Using Goi-Taikei as an Upper Ontology to Build a Large-Scale Japanese Ontology from Wikipedia

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

Yumi Shibaki and **Kazuhide Yamamoto**
Nagaoka University of Technology
{shibaki,yamamoto}@jnlp.org

Abstract

We present a novel method for building a large-scale Japanese ontology from Wikipedia using one of the largest Japanese thesauri, *Nihongo Goi-Taikei* (referred to hereafter as "Goi-Taikei") as an upper ontology. First, The leaf categories in the Goi-Taikei hierarchy are semi-automatically aligned with semantically equivalent Wikipedia categories. Then, their subcategories are created automatically by detecting *is-a* links in the Wikipedia category network below the junction using the knowledge defined in Goi-Taikei above the junction. The resulting ontology has a well-defined taxonomy in the upper level and a fine-grained taxonomy in the lower level with a large number of up-to-date instances. A sample evaluation shows that the precisions of the extracted categories and instances are 92.8% and 98.6%, respectively.

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. One promising approach involves automatically extracting a large comprehensive ontology from Wikipedia, a freely available online encyclopedia with a wide variety of information. One problem with previous such efforts is that the resulting ontology is either fragmentary or trivial.

Ponzetto and Strube (2007) presents a set of lightweight heuristics such as *head matching* and *modifier matching* for distinguishing between *is-a* and *not-is-a* links in the Wikipedia category network. The most powerful heuristics is head matching in which a category link is labeled as *is-a* if the two categories share the same head lemma, such as CAPITALS IN ASIA and CAPITALS. For Japanese, Sakurai et al. (2008) present a method equivalent to head matching in Japanese. As Japanese is a head final language, they introduced a heuristics called *suffix matching* in which a category link is labeled as *is-a* if one category is the suffix of the other category, such as 日本の空港 (airports in Japan) and 空港 (airports). The problem with the ontology extracted by these two methods is that it is not a single interconnected taxonomy, but a set of taxonomic trees.

One way to make a single taxonomy is to use an existing large-scale taxonomy as a core for the resulting ontology. In YAGO, Suchanek et al. (2007) merged English WordNet and Wikipedia by adding instances (namely Wikipedia articles) to the *is-a* hierarchy of WordNet. Of the categories assigned to a Wikipedia article, they regarded one with a plural head noun as the article's hypernym, which is called a *conceptual category*. They then linked the conceptual category to a WordNet synset by heuristic rules including head matching. For Japanese, Kobayashi et al. (2008) present an attempt equivalent to YAGO, where they merged Goi-Taikei and Japanese Wikipedia. The problem with these two methods is that the core taxonomy is extended only one level although many new instances are added. They cannot make the most of the fine-grained taxonomic

Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010), pages 11–18,
Beijing, August 2010

information contained in the Wikipedia category network.

In this paper, we present a novel method for building a single interconnected ontology from Wikipedia, with a fine-grained taxonomy in the lower level, by using a manually constructed thesaurus as its upper ontology. In the following sections, we first describe the language resources used in this work. We then describe a semi-automatic method for building the ontology and report our experimental results.

2 Language Resources

2.1 *Nihongo Goi-Taikei*

Nihongo Goi-Taikei (日本語語彙大系, 'comprehensive outline of Japanese vocabulary')[1] is one of the largest and best known Japanese thesauri (Ikehara et al., 1997). It was originally developed as a dictionary for a Japanese-to-English machine translation system in the early 90's. It was then published as a book in 5 volumes in 1997 and as a CD-ROM in 1999. It contains about 300,000 Japanese words and the meanings of each word are described by using 2,715 hierarchical semantic categories. Each word has up to 5 semantic categories in order of frequency in use, and each category is assigned with a unique ID number and category name such as 4:person and 388:place[2].

Goi-Taikei has different semantic category hierarchies for common nouns, proper nouns, and verbs, respectively. We used only the common noun category in this work. For simplicity, we mapped all proper nouns in the proper noun category to the equivalent common noun category using the category mapping table shown in the Goi-Taikei book.

