Why Wikipedia needs to make friends with WordNet

Kow Kuroda,[*] **Francis Bond,**[**,*] **Kentaro Torisawa**[*]
kuroda@nict.go.jp bond@ieee.org torisawa@nict.go.jp
[*]National Institute of Information and Communications Technology (NICT), Japan
3-5 Hikari-dai, Seika-cho, Sooraku-gun, Kyoto, 619-, Japan
[**]Linguistics and Multilingual Studies, Nanyang Technological University, Singapore

Abstract

This paper describes the compilation of hypernym hierachies from the Japanese Wikipedia (Sumida et al., 2008). It then compares the Wikipedia-derived hypernyms and the lemmas from the Japanese WordNet (Bond et al., 2008; Bond et al., 2009) by determining how many matches there are at which levels. The results show that the two data sources contain different information. This means that the Wikipedia-derived data and manually crafted data like WordNet (Fellbaum, 1998) are best understood as complementary to each other.

1 Does Wikipedia dispense with the need for WordNet?— Introduction

Data of various kinds acquired from Wikipedia[1)] is gaining popularity in NLP and related areas of research. One reason for this is that Wikipedia provides us with broad coverage. No other freely available linguistic resource can match its breadth. It is often claimed that this is evidence for the triumph of "collective intelligence."

Radical enthusiasts of Wikipedia even go on to claim that researchers in NLP and SemanticWeb no longer need WordNet (WN) (Fellbaum, 1998).[2)] They allude to the superiority of Wikipedia-derived data over manually crafted data like WN in terms of development ease, speed, and cost as well as coverage. WN comes with precision endorsed by psychological reality that most WWW-derived data lacks, but some people also tend to criticize the subjective nature of the word senses that WN specify, no matter how fine-grained its sense distinctions are. All in all, they seem to try to dismiss WN-like lexical resources by suggesting that they are outdated in the age of WWW. And here comes the crucial question, *Does Wikipedia dispense with the need for WordNet?*

In this paper, we argue that the answer is *No*, suggesting that we should make a good compromise. We show that lexical hierarchies derived from the Japanese Wikipedia are not as well articulated as the upper ontology of Japanese WordNet. This allows us to presume that the hypernym set of language L obtained from the Wikipedia of L is poor compared to the WN of L. Under this assumption, the WN and the Wikipedia of language L are best understood to be complementary in the following way: The WN of L specifies the mapping between an upper ontology to lexical items, w_1, w_2, ..., w_n, of L. The conceptual hierarchies distilled from the Wikipedia written in L specify links to named entities described in w_1, w_2, ..., w_n of L.

Organization

This paper is organized as follows. In §2 we describe how we processed the hypernym-hyponym pairs acquired from the Japanese Wikipedia by Sumida et al. (2008). In §3, we show how the hypernyms obtained in the way specified in §2 were linked to lemmas of Japanese WordNet (WN-Ja) (Bond et al., 2008; Bond et al., 2009). In §4 we discuss the implications. Finally, in §5, we state tentative conclusions

2 Acquiring taxonomic hierarchies from Wikipedia hypernyms

Recently, we finished the manual cleaning of approximately 67,000 Japanese hypernym hierarchies paired with roughly 900,000 hyponyms. We show some details of this process in §2.4. The original data, comprising roughly 2,400,000 hypernym-hyponym pairs, was automatically compiled from the Japanese Wikipedia (Sumida

[1)] *Wikipedia: The Free Encyclopedia*, http://wikipedia.org/.

[2)] In http://www.mkbergman.com/417/99-wikipedia-sources-aiding-the-semantic-web/ (retrieved on 2009/12/01) for example, you can find a bold claim like: "Wikipedia has arguably replaced WordNet as the leading lexicon for concepts and relations. Because of its scope and popularity, many argue that Wikipedia is emerging as the de facto structure for classifying and organizing knowledge in the 21st century.".

et al., 2008). They used Support Vector Machines (Vapnik, 1995) to classify the acquired data. The hypernym-hyponym pairs extracted by Sumida et al. (2008) consist not only of links between Wikipedia entries, but also consider noun phrases extracted from the text of the Wikipedia entries themselves.

