International Journal of Electronic Commerce Studies
Vol.4, No.2, pp.323-336, 2013
doi: 10.7903/ijecs.1125

A RULE SET TO SELECT REPRESENTATIVE NOUNS FROM A NOUN SYNONYM SET FOR A JAPANESE FISHING WEBSITE

Kenji Kawabata
Kyushu University
Motooka 744, Nishi-Ku, Fukuoka-Shi, 819-0395, Japan
te106109@s.kyushu-u.ac.jp

Kunihiko Kaneko
Kyushu University
Motooka 744, Nishi-Ku, Fukuoka-Shi, 819-0395, Japan
kaneko@ait.kyushu-u.ac.jp

ABSTRACT

Japanese documents have noun synonyms. These use *kanji* notation, *hiragana* notation, and *katakana* notation for words. Sometimes words have alternate *kanji* expressions: alternate names for an object, different suffixes for *kanji*, etc. This is why noun synonym sets are formed for Japanese nouns. Thesauruses and dictionaries can be used to select a representative expression from a noun synonym set. However, these references do not consider the type of document. Representative nouns are often different depending on the type of articles. For example, in articles in newspapers, *kanji* is preferred. In contrast, in articles in encyclopedias, *katakana* is preferred. The problem is to form a rule set to select a representative noun from a noun synonym set, and the rule set must consider the type of document. We propose a rule set arranged for the WEB Fish Encyclopedia (in Japanese, *Sakanazukan*). We introduce a keyword category in the rule set to increase the correctness of the selected representative noun. As a result, most of the representative expressions were selected appropriately from noun synonyms. We expressed these noun synonyms as feature vectors. By using three numerical values and four Boolean values, all noun synonyms were expressed.

Keywords: Noun Synonym, Japanese Syntax Analysis, Keyword Dictionary

1. INTRODUCTION

Selecting representative nouns from noun synonym sets is an important issue in the document database research area. We accomplish this by using thesauruses and dictionaries. However, we assume that existing dictionaries, such as the Japanese dictionary *Kojien*[1] or Wikipedia[2], cannot cover all nouns, especially nouns specific to a particular field. Therefore, we need to create a rule set to select the representative nouns from noun synonym sets. Representative nouns are often different depending on the types of articles. For example, in Japanese documents, nouns are represented using *kanji* notation, *hiragana* notation, and *katakana* notation. In articles in newspapers, *kanji* is preferred. In contrast, in articles in encyclopedias, *katakana* is preferred. Furthermore, for Japanese documents in general, the problem is complicated by the fact that sometimes words have alternate *kanji* expressions: alternate names for an object, different suffixes for *kanji*, etc. Until now, a rule set to select a representative noun from a noun synonym set for Japanese documents has not been studied by any researchers as far as we know. We analyze the orthographical variants of the nouns in the WEB Fish Encyclopedia[3], and we construct a customized rule set that works well for the WEB Fish Encyclopedia. The organization of this paper is as follows. Section 2 briefly explains a method to extract Japanese nouns from Japanese documents. The details of the customized rule set are explained in Section 3. Section 4 explains examples of usages of the representative nouns of Japanese documents.

2. EXTRACTION OF NOUNS

Japanese has three different types of characters. One is *hiragana* (cursive Japanese syllabary), another is *katakana* (angular Japanese syllabary), and the other is *kanji* (Chinese characters). For example, if we want to write "moss" in Japanese, the *hiragana* expression is "こけ", the *katakana* expression is "コケ", and the *kanji* expression is "苔". The exact meaning can be expressed by using *kanji*, and the sound of the word can be expressed by *hiragana* and *katakana*.

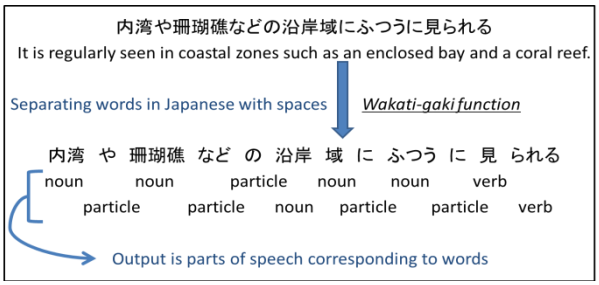

Figure 1. The WEB fish encyclopedia was separated into words using MeCab

We installed and operated the MeCabsoftware to extract nouns fromthe WEB Fish Encyclopedia(in Japanese, *Sakanazukan*)[3]. The WEB Fish Encyclopedia contains 4,668 files, contributed by many anonymous authors. Therefore, we cannot avoid the occurrence of orthographical variants caused by the method of contribution. The WEB Fish Encyclopedia was separated into words using MeCab. We obtained 1,365,149 words, including 9,545 nouns. Each noun is either *kanji*, *hiragana*, *katakana*, or a mixture.

