Constructing Large-Scale Person Ontology from Wikipedia

Yumi Shibaki
Nagaoka University of Technology
shibaki@jnlp.org

Masaaki Nagata
NTT Communication Science Laboratories
nagata.masaaki@labs.ntt.co.jp

Kazuhide Yamamoto
Nagaoka University of Technology
yamamoto@jnlp.org

Abstract

This paper presents a method for constructing a large-scale Person Ontology with category hierarchy from Wikipedia. We first extract Wikipedia category labels which represent person (hereafter, Wikipedia Person Category, WPC) by using a machine learning classifier. We then construct a WPC hierarchy by detecting *is-a* relations in the Wikipedia category network. We then extract the titles of Wikipedia articles which represent person (hereafter, Wikipedia person instance, WPI). Experiments show that the accuracy of WPC extraction is 99.3% precision and 98.4% recall, while that of WPI extraction is 98.2% and 98.6%, respectively. The accuracies are significantly higher than the previous methods.

1 Introduction

In recent years, we have become increasingly aware of the need for, up-to-date knowledge bases offering broad coverage in order to implement practical semantic inference engines for advanced applications such as question answering, summarization and textual entailment recognition. General ontologies, such as WordNet (Fellbaum et al., 1998), and *Nihongo Goi-Taikei* (Ikehara et al., 1997), contain general knowledge of wide range of fields. However, it is difficult to instantly add new knowledge, particularly proper nouns, to these general ontologies. Therefore, Wikipedia has come to be used as a useful corpus for knowledge extraction because it is a free and large-scale online encyclopedia that continues to be actively developed. For example, in DBpedia (Bizer et al. 2009), RDF triples are extracted from the Infobox templates within Wikipedia articles. In YAGO (Suchanek et al. 2007), an appropriate WordNet synset (most likely category) is assigned to a Wikipedia category as a super-category, and Wikipedia articles are extracted as instances of the category.

As a first step to make use of proper noun and related up-to-date information in Wikipedia, we focus on person names and the articles and categories related to them because it contains a large number of articles and categories that indicate person, and because large-scale person ontology is useful for applications such as person search and named entity recognition. Examples of a person article are personal name and occupational title such as "Ichiro" and "Financial planner," while an example of a person category is occupational title such as "Sportspeople."

The goal of this study is to construct a large-scale and comprehensive person ontology by extracting person categories and *is-a* relations[1] among them. We first apply a classifier based on machine learning to all Wikipedia categories to extract categories that represent person. If both of the linked Wikipedia categories are person categories, the category link is labeled as an *is-a* relation. We then use a heuristic-based rule to extract the title of articles that represent person as person instance from the person categories.

In the following sections, we first describe the language resources and the previous works. We then introduce our method for constructing the person ontology and report our experimental results.

[1] "*is-a* relation" is defined as a relation between A and B when "B is a (kind of) A."

Beijing, August 2010
Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources", Coling 2010, pages 1–9,

2 Language Resources

2.1 Japanese Wikipedia

Wikipedia is a free, multilingual, on-line encyclopedia that is being actively developed by a large number of volunteers. Wikipedia has articles and categories. The data is open to the public as XML files[2]. Figure 1 shows an example of an article. An article page has a title, body, and categories. In most articles, the first sentence of the body is the definition sentence of the title. Although the Wikipedia category system is organized in a hierarchal manner, it is a thematic classification, not a taxonomy. The relation between category and subcategory and that between a category and articles listed on it are not necessarily an *is-a* relation. A category could have two or more super categories and the category network could have loops.

Figure 1: Example of title, body (definition sentence), and categories for article page in Japanese Wikipedia (top) and its translation (bottom)

2.2 Nihongo Goi-Taikei

To construct the ontology, we first apply a machine learning based classifier to determine if a category label indicates a person or not. A Wikipedia category label is often a common compound noun or a noun phrase, and the head word of a Japanese compound noun and noun phrase is usually the last word. We assume the semantic category of the last word is an important feature for classification.

Nihongo Goi-Taikei (hereafter, *Goi-Taikei*) is one of the largest and best known Japanese thesauri. *Goi-Taikei* contains different semantic category hierarchies for common nouns, proper nouns, and verbs. In this work, we use only the common noun category (Figure 2). It consists of approximately 100,000 Japanese words (hereafter, instance) and the meanings of each word are described by using about 2,700 hierarchical semantic categories. Words (Instances) with multiple meanings (ambiguous words) are assigned multiple categories in *Goi-Taikei*. For example, the transliterated Japanese word (instance) *raita* (ライター) has two meanings of "writer" and "lighter," and so belongs to two categories, "353:author[3]" and "915:household."