While the data thus acquired has an impressive coverage, it is noisy and unreliable at least two ways: First, both hypernyms and hyponyms can be misparsed phrases, due to the low performances of a Japanese tokenizer[3]. Second, even correctly parsed phases can have hypernyms that are themselves relational nouns (such as *kind*, *member*): they suffer from **semantic unsaturatedness** in the sense of Kuroda et al. (2009) and fail to serve as good hypernyms.

2.1 Extracting base hypernyms

The data we processed resemble the following, where h is the hypernym and I is the instance (or hyponym):[4]

(1) a. h: famous British rock singer
I: Peter Gabriel
b. h: former member of Pink Floyd
I: Syd Barrett

The set of hypernyms extracted form Wikipedia consist mainly of complex NPs like *famous British rock singer* and *former member of Pink Floyd*. Clearly, terms like these are not ideal hypernyms. Thus, we tried to extract **base** hypernyms by gradually removing modifiers from the complex NPs. In this way, the two pairs in (1) are converted into the following hierarchies:

(2) a. h_1: singer
h_2: rock singer
h_3: British rock singer
h_4: famous British rock singer
I: Peter Gabriel
b. h_1: member
h_2: former member
h_3: former member of Floyd
h_4: former member of Pink Floyd
I: Syd Barrett

[3] The precision of Japanese tokenizers at the state-of-the-art come close to 98% against newspaper articles, but they show much lower precision against open text such as Wikipedia.

[4] Although we work on Japanese data, we present the English translations in this paper as the semantic phenomena are not language specific.

The pairs (H; I), where $H = h_1, \ldots, h_n$, are automatically generated from such pairs ($h_{\max}$; I). We refer to H as the **hypernym path** for I,[5] and to units like h_1, h_2, …, h_n as the **path elements** of H. A hypernym path may contain: (i) bare nouns (e.g., *singer*), (ii) modified nouns (e.g., *famous British rock singer*, *former member*), or (iii) noun phrases.

We are able to create these paths in this way because we are looking specifically at hypernym relations. Removing a modifier broadens the denotation, and thus gives a hypernym of the more restricted term.

2.2 Problems with automation

Paths like the ones above were constructed by automatically removing modifiers from h_n (in Japanese) one by one. This operation is not error-free. Manual cleaning was performed to eliminate unconventional and/or unacceptable units like *former member of Floyd*, h_4 of (2b) which was produced by the automatic simplification. We could not use a Named Entity tagger for this task as it performed too poorly on the isolated noun phrases.

2.3 What terms make good hypernyms?: Effect of semantic saturatedness

During the manual process of cleaning, it also became apparent that checking for the conventionality of path elements alone was not effective. We also needed a systematic treatment of composite units like *former member* to take care of the function of modifiers. However, **lexical** databases like WordNet are not guaranteed to contain composite phrasal units like *former X* (X={ member, president, … }). This means that we cannot rely on lexical resources to distinguish valid phrases from invalid ones which are only theoretically possible.

This was a problem because raters we hired showed confusion as to the conventionality of such terms and disagreed in their ratings. To them, terms like *member* made good terms even if they were presented in isolation, but terms like *former member* did not. When they were presented in isolation, most raters hesitated to rate them as good hypernyms.

Part of the reason for this disturbance can be attributed to the semantic unsaturatedness of units like *former member*, but the situation seems more

[5] For both practical and theoretical purposes, we did not distinguish between **instance** and **hyponym** relations.

complex. Interestingly, raters showed little disagreement on the goodness of *member*, which is also a semantically unsaturated noun. So, the real reason for rater's trouble in classification is not a term's semantic unsaturatedness alone. Consequent research suggests that frequency has an effect on this: frequent semantically unsaturated nouns tend to be classified as saturated nouns.

