Abstract
The Japanese language has various types of compound functional expressions, which are very important for recognizing the syntactic structures of Japanese sentences and for understanding their semantic contents. In this paper, we formalize the task of identifying Japanese compound functional expressions in a text as a chunking problem. We apply a machine learning technique to this task, where we employ that of Support Vector Machines (SVMs). We show that the proposed method significantly outperforms existing Japanese text processing tools.

1 Introduction
As in the case of other languages, the Japanese language has various types of functional words such as post-positional particles and auxiliary verbs. In addition to those functional words, the Japanese language has much more compound functional expressions which consist of more than one words including both content words and functional words. Those single functional words as well as compound functional expressions are very important for recognizing the syntactic structures of Japanese sentences and for understanding their semantic contents. Recognition and understanding of them are also very important for various kinds of NLP applications such as dialogue systems, machine translation, and question answering. However, 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, non-compositional) content word usage and a non-literal (in other words, non-compositional) functional usage.

For example, Table 1 shows two example sentences of a compound expression “に (ni) ついて (tsuite)”, which consists of a post-positional particle “に (ni)”, and a conjugated form “ついて (tsuite)” of a verb “つく (tsuku)”. In the sentence (A), the compound expression functions as a case-marking particle and has a non-compositional functional meaning “about”. On the other hand, in the sentence (B), the expression simply corresponds to a literal concatenation of the usages of the constituents: the post-positional particle “に (ni)” and the verb “ついて (tsuite)”, and has a content word meaning “follow”. Therefore, when considering machine translation of those Japanese sentences into English, it is necessary to precisely judge the usage of the compound expression “に (ni) ついて (tsuite)”, as shown in the English translation of the two sentences in Table 1.

There exist widely-used Japanese text processing tools, i.e., pairs of a morphological analysis tool and a subsequent parsing tool, such as JUMAN1+ KNP2 and ChaSen3+ CaboCha4. However, they process those compound expressions only partially, in that their morphological analysis dictionaries list only limited number of compound expressions. Furthermore, even if certain expressions are listed in a morphological analysis
Table 1: Translation Selection of a Japanese Compound Expression “に (ni) ついて (tsuite)”

<table>
<thead>
<tr>
<th></th>
<th>(A)</th>
<th>(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>私 (watashi) は (ha) 彼 (kare) に (ni) ついて (tsuite) 話した (hanashita)</td>
<td>私 (watashi) は (ha) 彼 (kare) に (ni) ついて (tsuite) 走った (hashitta)</td>
</tr>
<tr>
<td></td>
<td>(I) (TOP) (he) (about) (talked)</td>
<td>(I) (TOP) (he) (ACC) (follow) (ran)</td>
</tr>
<tr>
<td></td>
<td>(I talked about him.)</td>
<td>(I ran following him.)</td>
</tr>
</tbody>
</table>

Table 2: Classification of Functional Expressions based on Grammatical Function

<table>
<thead>
<tr>
<th>Grammatical Function Type</th>
<th># of major expressions</th>
<th># of variants</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>post-positional particle type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subsequent to predicate</td>
<td>36</td>
<td>67</td>
<td>となると</td>
</tr>
<tr>
<td>/ modifying predicate</td>
<td></td>
<td></td>
<td>(to-naru-to)</td>
</tr>
<tr>
<td>subsequent to nominal</td>
<td>45</td>
<td>121</td>
<td>にかけては</td>
</tr>
<tr>
<td>/ modifying predicate</td>
<td></td>
<td></td>
<td>(ni-kakete-ha)</td>
</tr>
<tr>
<td>subsequent to predicate, nominal</td>
<td>2</td>
<td>3</td>
<td>という</td>
</tr>
<tr>
<td>/ modifying nominal</td>
<td></td>
<td></td>
<td>(to-iu)</td>
</tr>
<tr>
<td>auxiliary verb type</td>
<td>42</td>
<td>146</td>
<td>っていい (te-ii)</td>
</tr>
<tr>
<td>total</td>
<td>125</td>
<td>337</td>
<td>—</td>
</tr>
</tbody>
</table>

dictionary, those existing tools often fail in resolving the ambiguities of their usages, such as those in Table 1. This is mainly because the framework of those existing tools is not designed so as to resolve such ambiguities of compound (possibly functional) expressions by carefully considering the context of those expressions.