Japanese WordNet (approximately 90,000 entries as of May 2010), which has recently been released to the public (Bonds et al., 2008), could be an alternative to *Goi-Taikei* as a large-scale Japanese thesaurus. We used *Goi-Taikei* in this work because Japanese WordNet was translated from English WordNet and it is not known whether it covers the concepts unique to Japanese.

3 Previous Works

3.1 Ponzetto's method and Sakurai's method

Ponzetto et al. (2007) presented a set of lightweight heuristics such as head matching and modifier matching for distinguishing *is-a* links from *not-is-a* links in the Wikipedia category network. The main heuristic, "Syntax-based methods" is based on head matching, in which a category link is labeled as *is-a* relation if the two categories share the same head lemma, such as CAPITALS IN ASIA and CAPITALS. Sakurai et al. (2008) presented a method equivalent to head matching for Japanese Wikipedia. As Japanese is a head final language, they introduced the heuristic called *suffix matching*; it labels a category link as a *is-a* relation if one category is the suffix of the other category, such as 日本の空港(airports in Japan) and 空港(airports). In the proposed method herein, if a Wikipedia category and its parent category are both person categories, the category link is labeled as *is-a* relation. Therefore, *is-a* relations, which cannot be extracted by Ponzetto's or Sakurai's method, can be extracted.

[2]http://download.wikimedia.org/jawiki

[3] The *Goi-Taikei* category is prefixed with ID number.

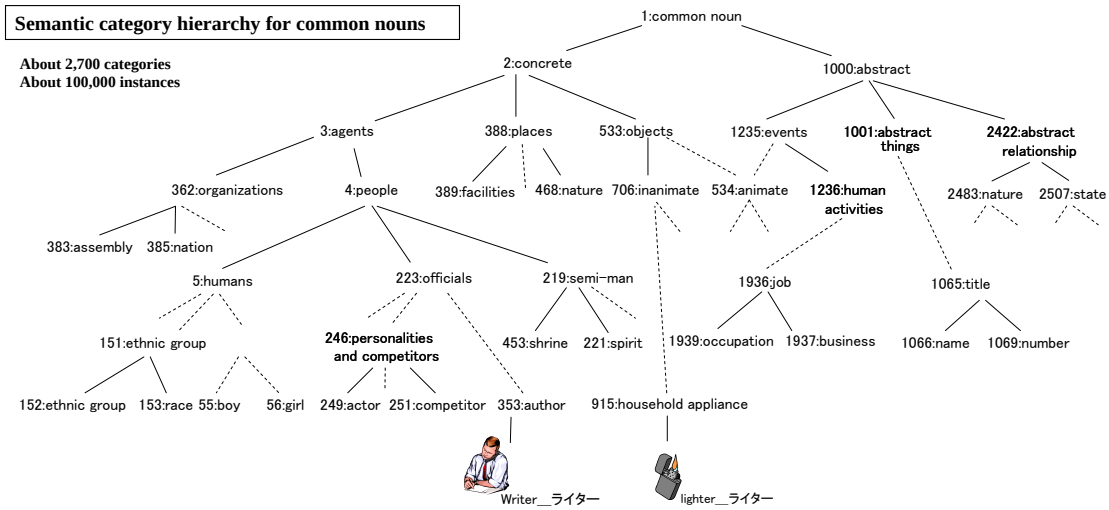

Figure 2: Part of a category hierarchy for common nouns in *Nihongo Goi-Taikei*

3.2 Kobayashi's method

Kobayashi et al. (2008) presented a technique to make a Japanese ontology equivalent to YAGO; it assigns *Goi-Taikei* categories to Japanese Wikipedia categories. These two methods and our method are similar in that a Wikipedia category and the title of an article are regarded as a category and an instance, respectively. Kobayashi et al. automatically extract hypernyms from the definition sentence of each article in advance (referred to hereafter as "*D-hypernym*.") They apply language-dependent lexico-syntactic patterns to the definition sentence to extract the *D-hypernym*. Here are some examples.

