Automatic Paraphrasing of Japanese Functional Expressions Using a Hierarchically Organized Dictionary

Suguru Matsuyoshi[†,‡] **Satoshi Sato**[‡]
[†] Graduate School of Informatics, Kyoto University, Japan
[‡] Graduate School of Engineering, Nagoya University, Japan
{s_matuyo,ssato}@nuee.nagoya-u.ac.jp

Abstract

Automatic paraphrasing is a transformation of expressions into semantically equivalent expressions within one language. For generating a wider variety of phrasal paraphrases in Japanese, it is necessary to paraphrase functional expressions as well as content expressions. We propose a method of paraphrasing of Japanese functional expressions using a dictionary with two hierarchies: a morphological hierarchy and a semantic hierarchy. Our system generates appropriate alternative expressions for 79% of source phrases in Japanese in an open test. It also accepts style and readability specifications.

1 Introduction

Automatic paraphrasing is a transformation of expressions into semantically equivalent expressions within one language. It is expected for various applications, such as information retrieval, machine translation and a reading/writing aid.

Automatic paraphrasing of Japanese text has been studied by many researchers after the first international workshop on automatic paraphrasing (Sato and Nakagawa, 2001). Most of them focus on paraphrasing of content words, such as noun phrases and verb phrases. In contrast, paraphrasing of *functional expressions* has less attention. A functional expression is a function word or a multi-word expression that works as a function word. For generating a wider variety of phrasal paraphrases in Japanese, as shown in Fig. 1, it is necessary to paraphrase functional expressions as well as content expressions, because almost all phrases in Japanese include one or more functional expressions. In this paper, we focus on paraphrasing of Japanese functional expressions.

In several applications, such as a reading aid, in paraphrasing of Japanese functional expressions, control of readability of generated text is important, because functional expressions are critical units that determine sentence structures and meanings. In case a reader does not know a functional expression, she fails to understand the sentence meaning. If the functional expression can be paraphrased into an easier one, she may know it and understand the sentence meaning. It is desirable to generate expressions with readability suitable for a reader because easier functional expressions tend to have more than one meaning.

A remarkable characteristic of Japanese functional expressions is that each functional expression has many different variants. Each variant has one of four styles. In paraphrasing of Japanese functional expressions, a paraphrasing system should accept style specification, because consistent use in style is required. For example, the paraphrase (b) in Fig. 1 is not appropriate for a document in normal style because the expression has polite style.

Paraphrasing a functional expression into a semantically equivalent one that satisfies style and readability specifications can be realized as a combination of the following two processes:

1. Transforming a functional expression into another one that is semantically equivalent to it, often with changing readability.

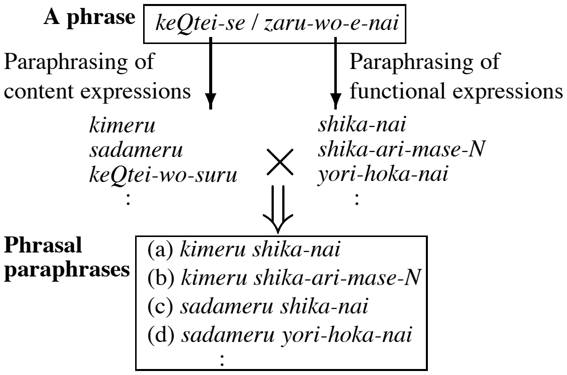
A phrase
keQtei-se / zaru-wo-e-nai
Paraphrasing of content expressions
Paraphrasing of functional expressions
kimeru
sadameru
keQtei-wo-suru
:
shika-nai
shika-ari-mase-N
yori-hoka-nai
:
×
⇓
Phrasal paraphrases
(a) kimeru shika-nai
(b) kimeru shika-ari-mase-N
(c) sadameru shika-nai
(d) sadameru yori-hoka-nai
:

Figure 1: Generation of a wider variety of phrasal paraphrases.

2. Rewriting a functional expression to a variant of it, often with changing style.

We propose a method of paraphrasing of Japanese functional expressions using a dictionary with two hierarchies: a *morphological hierarchy* and a *semantic hierarchy*. The former hierarchy provides a list of all variants specified with style for each functional expression, which is required for the above process 2. The latter hierarchy provides semantic equivalence classes of functional expressions and readability level for each functional expression, which are required for the above process 1.

