Wikipedia Mining for Triple Extraction Enhanced by Co-reference Resolution

Kotaro Nakayama

The Center for Knowledge Structuring
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
TEL: +81-3-5841-0462 FAX: +81-3-5841-0454
nakayama@cks.u-tokyo.ac.jp

Abstract. Since Wikipedia has become a huge scale database storing wide-range of human knowledge, it is a promising corpus for knowledge extraction. A considerable number of researches on Wikipedia mining have been conducted and the fact that Wikipedia is an invaluable corpus has been confirmed. Wikipedia's impressive characteristics are not limited to the scale, but also include the dense link structure, URI for word sense disambiguation, well structured Infoboxes, and the category tree. In previous researches on this area, the category tree has been widely used to extract semantic relations among concepts on Wikipedia. In this paper, we try to extract triples (Subject, Predicate, Object) from Wikipedia articles, another promising resource for knowledge extraction. We propose a practical method which integrates link structure mining and parsing to enhance the extraction accuracy. The proposed method consists of two technical novelties; two parsing strategies and a co-reference resolution method.

1 Introduction

Even though the importance of ontology construction is widely recognized and a considerable number of Semantic Web implementations based on standardized formats (such as RDF and OWL) are being built/published on the WWW, what seems lacking is the mapping of ontologies due to the nature of distributed environments. Since it is difficult to map local ontologies one by one, an approach based on the global ontology approach seems a solution having capability to intermediate local ontologies. However, previous methods for constructing huge scale ontologies faced technical difficulties, since it was impossible to manage a huge scale global ontology due to the lack of human resources.

Meanwhile, Wikipedia, a collaborative wiki-based encyclopedia, has become a phenomenon among Internet users. According to statistics of Nature, Wikipedia is about as accurate in covering scientific topics as the Encyclopedia Britannica. It covers concepts of various fields such as Arts, Geography, History, Science, Sports, Games. Wikipedia contains more than 2 million articles (Oct. 2007, English Wikipedia) and is becoming larger day by day while the largest paper-based encyclopedia Britannica contains only 65,000 articles. As a corpus for knowledge extraction, Wikipedia's impressive characteristics are not limited to the scale, but also include the dense link structure, sense disambiguation based on URL, brief link texts (a. k. a. anchor texts) and well structured sentences. The fact that these characteristics are valuable to extract accurate knowledge from Wikipedia is strongly confirmed by a number of previous researches on Wikipedia

Mining [1–8]. Besides, we proposed a scalable link structure mining method to extract a huge scale association thesaurus in a previous research [2]. In that research, we developed a huge scale association thesaurus dictionary extracting a list of related terms from any given term. Further, in a number of detailed experiments, we proved that the accuracy of our association thesaurus achieved notable results. However, association thesaurus construction is just the beginning of the next ambitious research on huge scale Web ontology construction from Wikipedia.

Semantic Wikipedia [9] is an impressive solution for developing a huge scale ontology on Wikipedia. Semantic Wikipedia is an extension of Wikipedia which allows editors to define semantic relations among concepts manually. Another major approach is to use Wikipedia's category tree as an ontology [7, 8]. These researchers proved that Wikipedia's categories are promising resources for ontology construction by showing significant results.

In contrast to these approaches, we propose a full-automated consistent approach for semantic relation extraction by mining Wikipedia articles. Since a Wikipedia article is a set of definitive sentences, the article text is yet another valuable resource for ontology construction. However, co-reference resolution will be one of the serious technical issues for this aim since a lot of abbreviations, pronouns and different expressions are used to point an entity in a Wikipedia article. Therefore, we propose a co-reference resolution method based on synonym information and an improvement method by using important sentence detection.

The rest of this paper is organized as follows. In section 2, we explain a number of researches on Wikipedia Mining for knowledge extraction in order to make our stance clear. In section 3, we describe our proposed integration method based on parsing and link structure mining. We describe the results of our experiments in section 4. Finally, we draw a conclusion in section 5.