は、[hypernym]の一つである <EOS>
one of [hypernym]

は、[hypernym]である<EOS>
is a [hypernym]

[hypernym] <EOS>
is a [hypernym] …

where <EOS> refer to the beginning of a sentence

For example, from the article in Figure 1, the words "ゴルフ選手 (golf player)" is extracted as the *D-hypernym* of the article "ミシェル・ウィー (Michelle Wie)."

Figure 3 outlines the Kobayashi's method. First, for a Wikipedia category, if its last word matches an instance of *Goi-Taikei* category, all such *Goi-Taikei* categories are extracted as a candidate of the Wikipedia category's super-class. If the last word of the *D-hypernym* of the Wikipedia article listed on the Wikipedia category matches an instance of the *Goi-Taikei* category, the *Goi-Taikei* category is extracted as the super-class of the Wikipedia category and its instances (Wikipedia articles) (Figure 3). Although the Kobayashi's method is a general one, it can be used to construct person ontology if the super-class candidates are restricted to those *Goi-Taikei* categories which represent person.

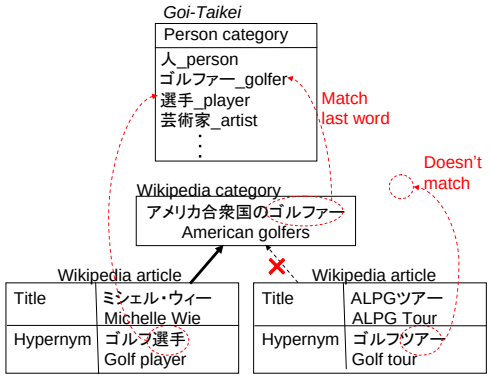

Figure 3: The outline of Kobayashi's method

3.3 Yamashita's method

Yamashita made an open source software which extracts personal names from Japanese Wikipedia[4]. He extracted the titles of articles listed on the categories ○年生(○ births) (e.g., 2000 births). As these categories are used to sort the names of people, horses, and dogs by born year, he used a simple pattern matching

[4]http://coderepos.org/share/browser/lang/perl/misc/wikipejago

rules to exclude horses and dogs. In the experiment in Section 5, we implemented his method by using not only "年生 (births)" but also "年没 (deaths)" and "世紀没 (th-century deaths)," "年代没 (s deaths)," "年代生 (s births)," and "世紀生 (th births)" to extract personal names. As far as we know, it is the only publicly available software to extract a large number of person names from the Japanese Wikipedia. For the comparison with our method, it should be noted that his method cannot extract person categories.

4 Ontology Building Method

4.1 Construction of Wikipedia person category hierarchy (WPC)

We extract the WPC by using a machine learning classifier. If a Wikipedia category and its parent category are both person categories, the category link is labeled as an *is-a* relation. This means that all *is-a* relations in our person ontology are extracted from the original Wikipedia category hierarchy using only a category classifier. This is because we investigated 1,000 randomly sampled links between person categories and found 98.7% of them were *is-a* relations. Figure 4 shows an example of the Wikipedia category hierarchy and the constructed WPC hierarchy.

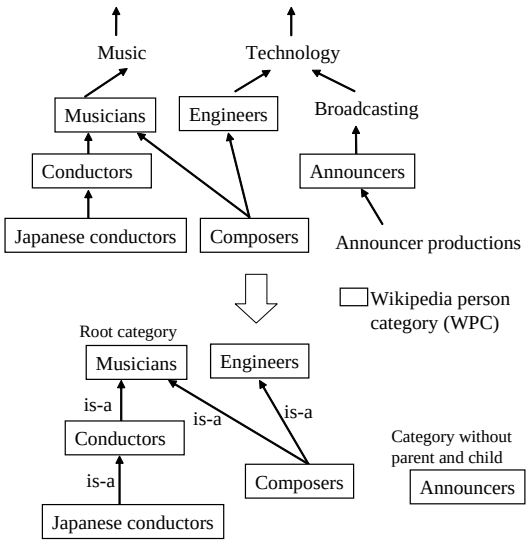

Figure 4: Example of Wikipedia category hierarchy (top) and constructed Wikipedia person category hierarchy (bottom)

We detect whether the Wikipedia category label represents a person by using Support Vector Machine (SVM). The semantic category of the words in the Wikipedia category label and those in the neighboring categories are used for the features. We use the following three aspects of the texts that exist around the target category for creating the features:

1. Structural relation between the target category and the text in Wikipedia. (6 kinds)
2. Span of the text. (2 kinds)
3. Semantic category of the text derived from *Goi-Taikei*. (4 kinds)

We examined 48 features by combining the above three aspects (6*2*4).