2 Related Work

A few studies on paraphrasing of Japanese functional expressions have been conducted. In order to implement automatic paraphrasing, some studies (Iida et al., 2001; Tsuchiya et al., 2004) use a set of paraphrasing rules, and others (Tanabe et al., 2001; Shudo et al., 2004) use semantic equivalence classes.

All of these studies do not handle variants in a systematic way. In case a system paraphrases a functional expression f into f', it also should generate all variants of f' in potential. However, any proposed system does not guarantee this requirement. Output selection of variants should be determined according to the given style specification. Any proposed system does not have such selection mechanism.

Controlling readability of generated text is not a central issue in previous studies. An exception is a study by Tsuchiya et al. (Tsuchiya et al., 2004). Their system paraphrases a functional expression into an easier one. However, it does not accept the readability specification, e.g. for learners of beginner course or intermediate course of Japanese.

Level		Num
L^1	Headword	341
L^2	**Headwords with unique meaning**	**435**
L^3	Derivations	555
L^4	Alternations of function words	774
L^5	Phonetic variations	1,187
L^6	Insertion of particles	1,810
L^7	Conjugation forms	6,870
L^8	Normal or *desu*/*masu* forms	9,722
L^9	Spelling variations	16,801

Table 1: Nine levels of the morphological hierarchy.

3 A Hierarchically Organized Dictionary of Japanese Functional Expressions

3.1 Morphological hierarchy

In order to organize many different variants of functional expressions, we have designed a morphological hierarchy with nine abstraction levels (Matsuyoshi et al., 2006). Table 1 summarizes these nine levels. The number of entries in L^1 (headwords) is 341, and the number of leaf nodes in L^9 (surface forms) is 16,801. For each surface form in the hierarchy, we specified one of four styles (normal, polite, colloquial, and stiff) and connectability (what word can be to the left and right of the expression).

3.2 Semantic hierarchy

There is no available set of semantic equivalence classes of Japanese functional expressions for paraphrasing. Some sets are described in books in linguistics (Morita and Matsuki, 1989; Tomomatsu et al., 1996; Endoh et al., 2003), but these are not for paraphrasing. Others are proposed for paraphrasing in natural language processing (Tanabe et al., 2001; Shudo et al., 2004), but these are not available in public.

For 435 entries in L^2 (headwords with unique meaning) of the morphological hierarchy, from the viewpoint of paraphrasability, we have designed a semantic hierarchy with three levels according to the semantic hierarchy proposed by a book (Morita and Matsuki, 1989). The numbers of classes in the top, middle and bottom levels are 45, 128 and 199, respectively. For each entry in L^2, we specified one of readability levels of A1, A2, B, C, and F according to proficiency level in a book (Foundation and of International Education, Japan, 2002), where A1 is the most basic level and F is the most advanced level.

3.3 Producing all surface forms that satisfy style and readability specifications

For a given surface form of a functional expression, our dictionary can produce all variants of semantically equivalent functional expressions that satisfy style and readability specifications. The procedure is as follows:

1. Find the functional expression in L^2 for a given surface form according to the morphological hierarchy.
2. Obtain functional expressions that are semantically equivalent to the functional expression according to the semantic hierarchy.
3. Exclude the functional expressions that do not satisfy readability specification.
4. Enumerate all variants (surface forms) of the remaining functional expressions according to the morphological hierarchy.
5. Exclude the surface forms that do not satisfy style specification.

4 Formulation of Paraphrasing of Japanese Functional Expressions

As a source expression of paraphrasing, we select a phrase (or *Bunsetsu*) in Japanese because it is a base unit that includes functional expressions. In this paper, we define a phrase as follows. Let c_i be a content word, and f_j a functional expression. Then, a phrase is formulated as the following:

$$\text{Phrase} = c_1 c_2 \cdots c_m f_1 f_2 \cdots f_n, \qquad (1)$$

where $c_1 c_2 \cdots c_m$ is the content part of the phrase and $f_1 f_2 \cdots f_n$ is the functional part of it.

Paraphrasing of a functional part of a phrase is performed as a combination of the following five types of paraphrasing:

1→1 Substituting a functional expression with another functional expression ($f \rightarrow f'$).

Paraphrasing type	Num
1→1 only	214 (61%)
1→N (and 1→1)	69 (20%)
N→1 (and 1→1)	18 (5%)
M→N (and 1→1)	8 (2%)
Otherwise	44 (12%)
Sum	353 (100%)

Table 2: Number of paraphrases produced by a native speaker of Japanese.

1→N Substituting a functional expression with a sequence of functional expressions ($f \rightarrow f'_1 f'_2 \cdots f'_N$).

