Integration of heterogeneous language resources: A monolingual dictionary and a thesaurus

Tokunaga Takenobu[†], **Syôtu Yasuhiro**[†], **Tanaka Hozumi**[†] and **Shirai Kiyoaki**[‡]

[†] *Tokyo Institute of Technology*
[‡] *Japan Advanced Institute of Science and Technology*
{take,shotsu,tanaka}@cl.cs.titech.ac.jp, kshirai@jaist.ac.jp

Abstract

Linguistic knowledge plays a crucial role in natural language processing. Constructing large linguistic knowledge bases requires a lot of human effort and much cost. There have been many attempts to construct linguistic knowledge automatically, based on two primary strategies: knowledge extraction from annotated corpora and the augmentation of existing knowledge bases using annotated corpora. This paper describes an algorithm to enlarge existing linguistic knowledge through integration with heterogeneous linguistic resources. Specifically, this algorithm links a word sense defined in a monolingual dictionary to semantic classes in a thesaurus. Experiments show that we achieve a linking precision of 85.5% and coverage of 61.4%.

1 Introduction

Linguistic knowledge plays a key role in natural language processing. However constructing large and precise linguistic knowledge bases requires much human effort and cost.

In order to overcome this problem, many attempts have been made to automatically construct various types of linguistic knowledge bases. In terms of resources from which linguistic knowledge bases have been constructed, such attempts can be classified into two basic approaches. In the first approach, linguistic knowledge is extracted from plain text corpora (Pereira et al., 1993; Tokunaga et al., 1995; Utsuro et al., 1998). This approach does not require structured or annotation-rich linguistic resources, and thus tends to be statistical oriented. In consequence, less human labor is necessary in construction but the preciseness of the resultant knowledge is diminished, and it is difficult to use such knowledge bases without manual correction.

The second approach assumes "core" knowledge which is usually compiled by humans and therefore of adequate precision. This core knowledge is expanded upon by adding information extracted from plain linguistic knowledge (Uramoto, 1996; Tokunaga et al., 1997). Since this approach utilizes precise core knowledge as a backbone, it has an advantage in terms of precision over the first approach. In both of these approaches, the constructed knowledge is homogeneous: for example, in the second approach, resultant knowledge would be of the same type as the core knowledge.

In other words, a homogeneous linguistic knowledge capture a certain aspect of language (words), even though a word has various aspects such as morphological, syntactic, semantic and so on. It is not, however, realistic to construct linguistic resources considering all aspects at the same time. Actually, linguistic resources have been constructed focusing on a certain kind of aspect, moreover in many cases, each resource has been constructed independently based on a different framework. Considering such background, it is worthwhile to explore a method to integrate heterogeneous resources to cover more aspects of language.

In this paper, we propose a method for integrating different kinds of manually-compiled linguistic knowledge, to enlarge the knowledge base at hand. Our approach is similar to the second approach above, but different in that we use only well-structured knowledge and produce heterogeneous linguistic knowledge. Specifically, our method automatically constructs links between word senses defined in a monolingual dictionary and the semantic classes of a thesaurus.

A monolingual dictionary defines the word senses of a word in the form of a set of natural language-based meaning descriptions. On the other hand, a thesaurus classifies words according to their semantic classes, often hierarchically. Unlike monolingual dictionaries, thesauri focus on relations between words rather than the actual definitions of those words. Thus monolingual dictionaries and thesauri describe the characteristics of words from different aspects. Integrating these different resource types provides heterogeneous word knowledge, in which the information provided for each word is enriched; as such, each type of linguistic knowledge is supplemented by other knowledge based on different aspects. Such kind of linguistic knowledge is rare, with the notable exception of WordNet (Fellbaum, 1998).

The structure of this paper is as follows. Section 2 describes an overview of linguistic resources discussed in this paper. Section 3 and 4 describe two different methods to linking different linguistic resources introduced in section 2. The method described in section 3 is based on superordinate/subordinate relations between words in definition sentences of the dictionary. On the other hands, the methods in section 4 uses common verbs appearing in definition sentences for linking. Section 3 and 4 describes preliminary experiments and the results as well. Section 5 compares the proposed methods with related work. And finally we conclude the paper and mention future work in section 6.