Considering such a situation, it is necessary to develop a tool which properly recognizes and semantically interprets Japanese compound functional expressions. In this paper, we apply a machine learning technique to the task of identifying Japanese compound functional expressions in a text. We formalize this identification task as a chunking problem. We employ the technique of Support Vector Machines (SVMs) (Vapnik, 1998) as the machine learning technique, which has been successfully applied to various natural language processing tasks including chunking tasks such as phrase chunking (Kudo and Matsumoto, 2001) and named entity chunking (Mayfield et al., 2003). In the preliminary experimental evaluation, we focus on 52 expressions that have balanced distribution of their usages in the newspaper text corpus and are among the most difficult ones in terms of their identification in a text. We show that the proposed method significantly outperforms existing Japanese text processing tools as well as another tool based on hand-crafted rules. We further show that, in the proposed SVMs based framework, it is sufficient to collect and manually annotate about 50 training examples per expression.

2 Japanese Compound Functional Expressions and their Example Database

2.1 Japanese Compound Functional Expressions

There exist several collections which list Japanese functional expressions and examine their usages. For example, (Morita and Matsuki, 1989) examine 450 functional expressions and (Group Jamashii, 1998) also lists 965 expressions and their example sentences. Compared with those two collections, Gendaigo Hukugouji YoureiShu (National Language Research Institute, 2001) (henceforth, denoted as GHY) concentrates on 125 major functional expressions which have non-compositional usages, as well as their variants5 (337 expressions in total), and collects example sentences of those expressions. As a first step of developing a tool for identifying Japanese compound functional expressions, we start with those 125 major functional expressions and their variants. In this paper, we take an approach of regarding each of those variants as a fixed expression, rather than a semi-fixed expression or a syntactically-flexible expression (Sag et al., 2002). Then, we focus on evaluating the effectiveness of straightforwardly applying a stan-

5For 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.
Table 3: Examples of Classifying Functional/Content Usages

<table>
<thead>
<tr>
<th>Expression</th>
<th>Example sentence (English translation)</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) なると (to-naru-to)</td>
<td>しかしこの病気に効果がない なると 事は重大だ。 (The situation is serious if it is not effective against this disease.)</td>
<td>functional (ととなる (to-naru-to) = if)</td>
</tr>
<tr>
<td>(2) なると (to-naru-to)</td>
<td>彼が社長になるための条件の一つ なると 考えられている。 (They think that it will become a requirement for him to be the president.)</td>
<td>content (〜となる (to-naru-to) = that (something) becomes ~)</td>
</tr>
<tr>
<td>(3) にかけては (ni-kakete-ha)</td>
<td>お金を儲けること にかけては 素晴らしい才能をもっている。 (He has a great talent for earning money.)</td>
<td>functional (〜にかけては (ni-kakete-ha) = for ~)</td>
</tr>
<tr>
<td>(4) にかけては (ni-kakete-ha)</td>
<td>あまり気 にかけてはいない。 (I do not worry about it.)</td>
<td>content (〜を 気にかけては ((~) wo-ki-ni-kakete-ha) = worry about ~)</td>
</tr>
<tr>
<td>(5) いう (to-iu)</td>
<td>彼は生きている いう 知らせを聞いた。 (I heard that he is alive.)</td>
<td>functional (〜という (to-iu) = that ~)</td>
</tr>
<tr>
<td>(6) いう (to-iu)</td>
<td>遊びに来て下さい いう 人もいる。 (Somebody says “Please visit us.”)</td>
<td>content (〜という (to-iu) = say (that) ~)</td>
</tr>
<tr>
<td>(7) でいい (te-ii)</td>
<td>この議論が終わったら休憩 でいい。 (You may have a break after we finish this discussion.)</td>
<td>functional (〜でいい (te-ii) = may ~)</td>
</tr>
<tr>
<td>(8) でいい (te-ii)</td>
<td>このかばんは大きくていい。 (This bag is nice because it is big.)</td>
<td>content (〜でいい (te-ii) = nice because ~)</td>
</tr>
</tbody>
</table>

standard chunking technique to the task of identifying Japanese compound functional expressions.