Table 1. List of orthographical variants

English	Japanese	Orthographical variants	
Coudal fin	尾鰭	尾鰭，尾びれ	Fully kanji or partly
Dorsal fin	背鰭	背鰭，背びれ	
Mustache	口髭	口髭，口ひげ	
All over the world	世界中	世界中，世界じゅう	
Boiled food	煮付け	煮付け，煮つけ	
Hatching	孵化	孵化，ふ化	
Snakehead	ライギョ	ライギョ，雷魚	Kanji or katakana
Wedge	くさび	くさび，クサビ	Hiragana or katakana
Fold	ひだ	ひだ，ヒダ	
Puffer fish	フグ	フグ，ふぐ	
Kusaya	くさや	くさや，クサヤ	
Oyster	カキ	カキ，かき	
Month	ヵ月	ヵ月，カ月	Different counter suffix
Shop counter	売場	売場，売り場	Different declensional hiragana ending
Treatment	取扱い	取扱い，取り扱い	
Fish paste	練製品	練製品，練り製品	
Yaeyama Islands	八重山諸島	八重山諸島，八重山列島	Alternate names
Radial	放射状	放射状，放射線状	

Table 1. List of orthographical variants (Cont.)

English	Japanese	Orthographical variants	
First	一番	一番, いちばん	Kanji or not
Part	一部	一部, いちぶ	
Now	今	今, いま	
Place	場所	場所, ばしょ	
Human	人	人, ヒト, ひと	
Territory	縄張り	縄張り, なわばり	
Almost	殆ど	殆ど, ほとんど	
All	全て	全て, すべて	
Loose	緩やか	緩やか, ゆるやか	
Mortar	すり鉢	すり鉢, すりばち	
Moss	苔	苔, コケ, こけ	
Shape	形	形, かたち	
Varied	様々	様々, さまざま	
Two	二つ	二つ, ふたつ	
Barbel	髭	髭, 鬚, ヒゲ, ひげ	
Other	その他	その他, そのた	
Roundness	丸み	丸み, まるみ	
Fin	鰭	鰭, ヒレ	
Category	カテゴリー	カテゴリー, カテゴリ	Existence of a long vowel
Superior	スペリオル	スペリオル, スペリオール	
Pair	ペア	ペア, ペアー	
Orange raffia	オレンジラフィ	オレンジラフィ, オレンジラフィー	
Manitoba	マニトバ	マニトバ, マニトーバ	
Laccadive	ラカディヴ	ラカディヴ, ラカディーヴ	
Bermuda	バミューダ	バミューダ, バーミューダ	
New South Wales	ニューサウスウェールズ	ニューサウスウェールズ, ニューサウスウェルズ	
Chao Phraya	チャオプラヤ	チャオプラヤ, チャオプラヤー	

Table 1. List of orthographical variants (Cont.)

English	Japanese	Orthographical variants	
Baja California	バハカリフォルニア	バハカリフォルニア, バハ・カリフォルニア	Existence of a bullet point
Papua New Guinea	パプアニューギニア	パプアニューギニア, パプア・ニューギニア	
Suffix use: when Spanish mackerel	ころ	ころ, ごろ	Difference caused by taking along to other nouns
	サワラ	サワラ, ザワラ	
Kamchatka	カムチャツカ	カムチャツカ, カムチャッカ	Different katakana expression
Leptocephalus	レプトセファルス	レプトセファルス, レプトケファルス, レプトケパルス	
Timor	チモール	チモール, ティモール	
Philipin	フィリピン	フィリピン, フィリンピン	
Diver	ダイバー	ダイバー, ダイヴァー	
Kermadec	ケルマデック	ケルマデック, ケルマディック	
Lord Howe	ロードハウ	ロードハウ, ロードホウ	
Tuamotu	トゥアモトゥ	トゥアモトゥ, ツアモツ	
Maldive	モルディブ	モルディブ, モルジブ, モルディヴ	
Loyalty	ロイヤルティ	ロイヤルティ, ロヤルティ	
Society	ソサイエティ	ソサイエティ, ソサエティ, ソサイティ	
Pitcairn	ピトケアン	ピトケアン, ピトカーン	
Nova Scotia	ノバスコシア	ノバスコシア, ノヴァスコシア	
Marquesas	マルケサス	マルケサス, マーケサス	