The following are the six structural relations in Wikipedia between the target category and the text information:

Structural relation

A. The target Wikipedia category label.

B. All parent category labels of the target category.

C. All child category labels of the target category.

D. All sibling category labels of the target category.

E. All *D-hypernym*[5] from each article listed on the target category.

F. All *D-hypernyms* extracted from the articles with the same name as the target category.

As for F, for example, when the article ベーシスト(bassist) is listed on the category: ベーシスト (bassist), we regard the *D-hypernym* of the article as the hypernym of the category.

As most category labels and *D-hypernyms* are common nouns, they are likely to match instances in *Goi-Taikei* which lists possible semantic categories of words.

[5]As for *D-hypernym* extraction patterns, we used almost the same patterns described in previous works on Japanese sources such as (Kobayashi et al. 2008; Sumida et al., 2008), which are basically equivalent to the works on English sources such as (Hearst, 1992).

After the texts located at various structural relations A-F are collected, they are matched to the instances of *Goi-Taikei* in two different spans:

Span of the text

Ⅰ. All character strings of the text

Ⅱ. The last word of the text

For the span Ⅱ, the text is segmented into words using a Japanese morphological analyzer. The last word is used because the last word usually represents the meaning of the entire noun phrase (semantic head word) in Japanese.

In the proposed method, hierarchical semantic categories of *Goi-Taikei* are divided into two categories; "*Goi-Taikei* person categories" and other categories. *Goi-Taikei* person category is defined as those categories that represent person, that is, all categories under "5:humans" and "223:officials," and "1939: occupation" and "1066:name" in *Goi-Taikei* hierarchy as shown in Figure 1.

For each structural relation A-F and span Ⅰ and Ⅱ, we calculate four relative frequencies a-d, which represents the manner in which the span of texts match the instance of *Goi-Taikei* person category. It basically indicates the degree to which the span of text is likely to mean a person.

Semantic type

a. The span of text matches only instances of *Goi-Taikei* person categories.

b. The span of text matches only instances of categories other than *Goi-Taikei* person categories.

c. The span of text matches both instances of *Goi-Taikei* person categories and those of other categories.

d. The span of text does not match any instances of *Goi-Taikei*.

For example, when the target category is "音楽家" (musicians) in Figure 5 and the feature in question is B-Ⅱ (the last word of its parent categories), the word "家" (whose senses are family and house) falls into semantic type c, and the word "音楽" (music) falls into semantic type b. Therefore, the frequency of semantic types a, b, c, d are 0, 1, 1, 0, respectively, in the features related to B-Ⅱ, and the relative frequencies used for the feature value related B-Ⅱ are 0, 0.5, 0.5, 0, respectively. In this way, we use 48 relative frequencies calculated from the combinations of structural relation A-F, span Ⅰ and Ⅱ, and semantic type a-d, as the feature vector for the SVM.

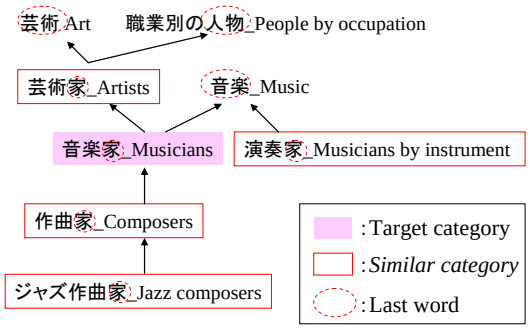

Figure 5: Example of Wikipedia category hierarchy when the target category is "音楽家"

4.2 *Similar category*

In Wikipedia, there are categories that do not have articles and those with few neighboring categories. Here, we define the neighboring categories for a category as those categories that can be reached through a few links from the category. In these cases, there is a possibility that there is not enough text information from which features (mainly semantic category of words) can be extracted, which could degrade the accuracy.

The proposed method overcomes this problem by detecting categories similar to the target category (the category in question) from its neighboring categories for extracting sufficient features to perform classification. Here, "s*imilar category*" is defined as parent, child, and sibling categories whose last word matches the last word of the target category. This is because there is a high possibility that the *similar categories* and the target category have similar meaning if they share the same last word in the category labels. If the parent (child) category is determined as a *similar category*, its parent (child) category is also determined as a *similar category* if the last word is the same. The procedure is repeated as long as they share the same last word.