2 Related Works

2.1 Relation Acquisition from Text Corpora

In the statistical NLP research area, a significant number of researches on relation acquisition from large scale text corpora have been conducted. For instance, Hearst [10] is one of the researchers who has pointed out that lexico-syntactic patterns (mainly for is-a relation) can be extracted from large scale corpora. Berland and Charniak [11] have proposed similar methods for part-whole relations. Kim and Baldwin [6] focused on nominal relations in compound nouns.

These researches are targeting ordinary text corpora or Web corpora, thus in order to apply these methods for Wikipedia, we need to consider about the characteristics and reconstruct the methods since Wikipedia has various unique characteristics compared with other corpora.

2.2 Wikipedia Mining

"Wikipedia mining" is a new research area that is recently addressed. Researches on semantic relatedness measurement are already well conducted [1–3]. WikiRelate [3] is one of the pioneers in this research area. The algorithm finds the shortest path between categories which the concepts belong to in a category tree. As a measurement method for two given concepts, it works well. However,

it is impossible to extract all related terms for all concepts because we have to search all combinations of category pairs of all concept pairs (2 million × 2 million). Therefore, in our previous research, we proposed $pfibf$ (Path Frequency - Inversed Backward Link Frequency)[1], a scalable association thesaurus construction method to measure relatedness among concepts in Wikipedia. The basic strategy of $pfibf$ is quite simple. The relativity between two articles v_i and v_j is assumed to be strongly affected by the following two factors:

- the number of paths from article v_i to v_j,
- the length of each path from article v_i to v_j.

The relativity is strong if there are many paths (sharing of many intermediate articles) between two articles. In addition, the relativity is affected by the path length. In other words, if the articles are placed closely together in the graph of the Web site, the relativity is estimated to be higher than that of farther ones. Therefore, by using all paths from v_i to v_j given as $T = \{t_1, t_2, ..., t_n\}$, the relativity pf (Path Frequency) between them is defined as follows:

$$pf(v_i, v_j) = \sum_{k=1}^{n} \frac{1}{d(|t_k|)}, \tag{1}$$

$$pfibf(v_i, v_j) = pf(v_i, v_j) \cdot \log \frac{N}{bf(v_j)}. \tag{2}$$

$d()$ denotes a function which increases the value according to the length of path t_k. N denotes the total number of articles and $bf(v_j)$ denotes the number of backward links of the page v_j. Wikipedia Thesaurus [2] [2] is an association thesaurus search engine that uses $pfibf$ in its behind. It provides over 243 million relations for 3.8 million concepts in Wikipedia.

2.3 Wikipedia and Web Ontology

Semantic Wikipedia [9] is one of the pioneers that remarked the effectiveness of Wikipedia style editing for making a huge ontology covering wide range topics. Semantic Wikipedia is an extension of Wikipedia which allows editors to define relations among concepts manually. The contribution of Semantic Wikipedia is that it showed a new direction to achieve the vision of the Semantic Web. While Semantic Wikipedia is a promising approach for a huge scale Web ontology construction, it needs human-effort. Therefore, we try to develop a completely-automated method without any additional human-effort since Wikipedia articles already include rich semantic relations.

Another interesting approach is to use Wikipedia's category tree as an ontology [7, 12]. In previous researches on Wikipedia mining, a large number of researches were based on category tree analysis since Wikipedia categories are a promising resource for ontology construction. For instance, DBPedia [5] uses several types of information on Wikipedia such as InfoBox, article texts, categories in order to extract structured knowledge and provide Web APIs.

[1] The method name was $lfibf$ in the past and was changed to $pfibf$

[2] http://wikipedia-lab.org:8080/WikipediaThesaurusV2

In this research, in contrast to these approaches, we developed a full-automated consistent approach for semantic relation extraction by mining Wikipedia article texts. Wikipedia article texts are promising resources to extract semantic relations but a small number of researches have been conducted in this area.

2.4 Characteristics of Wikipedia

As a Web corpus for knowledge extraction, URL for word sense disambiguation is one of the most notable characteristics of Wikipedia. In Wikipedia, almost every page (article) corresponds to exactly one concept and has an own URL respectively. For example, the concept apple as a fruit has a Web page and its own URL. Further, the computer company Apple also has its own URL and these concepts are semantically separated. This means that it is possible to analyze term relations avoiding ambiguous term problems or context problems.

