Detecting Japanese Compound Functional Expressions using Canonical/Derivational Relation

Takafumi Suzuki† **Yusuke Abe**† **Itsuki Toyota**† **Takehito Utsuro**†
Suguru Matsuyoshi‡ **Masatoshi Tsuchiya**††

†University of Tsukuba, Tsukuba, 305-8573, Japan
‡ University of Yamanashi, 4-3-11, Takeda, Kofu, Yamanashi, 400-8511, Japan
†† Toyohashi University of Technology, Toyohashi, 441-8580, Japan

Abstract

The Japanese language has various types of functional expressions. In order to organize Japanese functional expressions with various surface forms, a lexicon of Japanese functional expressions with hierarchical organization was compiled. This paper proposes how to design the framework of identifying more than 16,000 functional expressions in Japanese texts by utilizing hierarchical organization of the lexicon. In our framework, more than 16,000 functional expressions are roughly divided into canonical / derived functional expressions. Each derived functional expression is intended to be identified by referring to the most similar occurrence of its canonical expression. In our framework, contextual occurrence information of much fewer canonical expressions are expanded into the whole forms of derived expressions, to be utilized when identifying those derived expressions. We also empirically show that the proposed method can correctly identify more than 80% of the functional / content usages only with less than 38,000 training instances of manually identified canonical expressions.

Keywords: Japanese compound functional expressions, hierarchical lexicon, example-based disambiguation

1. Introduction

The Japanese language has many compound functional expressions which consist of more than one word including both content words and function words. Recognition and semantic interpretation of compound functional expressions are especially difficult because it often happens that one compound expression may have both a literal (in other words, compositional) *content word* usage and a non-literal (in other words, non-compositional) *functional* usage.

For example, Table 1 shows two example sentences with a compound expression "て (te) よい (yoi)", which consists of a conjunctive particle "て (te)", and a base form "よい (yoi)" of an adjective "よい (yoi)". In the sentence (1), the compound expression functions as an auxiliary verb and has a non-compositional functional meaning "*may*". On the other hand, in the sentence (2), the expression simply corresponds to a literal concatenation of the usages of the constituents: the conjunctive particle "て (te)" and the adjective "よい (yoi)", and has a content word meaning "*good* (∼) *because* ∼". Therefore, when considering machine translation of those Japanese sentences into English, it is necessary to precisely judge the usage of the compound expression "て (te) よい (yoi)", as shown in the English translation of the two sentences in Table 1.

Considering such a situation, it is necessary to develop a tool which properly recognizes and semantically interprets Japanese compound functional expressions. 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. As recently reported in Matsuyoshi et al. (2006), the Japanese language has a large number of variants of functional expressions, where their total number is counted as over 16,000. So far, it has not been proved that the technique of Tsuchiya et al. (2006) can be applied to the whole list of over 16,000 Japanese functional expressions.

Based on the argument above, this paper proposes how to design the framework of identifying more than 16,000 functional expressions in Japanese texts by utilizing the recently compiled large scale hierarchical lexicon of Japanese functional expressions (Matsuyoshi et al., 2006). In our framework, more than 16,000 functional expressions are roughly divided into about 1,300 canonical functional expressions and the remaining derived functional expressions. Based on a variant of example-based architectures, each derived functional expression is to be identified by referring to the most similar occurrence of its canonical expression included in the example database of manually identified canonical expressions. Contextual occurrence information of much fewer canonical expressions are expanded into the whole forms of derived expressions. We empirically show that the proposed method can correctly identify more than 80% of the functional/content usages only with less than 38,000 instances of manually identified canonical expressions.

2. 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[1]. Matsuyoshi et al. (2006) compiled the lexicon with 341 headwords and 16,801 surface forms. As shown in Table 2,

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

Table 1: Examples of Ambiguity of Functional/Content Usages

	Expression	Example sentence (English translation)	Usage
(1)	てよい (te-yoi)	申し込みが少ないので，価格を下げ てよい . (You *may* discount the price because we have a small number of applicants.)	functional (～ てよい (te-yoi) = *may* ～)
(2)	てよい (te-yoi)	この店は安く てよい 評判だ. (This store has a *good* reputation *because* it sells at low prices.)	content (～ てよい) (te-yoi)) = *good (～) because* ～)

