Composition and Decomposition of Japanese Katakana and Kanji Morphemes for Decision Rule Induction from Patent Documents

Michiko Yasukawa and Hidetoshi Yokoo
Department of Computer Science
Gunma University
1-5-1 Tenjin-cho, Kiryu, Gunma, 376-8515 Japan
{*michi, yokoo*}*@cs.gunma-u.ac.jp*

Abstract *We propose a new method to construct a word list for rule induction from Japanese patent documents. For word segmentation in Japanese, statistical morphological analyzers have been used in many applications. However, the output of these morphological analyzers presents defects when analyzing unknown words, specifically words that contain Kanji/Katakana morphemes. Some words are overly segmented, and their original meanings are obscured. Furthermore, boundaries between compound nouns are uncertain, which impedes investigation in the initial stages of the application. In our method, we first perform morphological analysis to segment sentences into morphemes. Second, segmented compound words are filtered by character types and Katakana/Kanji morphemes in the compound words are concatenated. Third, the concatenated morphemes are truncated to reduce verbosity. Then, words comprising Katakana/Kanji are retained for use in a word list for rule induction. The experiment results show that our method is effective for extracting decision rules for patent classification.*

Keywords Information Retrieval, Natural Language Techniques and Documents

1 Introduction

Because of the growing demand for protection of intellectual property, patent documents have been increasing numerically on a global scale. To manage many documents, document classification is conducted using decision rules[1].

During the process of decision rule induction, a word list is referred to as a vocabulary of terms. To build the word list, word segmentation[2] is prerequisite. Although Japanese is an *unsegmented* language wherein word boundaries in texts are not clear, it is also a strongly *agglutinative* language, wherein boundaries between the morphemes (units smaller than words) are clear[3]. Therefore, Japanese texts are usually segmented into morphemes; these morphemes are used as terms in information retrieval systems in Japanese.

To segment Japanese sentences into morphemes, dictionary-based statistical morphological analyzers[4, 5] have been used in various applications. These Japanese morphological analyzers have high precision, and they are effective in many cases. However, morphemes suggested by a morphological analyzer can have numerous defects, especially when documents include many unknown Katakana/Kanji words. In general, patent documents are written using uncommon Katakana/Kanji jargon. Specifically, patent documents contain words or expressions of foreign origin; newly coined compound nouns representing novel technologies; names of uncommon substances, raw materials, medicines, or chemical products; etc. Therefore, morphological analyzers tend to produce incorrect results: they separate words excessively or wrongly. Table 1 presents some examples of excessive word segmentation by ChaSen[4], which is a commonly used morphological analyzer in Japanese. The left column shows keywords in patent documents. The right column shows results of word segmentation by the morphological analyzer. In the table, adjacent morphemes are separated by a colon (:) and displayed in the Key Word In Context (KWIC) format. The first three rows show examples that include the Kanji morpheme "空"(sora/kū). This particular Kanji symbolizes the word *sky* in English, but it can also have variant meanings such as *air*, *idle*, or *waste* depending on the context. The next three rows show examples that include the Kanji morpheme "導"(dō). This Kanji symbolizes the word *leading* in English. It also yields derivative meanings such as *assistance*, *conduction*, or *derived* depending on the adjacent Kanji characters. The last four rows present examples that include the Katakana "レ"(re). This Katakana character transliterates the syllable "le" or "re" in foreign words. To convey a meaning, the sequence of Katakana characters in each keyword should not be separated. If the Katakana sequences are decomposed into fragmented morphemes of Katakana sequences, then identifying documents by keywords would be difficult. For example, a user who wants to search

Proceedings of the 15th Australasian Document Computing Symposium, Melbourne, Australia, 10 December 2010.