Table 1. List of orthographical variants (Cont.)

English	Japanese	Orthographical variants	
Monterey	モンタレイ	モンタレイ, モンタレー	Different katakana expression
Cobitis	スジシマドジョウ	スジシマドジョウ, スジシマドショウ	
Sardinella melanura	オグロイワシ	オグロイワシ, オグロオワシ	
Tetrodotoxin	テトロドトキシン	テトロドトキシン, テトロドドキシン	
Saskatchewan	サスカチュワン	サスカチュワン, サスカッチュュワン	
Georgia	ジョージア	ジョージア, ジュージア	
Great Barrier Reef	グレートバリアリーフ	グレートバリアリーフ, グレイトバリアリーフ, グレートバリアーリーフ	Different katakana expression and a long vowel
New Zealand	ニュージーランド	ニュージーランド, ニュージランド, ニュージーーランド	

3. SELECTING KEYWORDS

3.1 Orthographical Variants

As stated above, the noun list contains orthographical variants of noun synonyms depending on the content and the contributors. When a Japanese text describes English names and the names of other countries in *katakana*, it produces several other descriptions. Differences are also found if the text is in *kanji*. We found 72 noun synonym sets from the WEB Fish Encyclopedia, and we selected a representative noun fromeach set manually. We use each noun as the ground truth. Then, we use the term "orthographical variants" to refer to nouns other than the representative nouns in the noun synonym sets. In all, we categorized the 72 noun synonym sets into 11 types according to their differences. Table 2 shows the representative nouns, orthographical variants, and categorized results.

As we expected, most of the differences were caused by *kanji* and *katakana* expressions. Almost the same number of occurrences, such as "Kamchatka"

(a peninsula in Russia), appeared in the files. In Japanese *"*カムチャツカ*"* appeared 7 times and *"*カムチャッカ*"* appeared 9 times. Some words such as *"*Great Barrier Reef*"* were divided into a majority notation and a minority. *"*グレートバリアリーフ*"*, in Japanese, appeared 18 times and *"*グレイトバリアリーフ*"* appeared once.

Some cases were obviously caused by an error. Our goal is to form a rule set to select representative nouns automatically from the synonym sets. We formed six rules, as shown in the following bulleted list. In the rules, each noun has attributes. These are *Is_general_term*(*x*), *Is_proper_noun*(*x*), *Is_katakana*(*x*), *Is_common_noun*(*x*), *Is_kanji*(*x*), *Is_Kojien*(*x*), *Is_most_frequent*(*x*), *Is_Wikipedia_keyword*(*x*), and *Hit_Google_search_the_most*(*x*). We utilized the keyword entries of the Japanese dictionary*Kojien* fourth edition, Wikipedia, and the Google search engine.

$$Input\_words\{\exists x \mid x \in Is\_general\_term(x) \text{ or } (Is\_proper\_noun(x) \text{ and } Is\_katakana(x))\} \Rightarrow Is\_representative\_notation(x) \quad (1)$$

$$Input\_words\{\exists x \mid x \in Is\_common\_noun(x) \text{ and } Is\_\ kanji(x)\} \Rightarrow Is\_representative\_notation(x) \quad (2)$$

$$Input\_words\{\exists x \mid x \in Is\_Kojien\_keyword(x)\} \Rightarrow Is\_representative\_notation(x) \quad (3)$$

Here, Kojien is the Japanese dictionary published by Iwanami

$$Input\_words\{\exists x \mid x \in Is\_most\_frequent(x)\} \Rightarrow Is\_representative\_notation(x) \quad (4)$$

$$Input\_words\{\exists x \mid x \in Is\_Wikipedia\_keyword(x)\} \Rightarrow Is\_representative\_notation(x) \quad (5)$$

$$Input\_words\{\exists x \mid x \in Hit\_Google\_search\_the\_most(x)\} \Rightarrow Is\_representative\_notation(x) \quad (6)$$