As in Table 2, according to their grammatical functions, those 337 expressions in total are roughly classified into post-positional particle type, and auxiliary verb type. Functional expressions of post-positional particle type are further classified into three subtypes: i) those subsequent to a predicate and modifying a predicate, which mainly function as conjunctive particles and are used for constructing subordinate clauses, ii) those subsequent to a nominal, and modifying a predicate, which mainly function as case-marking particles, iii) those subsequent to a nominal, and modifying a nominal, which mainly function as adnominal particles and are used for constructing adnominal clauses. For each of those types, Table 2 also shows the number of major expressions as well as that of their variants listed in GHY, and an example expression. Furthermore, Table 3 gives example sentences of those example expressions as well as the description of their usages.

2.2 Issues on Identifying Compound Functional Expressions in a Text

The task of identifying Japanese compound functional expressions roughly consists of detecting candidates of compound functional expressions in a text and of judging the usages of those candidate expressions. The class of Japanese compound functional expressions can be regarded as closed and their number is at most a few thousand.
Table 4: Examples of Detecting more than one Candidate Expression

<table>
<thead>
<tr>
<th>Expression</th>
<th>Example sentence (English translation)</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>(9) とある (to-iu) とあるもの の難しいです。 (That’s why a match is not so easy.)</td>
<td>functional ( NP_1 ) とある (to-iu) ( NP_2 ) ( = NP_2 ) called as ( NP_1 )</td>
<td></td>
</tr>
<tr>
<td>(10) とあるもの とあるもの とあるもの とあるもの スコアは悪い。 (Although he won, the score is bad.)</td>
<td>functional ( \sim ) とあるもの とあるもの とあるもの とあるもの = although ( \sim )</td>
<td></td>
</tr>
</tbody>
</table>

Therefore, it is easy to enumerate all the compound functional expressions and their morpheme sequences. Then, in the process of detecting candidates of compound functional expressions in a text, the text are matched against the morpheme sequences of the compound functional expressions considered.

Here, most of the 125 major functional expressions we consider in this paper are compound expressions which consist of one or more content words as well as functional words. As we introduced with the examples of Table 1, it is often the case that they have both a compositional content word usage as well as a non-compositional functional usage. For example, in Table 3, the expression “となります (to-naru-to)” in the sentence (2) has the meaning “that (something) becomes ~”, which corresponds to a literal concatenation of the usages of the constituents: the post-positional particle “と”, the verb “なる”, and the post-positional particle “と”, and can be regarded as a content word usage. On the other hand, in the case of the sentence (1), the expression “となります (to-naru-to)” has a non-compositional functional meaning “if”. Based on this discussion, we classify the usages of those expressions into two classes: functional and content. Here, functional usages include both non-compositional and compositional functional usages, although most of the functional usages of those 125 major expressions can be regarded as non-compositional. On the other hand, content usages include compositional content word usages only.

More practically, in the process of detecting candidates of compound functional expressions in a text, it can happen that more than one candidate expression is detected. For example, in Table 4, both of the candidate compound functional expressions “とある (to-iu)” and “とあるもの の (to-iu-mono-no)” are detected in the sentence (9). This is because the sequence of the two morphemes “と (to)” and “ある (iu)” constituting the candidate expression “とある (to-iu)” is a subsequence of the four morphemes constituting the candidate expression “とあるもの の (to-iu-mono-no)” as below:

Morpheme sequence と (to) ある (iu) もの (mono) の (no)

Candidate expression とある (to-iu)

Candidate expression とあるもの の (to-iu-mono-no)

This is also the case with the sentence (10).