Table 1: Examples of excessive word segmentation

	Example keywords	Word segmentation by a morphological analyzer
1.	水陸空	**水陸**: 空 :兼用:輸送:機
	suirikukū	multi-use transport *on land, at sea, and in the air*
2.	空運転	ポンプ: 空 :**運転**:防止:用:強制:停止:信号
	karaunten	a forced stop signal for prevention of *idle running* of pumps
3.	空缶	リサイクル:用: 空 :**缶**:箱
	akikan	a box for *waste can* recycling
4.	聴導犬	**聴**: 導 :**犬**:用:警報:音:発生:回路
	chōdōken	an alarm-tone-generating circuit for *hearing assistance dog*
5.	骨導	**骨**: 導 :ヘッド:セット
	kotsudō	a *bone-conduction* headset
6.	導関数	二:次: 導 :**関数**
	dōkansū	a second order *derivative*
7.	ソレノイド	電磁:**ソ**: レ : **ノイ**:**ド**
	sorenoido	an electromagnetic *solenoid*
8.	レジン	義歯:用: レ :**ジン**
	rejin	*resin* for artificial teeth
9.	レコーダ	カセットテープ: レ :**コーダ**
	rekōda	a cassette tape *recorder*
10.	スフレ	**スフ**: レ :生地
	sufure	a mixture of ingredients for baking *souffle*

for "solenoid"[1] might also receive documents about "souffle"[2] because they have the common Katakana morpheme "レ" in "ソ:レ:ノイ:ド"(sorenoido) for *solenoid* and "スフ:レ"(sufure) for *souffle*.

When a morphological analyzer separates a long sequence of compound words into words, the situation becomes more complicated. The boundaries between compound words can be ambiguous. Therefore, word segmentation using only a statistical morphological analyzer is insufficient. Numerous combinations of shorter compound words can exist in a lengthy compound word. To select optimal words for the target application, character types and statistics might be used to determine the plausible word boundaries. In addition, a word list for the rule induction should contain multiple choices that can be informative for additional processes to meet the final objective of the application.

In Section 2, we describe our objective, i.e., to induce decision rules from patent documents. Furthermore, we address the problem of incorrectly segmented morphemes. In Section 3, we describe our proposed method to compose and decompose the segmented morphemes. In Section 4, we describe experiments that show that our method can be effective for improving the quality of the rules induced from patent documents.

[1] a current-carrying coil of wire
cf. http://www.thefreedictionary.com/solenoid

[2] light fluffy dish of eggs
cf. http://www.thefreedictionary.com/souffle

2 Decision Rules for Patent Classification

As described in this paper, our goal is to improve the performance of patent classification using decision rules. The objective of the decision rules is to acquire appropriate labels for newly arrived documents. The rules can also suggest useful keywords used to search in documents. The rules comprise words, which have meaning. Therefore, they can suggest reasons for reaching a conclusion. Decision rules can be more predictive and insightful than categorizers based on scores or measures of similarity[1].

A schematic of rule induction from documents is presented in Figure 1. The objective of decision rules is to distinguish one class from the other. Consequently, prediction is conducted using binary classification wherein the positive (interesting) documents are separated from the negative (not interesting) ones. In Figure 1, both labeled documents and unlabeled documents are transformed into a spreadsheet because the values in a spreadsheet are easier to handle than in an unstructured document. In the spreadsheets, rows represent documents and columns represent words, except the column on the extreme right that shows the labels for the documents. The values in the columns for words are one or zero, respectively indicating the presence or absence of each word in the documents. The values of labels are also one or zero, respectively denoting a positive document or a negative document. Once the spreadsheet of labeled documents is obtained, the decision rules can be induced. The induced rules