We created a flowchart to select representative nouns by using the above rules. The flowchart, arranged for *Sakanazukan*, is shown in Fig. 2. A single representative can be selected by following this flowchart. For the general terms of fish and overseas place names in the WEB Fish Encyclopedia, the representative noun should be*katakana*, and for the domestic place names and common nouns, the representative noun should be *kanji*. We used those rules in the flowchart.

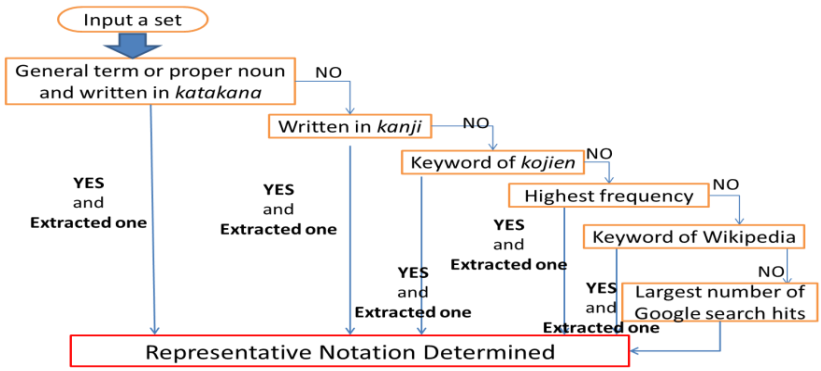

Figure 2. Flowchart of keyword selection

We also represented the noun attributes as noun feature vectors, as shown in Table 2. We selected 70 correct representative notations from the 72 noun synonym sets, and so the correctness is 70/72.

Table 2. Feature vectors of the nouns

Word (in Japanese)	Katakana	Number of kanji	Hiragana	Kojien	Wikipedia	Frequency	Google Hit
一番	F	2	F	T	F	2	616000000
いちばん	F	0	T	T	F	2	34900000
一部	F	2	F	T	T	50	526000000
いちぶ	F	0	T	T	F	1	270000
今	F	1	F	T	T	9	2310000000
いま	F	0	T	T	F	10	480000000
売場	F	2	F	T	F	3	8920000
売り場	F	2	T	F	F	1	21800000
オグロイワシ	T	0	F	F	F	1	89
オグロオワシ	T	0	F	F	F	2	1
尾鰭	F	2	F	T	T	1071	436000
尾びれ	F	1	T	F	T	5	485000
オレンジラフィ	T	0	F	F	F	2	1140
オレンジラフィー	T	0	F	F	F	1	16100
カキ	T	0	F	F	T	5	13500000
かき	F	0	T	T	T	1	34000000
ヵ月	F	1	F	F	T	5	54700000
カ月	T	1	F	F	T	1	58300000
形	F	1	F	T	T	222	706000000
かたち	F	0	T	T	F	2	15000000
カテゴリー	T	0	F	T	T	2	1180000000
カテゴリ	T	0	F	F	T	1	1250000000
カムチャツカ	T	0	F	T	T	7	222000
カムチャッカ	T	0	F	F	T	9	307000
くさや	F	0	T	T	T	1	1420000
クサヤ	T	0	F	F	T	3	125000
くさび	F	0	T	T	T	4	948000
クサビ	T	0	F	F	T	5	700000
口髭	F	2	F	T	T	6	441000

Table 2. Feature vectors of the nouns (Cont.)