Figure 5 shows an example of *similar categories* when the target category is "Musicians." In this case, features extracted from A-F of

similar categories are added to features extracted using A-F of the target category, "Musicians." For example, *similar category* "Artists" has "Art" and "People by occupation" as B (parent categories of the target category) in Figure 5, therefore "Art" and "People by occupation" are added to B of "Musicians."

4.3 Extracting Wikipedia person instance (WPI)

The proposed method extracts, as WPIs the titles of articles listed as WPCs that meet the following four requirements.

1. The last word of the *D-hypernym* of the title of the Wikipedia article matches an instance of *Goi-Taikei* person category.
2. The last word of the title of Wikipedia article matches an instance of Goi-Taike person category.
3. At least one of the Wikipedia categories assigned to the Wikipedia article matches the following patterns:

   (年没|世紀没|年代没|年生|世紀生|年代生)<EOS>
   (deaths | th-century deaths | 's deaths | births | th-births | 's births) <EOS>

   These categories are used to sort a large number of person names by year.

4. Wikipedia categories assigned to the Wikipedia article satisfy the following condition:

$$\frac{\text{Number of extracted WPCs in Section 4.1}}{\text{All number of Wikipedia categories}} > 0.5$$

   This condition is based on the observation that the more WPCs a Wikipedia article is assigned to, the more it is likely to be a WPI. We set the threshold 0.5 from the results of a preliminary experiment.

5 Experiments

5.1 Experimental setup

We used the XML file of the Japanese Wikipedia as of July 24, 2008. We removed irrelevant pages by using keywords (e.g., "image:," "Help:") in advance. This cleaning yielded 477,094 Wikipedia articles and 39,782 Wikipedia categories. We manually annotated each category to indicate whether it represents person (positive) or not (negative). For ambiguous cases, we used the following criteria:

- ＊Personal name by itself (e.g., Michael Jackson) is not regarded as WPC because usually it does not have instances. (Note: personal name as article title is regarded as WPI.)
- ＊Occupational title (e.g., Lawyers) is regarded as WPC because it represents a person.
- ＊Family (e.g., Brandenburg family) and Ethnic group (e.g., Sioux) are regarded as WPC.
- ＊Group name (e.g., The Beatles) is not regarded as WPC.

In order to develop a person category classifier, we randomly selected 2,000 Wikipedia categories (positive:435, negative:1,565) from all categories for training[6]. We used the remaining 37,767 categories for evaluation. To evaluate WPI extraction accuracy, we used Wikipedia articles not listed on the Wikipedia categories used for training. 417,476 Wikipedia articles were used in the evaluation.

To evaluate our method, we used TinySVM-0.09[7] with a linear kernel for classification, and the Japanese morphological analyzer JUMAN-6.0[8] for word segmentation. The comparison methods are Kobayashi's method and Yamashita's method under the same conditions as our method.

5.2 Experimental results

Table 1 shows the WPCs extraction accuracy. Precision and recall of proposed method are 6.5 points and 14.8 points better than those of Kobayashi's method, respectively.

	Precision	Recall	F-measure
Kobayashi's method	92.8% (6727/7247)	83.6% (6727/8050)	88.0%
Proposed method	**99.3%** (7922/7979)	**98.4%** (7922/8050)	**98.8%**

Table 1: The Wikipedia person categories (WPCs) extraction accuracy

[6]We confirmed that the accuracy will level off about 2, 000 training data by experiment. Details will be described in Section 6.
[7]http://chasen.org/~taku/software/TinySVM/
[8]http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

To confirm our assumption on the links between WPCs, we randomly selected 1,000 pairs of linked categories from extracted WPCs, and manually investigated whether both represented person and were linked by *is-a* relation. We found that precision of these pairs was 98.3%.

Errors occurred when the category link between person categories in the Wikipedia category network was not an *is-a* relation, such as 千葉氏(*Chiba* clan) – 大須賀氏(*Ohsuga* clan). However, this case is infrequent, because 98.7% of the links between person categories did exhibit an *is-a* relation (as described in Section 4.1).