Hyperlinks do not just provide a jump function between pages, but have more valuable information than we expect. There are two type of links; "forward links" and "backward links". A "forward link" is an outgoing hyperlink from a Web page, an incoming link to a Web page is called "backward link". Researches on Web structure mining, such as Google's PageRank [13] and Kleinberg's HITS [14], emphasize the importance of backward links in order to extract objective and trustful data. "Link texts" also contains valuable information.

Link texts in Wikipedia have a quite brief, clear and simple form compared with those of ordinary Web sites. Among the authors of Wikipedia, it is a common practice to use the title of an article for the link text but users also have the possibility to give other link texts to an article. This feature makes another important characteristic; the "variety of link texts," which can be used to extract valuable information. However, what seems interesting is that link texts do not contain any wordy information in most cases. Since no link text data is available on Wikipedia dump data, we customized the Wiki parser engine on Wikipedia to extract the link text data.

3 Proposed method

In order to extract semantic relations from Wikipedia, we propose a method that analyzes both the Wikipedia article texts and link structure. Basically, the proposed method extracts semantic relations by parsing texts and analyzing the structure tree generated by a parser. However, parsing all sentences in an article is not efficient since an article contains both valuable sentences and non-valuable sentences. We assume that it is possible to improve accuracy and scalability by analyzing only important sentences on the page. Furthermore, we use synonyms to enhance co-reference resolution. In a Wikipedia article, usually a number of abbreviations, pronouns and different expressions are used to point to an entity, thus co-reference resolution is one of the technical issues in order to make the parsing process accurate.

Figure 1 shows the whole flow of the proposed method. The method consists of three main phases; parsing, link (structure) analysis, and integration. First, for a given Wikipedia article, the method extracts a list of related terms for an article using $pfibf$ [2]. At the same time, it provides synonyms by analyzing the link texts of backward links of the article. Second, the method analyzes the article text to extract explicit semantic relations among concepts by parsing the

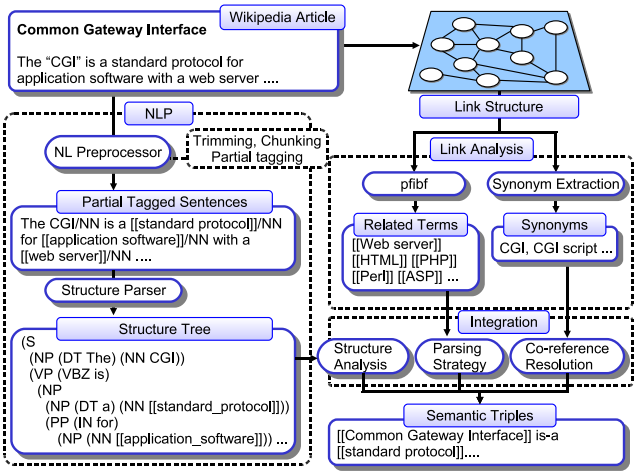

Fig. 1. Whole flow of the proposed method.

Table 1. Synonym extraction by link text analysis.

Concept	Synonyms
Apple Computer	'Apple' (736), 'Apple Computer, Inc.' (41), 'Apple Computers' (17)
Macintosh	'Apple Macintosh' (1,191), 'Mac' (301), 'Macs' (30),
Microsoft Windows	'Windows' (4,442), 'WIN' (121), 'MS Windows' (98)
Intl. Organization for Standardization	'ISO' (1026), 'international standard' (4), 'ISOs' (3)
Mobile phone	'mobile phones' (625), 'cell phone' (275), 'Mobile' (238)
United Kingdom	'United Kingdom' (50,195), 'British' (28,366), 'UK' (24,300)

(): Number of backward links
(Link texts corresponding to the title of an article are excluded).

sentences. Finally, in the integration phase, three steps for triple extraction are conducted; 1) analyzing the structure tree generated by the parser, 2) filtering important semantic information using parsing strategies, and 3) resolving co-references by using synonyms. The main steps of the proposed method are described as follows.