In passing, it deserves a brief mention that most linguists' tacit assumption that relational nouns are relatively rare and exceptional, and that their set is closed seems far from well grounded. The assumption would be true of simple nouns, but it is not true of composite nouns with modifiers. The source of unsaturatedness of composite nouns are their modifiers. For example, *former* caused the effect in the example above. It is easy to provide similar examples: the unsaturatedness in *sister city* comes from a metaphorical sense of *sister*: *city* is arguably a saturated noun, but *sister city* is unsaturated because *sister* adds unsaturatedness to it. This is why it is possible to say "*X* and *Y* are *sister cities*" or "*X* is the/a *sister city* of *Y*."

We can add examples with more complexity. For example, *disciple* is a semantically unsaturated noun. In the combination *fellow disciple(s)*, *fellow* adds further unsaturatedness. This is why it is possible to say "*X* and *Y* are *fellow disciples* under *Z*" or "*X* is the/a *fellow disciple* of *Y* (?under *Z*)" and why we infer that both *X* and *Y* have a common master, *Z* when we hear such expressions.

Notably, the unsaturatedness for *fellow* and *disciple* can co-exist.[6)] Cases like this show that unsaturatedness accumulates through modification.

Another class of cases show that unsaturatedness is composable, allowing the unsaturatedness of one noun to get **bridged** to another. In cases like *secretary of the Minister (of Foreign Affairs)*, unsaturatedness is reduced through variable-binding, because *secretary of X*, *X* is bound to *the Minister (of Y)*, and when *Y* is bound to *Foreign Affairs* (with the aid of *of*), it gets saturated; otherwise, it stays unsaturated.

Examples of the sort briefly mentioned above strongly suggest that the semantics of modifiers is rather complex and needs serious investigation. It is not guaranteed that proper analysis of modifiers is possible within a natural extension of the semantics of individual words, partly because analogy, metaphor and metonymy play a crucial role. Note that modifiers undergo (often very subtle) semantic extensions: at least, *sister* is not used in its literal sense in *sister cit(y|ies)*.[7)] With this fact in mind, it would be safe to assume that semantic unsaturatedness becomes more serious at the level of composite nouns or noun phrases. In fact, they posed a challenge in the cleaning process to be explained below.

2.4 Path element cleaning in some detail

We give an outline of the cleaning procedure below.

Step 1: Fully automatic generation of hypernym paths

First, all hypernyms were morphologically analyzed with a morphological analyzer/tokenizer for Japanese.[8)] This gave us a set of paths consisting of a series of morphemes coupled with part-of-speech (POS) and other information. Based on the POS information thus provided, we generate a series of terms that serve as a hypernym path.[9)]

During this process we excluded some problematic hypernyms. However, hypernyms with disjunctive semantics were not excluded.

Step 2: Manual evaluation of path elements

We asked four raters to evaluate each of the path elements for their conventionality and/or semantic saturatedness. The criteria used were:

(3) a. If a path element *X* is felt to be fully conventional and saturated, it should be classified as G[ood].

[6)] Note here that a mutuality interpretation of relational nouns (Eschenbach, 1993) seems to have an interesting effect on the construction and interpretation of *sister cit(y|ies)* and *fellow disciple(s)*.

[7)] For the case of *fellow disciples*, the meaning of *fellow* can be literal. We can say that two disciples of *Z*, *X* and *Y*, are in the relation of FELLOW-OF-THE-OTHER(X, Y). But this is not true of *sister cities*. We can only say that two cities, *X* and *Y*, are in **some relation analogous to the relation** of *sister-of-the-other* rather than they are properly in the relation of SISTER-OF-THE-OTHER(X, Y).

[8)] We used MeCab 0.95 (`http:///mecab.sourceforge.net/`) with its default dictionary IPA Dic. While our pilot study showed that the combination of MeCab with UniDic (`http://www.tokuteicorpus.jp/dist/`) developed by the National Institute for Japanese Language provided better results we could not in the end use it due to its restrictive license.