2 Dictionary and Thesaurus

In this paper, we use the Iwanami Japanese dictionary (5th ed.) (Nisio et al., 1994) as our monolingual dictionary, and *Nihongo Goi Taikei* (the NTT Japanese thesaurus) (Ikehara et al., 1997) as our thesaurus. We focus on finding the relations between nouns appearing in both of these resources. Since the number of nouns is larger than that of other parts of speech, and difference of difinitions is easier to identify.

The Iwanami dictionary has been annotated with various tags as part of the RWCP text database project (Hasida et al., 1998), including morphological tags, word sense tags, and syntactic relation tags. The dictionary has 51,438 noun entries, with more than one word sense for some entries.

Nihongo Goi Taikei classifies a total of 264,312 nouns into 2,710 semantic classes. Many words are classified into more than one semantic class. *Nihongo Goi Taikei* has a hierarchical structure, in which nodes correspond to semantic classes and links between nodes denote a IS-A or HAS-A relations between the linked semantic classes. Figure 1 shows a fragment of *Nihongo Goi Taikei*.

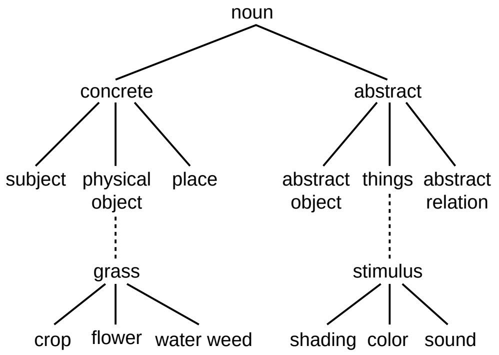

Figure 1: A fragment of *Nihongo Goi Taikei*

3 Linking by superordinate words

This section describes a method to identify the superordinate word of an entry word in the monolingual dictionary, and using this information, to find a link between a word sense in the dictionary and a semantic class in the thesaurus.

3.1 Identifying superordinate words

Much research has aimed at extracting relations between words in a monolingual dictionary (Tsurumaru et al., 1986; Nakamura and Nagao, 1988). In the manner of previous research, we analyze dictionary definitions, and extract patterns that suggest superordinate words of the entry word. When the dictionary definition ends with one of the following patterns, ⟨noun⟩ is identified as the superordinate word of the entry.

(a) ⟨noun⟩.

(b) ⟨noun⟩ すること. (*doing* ⟨noun⟩)

(c) ⟨noun⟩ をすること. (*doing* ⟨noun⟩)

(d) ⟨noun⟩ の一つ. (*a kind of* ⟨noun⟩)

(e) ⟨noun⟩ の一種. (*a kind of* ⟨noun⟩)

(f) ⟨noun⟩ の略. (*an abbreviation of* ⟨noun⟩)

(g) ⟨noun⟩ の〜称. (*an alias of* ⟨noun⟩)

There are several exceptions to ⟨noun⟩, such as "こと (*thing*)", "一種 (*a kind of*)", and " 方 (*way*)". These words are not identified as superordinate words.

3.2 Linking words based on thesaurus path length

Relations between word senses in the dictionary and semantic classes in the thesaurus are identified based on the path length between the entry word and its superordinate word, as identified in the previous step. For example, suppose a word w has two word senses ws_1 and ws_2 in a dictionary, and w is classified into the two thesaurus classes sc_1 and sc_2. Our goal is to identify the correct mappings between ws_i and sc_j. In order to achieve this goal, the path length between each sc_i and $sup(ws_j)$ is calculated, and that sc_i of shortest path length from $sup(ws_j)$ is selected as the semantic class of ws_j in the thesaurus. Here $sup(ws_j)$ denotes the superordinate word of word sense ws_j.