3.1 Synonym Extraction

We describe our co-reference resolution method by using synonyms extracted from anchor texts. A synonym word has one meaning but various expressions. Since backward links of a web page have a "variety of backward link texts," this variety can be used to extract synonyms of a concept (article). For instance, the computer company "Apple" is sometimes referred to as "Apple", but it is sometimes also written as "Apple Computer, Inc," "Apple Computers," etc. Table 1 shows a number of examples of randomly chosen synonym terms.

The article "Apple Computer" has 1,191 backward links with the link text "Apple Macintosh" and 301 backward links with the link text "Mac." This shows that both words are typical synonyms for the concept "Apple Computer." Statistical data unveiled that backward link texts analysis can extract high quality synonyms by specifying a threshold to filter noisy data such as 'international standard' and 'ISOs' for ISO.

Synonyms are helpful information to detect whether two sentences are describing the same subject. In other words, the information is needed for coreference resolution. For example, there is an article about "United Kingdom" in Wikipedia and it contains "UK" many times. However, if the machine does not know that "UK" is a synonym of "United Kingdom," it can not extract many relations on the topic. Therefore, we use the extracted synonyms in the following steps to improve the coverage.

For a given article a and the synonym candidate s, we define a simple scoring function $syn(a, s)$ as follows;

$$syn(a, s) = \frac{log\ num\_bk(a, s)}{log\ num\_bk(a, *)}. \tag{3}$$

$syn(a, s)$ basically measures the popularity of the label for the concept by calculating ratio of total backward links and the link texts. $num\_bk(a, s)$ is the number of backward links of a with link text s. $num\_bk(a, *)$ is the total number of backward links of a. We defined a threshold for $syn(a, s)$ to filter irrelevant synonyms by 200 training data evaluated by human effort.

3.2 Preprocessing

Since the structure and syntax of a Wiki is much different from natural languages, we need to modify and optimize the parser by considering special syntax composed of HTML tags to achieve better accuracy. Basically, special Wiki command tags such as triple quotation, brackets for hyperlinks and tables, prevent correct parsing. However, it is also true that this kind of information is helpful to analyze the content since it contains hyperlinks and helpful information to compound words into semantic chunks. Therefore, we constructed a preprocessor by ourselves to achieve better accuracy. The preprocessor trims the Wikipedia article to remove unnecessary information such as HTML tags and special Wiki commands first. It also removes table tags because contents in tables are usually not sentences. However, it does not remove link tags ("[[...]]") because links in Wikipedia are explicit relations to other pages and we use the link information in the following steps. Finally, phrases in quotations and link tags are tagged as nouns to help the following parsing step.

Parsing and Structure Tree Analysis After the preprocessing, it provides partially-tagged sentences. In this step, the method parses the sentences to get a structure tree and analyzes the structure tree to extract semantic relations. To parse sentences, we adopted a lexcicalized probabilistic parsing method based on the factored product model. We used the Stanford parser [15] for this purpose. It can parse a sentence accurately if the sentence is trimmed, chunked and tagged correctly by preprocessing. A list of main POS (Part Of Speech) tags used in this step is shown in Table 2 (Right).

Table 2. Wikipedia statistics and POS tags.

Statistics of Wikipedia articles.

# of concept pages (exc. redirect and category pages)	1,580,397
# of pages having more than 100 backward links: P_a	65,391
# of pages (in P_a) begin with is-a definition sentence: P_b	56,438
# of pages (in P_a) that the 1st sentence has links: P_c	62,642
# of $P_b \cap P_c$	56,411

POS Tags.