N→1 Substituting a sequence of functional expressions with one functional expression ($f_1 f_2 \cdots f_N \rightarrow f'$).

M→N Substituting a sequence of functional expressions with another sequence of functional expressions ($f_1 f_2 \cdots f_M \rightarrow f'_1 f'_2 \cdots f'_N$).

f⇒c Substituting a functional expression with an expression including one or more content words.

In a preliminary experiment, we investigated which type of the above a native speaker of Japanese tended to use in paraphrasing a functional part. Table 2 shows the classification result of 353 paraphrases produced by the subject for 238 source phrases.[1] From this table, it was found out that paraphrasing of "1→1" type was major in that it was used for producing 61% of paraphrases.

Because of dominance of paraphrasing of "1→1" type, we construct a system that paraphrases Japanese functional expressions in a phrase by substituting a functional expression with a semantically equivalent expression. This system paraphrases a phrase defined as the form in Eq. (1) into the following form:

$$\text{Alternative} = c_1 c_2 \cdots c_{m-1} c'_m w f'_1 f'_2 \cdots f'_n,$$

where c'_m is c_m or a conjugation form of c_m, f'_j is a functional expression that is semantically equivalent to f_j, and w is a null string or a function word that is inserted for connecting f'_1 to c'_m properly.

[1] These source phrases are the same ones that we use in a closed test in section 6.

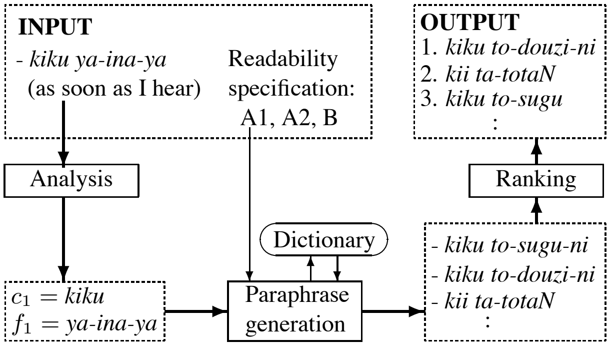

Figure 2: Overview of our system.

The combination of simple substitution of a functional expression and insertion of a function word covers 22% (15/69) of the paraphrases by paraphrasing of "1→N (and 1→1)" type in Table 2. Therefore, our system theoretically covers 65% (229/353) of the paraphrases in Table 2.

5 System

We have implemented a system that paraphrases Japanese functional expressions using a hierarchically organized dictionary, by substituting a functional expression with another functional expression that is semantically equivalent to it. The system accepts a phrase in Japanese and generates a list of ranked alternative expressions for it. The system also accepts style and readability specifications.

Fig. 2 shows an overview of our system. This system consists of three modules: analysis, paraphrase generation, and ranking.

5.1 Analysis

Some methods have been proposed for detecting Japanese functional expressions based on a set of detection rules (Tsuchiya and Sato, 2003) and machine learning (Uchimoto et al., 2003; Tsuchiya et al., 2006). However, because these methods detect only a limited number of functional expressions (and their variants), we cannot apply them to the analysis of a phrase. Another method is to add a list of about 17,000 surface forms of functional expressions to a dictionary of an existing morphological analyzer and determine connecting costs based on machine learning. However, it is infeasible because there is no large corpus in which all of these surface forms have been tagged.

Instead of these methods, we use a different method of decomposing a given phrase into a sequence of content words and functional expressions. Our method uses two analyzers.

We constructed a functional-part analyzer (FPA). This is implemented using a morphological analyzer MeCab[2] with a special dictionary containing only functional expressions. FPA can decompose a functional part (string) into a sequence of functional expressions, but fails to decompose a string when the string includes one or more content words. In order to extract a functional part from a given string, we use original MeCab.

First, original MeCab decomposes a given string into a sequence of morphemes $m_1 m_2 \cdots m_k$. Next, we suppose that m_1 is a content part and $m_2 m_3 \cdots m_k$ is a functional part. If FPA can decompose $m_2 m_3 \cdots m_k$ into a sequence of functional expressions $f_1 f_2 \cdots f_n$, then we obtain $c_1 f_1 f_2 \cdots f_n$ as shown in Eq. (1) as an analyzed result, where $c_1 = m_1$. Otherwise, we suppose that $m_1 m_2$ is a content part and $m_3 m_4 \cdots m_k$ is a functional part. If FPA can decompose $m_3 m_4 \cdots m_k$ into a sequence of functional expressions $f_1 f_2 \cdots f_n$, then we obtain $c_1 c_2 f_1 f_2 \cdots f_n$ as an analyzed result, where $c_1 = m_1$ and $c_2 = m_2$. This procedure is continued until FPA succeeds in decomposition.