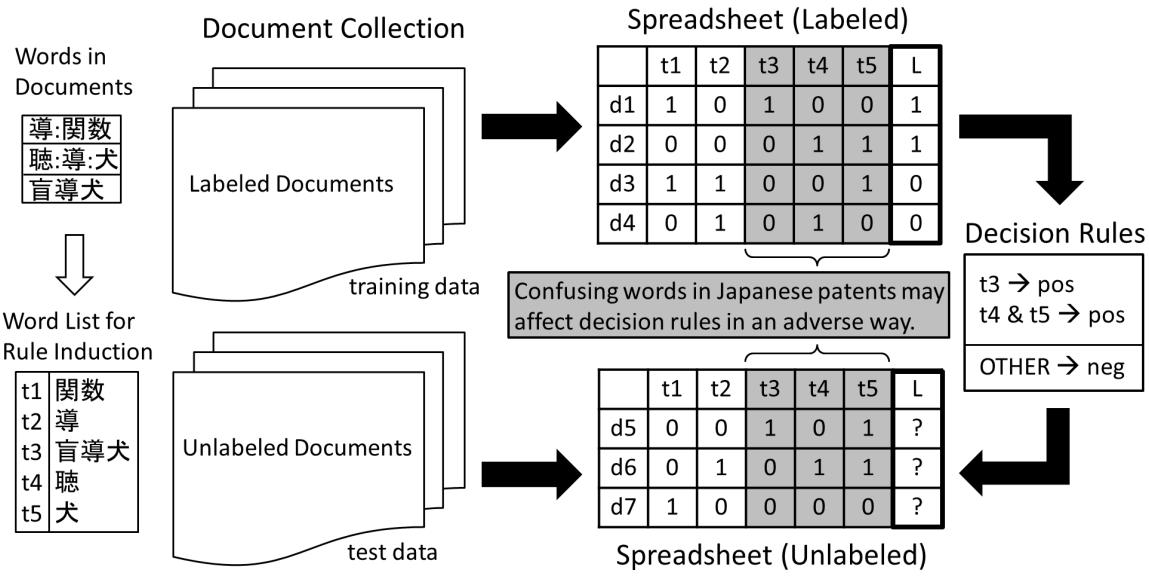

Figure 1: Rule Induction from Japanese Patent Documents

are applied to the unlabeled documents to predict the labels for those documents.

The primary steps in rule induction are the following. (1) Find a set of rules to separate the two classes, i.e. positive vs. negative. (2) Iteratively prune the rule set into simpler rule sets. (3) Select the best set of decision rules that are fairly simple and produce fewer errors.

Using the induced decision rules, document categorization with high precision is achievable for documents written in English[1]. However, when they are applied to Japanese patent documents, the categorization precision declines. This is true because word segmentation in Japanese is not successful, especially in patent documents. Japanese includes multiple intermingled writing systems for which the morphemes are not separated in written text. To obtain morphemes in sentences, morphological analysis is generally conducted.

During rule induction, the columns of the spreadsheet in Figure 1 do not represent mere words, but are presumed to serve as important keywords that represent concepts in the documents. Because dictionary-based statistical morphological analyzers refer to their own dictionaries and because dictionaries do not include all the words in Japanese, some words are excessively or wrongly separated, as shown in Table 1. When the word "盲導犬"(mōdōken; guide dog) is analyzed as a single word, native speakers of Japanese would expect that "聴導犬"(chōdōken; hearing assistance dog) is also analyzed as a single word because they are comparable types of the word *assistance dog*, which is represented by the Kanji sequence "導犬"(dōken). However, morphological analyzers separate the second one into three Kanji morphemes "聴"(chō) for *hearing*, "導"(dō) for *assistance*, and "犬"(ken) for *dog* because it is not included in their dictionaries. The first is not separated because it is specified as a single noun word in their dictionaries. In another example, morphological analyzers incorrectly separate the unknown word "スフレ"(sufure) into two morphemes "スフ"(sufu) and "レ"(re). Here, "スフ"(sufu) in Japanese is an abbreviation for "ステープルファイバー"(sutēpuru faibā), which means *staple fiber* in English, and and "レ"(re) is *re*, which is the second tone of the diatonic scale in solfeggio. When Japanese morphological analyzers process Japanese patent documents, such mistakes occur frequently. Therefore, it is necessary to amend the output of morphological analyzers when constructing a word list from patent documents.

3 Composition and Decomposition of Morphemes

In this section, a method for constructing a word list from Japanese patent documents is described. The method consists of two processes: (1) composition of Katakana/Kanji morphemes and (2) decomposition and re-composition of morphemes. In the following sections, these processes are explained in detail.

3.1 Composition of Katakana/Kanji Morphemes

In patent documents, Katakana/Kanji words are used widely to describe novel concepts in science and technology. Katakana are Japanese characters that are mainly used for spelling loan words. Many words used in terminology of science and technology are transliterated from foreign characters into Katakana according to their original pronunciation and spelling. Kanji are ideographic representations of objects and ideas. In general, nouns are written in Kanji, although verbs and adjectives are written using a combination of Kanji and Hiragana[2]. Although some nouns are

Table 2: Examples of Long Compound Words

1.	(a) を持たせたことを特徴とする油圧減速機常備小型自走式クレーン搭載小型杭打機に
	(b) 油圧:減速:機:常備:小型:自:走:式:クレーン:搭載:小型:杭:打:機
	(c) yuatsu gensoku ki jōbi kogata ji sō shiki kurēn tōsai kogata kui uchi ki
2.	(a) である、高域劣化補償ガードインターバル挿入式直交周波数分割多重変調装置と
	(b) 高域:劣化:補償:ガード:インターバル:挿入:式:直交:周波数:分割:多重:変調:装置
	(c) kōiki rekka hoshō gādo intābaru sōnyū shiki chokkō shūhasū bunkatsu tajū henchō sōchi

(a): compound words in context, (b): extracted morphemes, (c): reading of the compound words

written in Hiragana or Katakana as well as Kanji, adults prefer Kanji to Hiragana or Katakana to write formal documents. Patent documents are written in a formal tone by adults. Consequently, Kanji are in heavy usage.