Figure 1 shows the top three layers for common nouns[3]. For example, the transliterated Japanese word *raita* (ライター) has two semantic categories 353:author and 915:household appliance. The former originates with the English word "writer" while the latter originates with English word "lighter". By climbing up the Goi-Taikei category hierarchy, we can infer that the former refers to a human being (4:person) while the latter refers to a physical object (533:concrete object).

2.2 Japanese Wikipedia

Wikipedia is a free, multilingual, on-line encyclopedia actively developed by a large number of volunteers. Japanese Wikipedia now has about 500,000 articles. Figure 2 shows examples of an article page and a category page. An article page has a title, body, and categories. In most articles, the first sentence of the body gives the definition of the title. A category also has a title, body, and categories. Its title is prefixed with "Category:" and its body includes a list of articles that belong to the category.

Although the Wikipedia category system is organized in a hierarchal manner, it is not a taxonomy but a thematic classification. An article could belong to many categories and the category network has loops. The relations between linked categories are chaotic, but the lower the category link is in the hierarchy, the more it is likely to be an *is-a* relation. For example, the category link between カクテル (COCKTAIL) and 酒 (ALCOHOLIC BEVERAGE) is an *is-a* relation. Although the article シェイカー (shaker) is in the category カクテル (COCKTAIL), a shaker is not a cocktail but an appliance. Extracting a taxonomy from the Wikipedia category network is not trivial.

3 Ontology Building Method

Figure 3 shows an outline of the proposed ontology building method. We first semi-automatically align each leaf category in the Goi-Taikei category hierarchy with one or more Wikipedia categories. We call a Wikipedia category aligned with a Goi-Taikei category a *junction category*. We then extend each Goi-Taikei leaf category by detecting the *is-a* links below the junction category in the Wikipedia category network using the knowledge defined above the junction category in Goi-Taikei

.

[1] Referred to as "Goi-Taikei" unless otherwise noted.

[2] We use Sans Serif for the Goi-Taikei category and SMALL CAPS for the Wikipedia category. The Goi-Taikei category is prefixed with ID number.

[3] The maximum depth of the common noun hierarchy is 12. Most links are *is-a* relations, but some are *part-of* relations, which are explicitly marked

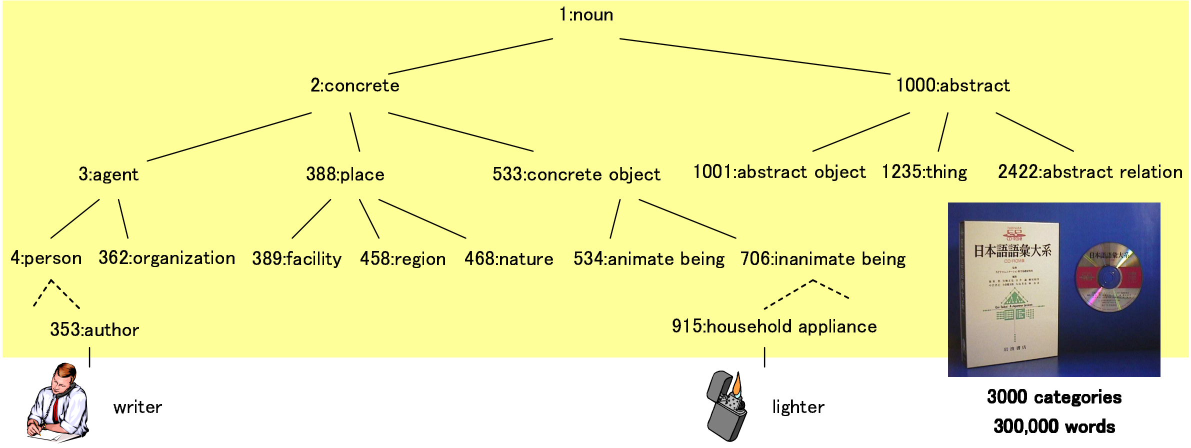

Figure 1: Top three layers of the common noun semantic category hierarchy in *Nihongo Goi-Taikei*