5.2 Paraphrase generation

This module accepts an analyzed result $c_1 c_2 \cdots c_m f_1 f_2 \cdots f_n$ and generates a list of alternative expressions for it.

First, the module obtains a surface form f_1' that is semantically equivalent to f_1 from the dictionary in section 3. Next, it constructs $c_1 c_2 \cdots c_{m-1} c_m' w f_1'$ by connecting f_1' to $c_1 c_2 \cdots c_m$ by the method described in section 4. Then, it obtains a surface form f_2' that is semantically equivalent to f_2 and constructs $c_1 c_2 \cdots c_{m-1} c_m' w f_1' f_2'$ in similar fashion. This process proceeds analogously, and finally, the module constructs $c_1 c_2 \cdots c_{m-1} c_m' w f_1' f_2' \cdots f_n'$ as an alternative expression.

Because in practice the module obtains more than one surface form that is semantically equivalent to

[2] http://mecab.sourceforge.net/

	Top 1	Top 1 to 2	Top 1 to 3	Top 1 to 4	Top 1 to 5
Closed	177 (74%)	197 (83%)	**210 (88%)**	213 (90%)	213 (90%)
Closed (Perfect analysis)	196 (82%)	211 (89%)	**219 (92%)**	221 (93%)	221 (93%)
Open	393 (63%)	461 (73%)	**496 (79%)**	500 (80%)	501 (80%)
Open (Perfect analysis)	453 (72%)	508 (81%)	**531 (85%)**	534 (85%)	534 (85%)

Table 3: Evaluation of paraphrases generated by the paraphrasing system

f_j by the method described in subsection 3.3, it generates more than one alternative expression by considering all possible combinations of these surface forms and excluding candidates that include two adjacent components that cannot be connected properly.

If the module generates no alternative expression, it uses the semantic equivalence classes in the upper level reluctantly.

5.3 Ranking

Because a functional expression seems to be more standard and common as it appears more frequently in newspaper corpus, we use frequencies of functional expressions (strings) in newspaper corpus in order to rank alternative expressions. We define a scoring function as the product of frequencies of functional expressions in a phrase.

6 Evaluation

We evaluate paraphrases generated by our paraphrasing system for validating our semantic equivalence classes, because the dictionary that the system uses guarantees by the method described in subsection 3.3 that the system can generate all variants of a functional expression and accept style and readability specifications.

6.1 Methodology

We evaluated paraphrases generated by our paraphrasing system from the viewpoint of an application to a writing aid, where a paraphrasing system is expected to output a few good alternative expressions for a source phrase.

We evaluated the top 5 alternative expressions generated by the system for a source phrase by classifying them into the following three classes:

Good Good alternative expression for the source phrase.

Intermediate Expression that keeps the meaning roughly that the source phrase has.

Bad Inappropriate expression.

Then, we counted source phrases for which at least one of the alternative expressions of the top 1 to n was judged as “Good”. One of the authors performed the judgment according to books (Morita and Matsuki, 1989; Endoh et al., 2003).

As a closed test set, we used 238 example phrases for 140 functional expressions extracted from a book (Foundation and of International Education, Japan, 2002), which we had used for development of our semantic equivalence classes. As an open test set, we used 628 example phrases for 184 functional expressions extracted from a book (Tomomatsu et al., 1996). We used the Mainichi newspaper text corpus (1991-2005, about 21 million sentences, about 1.5 gigabytes) for ranking alternative expressions.

6.2 Results

Table 3 shows the results. The rows with “Perfect analysis” in the table show the results in analyzing source phrases by hand. Because the values in every row of the table are nearly saturated in “Top 1 to 3”, we discuss the results of the top 1 to 3 hereafter.

Our system generated appropriate alternative expressions for 88% (210/238) and 79% (496/628) of source phrases in the closed and the open test sets, respectively. We think that this performance is high enough.

We analyzed the errors made by the system. In the closed and the open tests, it was found out that paraphrasing of “1→1” type could not generate alternative expressions for 7% (16/238) and 7% (41/628) of source phrases, respectively. These values define the upper limit of our system.