Although e-mail messages contain informal spelling alternatives[7] and web texts contain slang words[6], patent documents are written in a distinctive style. Therefore, slang words do not present important issues in patent documents. However, extremely lengthy compound words are frequently used in patent documents. They are difficult to handle during word segmentation by morphological analyzers because these words are not learned from training data. These compound words are coined frequently by the authors of patent documents to describe novel technologies. Table 2 shows some difficult compound words. The first one includes 23 characters. The second one, in fact, includes 31 characters.

In the composition process, we first let morphological analyzers break sentences into morphemes. Generally speaking, morphological analyzers try to carry out word segmentation to the greatest possible extent when they encounter unknown words. Then, we identify certain morphemes according to the types of characters they include. In particular, the following morphemes are identified and extracted.

- A morpheme that is made up only of Kanji.
- A morpheme that begins only with Katakana and for which the latter part, if any, is made up of Katakana or circumflexes.

Finally, we extract Katakana/Kanji morphemes and compound them with an interposing colon (:) between adjacent morphemes, as shown in the (b) parts in Table 2. All Hiragana, Latin alphabet characters, numbers, and punctuation marks are filtered out during this process.

3.2 Decomposition and Re-composition of Morphemes

In the previous process, a sequence of morphemes is composed to be included in the word list for rule induction. Although many of the composed morphemes are very good restorations of original compound words, some inlcude attached insignificant morphemes in front or behind them.

In patent documents, most important morphemes tend to be located in the middle of a word; ancillary morphemes are adhered to them. Some morphemes are added only as a matter of form, or as stereotypical expressions in patent documents. Others are attached to rephrase the word, or to broaden the extent of the patent. Japanese is an agglutinative language. Therefore, additional morphemes are adherent to the words, apparently forming parts of the compound words. However, if a morpheme is attached to any word repeatedly in the same patent, then it is presumably a trivial affix. Some morphemes might be repeated in the same category of patents, or in the whole collection of patent documents. The most typical affixes in Japanese patent documents are morphemes representing "上記" for *aforementioned*, "等" for *such as*, "装置" for *apparatus*, or "手段" for *means*. Although some affixes have a significant number of document frequencies, others are not always outstanding. A repetitive morpheme that is meaningless in many documents can be a part of important compound words in some documents. Therefore, it is necessary to extract plausible important keywords in the pre-process and let the main process decide which one to choose.

To truncate lengthy compound words, the first and the last morphemes should be removed from the composed morphemes. The composed morphemes are decomposed and re-composed as shown in Figure 2. When a composed morpheme includes only two morphemes, it is not truncated. When a composed morpheme includes three morphemes, the first morpheme is truncated and the remainder of morphemes are re-composed. Successively, the last one is truncated and the remaining morphemes are re-composed. Then, two re-composed morphemes are obtained. When a composed morpheme includes four or more morphemes, the truncation is performed, respectively, at the first, the last, and the first and the last morphemes. Then, three re-composed morphemes are obtained. Finally, the truncated morphemes, namely, re-composed morphemes, are added to the word list as well as the original ones. The procedure of our method, including decomposition and re-composition of morphemes, is described in the following algorithm.