<title> カクテル </title> カクテル (英語:Cocktail) とは、主にベースとなる酒に、他の酒またはジュースなどを混ぜて作るアルコール飲料 . . . <Category> カクテル </Category>	<title>cocktail</title> A cocktail (English:Cocktail) is an alcoholic beverage made by mixing a base liquor with other liquor or juice. . . . <Category>cocktail</Category>
<title>Category:カクテル </title> [[カクテル]] に関するカテゴリ . . . <Category> 酒 </Category>	<title>Category:Cocktails</title> Category on [[cocktails]] . . . <Category>alcoholic beverages</Category>

Figure 2: Examples of title, body (definition sentence), and category for article page and category page in Japanese Wikipedia (left) and their translation (right)

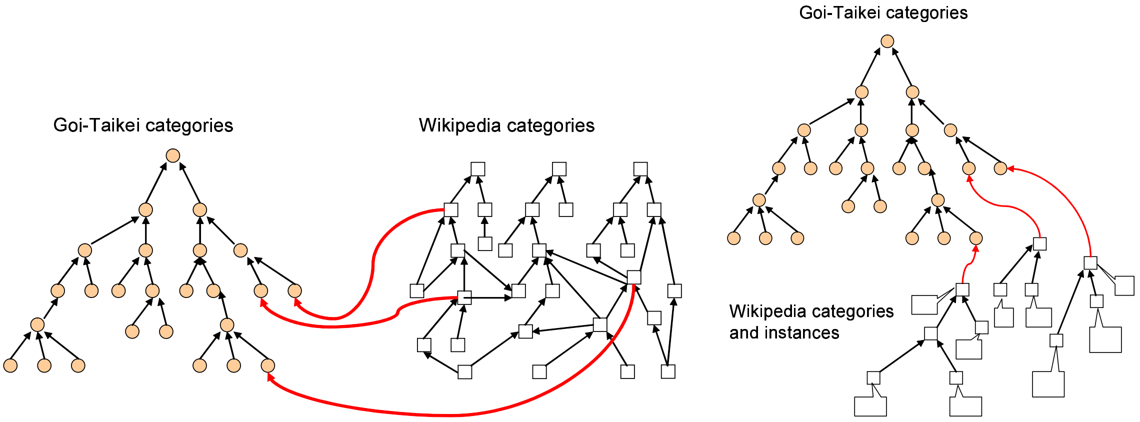

Figure 3: The ontology building method: First, Goi-Taikei leaf categories are aligned with Wikipedia categories (left), then each leaf category is extended by detecting *is-a* links in Wikipedia (right).

3.1 Category Alignment

For each leaf category in Goi-Taikei, we first make a list of junction category candidates. Wikipedia categories satisfying at least one of the following three conditions are extracted as candidates:

- The Goi-Taikei category name exactly matches the Wikipedia category name.
- One of the instances of the Goi-Taikei category exactly matches the Wikipedia category name.
- More than two instances of the Goi-Taikei category exactly match either instances or subcategories of the Wikipedia category.

Here, an instance of a Goi-Taikei category refers to words belonging to the Goi-Taikei category while that of a Wikipedia category refers to the title (name) of articles belonging to the Wikipedia category.

If a Goi-Taikei category and a Wikipedia category refer to the same concept, we regard them as semantically equivalent. If an instance of a Goi-Taikei category and a Wikipedia category refer to the same concept, we regard the name of the Goi-Taikei instance as a subcategory of the Goi-Taikei category and regard the subcategory and the Wikipedia category as semantically equivalent.

This is a sort of word sense disambiguation problem. For example, Wikipedia category ロケット (ROCKET) exactly matches the word ロケット in Goi-Taikei, which has two semantic categories, 990:aircraft (rocket) and 834:accessories (locket). Only the 990:aircraft sense of the word in Goi-Taikei matches the Wikipedia category.

We performed manual alignment because the accuracy of this category alignment is very important as regards the subsequent steps. Manual alignment is feasible and cost effective since there are only 1,921 leaves in the Goi-Taikei category hierarchy. However, we also report the result of automatic alignment in the experiment.

3.2 Hypernym Extraction

As preparation for detecting *is-a* links in the Wikipedia category network, we automatically extract a *hypernym* of the name of each article and category in advance.