Here, however, as indicated in Table 4, the sentence (9) is an example of the functional usage of the compound functional expression “とある (to-iu)”, where the sequence of the two morphemes “と (to)” and “ある (iu)” should be identified and chunked into a compound functional expression. On the other hand, the sentence (10) is an example of the functional usage of the compound functional expression “とあるもの の (to-iu-mono-no)”, where the sequence of the four morphemes “と (to)”, “ある (iu)”, “もの (mono)”, and “の (no)” should be identified and chunked into a compound functional expression. Actually, in the result of our preliminary corpus study, at least in about 20% of the occurrences of Japanese compound functional expressions, more than one candidate expression can be detected. This result indicates that it is necessary to consider more than one candidate expression in the task of identifying a Japanese compound functional expression, and also in the task of classifying the functional/content usage of a candidate expression. Thus, in this paper, based on this observation, we formalize the task of identifying Japanese compound functional expressions as a chunking problem, rather than a classification problem.
usages. Thus, this paper focuses on those 52 expressions in the training/testing of chunking compound functional expressions. We extract 2,600 sentences (= 52 expressions × 50 sentences) from the whole example database and use them for training/testing the chunker. The number of the morphemes for the 2,600 sentences is 92,899. We ignore the chunk labels for the expressions other than the 52 expressions, resulting in 2,482/701 chunk labels for the functional/content usages, respectively.

3 Chunking Japanese Compound Functional Expressions with SVMs

3.1 Support Vector Machines

The principle idea of SVMs is to find a separate hyperplane that maximizes the margin between two classes (Vapnik, 1998). If the classes are not separated by a hyperplane in the original input space, the samples are transformed in a higher dimensional features space.

Giving x is the context (a set of features) of an input example; x_i and y_i(i = 1, ..., l, x_i ∈ R^n, y_i ∈ {1, -1}) indicate the context of the training data and its category, respectively; The decision function f in SVM framework is defined as:

\[ f(x) = \text{sgn} \left( \sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b \right) \]  (1)

where \(K\) is a kernel function, \(b \in \mathbb{R}\) is a threshold, and \(\alpha_i\) are weights. Besides, the weights \(\alpha_i\) satisfy the following constraints:

\[ 0 \leq \alpha_i \leq C \ (i = 1, ..., l) \]  (2)
\[ \sum_{i=1}^{l} \alpha_i y_i = 0 \]  (3)

where C is a misclassification cost. The \(x_i\) with non-zero \(\alpha_i\) are called support vectors. To train an SVM is to find the \(\alpha_i\) and the \(b\) by solving the optimization problem; maximizing the following under the constraints of (2) and (3):

\[ L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \]  (4)

The kernel function \(K\) is used to transform the samples in a higher dimensional features space. Among many kinds of kernel functions available, we focus on the \(d\)-th polynomial kernel:

\[ K(x, y) = (x \cdot y + 1)^d \]  (5)
Through experimental evaluation on chunking Japanese compound functional expressions, we compared polynomial kernels with \( d = 1, 2, \) and 3. Kernels with \( d = 2 \) and 3 perform best, while the kernel with \( d = 3 \) requires much more computational cost than that with \( d = 2 \). Thus, throughout the paper, we show results with the quadratic kernel \( (d = 2) \).

### 3.2 Chunking with SVMs

This section describes details of formalizing the chunking task using SVMs. In this paper, we use an SVMs-based chunking tool YamCha\(^8\) (Kudo and Matsumoto, 2001). In the SVMs-based chunking framework, SVMs are used as classifiers for assigning labels for representing chunks to each token. In our task of chunking Japanese compound functional expressions, each sentence is represented as a sequence of morphemes, where a morpheme is regarded as a token.

#### 3.2.1 Chunk Representation

For representing proper chunks, we employ IOB2 representation, one of those which have been studied well in various chunking tasks of natural language processing (Tjong Kim Sang, 1999; Kudo and Matsumoto, 2001). This method uses the following set of three labels for representing proper chunks.

- **I**: Current token is a middle or the end of a chunk consisting of more than one token.
- **O**: Current token is outside of any chunk.
- **B**: Current token is the beginning of a chunk.