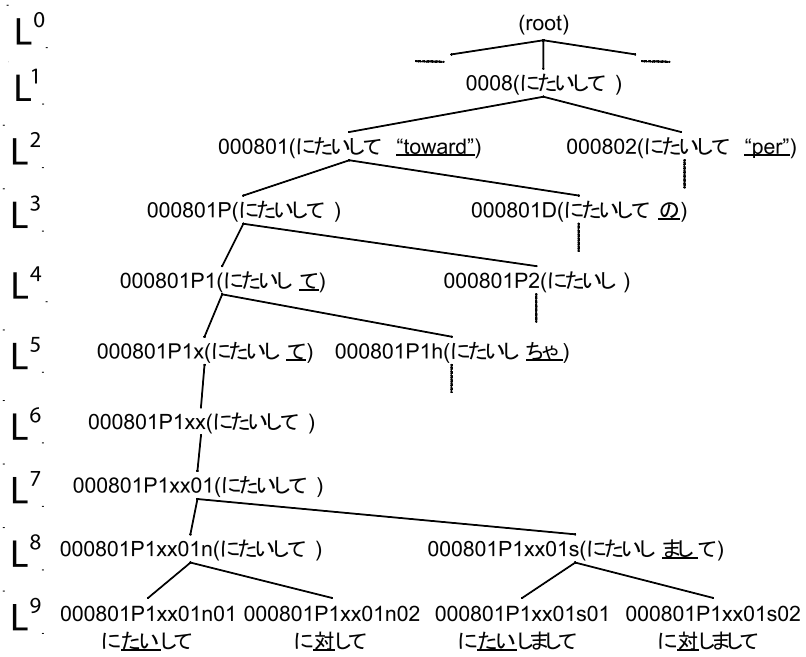

Figure 1: A Part of the Hierarchical Lexicon of Japanese Functional Expressions

the hierarchy of the lexicon has nine abstraction levels and Figure 1 shows a part of the hierarchy[2]. 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. 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]In this lexicon, following Sag et al. (2002), each functional expression is regarded as a fixed expression, rather than a semi-fixed expression or a syntactically-flexible expression.

3. Disambiguation of Functional/Content Usages

3.1. Canonical/Derived Expressions

The underlying motivation of the proposed framework is to divide the whole list of more than 16,000 functional expressions into about 1,300 canonical functional expressions and the remaining derived functional expressions. When automatically identifying an occurrence of derived functional expressions, we refer to the most similar occurrence of its canonical expression included in the example database of manually identified canonical expressions.

In the process of dividing the whole list of more than 16,000 functional expressions into canonical and derived expressions, based on our preliminary analysis, we first select 774 expressions at the level L^4 as the canonical expressions. In this analysis, we discovered that expressions which share an identical L^4 level ID have mostly similar contextual occurrences and distinction of functional/content usages, while those which do not share an identical L^4 level ID have relatively dissimilar contextual occurrences and distinction of functional/content usages. We then further distinguish expressions which have distinct phonetic variations at L^5 level and spelling variations between hiragana and kanji characters at L^9 level as having distinct canonical expressions, re-

Table 2: Nine Abstraction Levels of the Morphological Hierarchy

Abstraction Levels		# of entries
L^1	Headword	341
L^2	Headword with unique meaning	435
L^3	Grammatical functions	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