Tag	Description
NN	Singular or mass noun
NNS	Plural noun
NNP	Singular proper noun
NNPS	Plural proper noun
NP	Noun phrase
VB	Base form verb
VBD	Past tense
VBZ	3rd person singular
VBP	Non 3rd person singular present
VP	Verb phrase
JJ	Adjective
CC	Conjunction, coordinating
IN	Conjunction, subordinating

For example, for a sentence "Lutz_D._Schmadel is [[Germany]] [[astronomer]]." about the person with the name "Lutz_D._Schmadel," the parser generates a structure tree like this;

```
(S (NP (NN Lutz_D._Schmadel) (VP (VBZ is) (NP (NN [[Germany]]) (NN [[astronomer]])))))
```

In our proposed method, the parser takes a partially tagged sentence made by preprocessing and generates a structure tree from the sentence. After that, the structure tree is analyzed in order to extract triples (Subject, Predicate, Object) in the following steps:

1. Extract "(NP ...) (VP (VBZ/VBD/VBP ...) (NP ...))" pattern from the parsed sentence.
2. For both NP, replace the NP by the last NN/NNS in the NP if the NP parts consist of JJ and NN/NNS.
3. For both NP, split the NP into two NP parts if the NP contains CC. After that, perform step 2 again.
4. If the 1st NP is a synonym of the concept representing the article, replace the NP part by the title of the main subject.
5. Finally, extract the 1st NP part as a subject, VB part as a predicate, the 2nd NP part as an object.

In the first step, we extract "(NP ...) (VP (VBZ/VBD/VBP ...) (NP ...))" and assume that the 1st NP part is the subject, the VB part is the predicate, the 2nd NP part is the object respectively.

In the second step, for both NP parts, we replace NP by the last NN/NNS term (or hyperlink) because the last term is the mainstay of the phrase. For instance, the 2nd NP in the sentence about "Lutz_D._Schmadel" consists of two NN and both of them have a hyperlink to other pages and the 1st NN has a link to a country "Germany". So in this case, it obtains "[[astronomer]]" as the mainstay of the object part.

In the third step, NP will be separated if it contains CC such as "and" and "or". In the fourth step, if the 1st NP is a literal and it is a synonym of the concept representing the article, then the NP is replaced by the concept of the article. Finally, the first NP part is extracted as a subject, the VB part as a predicate, the 2nd NP part as an object.

The first step's POS tag pattern can be replaced by other alternatives. Currently, we prepared following three for the first step.

1. (NP ...) (VP (VBZ/VBD/VBP ...) (NP ...))
   Normal pattern. E. g. "is-a"
2. (NP ...) (VP (NP (NP ...) (PP (IN ...) ...))
   Subordinating pattern. E. g. "is-a-part-of"
3. (NP ...) (VP (VBZ ...) (VP (VPN ...) ...))
   Passive pattern. E. g. "was-born-in"

We can prepare further POS tag patterns to improve the coverage of triples. However, in this research, we applied these three basic patterns to confirm the capability of this direction of research.

In this research, we also extract a relation if the object part does not contain any hyperlinks to other pages. We call it "literal" object. For example, assume that there is a sentence "Brescia is a city" with the following structure tree;

```
(S (NP (NNP [[Brescia]])) (VP (VBZ is) (NP (DT a) (NN city))))
```

The subject part is "a city" but it is not a hyperlink to an article about "city" but it is just a literal. Literal objects are not machine understandable but the literal information is useful depending on the application even if the meaning of the term can not be specified. So we extract the literal information as well.

Co-reference Resolution In a Wikipedia article, usually a number of abbreviations, pronouns and different expressions are used to point an entity, thus co-reference resolution is one of the technical issues in order to make the parsing process accurate. In several previous researches on Wikipedia Mining, co-reference resolution methods optimized for Wikipedia article are proposed [4, 16]. Gang mentioned that emphasized words are likely

Let us assume that there is a Wikipedia article A_t which is describing the topic t (the main subject of the article). A_t is a set of sentences and each sentence a has triple; subject s_a, predicate p_a, and object o_a. Co-reference resolution is a procedure that judges whether s_a is describing about same topic as the main subject t or not. We use three co-reference resolution approaches (included one novel approach) considering following three factors; article title ($C1$), frequent pronouns ($C2$) and synonyms ($C3$).

$C1$ is an approach to detect co-references if the terms used in s_a are all contained in the title of A_t. $C2$ uses pronouns for the judgment. It judges s_a as a co-reference to t if s_a is the most frequently used pronoun in A_t. $C1$ and $C2$ were proposed in previous research [16], but $C3$ is a novel approach proposed by us. The main idea of the approach is to detect co-references if the s_a is a synonym of t. In addition, we investigated the effectiveness of combining these three approaches in detail.