[9)] We are afraid the the same automation would not be possible for nonhead-final languages because the criteria used are specific to modifiers that appear before the head noun.

b. If *X* is felt to be incomplete for less conventionality or strong unsaturatedness, it should be classified as L[ess Good].
c. If *X* is felt not to be rather unconventional but the rater cannot be sure if it is really a nonword, it should be classified as D[ubious].
d. If *X* is felt to be fully unconventional and nonsensical, it should be classified as B[ad].

We collected the ratings thus created and selected the most appropriate class.[10] Admittedly, L and D are mixtures of different subclasses. But we did not attempt to create proper labels for subclasses.[11]

Step 3: Automatic reconstruction of hypernym paths and finalization

Because relevant information is distributed over different paths, they are reconstructed from scratch for canonicalization. After this, the first author edited the results based on his intuition. He edited the paths and even added missing intermediate terms and some abstract (super)hypernyms with D status at the root (such as 者*, 手*, and 校* to be discussed in §3.3).

In the process of this cleaning, the original set of roughly 95,000 hypernyms was reduced to the set of 67,000.

3 Linking Wikipedia hypernyms to WordNet

In this section, we describe how compared the hypernym hierachies constructed in the method above with the Japanese lemmas in the Japanese WordNet (WN-Ja) (Bond et al., 2008; Bond et al., 2009).

3.1 Nature of hypernyms and hyponyms in Wikipedia

Recall that we processed hypernym-hyponym pairs automatically acquired from the Japanese Wikipedia. The data consists of roughly 67,000 hypernym hierarchies paired with roughly 900,000 hyponyms.

We cleaned up all the hypernyms of the data, but we did not process the hyponyms for the following reason: nearly 2/3 of the hyponyms are proper names or named entities. The amount of knowledge required to determine if such pairings are valid or not goes well beyond the personal knowledge of an average person. At the time we started the cleaning, it was unclear how to deal with them.

This, on the other hand, suggests an interesting possibility: if pairings of cleaned-up hypernyms with hyponyms turn out to be valid, the huge database of such pairings should complement traditional thesauri as WordNet (Fellbaum, 1998) which mainly consist of upper level concepts (by its very design). With this hypothetical mapping between coarse-grained concepts in the upper ontology and finer-grained concepts in the lower ontology, we can specify the linkage from named entities to upper ontology. If this is possible, it is very promising.

With his hope in mind, we linked the roots of the hypernym hierarchies cleaned in the way illustrated above to nodes in the WN-Ja.

WN-Ja is a Japanese translation of WordNet 3.0 developed and maintained at the National Institute for Information and Communications Technology (NICT). After the first public release in 2009, WN-Ja underwent several updates. We used versions 0.80 and 0.90 for this study.

3.2 Current status

Currently, 95% of the hypernym hierarchies are linked to WN-Ja. Crude statistics are given in Table 1. A sample of matches are shown in (4).

(4) shows sample matches of WN-Ja lemmas (in bold) against Wikipedia hypernym path elements (all in Japanese): Terms are separated by ":" and matched terms are underlined.

(4) a. **校**: **学校**: **大学**: 締結大学: 協定締結大学: 交換留学協定締結大学: 国内交換留学協定締結大学
b. **機**: コミューター機
c. **ジャーナリスト**: 経済ジャーナリスト
d. **病**: 消化器病
e. **大会**: 選手権大会: 日本選手権大会
f. **船**: 艦船: 海軍の艦船: イギリス海軍の艦船

[10] The final class selection was done by the first author on the basis of their intuition guided by the information about distribution. This means that the winners were not always the ones that had acquired the most votes. One reason for this was that some of the raters committed systematic errors.