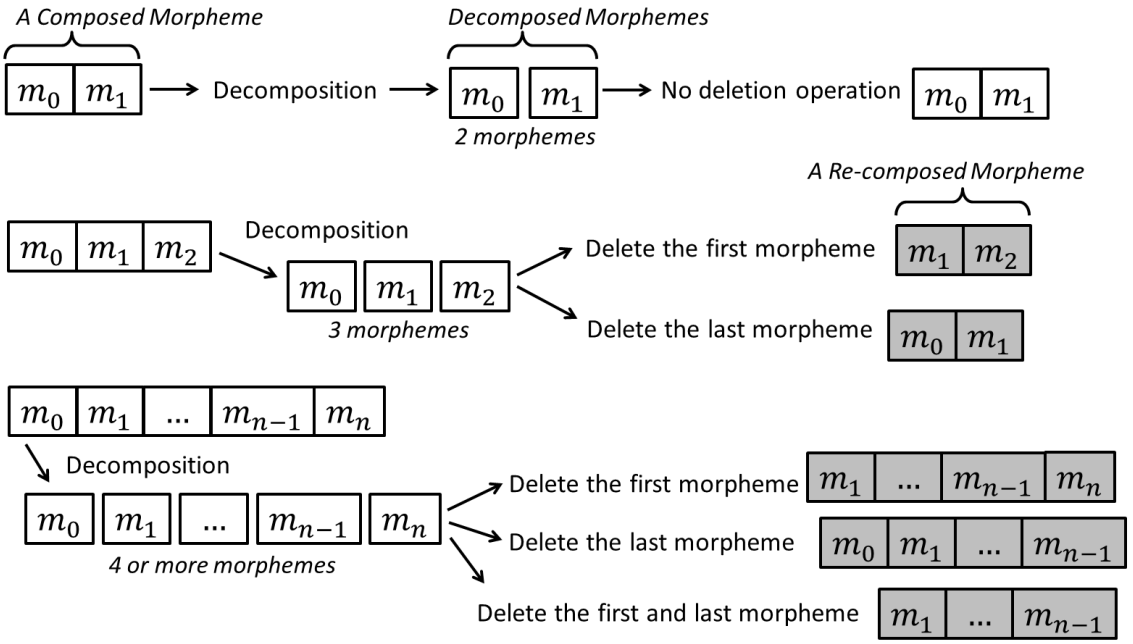

Figure 2: Composed, Decomposed and Re-composed Morphemes

Algorithm— Word List Construction

01: Initialize: L = word list;

02: for every morpheme M_i in documents **do**

03: **if** M_i is noun, verb, or adjective **then**

04: add M_i to L

05: **endif**

06: endfor

07: for every composed morpheme C_j **do**

08: add C_j to L

09: **if** C_j has three or more morphemes **then**

10: decompose and re-compose morphemes in C_j without the first morpheme and add the re-composed morpheme to L

11: decompose and re-compose morphemes in C_j without the last morpheme and add the re-composed morpheme to L

12: **endif**

13: **if** C_j has four or more morphemes **then**

14: decompose and re-compose morphemes in C_j without the first and the last morphemes and add the re-composed morpheme to L

15: **endif**

16: endfor

4 Experiment

4.1 Patent Classification by Decision Rules

An implementation of the text categorization method proposed in [1], which is a software tool kit called RIKTEXT[11], is available. We used this software to evaluate our proposed method in a patent classification application. The classifier performance is assessed using three ratios: precision, recall, and the F-measure[12]. Precision is the ratio of the number of correct positive predictions to the number of positive predictions. Recall is the number of correct positive predictions to the number of positive class documents. The F-measure F is derived using the following equation.

$$F = \frac{2}{1/P + 1/R} \tag{1}$$

where P is the precision and R denotes the recall.

4.2 Datasets for Experiments

For the experiment, we must prepare document collection that consists of training data and test data, as shown in Figure 1. Both training and test data must be labeled respectively with a one or zero, indicating a positive document or a negative document. In the past NTCIR Patent Retrieval Task, the Classification Subtask was conducted. A test collection for patent classification was released in the subtask. This test collection is well designed, but it does not meet the requirements for our experiment. In the test collection, the system must determine one or more patent categories for each patent document. In our experiment, the system must determine one or zero for the label of each patent document

in the test data. Therefore, we construct the test collection that satisfies the experimental requirements.

For the experiment, we use a document collection constructed in NTCIR-6 Patent Retrieval Task[8]. The document collection consists of 3,496,352 Japanese patent applications published during 1993–2002. The number of search topics is 2,908. Each document is given one or more International Patent Classification (IPC) codes[9]. We used these IPC codes to assign labels for positive/negative documents. For each session of classification, the positive documents have the same IPC codes although the negative documents have different IPC codes from positive ones.