Table 2 shows the WPIs extraction accuracy. We randomly selected 1,000 Wikipedia articles from all categories in Wikipedia, and manually created evaluation data (positive:281, negative:719). The recall of the proposed method was 98.6%, 21.0 points higher than that of Yamashita's method. Our method topped the F-measure of Kobayashi's method by 3.4 points. Among 118,552 extracted as WPIs by our method, 116,418 articles were expected be correct. In our method, errors occurred when WPI was not listed on any WPCs. However, this case is very rare. Person instances are almost always assigned to at least one WPC. Thus, we can achieve high coverage for WPIs even if we focus only on WPCs. We randomly selected 1,000 articles from all articles and obtained 277 person instances by a manual evaluation. Furthermore, we investigated the 277 person instances, and found that only two instances were not classified into any WPCs (0.7%).

	Precision	Recall	F-measure
Yamashita's method	**100.0%** (218/218)	77.6% (218/281)	87.4%
Kobayashi's method	96% (264/275)	94.0% (264/281)	95.0%
Proposed method	98.2% (277/282)	**98.6%** (277/281)	**98.4%**

Table 2: The Wikipedia person instance (WPIs) extraction accuracy

Table 3 shows the extracted WPC-WPI pairs (e.g., American golfers-Michelle Wie, Artists-Meritorious Artist) extraction accuracy. We randomly selected 1,000 pairs of Wikipedia category and Wikipedia article from all such pairs in Wikipedia, and manually investigated whether both category and article represented a person and whether they were linked by an *is-a* relation (positive:296, negative:704). Precision and recall of proposed method are 2.1 points and 11.8 points higher than those of Kobayashi's method, respectively. Among all 274,728 extracted as WPC-WPI pairs by our method, 269,233 was expected be correct.

	Precision	Recall	F-measure
Kobayashi's method	95.9% (259/270)	87.5% (259/296)	91.5%
Proposed method	**98.0%** (294/300)	**99.3%** (294/296)	**98.7%**

Table 3: The extraction accuracy of the pairs of Wikipedia person category and person instance (WPC-WPI)

6 Discussions

We constructed a WPC hierarchy using the 8,357 categories created by combining extracted categories and training categories. The resulting WPC hierarchy has 224 root categories (Figure 4). Although the majority of the constructed ontology is interconnected, 194 person categories had no parent or child (2.3 % of all person categories). In rare cases, the category network has loops (e.g., "Historians" and "Scholars of history" are mutually interlinked).

Shibaki et al. (2009) presented a method for building a Japanese ontology from Wikipedia using *Goi-Taikei*, as its upper ontology. This method can create a single connected taxonomy with a single root category. We also hope to create a large-scale, single-root, and interconnected person ontology by using some upper ontology.

Our method is able to extract WPCs that do not match any *Goi-Taikei* instance (e.g., Violinists and Animators). Furthermore, our method is able to detect many ambiguous Wikipedia category labels correctly as person category. For example, "ファッションモデル (fashion model)" is ambiguous because the last word "モデル (model)" is ambiguous among three senses: person, artificial object, and abstract relation. Kobayashi's method cannot extract a WPC if the last word of the category label does not match any instance in *Goi-Taikei*. Their method is error-prone if the last word has mul-

tiple senses in *Goi-Taikei* because it is based on simple pattern matching. Our method can handle unknown and ambiguous category labels since it uses machine learning-based classifiers whose features are extracted from neighboring categories.

Our method can extract *is-a* person category pairs that could not be extracted by Ponzetto et al. (2007) and Sakurai et al. (2008). Their methods use head matching in which a category link is labeled as an *is-a* relation only if the head words of category labels are matched. However, our method can extract *is-a* relations without reference to surface character strings, such as "ジャーナリスト(Journalists)" and "スポーツライター(Sports writers)." Among all 14,408 Wikipedia category pairs extracted as *is-a* relations in our method, 5,558 (38.6%) did not match their head words.

We investigated the learning curve of the machine learning-based classifier for extracting WPCs, in order to decide the appropriate amount of training data for future updates.

As we have already manually tagged all 39,767 Wikipedia categories, we randomly selected 30,000 categories and investigated the performance of our method when the number of the training data was changed from 1,000 to 30,000. The evaluation data was the remaining 9,767 categories.