We regard the first sentence of each article page as the definition of the concept referred to by the title. We applied language dependent lexico-syntactic patterns to the definition sentence to extract the hypernym. The hypernym of the category name is extracted from the definition sentence if it exists. If there is an article whose title is the same as its category, the hypernym of the article is used as that of the category.

As for lexico-syntactic patterns, we used almost the same patterns described in previous work related to Japanese such as (Kobayashi et al., 2008; Sumida et al., 2008), which is basically equivalent to work related to English such as (Hearst, 1992). Here are some examples.

[hypernym] の (一つ | 一種 | 名称 |. . .)
(one|kind|name|. . .) of [hypernym]

[hypernym](をいい | である |. . .)
(is_a|refers_to|. . .) [hypernym]

[hypernym]<EOS>
<BOS>[hypernym]

where <BOS> and <EOS> refer to the beginning and the end of a sentence.

For example, from the first article in Figure 2, the words アルコール飲料 (alcoholic beverage) are extracted as the hypernym of the article カクテル (cocktail), using the third lexico-syntactic pattern above. Since the title of the article is the same as the category name, アルコール飲料 (alcoholic beverage) is regarded as the hypernym of the category カクテル (COCKTAIL).

3.3 *Is-a* Link Detection

We automatically detect *is-a* links in the Wikipedia category network to extend the original Goi-Taikei category hierarchy. Starting from a junction category, we recursively traverse the Wikipedia category network if the link from the current category to the child category is regarded as an *is-a* link.

We regard a link between a parent category and a child category as an *is-a* link if the suffix of the child category name matches one of the *hypernym*

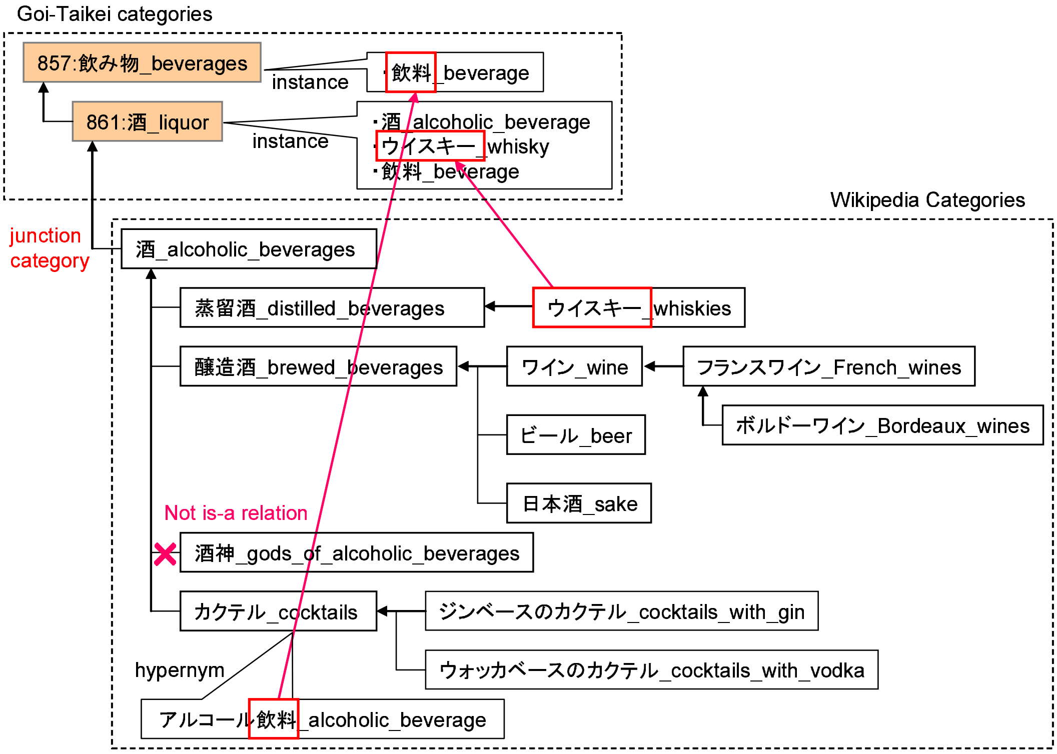

Figure 4: Extending Goi-Taikei leaf categories using the Wikipedia category network

candidates for the child category. We define the hypernym candidates for a category as the union of the following words:

- The names of three super categories in Goi-Taikei from the junction category, namely the leaf category, its parent, and its grandparent.
- All instance names belonging to the above three categories in Goi-Taikei.
- The names of all super categories in Wikipedia from the current category to the junction category.