The IPC codes consist of five level layers, but the lowest level is too specific and the upper level is too general. Therefore, we focused on the middle level layer: the 3rd level. For example, a document that is given the IPC code "A01M 21/00" is a positive document for the 3rd level IPC code "A01M." Any documents that are given different IPC codes from "A01M" are negative documents in this case. More specifically, if a document attached "A01M 21/00," which represents "Apparatus for destruction of unwanted vegetation, e.g. weeds," then this document is a positive document for the category "A01M," which represents "Catching, trapping or scaring of animals. Apparatus for the destruction of noxious animals or noxious plants." A document attached "A01H 3/00," which represents "Processes for modifying phenotypes," is a negative document.

Regarding pre-processing, first, the IPC codes were extracted from all documents. Then, documents were analyzed using the morphological analyzer ChaSen[4] to extract morphemes. After pre-processing, patent documents that satisfy the following conditions were collected.

- Only one IPC code is given: no multiple IPC codes are given for each document in the experimental document collection.
- The 5th level of the given IPC code is '00' that means the main group of each category.
- The number of morphemes in the document is not extreme. We used documents that contain 100 or more, and 10000 or fewer morphemes.

In this way, 148,892 documents were collected. From the collected documents, datasets of two types were produced: positive/negative sets per IPC (dataset-1) and positive/negative sets per search topic (dataset-2).

For dataset-1, IPC codes that include a moderate number of documents were selected. Specifically, we selected IPC codes that include 500 or more documents and 1000 or fewer documents. In this way, 51 IPC codes and 60,554 documents were aggregated. For every IPC code, 300 positive and 1,200 negative documents were randomly picked out from the aggregated documents. Then, 200 positive and 800 negative documents were used as training data. Furthermore, 100 positive and 400 negative documents were used as test data.

For dataset-2, document search was conducted for every search topic. For document search, GETA[10] was used. For every document search, the search query consisted of nouns, verbs, and adjectives in the claim part of a search topic. As for term weighting in the document search, the following pivoted normalization of TF-IDF weight, which is proposed in [13], was used.

$$w_{d,t} = \frac{1+\log(f_{d,t})}{1+\log(avef_d)} \times \frac{(1+\log(f_{q,t}))\times idf_t}{avedlb + S \times (dlb_d - avedlb)} \quad (2)$$

In that equation, d represents a unique document, t represents a unique term, $f_{x,t}$ is the frequency of term t in x, $idf_t = 1 + \log(N/n_t)$ is the inverse document frequency of term t, $avef_x$ denotes the average frequency of each term in x, dlb_x is the number of unique terms in x, $avedlb$ represents the average of dlb_x in the collection, and $S = 0.2$ is a constant.

From every search result, 500 positive and 1,000 negative documents were obtained. Then, the top 200 positive and top 300 negative documents are used as test data. The next 300 positive and the next 700 negative documents are used as training data. Search topics that have fewer than 500 positive documents were discarded. In this way, 1,129 search topics and 137,752 documents were aggregated.

All labeled documents for dataset-2 were obtained through document searching. Although no IPC codes are common between positive and negative classes, both positive and negative documents are controlled to become similar in dataset-2. For this reason, rule induction from dataset-2 is expected to be difficult.

4.3 Experimental Results

For comparison, components of the word list for rule induction were varied. Specifically, for components of the word list, we used (1) only nouns; (2) nouns, verbs, and adjectives; (3) nouns, verbs, adjectives, and compound morphemes; (4) only compound morphemes; (5) compound and re-compound morphemes; (6) nouns, verbs, adjectives, compound morphemes, and re-compound morphemes. The last one is our proposed method. For every condition and dataset, experimental patent classification was performed. Table 3 and Table 4 respectively present the classification performance in dataset-1 and dataset-2. In the tables, Prec and Rec respectively represent precision and recall. As shown in the tables, the proposed method has the highest average F-measure in both dataset-1 and dataset-2. The number of induced decision rules is shown in Figure 3. When the word list consists of compound morphemes, many decision

Table 3: Patent Classification (dataset-1)

component of word list	min.			max.			avg. per IPC code		
	Prec	Rec	F-measure	Prec	Rec	F-measure	Prec	Rec	F-measure
(1) n	72.53	36.00	52.94	100.0	95.00	97.44	91.50	78.33	83.84
(2) n+va	73.12	36.00	52.94	100.0	96.00	97.44	91.61	78.75	84.10
(3) n+va+comp	71.84	36.00	52.94	100.0	96.00	97.44	91.87	78.73	84.17
(4) comp	60.00	2.00	3.92	100.0	80.00	87.91	92.12	50.49	63.30
(5) comp+re	64.71	8.00	14.29	100.0	90.00	93.62	91.95	59.12	70.60
(6) n+va+comp+re	73.68	36.00	52.94	100.0	96.00	97.44	91.50	79.29	**84.41**

n: nouns, va: verbs and adjectives, comp: compound morphemes, re: re-compound morphemes