However, utilizing simple path length does not work for calculating the distance between two semantic classes due to the following problems. First, in *Nihongo Goi Taikei*, the path length from the root node to each leaf node is not uniform, which means that the semantic similarity between a node and its parent node depends on their relative location in the thesaurus. For example, pathes from the two leaf nodes "villain" and "monster" to the root node are shown below:

- "villain" → "villain etc." → "good/bad person" → "human (nature)" → "human (ability, nature)" → "human" → "man" → "subject" → "physical object" → "noun"
- "monster → "pseudo human" → "man" → "subject" → "physical object"

In this example, the semantic difference between "villain" and "villain etc." is more subtle than that between "monster" and "pseudo human". Since our method links word senses and semantic classes based on path length in the thesaurus, this feature of the thesaurus could potentially cause problems.

Another problem occurs when a corresponding semantic class does not exist in the thesaurus. For example, the Iwanami dictionary defines two word senses for "藍 (*indigo*)": "plant" and "dye". On the other hand, *Nihongo Goi Taikei* classifies "藍 (*indigo*)" into the two classes "dye" and "color", but not "plant". Based on the simple shortest path length criteria, the "plant" word sense would be linked to the wrong semantic class.

In order to solve these problems, we take the following approach. For the first problem, we weight each node in the thesaurus, and calculate the distance between two nodes by adding all weights of nodes on the path between them. The weight of node c is calculated by the following formula.

$$W(c) = \begin{cases} 100 & (d < 4) \\ \frac{1}{d} & (d \geq 4, C = \phi) \\ \dfrac{|C|}{\displaystyle\sum_{c_i \in C} \frac{1}{W(c_i)}} & (d \geq 4, C \neq \phi) \end{cases} , \tag{1}$$

where d denotes a depth of semantic class c from the root node, C denotes a set of children class of c. The depth of the root node is defined as 0. Equation (1) assign a weight inversely proportional to its depth to a leaf node. Otherwise, an average of its children weights is assigned when its depth is deeper or equal to 4. When the depth is shallower than 4, the node is put heavy weight 100. This value works as threshold. Figure 2 illustrates an example of this weighting scheme. A distance of two nodes is calculated by summing weights of all nodes on the path between the nodes.

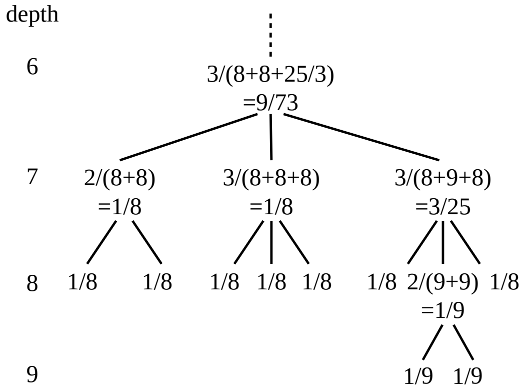

Figure 2: An example of path weighting

For the second problem, we set a (minimum) threshold on plausible links between word senses and semantic classes. This value was set to 100 in the experiments described below. This threashold is reflected in the first case in equation (1).

3.3 Normalizing orthographical variants

There are orthographical differences between the Iwanami dictionary and *Nihongo Goi Taikei*. In order to resolve such differences, entry words are preprocessed before linking in the following way.

- When a word can be written in several forms, the Iwanami dictionary denotes the word in a particular way. For example, characters enclosed in parentheses are optional, such as in "明 (か) り (*light*)". All possible lexicalisations are generated is such cases. With this example, therefore, we would generate "明かり" and "明り". When finding the corresponding semantic classes, all expanded forms are used.
- The Iwanami dictionary tends to use hiragana (phonograms) in definition sentences, which causes mismatches when finding the semantic class of the superordinate word. On failing to find a semantic class, hiragana strings are converted into kanzi strings by referring to the Iwanami dictionary, and the search is retried.
- When the idetified superordinate word is a compound noun and not found in the thesaurus, its head noun (the rightmost noun) is searched for. For example, " 一年生植物 (*annual plant*)" is not found in the thesaurus, but a search for "植物 (*plant*)" succeeds.

3.4 Experiments

We first evaluated our method of identifying superordinate words, as described in Section 3.1. Table 1 shows the results of our experiments. We categorized the seven patterns (a) through (g) in Section 3.1 into three groups, and evaluated the performance of each group. Precision is calculated based on human inspection of 300 sample entries (100 for each group). There are 64,082 word senses defined for nouns in the Iwanami dictionary, and hence the coverage of our method is 64.1%.