We also regard a link as being *is-a* if the suffix of the hypernym (defined in Sec 3.2) of the child category name matches one of the hypernym candidates for the child category.

Figure 4 shows examples. The link between the category 蒸留酒 (DISTILLED BEVERAGES) and the category ウイスキー (WHISKIES) in Wikipedia is regarded as *is-a* because the word ウイスキー (whisky) is an instance of Goi-Taikei category 861:liquor just above the junction category 酒 (ALCOHOLIC BEVERAGES). The link between the category 酒 ALCOHOLIC BEVERAGES and the category カクテル (COCKTAILS) in Wikipedia is regarded as *is-a* because the suffix of アルコール飲料 (alcoholic beverage), the hypernym of the category カクテル (COCKTAILS), matches 飲料 (beverage), an instance of the category 857:beverages in Goi-Taikei. However, the link between the category 酒 (ALCOHOLIC BEVERAGES) and the category 酒神 (GODS OF ALCOHOLIC BEVERAGES) in Wikipedia is not *is-a* because the two Japanese strings do not have a common suffix.

3.4 Instance Extraction

For each Wikipedia category included in the *is-a* hierarchy constructed by the procedure described in the previous subsection, we extract the title of Wikipedia articles listed on the category page as an instance. The instance extraction method is basically the same for *is-a* category detection. We regard the link between a category and an article

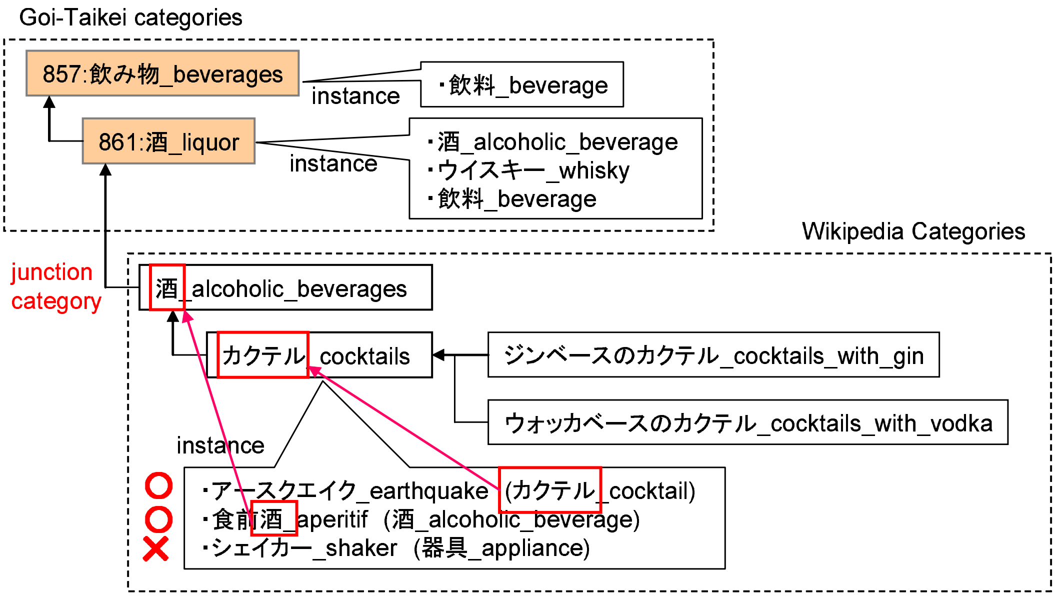

Figure 5: Extracting instances from Wikipedia category pages

as *is-a* if the suffix of either the article name or its hypernym (defined in Sec 3.2) matches one of the hypernym candidates (defined in Sec 3.3) of the article.