[11] The agreement rate in terms of Fleiss' kappa against a sample of 2000 cases was 0.492 under the distinction among G, L, D, and B. This is not so good, but it increased to 0.759 if class L was discarded, and it increased up to 0.916 if classes L and D were unified as one. This suggests that rater's classification is highly stable over the identification of G and B, and that raters were confused between L and D.

g. **地理**: 府の地理: 京都府の地理
h. **違反**: 交通違反
i. **手**: **選手**: スポーツ選手: グルジアのスポーツ選手
j. **遺産**: 世界遺産: ベトナムの世界遺産
k. **企業**: 親密な企業: グループと親密な企業: 三井グループと親密な企業
l. **シングル**: 未歩のシングル: 小松未歩のシングル
m. **員**: **委員**: 専門委員
n. **ソフト**: 書き換えソフト: ニンテンドウパワー書き換えソフト

Table 1: Number and ratio of matches of WN-Ja lemmas over Wikipedia-derived hypernym

Depth	# of Covered	Ratio	# of Types
1	64,412	0.9592	3,272
2	24,554	0.3657	2,447
3	2,804	0.0418	465
4	53	0.0008	30

Depth in Table 1 refers to the levels of hypernym hierarchy (measured from the root) at which WN-Ja lemmas have matches. For example, 64,400 root hypernyms out of 67,000 (tokens) have matches with WN-Ja, consisting of 3,272 unique types.

In this linkage process, however, we did not take into account the effect of word sense disambiguation. This suggests that we have fewer correct matches than the figures in Table 1 indicate.

As Table 1 suggests, WN-Ja hypernyms and Wikipedia-derived hypernyms have matches at very shallow levels (the average is nearly 3). More specifically, lower level nodes of WN-Ja match the upper level nodes of Wikipedia-derived hypernyms. This forms the strongest support for our suggestion that Wikipedia-derived hypernyms cannot do without WN. Rather, the two kinds of resource enhance each other.

3.3 Details of the hypernym paths

In Table 2, we show some examples with relevant details. The most common 12 root hypernyms were picked with example paths. In most cases, the lowermost elements of the hypernym paths are hypernyms for named entities. This tendency is obvious when they are at the bottoms of the long paths with more than one modifiers. All in all, the result suggests that the structure of modification needs to be carefully examined to have effective links between named entities and categories/classes of upper ontology.

3.4 Prospects for sense matches

In the example above, the hypernym matches against WN-Ja are simple string-matches and are not sense-matches, because sense disambiguation is not performed on any of the path elements.

This is regrettable. Fortunately, the co-occurrence information required for sense disambiguation on the upper-ontological elements, which have WN-Ja matches, is already available in the paths as long as they are long enough. Actually, it is intuitively obvious that the terms with WN-Ja matches have sufficiently specific senses, unless they are too short. For example, 手 in 手:選手:スポーツ選手:グルジアのスポーツ選手 of (4i) corresponds to agent-denoting suffix *-er* of English, though it means "hand" when it is used us an independent word. For, the English translation of the path would be: *-er: player: sport(s) player: Georgian sport(s) player*.[12] This can be contrasted with cases like 手:禁じ手:相撲の禁じ手 which can be translated into *technique(s): prohibited technique(s): prohibited technique(s) in Sumo wrestling*.

In Japanese analysis, recognition of sub-lexical units like 手 in 選手 and 機 in コミューター機 is unavoidable, because they are bound morphemes that play a role in basic word-formation. To our great annoyance, they are not always properly recognized in the analysis using Japanese tokenizers because they tend to treat them as single units when combinations become conventional. For example, there is no tokenizer that separates 手 from 選手.[13]

This implies that comparison of daughter terms on the WN-Ja side would enable sense matches; and that sense disambiguation is easy to do if (i) enough positive examples of specific senses are provided in composite form and (ii) similarity of a target term against the composite positive examples can be calculated. Thus, the only barrier is

[12] Incidentally, 手 is not the only morpheme that corresponds to *-er*. 者 and 人 are other major possibilities.

