translated into an English conjunctive phrase "if only".
  - Another Japanese functional expression "demo" is translated into an English adverb "even".

10 Concluding Remarks

In the "Sandglass" MT architecture (Yamamoto, 2002), we identified the class of monosemous Japanese functional expressions and utilized it in 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 then studied whether functional expressions within a class can be translated into a single canonical English expression. Next, we introduced two types of ambiguities of functional expressions and identified monosemous functional expressions. In the evaluation of our translation rules for Japanese functional expressions, we directly applied those rules to monosemous functional expressions, and showed that the proposed framework outperforms commercial machine translation software products. We further studied how to extract rules for translating functional expressions in Japanese patent documents into English. In the result of this study, we showed that translation rules manually developed based on the corpus for Japanese language grammar learners is reliable also in the patent domain. Future work includes scaling up the procedure of empirical examination on discovering "single translation" semantic equivalence classes into the whole 199 classes.

References

Fujii, A., M. Utiyama, M. Yamamoto and T. Utsuro. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and and Cross-Lingual Information Access*, 389–400.

Group Jamashii., ed. 1998. *Nihongo Bunkei Jiten*. Kuroshio Publisher. (in Japanese).

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 177–180.

Matsuyoshi, S. and S. Sato. 2008. Automatic paraphrasing of Japanese functional expressions using a hierarchically organized dictionary. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, 691–696.

Matsuyoshi, S., S. Sato and T. Utsuro. 2006. Compilation of a dictionary of Japanese functional expressions with hierarchical organization. In Y. Matsumoto, R. Sproat, K.-F. Wong and M. Zhang, eds., *Computer Processing of Oriental Languages: Beyond the Orient: The Research Challenges Ahead*, Lecture Notes in Artificial Intelligence: Vol. 4285. Springer. 395–402.

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

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

Shirai, S., S. Ikehara and T. Kawaoka. 1993. Effects of automatic rewriting of source language within a Japanese to English MT system. In *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation*, 226–239.

Tsuchiya, M., T. Utsuro, S. Matsuyoshi, S. Sato and S. Nakagawa. 2005. A corpus for classifying usages of Japanese compound functional expressions. In *Proceedings of the Pacific Association for Computational Linguistics*, 345–350.

Tsuchiya, M., T. Shime, T. Takagi, T. Utsuro, K. Uchimoto, S. Matsuyoshi, S. Sato and S. Nakagawa. 2006. Chunking Japanese compound functional expressions by machine learning. In *Proceedings of the Workshop on Multi-Word-Expressions in a Multilingual Context (EACL (European Chapter of the Association for Computational Linguistics)-2006 Workshop)*, 25–32.

Yamamoto, K. 2002. Machine translation by interaction between paraphraser. In *Proceedings of the 19th International Conference on Computational Linguistics*, 1107–1113.