As we described in section 2.2, given a candidate expression, we classify the usages of the expression into two classes: functional and content. Accordingly, we distinguish the chunks of the two types: the functional type chunk and the content type chunk. In total, we have the following five labels for representing those chunks: B-functional, I-functional, B-content, I-content, and O. Table 6 gives examples of those chunk labels representing chunks.

Finally, as for extending SVMs to multi-class classifiers, we experimentally compare the pairwise method and the one vs. rest method, where the pairwise method slightly outperformed the one vs. rest method. Throughout the paper, we show results with the pairwise method.

---

\(^8\)http://chasen.org/~taku/software/yamcha/

### 3.2.2 Features

For the feature sets for training/testing of SVMs, we use the information available in the surrounding context, such as the morphemes, their parts-of-speech tags, as well as the chunk labels. More precisely, suppose that we identify the chunk label \( c_i \) for the \( i \)-th morpheme:

\[
\begin{array}{c}
\text{Label} \quad c_{i-2} \quad c_{i-1} \quad c_i \quad c_{i+1} \quad c_{i+2} \\
\text{Feature set} \quad F_{i-2} \quad F_{i-1} \quad F_i \quad F_{i+1} \quad F_{i+2} \\
\text{at a position} \quad \rightarrow \text{Parsing Direction} \quad \rightarrow \\
\end{array}
\]

Here, \( m_i \) is the morpheme appearing at \( i \)-th position, \( F_i \) is the feature set at \( i \)-th position, and \( c_i \) is the chunk label for \( i \)-th morpheme. Roughly speaking, when identifying the chunk label \( c_i \) for the \( i \)-th morpheme, we use the feature sets \( F_{i-2}, F_{i-1}, F_i, F_{i+1}, F_{i+2} \) at the positions \( i - 2, i - 1, i, i + 1, i + 2 \), as well as the preceding two chunk labels \( c_{i-2} \) and \( c_{i-1} \).

The detailed definition of the feature set \( F_i \) at \( i \)-th position is given below. The feature set \( F_i \) is defined as a tuple of the morpheme feature \( MF(m_i) \) of the \( i \)-th morpheme \( m_i \), the chunk candidate feature \( CF(i) \) at \( i \)-th position, and the chunk context feature \( OF(i) \) at \( i \)-th position.

\[
F_i = (MF(m_i), CF(i), OF(i))
\]

The morpheme feature \( MF(m_i) \) consists of the lexical form, part-of-speech, conjugation type and form, base form, and pronunciation of \( m_i \).

The chunk candidate feature \( CF(i) \) and the chunk context feature \( OF(i) \) are defined considering the candidate compound functional expression, which is a sequence of morphemes including the morpheme \( m_i \) at the current position \( i \). As we described in section 2, the class of Japanese compound functional expressions can be regarded as closed and their number is at most a few thousand. Therefore, it is easy to enumerate all the compound functional expressions and their morpheme sequences. Chunk labels other than O should be assigned to a morpheme only when it constitutes at least one of those enumerated compound functional expressions. Suppose that a sequence of morphemes \( m_j \ldots m_k \) including \( m_i \) at the current position \( i \) constitutes a candidate functional expression \( E \) as below:

\[
\begin{array}{c}
\text{Candidate } E \text{ of a compound } \\
\text{functional expression} \quad m_{j-2} \quad m_{j-1} \quad m_j \ldots m_k \quad m_{k+1} \quad m_{k+2}
\end{array}
\]

where the morphemes \( m_{j-2}, m_{j-1}, m_{k+1}, \) and \( m_{k+2} \) are at immediate left/right contexts of \( E \). Then, the chunk candidate feature \( CF(i) \) at \( i \)-th position is defined as a tuple of the number of morphemes constituting \( E \) and the position of \( m_i \) in \( E \). The chunk context feature \( OF(i) \) at \( i \)-th position is defined as a tuple of the morpheme features
4 Experimental Evaluation

The details of the data set we used in the experimental evaluation was presented in section 2.3. As we

In such cases, we prefer the one starting with the leftmost morpheme. If more than one candidate

expressions with the leftmost morpheme, we prefer the longest one. In the example above, we

prefer the candidate $E_1$ and construct the chunk candidate features and chunk context features con-
sidering $E_1$ only.