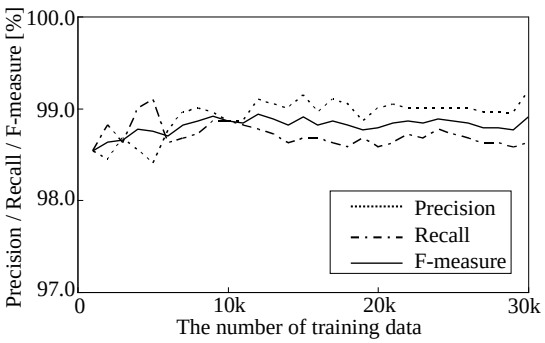

Figure 6: The effect of training data size to WPC extraction accuracy

Figure 6 shows the precision, recall, and F-measure for different training data sizes. F-measure differed only 0.4 points from 1,000 samples (98.5%) to 30,000 samples (98.9%). Figure 6 shows that the proposed method offers high accuracy in detecting WPCs with only a few thousand training examples.

Our method uses *similar categories* for creating features as well as the target Wikipedia category (Section 4.1). We compared the proposed method to a variant that does not use *similar categories* to confirm the effectiveness of this technique. Furthermore, our method uses the Japanese thesaurus, *Goi-Taikei*, to look up the semantic category of the words for creating the features for machine learning. We also compared the proposed method with the one that does not use semantic category (derived from *Goi-Taikei*) but instead uses word surface form for creating features (This one uses *similar categories*).

Figure 7 shows the performance of the classifiers for each type of features. We can clearly observe that using *similar categories* results in higher F-measure, regardless of the training data size. We also observe that when there is little training data, the method using word surface form as features results in drastically lower F-measures. In addition, its accuracy was consistently lower than the others even if the training data size was increased. Therefore, we can conclude that using *similar category* and *Goi-Taikei* are very important for creating good features for classification.

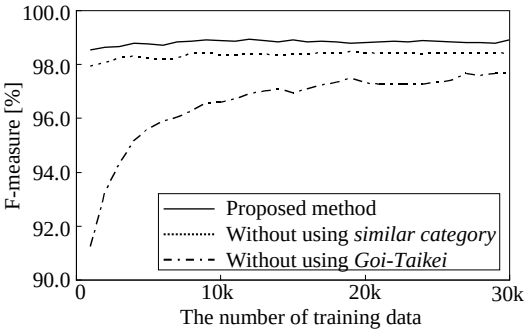

Figure 7: The effects of using *similar categories* and *Goi-Taikei*

In future, we will attempt to apply our method to other Wikipedia domains, such as organizations and products. We will also attempt to use other Japanese thesauri, such as Japanese WordNet. Furthermore, we hope to create a large-scale and single connected ontology. As a final note, we plan to open the person ontology constructed in this paper to the public on Web in the near future.

References

Bizer, C., J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. 2009. "DBpedia - A crystallization point for the web of data," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 7, No.3, pages 154-165.

Bond, Francis, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Boot-strapping a wordnet using multiple existing wordnets. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 28-30.

Fellbaum, Christiane. 1998. WordNet: An Electronic Lexical Database, Language, Speech, and Communication Series. MIT Press.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th Conference on Computational Linguistics (COLING), pages 539-545.

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai,Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi, editors. 1997. *Nihongo Goi-Taikei – a Japanese Lexicon*. Iwanami Shoten. (in Japanese).

Kobayashi, Akio, Shigeru Masuyama, and Satoshi Sekine. 2008. A method for automatic construction of general ontology merging goitaikei and Japanese Wikipedia. In *Information Processing Society of Japan (IPSJ) SIG Technical Report2008-NL-187 (in Japanese)*, pages 7-14.

Ponzetto, S. P. and Michael Strube. 2007. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence (AAAI)*, pages 1440–1445.

Sakurai, Shinya, Takuya Tejima, Masayuki Ishikawa, Takeshi Morita, Noriaki Izumi, and Takahira Yamaguchi. 2008. Applying Japanese Wikipedia for building up a general ontology. In *Japanese Society of Artificial Intelligence (JSAI) Technical Report SIG-SWO-A801-06 (in Japanese)*, pages 1-8.

Shibaki, Yumi, Masaaki Nagata and Kazuhide Yamamoto. 2009. Construction of General Ontology from Wikipedia using a Large-Scale Japanese Thesaurus. In *Information Processing Society of Japan (IPSJ) SIG Technical Report2009-NL-194-4. (in Japanese).*

Suchanek, Fabian M., Gjergji Kasneci, and GerhardWeikum. 2007. Yago: A core of semantic knowledge unifying wordnet and Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pages 697-706.

Sumida, Asuka, Naoki Yoshinaga, and Kentaro Torisawa. 2008. Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in Wikipedia. In *Proceedings of the Sixth Language Resources and Evaluation Conference(LREC)*, pages 28–30.