Word (in Japanese)	Katakana	Number of kanji	Hiragana	Kojien	Wikipedia	Frequency	Google Hit
口ひげ	F	1	T	F	T	55	374000
グレートバリアリーフ	T	0	F	F	T	18	1410000
グレイトバリアリーフ	T	0	F	F	F	1	13800
グレートバリアーリーフ	T	0	F	F	T	1	19100
ケルマデック	T	0	F	F	T	3	12100
ケルマディック	T	0	F	F	T	1	44900
苔	F	1	F	T	T	1	24200000
コケ	T	0	F	F	T	1	11600000
こけ	F	0	T	T	T	19	8730000
ころ	F	0	T	T	F	1	103000000
ごろ	F	0	T	F	F	2	66900000
サスカチュワン	T	0	F	F	T	1	207000
サスカッチュュワン	T	0	F	F	F	1	2
様々	F	2	F	F	F	14	193000000
さまざま	F	0	T	T	F	8	107000000
サワラ	T	0	F	F	T	10	1570000
ザワラ	T	0	F	F	F	1	6130
ジョージア	T	0	F	T	T	4	5600000
ジュージア	T	0	F	F	F	1	6810
スジシマドジョウ	T	0	F	F	F	6	108000
スジシマドショウ	T	0	F	F	F	1	278
全て	F	1	T	T	T	14	617000000
すべて	F	0	T	T	F	7	2440000000
スペリオル	T	0	F	T	T	1	618000
スペリオール	T	0	F	F	F	1	1120000
すり鉢	F	1	T	F	T	5	1760000
すりばち	F	0	T	T	T	2	133000
世界中	F	3	F	F	F	60	230000000
世界じゅう	F	2	T	F	F	2	333000
背鰭	F	2	F	T	T	2189	372000
背びれ	F	1	T	F	T	1	522000
ソサイエティ	T	0	F	F	F	3	627000
ソサエティ	T	0	F	F	F	1	1860000
ソサイティ	T	0	F	F	F	1	12300
その他	F	1	T	F	T	1143	1670000000
そのほか	F	0	T	F	F	1	45800000
ダイバー	T	0	F	T	T	14	9390000
ダイヴァー	T	0	F	F	F	4	78800
チモール	T	0	F	T	T	6	272000
ティモール	T	0	F	F	T	2	3370000
チャオプラヤ	T	0	F	T	F	1	925000
チャオプラヤー	T	0	F	F	T	1	490000
テトロドトキシン	T	0	F	T	T	48	92700
テトロドドキシン	T	0	F	F	F	1	2850

Table 2. Feature vectors of the nouns (Cont.)

Word (in Japanese)	Katakana	Number of kanji	Hiragana	Kojien	Wikipedia	Frequency	Google Hit
トゥアモトゥ	T	0	F	F	F	3	6410
ツアモツ	T	0	F	F	F	1	25800
取扱い	F	2	T	T	F	1	66200000
取り扱い	F	2	T	F	F	1	174000000
縄張り	F	2	T	F	T	21	2490000
ニューサウスウェールズ	T	0	F	T	T	4	939000
ニューサウスウェルズ	T	0	F	F	F	7	11500
ニュージーランド	T	0	F	T	T	108	24100000
ニュージランド	T	0	F	F	F	4	529000
ニュージーーランド	T	0	F	F	F	1	5440
煮付け	F	2	T	F	T	3	5020000
煮つけ	F	1	T	F	F	4	1210000
練製品	F	3	F	T	F	31	285000
練り製品	F	3	T	F	T	4	319000
ノバスコシア	T	0	F	F	T	5	253000
ノヴァスコシア	T	0	F	F	T	2	26200
場所	F	2	F	T	T	175	583000000
ばしょ	F	0	T	T	F	1	2520000
バハカリフォルニア	T	0	F	F	T	29	342000
バハ・カリフォルニア	T	0	F	F	T	4	342000
パプアニューギニア	T	0	F	F	T	33	5000000
パプア・ニューギニア	T	0	F	F	T	2	5010000
バミューダ	T	0	F	F	T	26	2840000
バーミューダ	T	0	F	F	T	2	237000
髭	F	1	F	T	T	69	16500000
鬚	F	1	F	T	T	12	4760000
ヒゲ	T	0	F	F	T	49	17700000
ひげ	F	0	T	T	T	29	19100000
ひだ	F	0	T	T	T	2	6840000
ヒダ	T	0	F	F	F	1	2230000
人	F	1	F	T	T	105	12280000000
ヒト	T	0	F	F	T	2	34900000
ひと	F	0	T	T	T	1	207000000
ピトケアン	T	0	F	F	T	1	788000
ピトカーン	T	0	F	F	F	1	1630
鰭	F	1	F	T	T	2705	2870000
ヒレ	T	0	F	F	T	21	6000000
フィリピン	T	0	F	T	T	438	28400000
フィリンピン	T	0	F	F	F	3	8430
孵化	F	2	F	T	T	18	25600000
ふ化	F	1	T	F	T	3	792000
フグ	T	0	F	F	T	371	6230000
ふぐ	F	0	T	T	T	99	24200000
二つ	F	1	T	T	F	4	76100000
ふたつ	F	0	T	T	F	1	16200000
ペア	T	0	F	T	T	28	54400000

Table 2. Feature vectors of the nouns (Cont.)