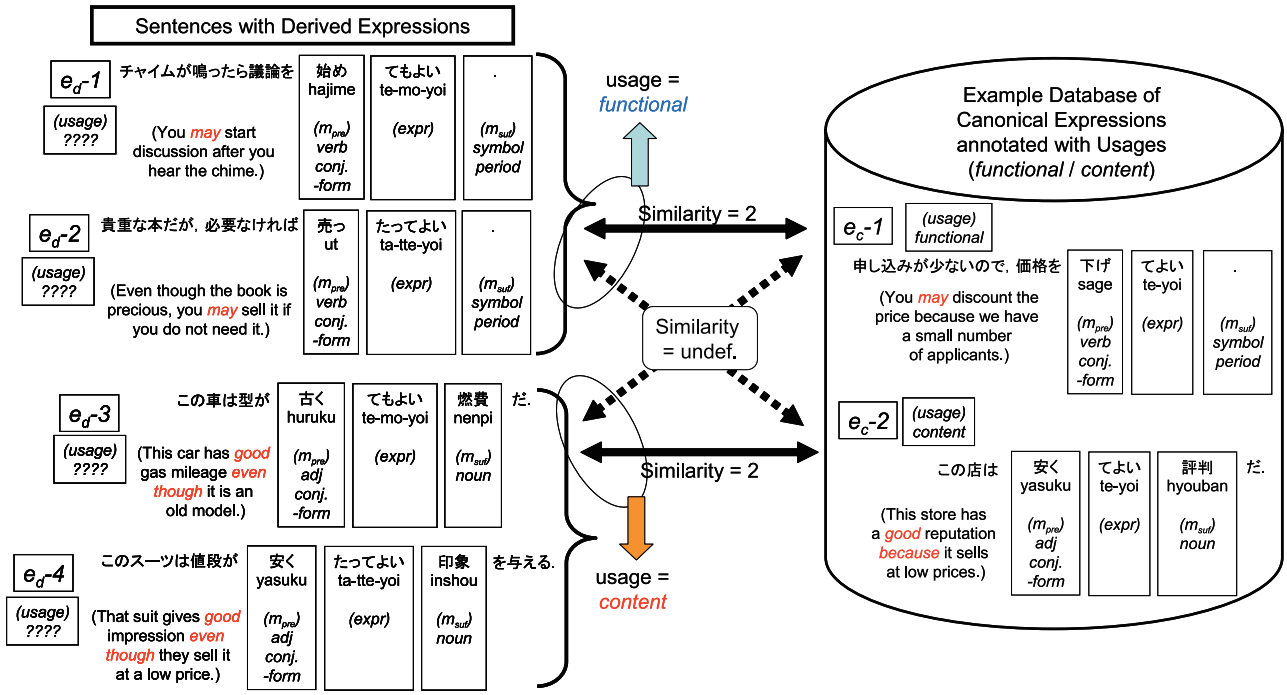

Figure 2: Example-based Disambiguation of Functional/Content Usages using Canonical/Derivational Relation

sulting in 1,302 canonical expressions in total.

For example, in the case of the canonical expression "て (te) よい (yoi)" in Figure 2, it has derived expressions such as "て (te) も (mo) よい (yoi)" having insertion of a particle at L^6 level and "た (ta) って (tte) よい (yoi)" having phonetic variations at L^5 level. In total, "て (te) よい (yoi)" has 36 derived expressions.

3.2. Example-based Disambiguation

j In the proposed framework, we employ an example-based architecture for disambiguation of functional/content usages. In the example-based architecture, an occurrence e_d of a derived expression $expr_d$ is denoted as a tuple $\langle m^d_{pre}, expr_d, m^d_{suf} \rangle$ where m^d_{pre} and m^d_{suf} denote the morpheme preceding the expression $expr_d$ and the one subsequent to $expr_d$. Similarly, an occurrence e_c of a canonical expression $expr_c$ in the example database of manually identified canonical expressions is denoted as a tuple $\langle m^c_{pre}, expr_c, m^c_{suf}, usage \rangle$ where m^c_{pre} and m^c_{suf} denote the morpheme preceding the expression $expr_c$ and the one subsequent to $expr_c$, and $usage$ denotes the manually annotated usage as "*functional*" or "*content*". Similarity of e_c and e_d is defined only when the canonical expression $expr_c$ and the derived expression $expr_d$ satisfy the canonical/derivational relation. Their similarity is defined as 2 when m^c_{pre} and m^d_{pre} share fine-grained parts-of-speech (e.g., case-marking particle) and conjugated forms of the morpheme lexicon IPAdic[3] annotated by the Japanese morphological analyzer MeCab[4], and m^c_{suf} and m^d_{suf} also share fine-grained parts-of-speech and conjugated forms. Otherwise, their similarity is defined as 1 when m^c_{pre} and m^d_{pre} share coarse-grained parts-of-speech (e.g., particle) of IPAdic annotated by the Japanese morphological analyzer MeCab, and m^c_{suf} and m^d_{suf} also share coarse-grained parts-of-speech.

Finally, if examples of both *functional* and *content* usages with the same similarity value are found in the database, we judge the output to be an error. Also, if any example with

[3] http://sourceforge.jp/projects/ipadic/

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

Table 3: Evaluation Results (%)

(a) correct / error rate of the proposed method

correct rate		**82.0**	
error rate	correct if an oracle example is in the example database of canonical expressions	12.7	18.0
	not correct even with an oracle example in the example database of canonical expressions	5.3	

(b) correct rate of the baseline

preferring the longest morpheme sequence and judging as *functional* usage	77.2

the similarity value 1 or 2 is not found in the database, we judge the output to be *content* usage[5].

For example, Figure 2, shows an example of a canonical expression "て (te) よい (yoi)" and its derived expressions "て (te) も (mo) よい (yoi)" and "た (ta) って (tte) よい (yoi)". As shown in Figure 2, in the example e_c-1 of the canonical expression "て (te) よい (yoi)", the canonical expression has the *functional* usage, and functions as an auxiliary verb and has a non-compositional functional meaning "*may*". On the other hand, in the example e_c-2 of the canonical expression "て (te) よい (yoi)", it has the *content* usage, and literally means as "*good* (∼) *because* ∼". Then, both of the examples e_d-1 and e_d-2 of derived expressions "て (te) も (mo) よい (yoi)" and "た (ta) って (tte) よい (yoi)" have similarity values as 2 with the example e_c-1 of the canonical expression "て (te) よい (yoi)", while they do not have similarity values defined against the example e_c-2. Thus, they are judged as having the *functional* usage. Next, in the case of both the examples e_d-3 and e_d-4 of derived expressions "て (te) も (mo) よい (yoi)" and "た (ta) って (tte) よい (yoi)", on the other hand, they have similarity values as 2 with the example e_c-2 while they do not have similarity values defined against the example e_c-1. Thus, they are judged as having the *content* usage.