Table 4: Patent Classification (dataset-2)

component of word list	min.			max.			avg. per search topic		
	Prec	Rec	F-measure	Prec	Rec	F-measure	Prec	Rec	F-measure
(1) n	54.90	19.00	29.92	100.0	100.0	99.75	85.52	73.60	78.20
(2) n+va	54.72	19.00	30.52	100.0	100.0	99.75	85.90	73.92	78.54
(3) n+va+comp	61.54	24.00	35.96	100.0	100.0	99.75	86.49	74.39	79.12
(4) comp	60.19	2.50	4.83	100.0	99.50	99.50	87.14	58.61	68.46
(5) comp+re	59.84	18.00	28.91	100.0	99.50	98.23	87.37	63.52	72.25
(6) n+va+comp+re	61.41	14.00	23.43	100.0	100.0	99.75	86.82	74.37	**79.26**

n: nouns, va: verbs and adjectives, comp: compound morphemes, re: re-compound morphemes

rules are generated because compound morphemes are conjunction of words and they are more specific than single morphemes. For this reason, without single morphemes ((4), (5) in Table 3, 4), patent classification tends to produce low recall. However, without considering compound morphemes ((1), (2) in Table 3, 4), patent classification tends to produce low precision. The number of extracted keywords is presented in Figure 4. As shown in the figure, our proposed method can retain the largest size of vocabulary to the classifier because it makes the best of components of all types. Therefore, the proposed method is considered to be optimal when extracting decision rules from Japanese patent documents.

5 Related Work

Although most European languages are space-delimited languages, Asian languages such as Chinese and Japanese are unsegmented languages[3]. In both unsegmented and space-delimited languages, specific challenges are posed by word segmentation. In Chinese, word segmentation methods based on Conditional Random Field (CRF) have been proposed[14, 15]. In German, a method for splitting compound words using a Support Vector Machine (SVM) has been proposed[16]. In general, such dictionary-based statistical methods are effective. However, exceptional compound words are not analyzed correctly by those methods because these words are not learned from training data. Regarding Japanese studies, some previous works have described extraction of unknown words: methods particularly addressing Kanji[17, 18], methods particularly addressing Katakana[19, 20], and a method particularly addressing morphological aspects[6]. Our method particularly addresses both Kanji and Katakana, and investigates character types rather than morphological aspects of compound words.

6 Conclusion

We proposed a method to construct a word list for rule induction from patent documents. Patent documents include unusual lengths of compound words. Consequently, morphological analyzers tend to produce incorrect results during word segmentation processing. When excessive word segmentation is performed on Katakana/Kanji words, their original meanings are obscured. They are adversely affected by such segmentation in applications such as information retrieval and text categorization.

Our method specifically addresses Katakana/Kanji that are used widely in Japanese patent documents. First, word segmentation is performed using a morphological analyzer. The resultant morphemes are examined using character types. Second, Katakana/Kanji morphemes are identified, extracted, and composed. Third, the composed morphemes are decomposed and re-composed to remove ancillary