Word (in Japanese)	Katakana	Number of kanji	Hiragana	Kojien	Wikipedia	Frequency	Google Hit
ペアー	T	0	F	F	F	1	1570000
放射状	F	3	F	T	F	12	2840000
放射線状	F	4	F	F	F	4	462000
殆ど	F	1	T	T	F	14	50800000
ほとんど	F	0	T	T	F	82	257000000
マニトバ	T	0	F	F	T	1	259000
マニトーバ	T	0	F	F	F	1	289
マルケサス	T	0	F	T	F	1	37100
マーケサス	T	0	F	F	F	1	4820
丸み	F	1	T	T	F	45	6170000
まるみ	F	0	T	T	F	3	1840000
モルディブ	T	0	F	F	T	22	5030000
モルジブ	T	0	F	F	F	4	1830000
モルディヴ	T	0	F	F	F	2	152000
モンタレイ	T	0	F	F	F	4	6420
モンタレー	T	0	F	F	T	1	148000
八重山諸島	F	5	F	T	T	43	3670000
八重山列島	F	5	F	F	T	6	183000
緩やか	F	1	T	T	F	41	7670000
ゆるやか	F	0	T	T	F	2	4440000
ライギョ	T	0	F	F	T	2	507000
雷魚	F	2	F	T	T	2	1030000
ラカディヴ	T	0	F	F	F	1	26
ラカディーヴ	T	0	F	F	F	1	6
レプトセファルス	T	0	F	T	T	5	11900
レプトケファルス	T	0	F	F	T	7	33400
レプトケパルス	T	0	F	F	T	4	771
ロードハウ	T	0	F	F	F	8	70500
ロードホウ	T	0	F	F	F	1	32
ロイヤルティ	T	0	F	F	T	1	4870000
ロヤルティ	T	0	F	F	F	1	385

3.2 Keyword Categories

Since keywords should help to identify information, some words are unsuitable as keywords. One group of unsuitable words is the group of nouns in common use, and the other group is words appropriate in other articles but unsuitable in the WEB Fish Encyclopedia. Examples of the group of common nouns are “thing”, “body”, “object”, “sake”, “diver”, and “fin”, and examples of words unsuitable in the WEB Fish Encyclopedia are “Ministry of the Environment”, “animal”, “liver”, “distribution”, “pebbles”, and “muscle”. We obtained 4,946 keywords by sorting these nouns. We created seven categories to reveal the types of selected keywords. Table 3 shows the results. We used “General term of fish” and “Distribution” as rule (1) “General term or proper noun.”

Table 3. Keyword categories

	Examples in Japanese	Examples in English	Total
General term of fish	スズキ ハゼ カサゴ	Perch Goby Scorpionfish	3983
Distribution	日本 インド 沖縄	Japan India Okinawa	613
Living place	大陸棚 河口 深海	Continental shelf Estuary Deep sea	80
Feature	軟体動物 毒性 斑点	Mollusk Toxicity Spot	57
Body color	紫色 茶色 ウグイス色	Violet Brown Greenish-brown	33
Part of general term of fish	ヨウジ レッド ボウ	Pipe Red Robin	93
Others	南方 デトリタス 模様	South Detritus Pattern	87

4. FURTHER DISCUSSION

4.1 Highlighted Display

We used the obtained representative nouns for a highlighted display of documents. For example, we can highlight nouns in the WEB Fish Encyclopedia files by tagging them in HTML format, as shown in Figure 3. We put together all the representative nouns in a file and mapped them one by one to each of the WEB Fish Encyclopedia files. We were able to distinguish the types of keywords by tags such as <place

name>Kamchatka</place name> and to characterize the keywords individually. This procedure makes it easier to link to external open dictionaries such as Wikipedia.