3.3 Parsing Strategies

LSP: Lead Sentence Parsing LSP is a strategy that parses only the lead sentences (first n sentences). After a simple inspection, we realized that a considerable number of Wikipedia articles begin with definitive sentences containing relations (hyperlinks) to other articles (concepts). Especially, the first sentence often defines "is-a" relation to other article. We took detailed statistics (Table 2 Left) from the English Wikipedia (Sept. 2006) to confirm this phenomenon.

First, we removed all redirect pages and category pages from the target of the statistics because these pages are not concept pages but navigational pages. After that, we removed all pages having only few backward links (less than 100) because such pages often contain noisy information and are not structured well [1]. Then, we investigated how many articles begin with a definitive sentence (contain is/are/was/were). The result showed that over 86.3% (P_b/P_a) of all pages begin with a definitive sentence.

We also investigated whether the first sentences have hyperlinks to other pages. The results showed that over 95.7 % (P_c/P_a) of all pages begin with a sentence having hyperlinks to other pages. Further, over 85.5 % ($(P_b \cap P_c)/P_a$) of pages begin with a definitive sentence having hyperlinks.

To conclude this, the statistics unveiled that a large number of pages in Wikipedia has a high potential for extracting "is-a" relations to other concepts thus the first sentence analysis seems a promising approach.

ISP: Important Sentence Parsing ISP detects important sentences in a page if the sentence contains important words/phrases for the page. Our assumption is that the sentences containing important words/phrases are likely to define valuable relations to the main subject of the page, thus we can make the co-reference resolution accurate even if the subject of the sentence is a pronoun or another expression for the main subject. We use $pfibf$ to detect important sentences. By using $pfibf$, a set of important links for each article (concept) in Wikipedia can be extracted. ISP detects important sentences in a page from sentences containing important words/phrases for the page. It crawls all sentences in the article to extract sentences containing links to the associated concepts. The extracted sentences are then parsed as the important sentences in the article. For each links in a sentence, the parser calculates $pfibf$ and the max value denotes the importance of the sentence. The importance can be used for filtering unimportant sentences by specifying thresholds.

For example, when analyzing the article about "Google," associated concepts such as "Search engine", "PageRank" and "Google search" are extracted from the association thesaurus. Therefore, ISP crawls all sentences in the article to extract sentences containing links to the associated concepts.

4 Evaluation

To prove the effectiveness of our proposed method, we conducted two experiments. The first experiment was conducted to measure the co-reference resolution accuracy. The second experiment was conducted to measure the accuracy of the extracted triples. We describe these experiments in detail as follows.

4.1 Experiment 1: Co-reference resolution

In this experiment, we first filtered noisy pages by checking the number of backward links of the articles and extracted 65,391 pages as a test collection. After that, we parsed 2,508 sentences in 52 articles chosen randomly from the test collection. Then, totally 1,002 triples were extracted by parsing patterns described before. A list of term examples used in this experiment is shown as follows; Niagara Falls, Root beer, Deer, Arrow, Odonata, Marie Antoinette, Germany, Colorado, and Blizzard.

Table 3. Evaluation results.

Co-reference resolution.

Methods	Precision	Recall	F-Measure
$C1$	99.22%	59.26%	74.20%
$C2$	65.00%	18.06%	28.26%
$C3$	89.04%	60.19%	71.82%
$C1 \cup C2$	81.78%	81.02%	81.40%
$C1 \cup C3$	89.94%	70.37%	78.96%
$C2 \cup C3$	81.99%	80.09%	81.03%
$C1 \cup C2 \cup C3$	82.33%	81.94%	82.13%

Important sentence selection.