Figure 5 (a) shows examples. The link between the article アースクエイク (earthquake) and the category カクテル (COCKTAILS) is *is-a* because カクテル (cocktail), the hypernym of the article name アースクエイク (earthquake), exactly matches the parent category name. The link between the article 食前酒 (aperitif) and the category カクテル (COCKTAILS) is *is-a* because the suffix of 食前酒 (aperitif) matches the junction category 酒 (ALCOHOLIC BEVERAGES). The link between the article シェイカー (shaker) and the category カクテル (COCKTAILS) is not *is-a* because neither the suffix of the category name シェイカー (shaker) nor that of its hypernym 器具 (appliance) matches any hypernym candidates of the article シェイカー (shaker).

4 Experimental Result and Discussion

4.1 Category Alignment

We used the XML file of the Japanese Wikipedia as of July 24, 2008[4]. There are 49,543 category pages and 479,231 article pages in the file. For each of the 1,921 Goi-Taikei leaf categories with the total of 108,247 instances, we applied the three conditions described in Sec 3.1 and obtained 6,301 Wikipedia categories as junction category candidates. We then manually selected 2,477 categories as the junction categories. The number of Goi-Taikei leaf categories with one or more junction categories is 719 (719/1921=38.4%).

We performed some preliminary experiments on the automatic selection of junction categories. We trained an SVM classifier using the above junction category candidates and manual judgement results. Given a pair consisting of a Goi-Taikei category and a Wikipedia category, the SVM classifier predicts whether or not the two categories should be aligned. We used standard ontology mapping features (Euzenat and Shavaiko, 2007) such as whether the (class|instance) name of the (self|parent|children|siblings) match one and the other. We undertook a fivefold cross validation and obtained about 90% precision and 70% recall. The results were encouraging but we decided to use the manual alignment results for subsequent experiments.

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

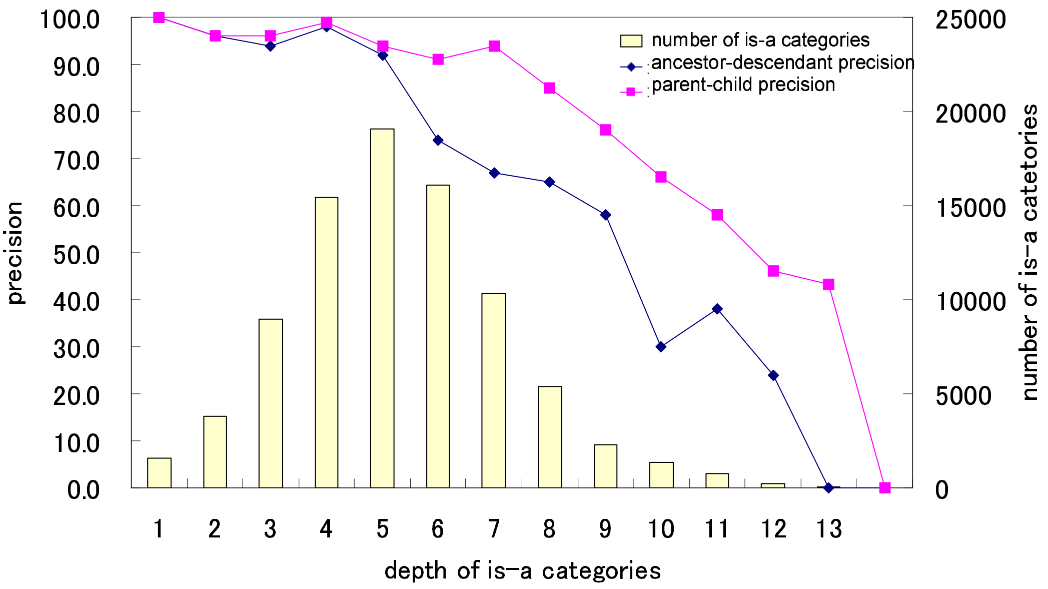

Figure 6: The precision of *is-a* links classified by the depth in the constructed category hierarchy

4.2 *Is-a* Link Detection

We extracted 23,289 categories from 49,543 categories in Wikipedia (47%) to extend the Goi-Taikei category hierarchy. We evaluated the *Is-a* link detection accuracy for the Wikipedia category network by employing the following two criteria:

- parent-child precision: whether the link between the current category and its immediate parent is an *is-a* relation.
- ancestor-descendant precision: whether all the links from the current category to the root are *is-a* relations.