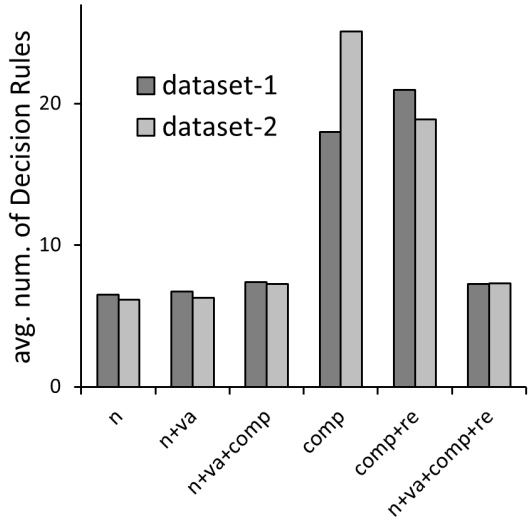

Figure 3: Induced Decision Rules

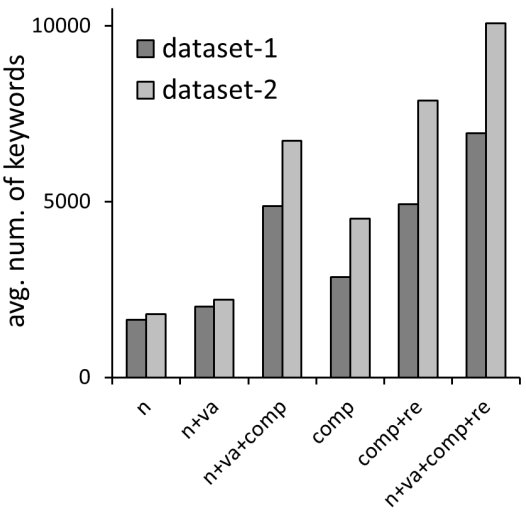

Figure 4: Extracted Keywords

morphemes. In the evaluation experiment, we applied our method to patent classification using decision rules. The re-composed morphemes and the original composed morphemes are added to the word list for rule induction to increase the number of extracted keywords. Experimental results show that our method increases the text categorization precision.

Nevertheless, there is room for additional truncation and normalization for extremely lengthy compound words. To cope with this problem, we aim to truncate composed morphemes iteratively in an efficient way. We are also planning to produce a stop word list to facilitate construction of a word list.

Acknowledgements We are grateful to the NTCIR project for providing NTCIR-6 patent collections.

References

[1] Apté, C., Damerau, F., Weiss, S.M.: Automated Learning of Decision Rules for Text Categorization. ACM Trans. Inf. Syst. 12(3): 233–251 (1994)

[2] Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press. (2008)

[3] Dale, R., Moisl, H., Somers, H.: Handbook of Natural Language Processing. CRC Press. (2000)

[4] ChaSen: http://chasen-legacy.sourceforge.jp/

[5] Mecab: http://mecab.sourceforge.net/

[6] Murawaki, Y., Kurohashi, S.: Online acquisition of Japanese unknown morphemes using morphological constraints. EMNLP 08, pp. 429–437 (2008)

[7] Nishimura, Y.: Linguistic Innovations and Interactional Features of Casual Online Communication in Japanese. JCMC, vol.9, no.1 (2003)

[8] Fujii, A., Iwayama, M., Kando, N.: Overview of the Patent Retrieval Task at the NTCIR-6. Proc. NTCIR-6 Workshop Meeting, pp. 359–365 (2007)

[9] WIPO: International Patent Classication (IPC). http://www.wipo.int/classifications/ipc/en/

[10] Generic Engine for Transposable Association (GETA). http://geta.ex.nii.ac.jp/e/

[11] Indurkhya, N.: RIKTEXT: Rule Induction Kit for Text. http://www.data-miner.com/riktext.pdf (2004)

[12] Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F.J.: Text Mining Predictive Methods for Analyzing Unstructured Information. Springer. (2005)

[13] Singhal, A., Buckley, C., Mitra, M.: Pivoted Document Length Normalization. SIGIR 96, pp. 21–29 (1996)

[14] Tseng, H., Chang, P., Galen, A., Jurafsky, D., Manning, C.: A Conditional Random Field Word Segmenter. 4th SIGHAN Workshop on Chinese Language Processing. pp. 168–171 (2005)

[15] Fuchun, P., Feng, F., McCallum, A.: Chinese segmentation and new word detection using conditional random fields, COLING 04. pp. 562–568 (2004)

[16] Alfonseca, E., Bilac, S., Pharies, S.: German Decompounding in a Difficult Corpus, CICLing 08, pp. 128–139 (2008)

[17] Watanabe, Y., Murata, M., Takeuchi, M., Nagao, M.: Document Classification Using Domain Specific Kanji Characters Extracted by X2 Method. COLING 96, pp. 794–799 (1996)

[18] Ando, R. K. Lee, L.: Mostly-Unsupervised Statistical Segmentation of Japanese Kanji Sequences. Natural Language Engineering 9 (2): 127–149 (2002)

[19] Seki, K., Hattori, H., Uehara, K.: Generating diverse katakana variants based on phonemic mapping. SIGIR 08, pp. 793-794 (2008)

[20] Nakazawa, T., Kawahara, D., Kurohashi, S.: Automatic Acquisition of Basic Katakana Lexicon from a Given Corpus. IJCNLP 05, pp. 682–693 (2005)