In the closed and the open tests, it was found out that the system failed to analyze 3% (8/238) and 3% (21/628) of source phrases, respectively, and that ambiguity in meaning caused inappropriate candidates to be ranked higher for 1% (2/238) and 4% (23/628) of source phrases, respectively. The rows with "Perfect analysis" in Table 3 show that almost all of these problems are solved in analyzing source phrases by hand. Improvement of the analysis module can solve these problems.

In the open test, insufficiency of semantic equivalence classes and too rigid connectability caused only 3% (19/628) and 3% (16/628) of source phrases to have no good candidates, respectively. The smallness of the former value validates our semantic equivalence classes.

The remaining errors were due to low frequencies of good alternatives in newspaper corpus.

7 Conclusion and Future Work

We proposed a method of paraphrasing Japanese functional expressions using a dictionary with two hierarchies. Our system can generate all variants of a functional expression and accept style and readability specifications. The system generated appropriate alternative expressions for 79% of source phrases in an open test.

Tanabe et al. have proposed paraphrasing rules of "1→N", "N→1", and "M→N" types (Tanabe et al., 2001). For generating a wider variety of phrasal paraphrases, future work is to incorporate these rules into our system and to combine several methods of paraphrasing of content expressions with our method.

References

Orie Endoh, Kenji Kobayashi, Akiko Mitsui, Shinjiro Muraki, and Yasushi Yoshizawa, editors. 2003. *A Dictionary of Synonyms in Japanese (New Edition)*. Shogakukan. (in Japanese).

The Japan Foundation and Association of International Education, Japan, editors. 2002. *Japanese Language Proficiency Test: Test Content Specifications (Revised Edition)*. Bonjinsha. (in Japanese).

Ryu Iida, Yasuhiro Tokunaga, Kentaro Inui, and Junji Etoh. 2001. Exploration of clause-structural and function-expressional paraphrasing using KURA. In *Proceedings of the 63rd National Convention of Information Processing Society of Japan*, volume 2, pages 5–6. (in Japanese).

Suguru Matsuyoshi, Satoshi Sato, and Takehito Utsuro. 2006. Compilation of a dictionary of Japanese functional expressions with hierarchical organization. In *Proceedings of the 21st International Conference on Computer Processing of Oriental Languages (ICCPOL), Lecture Notes in Computer Science*, volume 4285, pages 395–402. Springer.

Yoshiyuki Morita and Masae Matsuki. 1989. Nihongo Hyougen Bunkei*, volume 5 of* NAFL Sensho *(Expression Patterns in Japanese)*. ALC Press Inc. (in Japanese).

Satoshi Sato and Hiroshi Nakagawa, editors. 2001. *Automatic Paraphrasing: Theories and Applications*, The 6th Natural Language Processing Pacific Rim Symposium (NLPRS) Post-Conference Workshop.

Kosho Shudo, Toshifumi Tanabe, Masahito Takahashi, and Kenji Yoshimura. 2004. MWEs as nonpropositional content indicators. In *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing (MWE-2004)*, pages 32–39.

Toshifumi Tanabe, Kenji Yoshimura, and Kosho Shudo. 2001. Modality expressions in Japanese and their automatic paraphrasing. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS)*, pages 507–512.

Etsuko Tomomatsu, Jun Miyamoto, and Masako Wakuri. 1996. *500 Essential Japanese Expressions: A Guide to Correct Usage of Key Sentence Patterns.* ALC Press Inc. (in Japanese).

Masatoshi Tsuchiya and Satoshi Sato. 2003. Automatic detection of grammar elements that decrease readability. In *Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pages 189–192.

Masatoshi Tsuchiya, Satoshi Sato, and Takehito Utsuro. 2004. Automatic generation of paraphrasing rules from a collection of pairs of equivalent sentences including functional expressions. In *Proceedings of the 10th Annual Meeting of the Association for Natural Language Processing*, pages 492–495. (in Japanese).

Masatoshi Tsuchiya, Takao Shime, Toshihiro Takagi, Takehito Utsuro, Kiyotaka Uchimoto, Suguru Matsuyoshi, Satoshi Sato, and Seiichi Nakagawa. 2006. Chunking Japanese compound functional expressions by machine learning. In *Proceedings of the workshop on Multi-word-expressions in a multilingual context, EACL 2006 Workshop*, pages 25–32.

Kiyotaka Uchimoto, Chikashi Nobata, Atsushi Yamada, Satoshi Sekine, and Hitoshi Isahara. 2003. Morphological analysis of a large spontaneous speech corpus in Japanese. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 479–488.