We randomly selected 100 categories at each depth from the constructed hierarchy and manually evaluated the parent-child precision and the ancestor-descendant precision. Figure 6 shows the precisions of *is-a* links classified by the depth in the constructed category hierarchy. It also shows the number of categories at each depth.

The parent-child precision is more than 90% from depths 1 to 7, while the ancestor-descendant precision is more than 90% froms depth 1 to 5. After excluding depth 1 categories (junction categories whose precision is 100%), the average parent-child precision is 92.8% and the average ancestor-descendant precision is 82.6%.

4.3 Instance Extraction

We extracted 263,631 articles from 479,231 articles in Wikipedia (55%) as instances of the constructed category hierarchy. The category with the largest number of instances is 日本の俳優 (JAPANESE ACTORS) with 5,632 instances. The average number of instances for a category is 17.8.

We evaluate the accuracy of instance extraction as follows: For each category in the constructed hierarchy, we list all its articles, and construct a pair consisting of a category and an article. We randomly sample these pairs and leave only the pairs in which all the links from its category to the root are *is-a* relations by manual inspection. For 319 category-article pairs obtained by this procedure, 247 articles are manually classified as instances of the category, while 208 articles are automatically classified as instances. The intersection of the two is 205. Thus, the precision and recall of instance extraction are 98.6%(205/208) and 83.0%(205/247), respectively.

4.4 Comparison to Previous Methods

Sakurai et al. (2008) reported the parent-child precision of their suffix matching-based method was 91.2% and 6,672 Wikipedia categories are used to construct their (fragmentary) hierarchy. We used a much larger set of Wikipedia categories (23,239) to extend the Goi-Taikei to form a single unified hierarchy with a comparable parent-child precision (92.8%). Kobayashi et al. (2008) reported their alignment accuracy (parent-child precision) was 93% and 19,426 Wikipedia categories are directly aligned with Goi-Taikei categories. We used a significantly larger set of Wikipedia categories (19426/23239=0.84) to extend the Goi-Taikei with retaining the *is-a* relations included in the Wikipedia category network.

5 Conclusion

In this paper, we presented a method for building a large-scale, Japanese ontology from Wikipedia using one of the most popular Japanese thesauri, *Nihongo Goi-Taikei*, as its upper ontology. Unlike previous methods, it can create a single connected taxonomy with a well-defined upper level taxonomy inherited from Goi-Taikei, as well as a fined-grained and up-to-date lower level taxonomy with broad-coverage extracted from Wikipedia.

Future work will include automatic category alignment between Goi-Taikei and Wikipedia to fully automate the ontology building. It would be interesting to use another Japanese thesaurus, such as the recently released Japanese WordNet (Bond et al., 2008), as an upper ontology for the proposed method.

One of the problems with the proposed method is that it only uses about half of the knowledge (categories and articles) in Wikipedia. This is because we restricted the alignment points in Goi-Taikei category hierarchy to its leaves. In Ponzetto and Navigli (2009), they present a method for aligning WordNet and Wikipedia fully at many levels with both of them retaining a hierarchal structure. However, their method does not integrate the two hierarchies into a single taxonomy. We think that developing a method for merging the two hierarchies into one taxonomy is the key to extracting more information from Wikipedia.

References

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.

Euzenat, Jérôme and Pavel Shavaiko. 2007. *Ontology Matching*. Springer.

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 goi-taikei and japanese wikipedia. In *Information Processing Society of Japan (IPSJ) SIG Technical Report 2008-NL-187 (in Japanese)*, pages 7–14.

Ponzetto, Simone Paolo and Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating wikipedia. In *Proceedings of the 21st International Joint Conference of Artificial Intelligence (IJCAI)*, pages 2083–2088.

Ponzetto, Simone Paolo 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.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. 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.