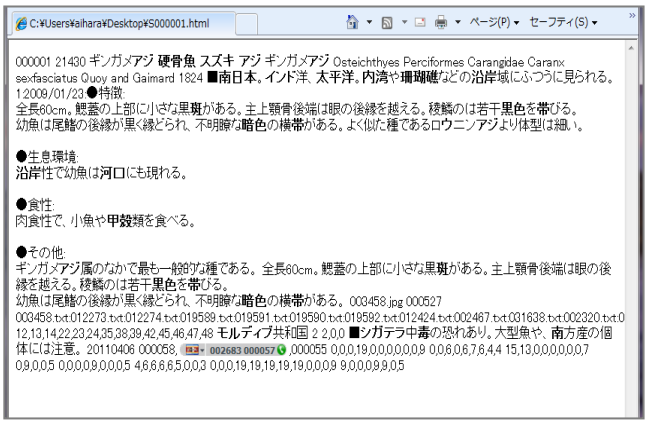

Figure 3. Example of tagging in a document

4.2 Necessity for a Personal Dictionary

We counted the number of keywords contained in the Japanese dictionary *Kojien* fourth edition and Wikipedia. The Japanese dictionary *Kojien* fourth edition contains only frequently used proper nouns. Also, the names of broad categories of fish were found. Furthermore, common nouns, such as nouns of habitats and features, were found. In Wikipedia, many proper nouns were found, but not many common nouns. According to the result of these observations, the Japanese dictionary *Kojien* fourth edition had 949 keywords out of 4,946, and Wikipedia had 1,238 keywords. As expected, the existing dictionaries did not cover all of the technical keywords, so we were able to prove the validity of making personal dictionaries.

5. ACKNOWLEDGMENT

The 4,668 text files of the WEB Fish Encyclopedia were provided by Mr. Kenichi Naoe, who is a student of the Graduate School of Information Science and Electrical Engineering, Kyushu University, and also a member of the Fishing Forum. Also, Prof. Masayoshi Aritsugi and Associate Prof. Teruaki Kitasuka of the Graduate School of Science and Technology, Kumamoto University gave valuable advice for this study.

6. REFERENCES

[1] I. Shinmura, *Kojien fourth edition (Japanese Dictionary)*. Tokyo:

Iwanami, 1991.

[2] Wikimedia Foundation, Inc. *Wikipedia*. Retrieved on January 31, 2012, from http://ja.wikipedia.org/.

[3] Fishing-Forum, *The WEB fish encyclopedia (in Japanese, Sakanazukan)*. Retrieved on January 31, 2012, from http://www.fishing-forum.org/zukan/.

[4] D. Kawahara, and S. Kurohashi, Case frame compilation from the web using high-performance computing. In Nicoletta Calzolari (Ed.), *Proceedings of the 5th International conference on Language Resource and Evaluation* (p1344-1347). Italy: LREC 2006 Committees, 2006.

[5] S. Ravi, and K. Knight, Minimized models for unsupervised part-of-speech tagging. In Jian Su and Janyce Wiebe (Eds.), *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language* (p504-512). Singapore: Curran Associates, 2009.http://dx.doi.org/10.3115/1687878.1687950.

[6] Z. Huang, V. Eidelman, and M. Harper, Improving a simple bigram hmm part-of-speech tagger by latent annotation and self-Training. In Mari Ostendorf (Ed.), *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (p213-216). Stroudsburg: Association for Computational Linguistics, 2009.http://dx.doi.org/10.3115/1620853.1620911.

[7] J. Lafferty, A. McCallum, and F. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, In Carla E. Brodley and Andrea Pohoreckyj Danyluk (Eds.), *Proceeding of the 18th International Conference on Machine Learning* (p282-289). San Francisco: Morgan Kaufmann, 2001.

[8] J. Halpern, Lexicon-based Orthographic Disambiguation in CJK Intelligent Information Retrieval. In Nicoletta Calzolari, Key-Sun Choi, Asanee Kawtrakul, Alessandro Lenci, and Tokunaga Takenobu (Eds.), *Proceedings of the 3rd workshop on Asian language resources and international standardization* (p1-7). Stroudsburg: Association for Computational Linguistics, 2002. http://dx.doi.org/10.3115/1118759.1118765.