As shown in Table 7, our SVMs-based chunkers

significantly outperform those baselines both in

terms of precision. Our SVMs-based chunkers

always assigns functional usage to the detected

functional expressions as well as those at imme-

diate functional contexts. Performance of our SVMs-

based chunkers is measured through 10-fold cross

validation.

Our main result is that existing Japanese text processing tools for functional expressions were not

capable of classifying detected expressions in general, nor did they perform well in terms of classifying detected functional expressions. "Hand-crafted rules", the

regularities in existing Japanese text processing tools, were compared with the full SVM-based chunkers.

Table 6 gives examples of chunk representation and chunk candidate/context features.

<table>
<thead>
<tr>
<th>Character</th>
<th>Morpheme (Morphological)</th>
<th>Chunk Feature (Functional)</th>
</tr>
</thead>
<tbody>
<tr>
<td>かける (toward)</td>
<td>(period)</td>
<td>(period)</td>
</tr>
<tr>
<td>えて (after)</td>
<td>(period)</td>
<td>(period)</td>
</tr>
<tr>
<td>たる (there)</td>
<td>(period)</td>
<td>(period)</td>
</tr>
<tr>
<td>て (to)</td>
<td>(period)</td>
<td>(period)</td>
</tr>
<tr>
<td>た (to)</td>
<td>(period)</td>
<td>(period)</td>
</tr>
</tbody>
</table>

Table 7 shows the accuracy of our SVMs-based chunkers compared with the existing tools for functional expressions. Our SVMs-based chunkers significantly outperform the existing tools in terms of accuracy of classifying detected functional expressions. Performance is evaluated also in terms of accuracy of classifying detected compound functional expressions.

As shown in Table 7, our SVMs-based chunkers

always assign compound functional usage to the detected

functional expressions as well as those at immedi-

date compound functional contexts. Performance of our SVMs-

based chunkers is measured through 10-fold cross

validation.
Table 7: Evaluation Results (%)

<table>
<thead>
<tr>
<th>Identifying functional chunks</th>
<th>Acc. of classifying functional/content chunks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prec.</td>
<td>Rec.</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Baselines</td>
<td></td>
</tr>
<tr>
<td>majority ( = \textit{functional})</td>
<td>78.0</td>
</tr>
<tr>
<td>Juman/KNP</td>
<td>89.2</td>
</tr>
<tr>
<td>ChaSen/CaboCha</td>
<td>89.0</td>
</tr>
<tr>
<td>hand-crafted rules</td>
<td>90.7</td>
</tr>
<tr>
<td>SVM (feature set)</td>
<td></td>
</tr>
<tr>
<td>morpheme</td>
<td>88.0</td>
</tr>
<tr>
<td>morpheme + chunk-candidate</td>
<td>91.0</td>
</tr>
<tr>
<td>morpheme + chunk-candidate/context</td>
<td>91.1</td>
</tr>
</tbody>
</table>

Figure 1: Change of \(F_{\beta=1}\) with Different Number of Training Instances

For the SVMs-based chunker with the chunk candidate feature with/without the chunk context feature, Figure 1 plots the change of \(F_{\beta=1}\) when training with different number of labeled chunks as training instances. With this result, the increase in \(F_{\beta=1}\) seems to stop with the maximum number of training instances, which supports the claim that it is sufficient to collect and manually annotate about 50 training examples per expression.

5 Concluding Remarks

The Japanese language has various types of compound functional expressions, which are very important for recognizing the syntactic structures of Japanese sentences and for understanding their semantic contents. In this paper, we formalized the task of identifying Japanese compound functional expressions in a text as a chunking problem. We applied a machine learning technique to this task, where we employed that of Support Vector Machines (SVMs). We showed that the proposed method significantly outperforms existing Japanese text processing tools. The proposed framework has advantages over an approach based on manually created rules such as the one in (Shudo et al., 2004), in that it requires human cost to manually create and maintain those rules. On the other hand, in our framework based on the machine learning technique, it is sufficient to collect and manually annotate about 50 training examples per expression.

References