[13] Another complication for composite terminology is obvious here. The most appropriate English translation of コミューター機 of (4b) would be "commuter type" (of aircraft) or "commuter model" (of aircraft) rather than "commuter apparatus" or "commuter machine" even if the most straightforward translation of 機 would be "apparatus" or "machine."

Table 2: Most common 12 path elements (including unsaturated (L) and dubious (D) ones): terms with asterisk (e.g., 者*, 品*, 社*, 家*) are bound morphemes whose hypernym status are dubious.

Rank	Term	Count	Sample Hypernym Path(s)
1	者*	2,396	者* (person): 首謀者 (mastermind): 直接首謀者 (active mastermind): 事件の直接首謀者 (active mastermind of (the) affair): 爆破事件の直接首謀者 (active mastermind of (the) bombing affair)
2	品*	2,115	(1) 品* (item): 製品 (product): ドイツの製品 (products of Germany) (2) 品 (item): 用品 (item(s) for . . .): 園芸用品 (gardening supply)
3	社*	1,973	(1) 社* (company): 出版社 (publisher): 音楽出版社 (music publisher): 日本の音楽出版社 (music publisher in Japan) (2) 社* (place for sacred activity): 神社 (shrine): 市の神社 (shrine of (a) city): 鎌倉市の神社 (shrine of Kamakura City)
4	会社	1,881	社* (company): 会社 (company): 食品会社 (food company): 大手食品会社 (major food company)
5	番組	1,758	番組 (program): 音楽番組 (music program): クラシック音楽番組 (classical music program)
6	作品	1,630	品* (item): 作品 ((piece of) work): 題材にした作品 ((piece of) work on . . .): 吸血鬼を題材にした作品 ((piece of) work on vampires)
7	家*	1,615	(1) 家* (family): 五家 ((major) five schools): 禅宗五家 ((major) five schools of Zen): 中国禅宗五家 ((major) five schools of Chinese Zen) (2) 家* (-ist): 運動家 (activist): フェミニズム運動家 (feminism activist)
8	人*	1,496	人* (person): 料理人 (cook): フランス料理人 (French cook)
9	校*	1,482	校* (school): 学校 (school): 高校 (high school): 女子高校 (girl's high school): 公立女子高校 (public girl's high school)
10	手*	1,425	(1) 手* (-er): 騎手 (jockey): イギリスの騎手 (British jockey) (2) 手 (technique(s)): 禁じ手 (prohibited technique(s), foul): 相撲の禁じ手 (prohibited technique(s) in Sumo wrestling)
11	人物	1,356	人物 (person): 長寿人物 (longevity person): 最長寿人物 (the oldest person): 世界最長寿人物 (world's oldest person): 元世界最長寿人物 (former world's oldest person)
12	選手	1,242	手* (-er): 選手 (player): 野球選手 (baseball player): プエルトリコの野球選手 (baseball player of Puerto Rico)

that we do not have enough positive examples for word senses in composite forms, arguing for the building of sense tagged data with broad coverage. In other words, if we build sense tagged data based on Wikipedia, it would be quite beneficial. We will try on this in future using the method described in Toral et al. (2009).

4 Discussion

The WN-Ja coverage over the original hypernym-hyponym pairs was only 8%: that is 8% of the extracted pairs were already found within the Japanese WordNet.[14] This means that most of the pairs extracted in §2.4 are new additions to WordNet. We are adding a great deal of new information to the Japanese WordNet.

Looking at named entities specified as hyponyms in the Wikipedia data and entities in WN-Ja, there are a lot of **missing links** with which various intermediate concepts can be specified. Our impression is that these intermediate, concrete enough concepts are exactly the concepts that people use to conceptualize the world around them. For example, *famous rock singer* (of a country) and *former member* (of a group) in (1). We may assume that they are building blocks in their mental models. If this is correct, filling the missing links would be very rewarding for NLP applications and related fields such as the SemanticWeb. Admittedly, it needs more research to validate this hypothesis.