Method	Literal	Extracted Relations	Correct Relations	Precision
ASP	Includes	458	285	62.22 %
	Excludes	162	133	82.09 %
LSP	Includes	101	91	90.09 %
	Excludes	54	52	96.30 %
ISP	Includes	67	54	80.59 %
	Excludes	59	51	86.44 %
$LSP \cup ISP$	Includes	153	130	84.96 %
	Excludes	99	88	88.88 %

We manually checked whether the subject of each sentence is a co-reference of the main subject of the article. Totally 216 subjects of sentences were co-references of the article subject. We used the data set to calculate precision, recall and f-measure. The result is shown in table 3.

As we can see, not surprisingly, $C1$ (article title approach) achieved quite high precision. However, the precision of $C2$ (frequent pronouns approach) was rather low. We investigated the reason and realized that the approach to use frequent pronouns is an error prone strategy. In particular, the pronouns "it" and "he/she" are not used for representing the main subject of an article but for different meanings. We tried all combinations and realized that the combination of all methods achieved the highest f-measure. This means that the combination of these three methods compensates for the weak points of each method, and is therefore helpful to achieve a higher coverage.

4.2 Experiment 2: Triple extraction

In this experiment, we first randomly selected 110 articles and totally 1,016 sentences were extracted as a test set. After that, we applied the proposed method to extract triples. We used LSP and ISP to improve the accuracy of triples. As a baseline, we also parsed all sentences and call it "All Sentence Parsing (Hence ASP)" method. Table 3 shows the result of the experiment.

First of all, we would like to mention that the accuracy of the LSP method is quite high. It achieved high quality relation extraction for both literal objects and non-literal objects. This means that our conviction that the first sentence is useful information is strongly confirmed. We have no strong evidence but we think that this is because of the reliability of the sentences. Usually, the top part of a page attracts much more attention than the bottom part. Thus, the top part is edited by many authors and structured well in most cases. Several parsing misses happened when the sentence is too complicated which was the cause of accuracy loss.

Second, the ISP method also achieved better results than ASP. In particular for literal objects, the accuracy significantly improved. Furthermore, by using the ISP method, we can determine whether a sentence contains important concepts before parsing it, decreasing the analysis time significantly. We also believe that the combination of LSP and ISP is a balanced method because it achieves high coverage and high precision at the same time.

Table 4 shows some examples of explicit relations extracted by LSP. "Explicit relation" means a relation where the object part is a hyperlink to another article. As we can see, the extracted relations are very accurate. As we mentioned

Table 4. Examples of the results.

Extracted explicit relations by LSP samples.

Subject	Predicate	Object
Apple	is-a	Fruit
Bird	is-a	Homeothermic
Cat	is-a	Mammal
Computer	is-a	Machine
Isola d'Asti	is-a	Comune
Jimmy Snuka	is-a	Pro. wrestler
Karwasra	is-a	Gotra
Mineral County	is-a	County
Sharon Stone	is-a	Model
Sharon Stone	is-a	Film producer

Extracted explicit relations by ISP samples.

Subject	Predicate	Object
Odonata	is an order of	Insect
Clarence Thomas	was born in	Pin Point, GA
Dayton, Ohio	is situated	Miami Valley
Germany	is bordered on	Belgium
Germany	is bordered on	Netherlands
Mahatma Gandhi	founded	N. Indian Congress
Mahatma Gandhi	established	Ashram
Rice	has	Leaf
Rice	is cooked by	Boiling
Rice	is cooked by	Steaming

before, almost all articles of Wikipedia begin with a definitive sentence, so LSP extracted mainly "is-a" relations. While is-a relation is one of the most basic (and important) relations in Semantic Web, the result shows the capability of this approach for ontology construction and the possibility for making practical approach to achieve next generation WWW technologies.

Table 4 shows some examples of explicit relations extracted by ISP. Since ISP analyzes important sentences in the article, it extracts various relations such as "was born in," "founded" and "has". However, machines cannot understand the meaning "was born in" without any instruction from humans. So, in order to make the predicate part machine understandable, we have to define the relation between predicates. For example, "is" and "was" have the same meaning but the tense is different. By giving this kind of knowledge, machines can infer semantic relations between two concepts. We believe that the relations among verbs are quite limited compared with relations between nouns, thus do not cause enormous workload.