4. Evaluation

For evaluation, we collect 37,761 example sentences of 496 canonical expressions from the 1995 Mainichi newspaper text corpus and manually annotate the usages of canonical expressions as *functional* or *content*. From the 1995 Mainichi newspaper text corpus, we also collect 2,832 examples of 248 derived expressions for evaluation. Out of the evaluation instances, about 80% are annotated as *functional* usage.

As in Table 3 (a), the proposed method achieved 82.0% correct rate. As with the case of usual example-based methods, the performance of the proposed method depends on the scale of the example database of canonical expressions. If we assume that we add an oracle example of the canonical expression to the database for each of the evaluation instances of derived expressions, we improve 12.7% of the whole evaluation instances, which amount to almost 95% correct rate in total. For the remaining 5.3% of the evaluation instances, examples of canonical expressions with the usage other than the reference one have the similarity value larger than those with reference usage. Table 3 (b) also shows the correct rate of a baseline as 77.2%, where it prefers the longest morpheme sequence and judges the usage of the evaluation instance as *functional*.

[5]More specifically, the task of identifying Japanese compound functional expressions is actually formalized as the task of chunking a morpheme sequence into a *functional* chunk or a *content* chunk. In this formalization, we prefer the longest morpheme sequence which represents a derived expression and satisfies the similarity value 1 or 2 against an example of canonical expression in the example database.

5. Related Works

Ambiguities of functional/content usages has been well studied in Tsuchiya et al. (2005), Tsuchiya et al. (2006), and (Shudo et al., 2004). 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. (Shudo et al., 2004) 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.

Utsuro et al. (2007) and (Nivre and Nilsson, 2004) studied syntactic analysis of functional expressions in sentences. Utsuro et al. (2007) studied how to incorporate the process of analyzing compound non-compositional functional expressions into the framework of Japanese statistical dependency parsing. (Nivre and Nilsson, 2004) also reported improvement of Swedish parsing when multi word units are manually annotated.

In the area of machine translation, Sakamoto et al. (2009) and Nagasaka et al. (2010) applied the "Sandglass" machine translation architecture (Yamamoto, 2002) to the task of translating Japanese functional expressions into English. Unlike Sakamoto et al. (2009) and Nagasaka et al. (2010), in order to address the issue of resolving various ambiguities of a compound expression in machine translation of Japanese functional expressions, Abe et al. (2011) took the approach of example-based machine translation (Sommers, 2003).

6. Concluding Remarks

We design the framework of identifying more than 16,000 functional expressions in Japanese texts by utilizing the large scale hierarchical lexicon of Japanese functional expressions. In our framework, each derived functional expression is to be identified by referring to the most similar occurrence of its canonical expression.

7. References

Y. Abe, T. Suzuki, B. Liang, T. Utsuro, M. Yamamoto, S. Matsuyoshi, and Y. Kawada. 2011. Example-based translation of Japanese functional expressions utilizing semantic equivalence classes. In *Proc. MT Summit XIII 4th Workshop on Patent Translation*, pages 91–103.

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

T. Nagasaka, R. Shimanouchi, A. Sakamoto, T. Suzuki, Y. Morishita, T. Utsuro, and S. Matsuyoshi. 2010. Utilizing semantic equivalence classes of Japanese functional expressions in translation rule acquisition from parallel patent sentences. In *Proc. 7th LREC*, pages 1778–1785.

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

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

A. Sakamoto, T. Nagasaka, T. Utsuro, and S. Matsuyoshi. 2009. Identifying and utilizing the class of monosemous Japanese functional expressions in machine translation. In *Proc. 23rd PACLIC*, pages 803–810.

K. Shudo, T. Tanabe, M. Takahashi, and K. Yoshimura. 2004. MWEs as non-propositional content indicators. In *Proc. 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pages 32–39.

H. Sommers. 2003. An overview of EBMT. In M. Carl and A. Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 3–57. Kluwer Academic.

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

M. Tsuchiya, 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 *Proc. Workshop on Multi-Word-Expressions in a Multilingual Context*, pages 25–32.

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

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