Hypernym-hyponym pairs automatically acquired from Wikipedia cannot be linked fully automatically. We required manual processing for the hypernym cleaning. With current extraction techniques lexical hierarchy data constructed fully automatically from Wikipedia is very unlikely to be as precise as WordNet's synset hierarchies.

Finally, we would like to also note that the kind of upper ontology specified in the form of WordNet and similar lexical resources would not be enough to cover the incredible variety of ontological entities that appear in Wikipedia. In particular, it contains quite a lot of imaginary entities — most notably, a full range of characters that appear in books, movies, legends, and folk tales. It is understandable, however, that they are not just components of people's fantasies but are actual elements of people's realities. Sometimes, it becomes quite hard to tell if they are real or unreal. Lexical resources like WordNet do not currently provide a proper place to hold them all. We may need to broaden the standard upper ontologies to meet the specification requirements by Wikipedia that seems to describe people's realities without categorically distinguishing between fact and fictions, between true and untrue facts, and between scientific and unscientific knowledge. Wikipedia can be a challenge for scientific categorization because pieces of knowledge of all kinds are mixed in it together. It would not be surprising if no single upper ontology can successfully handle it.

There has been much work on linking the English Wikipedia to WordNet, with YAGO (Suchanek et al., 2007) being a good example. Our work differs in several ways. Trivially, we are looking at Japanese, rather than English. More interestingly, we only consider only hypernym relations, while YAGO considers a wide range of relations, such as `BornInYear` and `LocatedIn`. On the other hand, we consider a wider range of possible entities: YAGO only looks at Wikipedia entries and their categories while Sumida et al. (2008) considers the text within the entry. Because of this, there is no guarantee that the terms we link are unambiguous entities, in fact we collapse even Wikipedia disambiguation pages. In future work, we hope to disambiguate these again, perhaps using automatic methods such as Toral et al. (2009).

In future work, we hope to extend these links to English, exploiting the multilingual links in both WordNet and Wikipedia, in cooperation with ongoing work on hyponymy extraction in both languages (Oh et al., 2009).

5 Conclusion

This paper described base hypernym extraction from the hypernym-hyponym pairs automatically acquired from the Japanese Wikipedia. It then compared Wikipedia-derived hypernyms and the lemmas of WordNet-Ja by determining how many matches there are at which levels. The results suggest that neither of the two data sources are redundant. This means that we cannot fully dispense with WordNet-like, manually developed high-precision lexical resources even if we have Wikipedia. Thus, the two kinds of resources are best understood as complementary to each other. In fact, if they are successfully coupled, we can finally have links from named entities to abstract entities in the upper ontology. The links help to

[14] This comparison was made using WN-Ja 0.8.

form the set of all encompassing, all inclusive hierarchies that we long for.

References

Francis Bond, 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-2008)*.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the Japanese WordNet. In *The 7th Workshop on Asian Language Resources*, pages 1–8, Singapore. ACL-IJCNLP 2009.

P. Eschenbach. 1993. Semantics of number. *Journal of Semantics*, 10(1):1–31.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Kow Kuroda, Masaki Murata, and Kentaro Torisawa. 2009. When nouns need co-arguments: A case study of semantically unsaturated nouns. In *Proceedings of the 5th International Workshop on Generatvei Approaches to the Lexicon, Sep. 17-19, 2009, Pisa, Italy*, pages 193–200.

Jong-Hoon Oh, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. Bilingual co-training for monolingual hyponymy-relation acquisition. In *Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on NLP of the Asian Federation of NLP*, Singapore.

F. M. Suchanek, G. Kasneci, and G. Weikum. 2007. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International World Wide Web Conference (WWW 2007)*, New York, NY, USA. ACM Press.

Asuka Sumida, N. Yoshinaga, and Kentaro Torisawa. 2008. Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in Wikipedia. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*.

A. Toral, Ó. Ferrández, E. Agirre, and R. Muñoz. 2009. A study on linking Wikipedia categories to Wordnet synsets using text similarity. In *RANLP 2009*.

V. N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.