5 Conclusion

In this paper, we showed that Wikipedia article is yet another invaluable corpus for ontology extraction by showing both detailed statistics and the effectiveness of integrating parsing and link structure mining methods. The experimental results showed that the integration method and co-reference resolution significantly improves the accuracy of triple extraction. Especially, the conviction that lead sentences have rich semantic information is strongly confirmed. Furthermore, important sentence detection by using link structure analysis was helpful to filter inaccurate results.

More than anything else, what we are trying to show in this paper is the possibility and capability of semantic relation extraction using Wikipedia knowledge. We believe that this direction will be an influential approach for Semantic Web in near future since Wikipedia has great capability for constructing a global ontology. The extracted association thesaurus and semantic relations are available on our Web site.

Wikipedia Lab : http://wikipedia-lab.org
Wikipedia Thesaurus : http://wikipedia-lab.org:8080/WikipediaThesaurusV2
Wikipedia Ontology : http://wikipedia-lab.org:8080/WikipediaOntology

We hope the concrete results will be a helpful information to judge the capability of this approach. Our next step is to apply the extracted semantic relations to Semantic Web applications (esp. Semantic Web search). To do that, we need further coverage of relations by enhancing the POS tag analysis patterns and mappings among relations.
Acknowledgment: This research was supported in part of the Microsoft Research IJARC Core Project. We appreciate helpful comments and advices from Prof. Yutaka Matsuo at the University of Tokyo as well as from Prof. Takahiro Hara and Prof. Shojiro Nishio at Osaka University.

References

1. E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis.," in *Proc. of International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pp. 1606–1611, 2007.
2. K. Nakayama, T. Hara, and S. Nishio, "Wikipedia mining for an association web thesaurus construction," in *Proc. of IEEE International Conference on Web Information Systems Engineering (WISE 2007)*, pp. 322–334, 2007.
3. M. Strube and S. Ponzetto, "WikiRelate! Computing semantic relatedness using Wikipedia," in *Proc. of National Conference on Artificial Intelligence (AAAI-06)*, pp. 1419–1424, July 2006.
4. G. Wang, Y. Yu, and H. Zhu, "Pore: Positive-only relation extraction from wikipedia text," in *International Semantic Web Conference, Asian Semantic Web Conference (ISWC/ASWC)*, pp. 580–594, 2007.
5. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, "Dbpedia: A nucleus for a web of open data," in *International Semantic Web Conference, Asian Semantic Web Conference (ISWC/ASWC)*, pp. 722–735, 2007.
6. S. N. Kim and T. Baldwin, "Interpreting semantic relations in noun compounds via verb semantics," in *Proc. of Conference on Applied Computational Linguistics (ACL)*, 2006.
7. F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proc. of International Conference on World Wide Web*, pp. 697–706, 2007.
8. D. N. Milne, O. Medelyan, and I. H. Witten, "Mining domain-specific thesauri from wikipedia: A case study," in *Proc. of ACM International Conference on Web Intelligence (WI)*, pp. 442–448, 2006.
9. M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer, "Semantic wikipedia," in *Proc. of International Conference on World Wide Web (WWW 2006)*, pp. 585–594, 2006.
10. M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proc. of COLING*, pp. 539–545, 1992.
11. M. Berland and E. Charniak, "Finding parts in very large corpora," in *Proc. of Conference on Applied Computational Linguistics (ACL)*, 1999.
12. S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou, "Extracting semantics relationships between wikipedia categories," in *Proc. of Workshop on Semantic Wikis (SemWiki 2006)*, 2006.
13. P. Lawrence, B. Sergey, M. Rajeev, and W. Terry, "The pagerank citation ranking: Bringing order to the web," *Technical Report, Stanford Digital Library Technologies Project*, 1999.
14. J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, no. 5, pp. 604–632, 1999.
15. D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proc. of Meeting of the Association for Computational Linguistics (ACL 2003)*, pp. 423–430, 2003.
16. D. P. T. Nguyen, Y. Matsuo, and M. Ishizuka, "Relation extraction from wikipedia using subtree mining," in *Proc. of National Conference on Artificial Intelligence (AAAI-07)*, pp. 1414–1420, 2007.