Group	# word senses	Precision
(a)	36,485	92%
(b), (c)	1,901	86%
(d), (e), (f), (g)	2,692	93%
Total/weighted ave.	41,078	91.8%

Table 1: Precision of superordinate word identification

Using these results, we linked word senses in the Iwanami dictionary with semantic classes in *Nihongo Goi Taikei* by the method described in Sections 3.2 and 3.3. 43,884 words are included in both the Iwanami dictionary and *Nihongo Goi Taikei*, accounting for a total of 55,875 word senses. The proposed method found semantic classes for 27,853 word senses out of 55,875. Therefore we achieved a coverage of 49.8%. Based on manual inspection of 100 sample entries, the precision was 84.5%[1]

4 Linking by common verbs

We achieved relatively high precision in linking word senses with semantic classes based simply on the superordinate words of entry words. However coverage is far from satisfactory. In order to improve coverage, in this section we introduce a new linking method based on common verbs shared by the dictionary definitions of different entry words.

When an entry word means a certain action, its dictionary definition often corresponds to the pattern "⟨verb⟩ こと (*to* ⟨verb⟩)", as in "飲酒: 酒を飲むこと (*to drink (alcohol)*)". Entries defined using the same verb tend to share the same semantic class. Using the distribution of semantic classes of linked entries, we make links between word senses and semantic classes. Links are established by the following procedure.

1. Calculate the frequency $fr(v,c)$ of entry w with definition sentence ending in "⟨verb v⟩ こと" being classified as semantic class c in the thesaurus.

2. Calculate the probability $P(c|v)$ for all pairs v and c where $fr(v,c) \geq 2$.

$$P(c|v) = \frac{fr(v,c)}{\sum\limits_{c_i \in C_v} fr(v,c_i)},$$

   where $C_v = \{c | fr(v,c) \geq 2\}$.

3. If the candidate semantic classes of an entry w are $c_{w_1}, \ldots, c_{w_n}$, and those of entries with definitions involving verb v are $c_{v_1}, \ldots, c_{v_m}$, calculate the distance $dist(c_{w_i}, c_{v_j})$ between every c_{w_i} and c_{v_j} based on the method described in Section 3.2.

4. Find the pair (c_{w_i}, c_{v_j}) giving the minimum value of $\frac{dist(c_{w_i}, c_{v_j})}{P(c_{v_j}|v_j)}$, and link the word sense to semantic class c_{w_j}.

In step 4 above, dividing the distance by $P(c_v|v)$ increases the distance when verb v of entry w does not strongly suggest semantic class c_v. Prior to this step, the orthography-based preprocessing described in Section 3.3 is also done.

In addition to the links obtained by the method based on superordinate words, we obtained a further 6,486 links by this method. In total, we have thus obtained 34,339 links, improving coverage from 49.8% to 61.4%. The precision of the new links is 90.0%, based on manual inspection of 100 sample entries. The combined precision is thus 85.5%.

5 Related work

Chen and Chang took a different appoach to achieve the same goal as us (Chen and Chang, 1998). Their method, the Linksense algorithm, utilizes an information retrieval technique, that is, the dictionary definition of each word sense is treated as a query, and semantic classes in the thesaurus are treated as documents for retrieval. In this section, we apply the Linksense algorithm to the Iwanami dictionary and *Nihongo Goi Taikei*, and compare these two algorithms.

The Linksense algorithm calculates the similarity between word sense D and semantic class C based on the following formula:

$$Sim(D,C) = \frac{\sum_{d \in \mathrm{KEY}_D} 2 \cdot w_d \cdot In(d,C)}{|\mathrm{KEY}_D| + 1},$$

where KEY_D is the set of keywords appearing in definition sentences of word sense D, w_k is the inverse of the number of semantic classes into which word d can be classified, and $In(d,C)$ is a binary function, returning 1 if d is classified into D or its ancestors or children, and 0 otherwise. The definition of $In(d,C)$ has been modified from Chen and Chang's original definition in order to reflect the difference in language resources.

[1] The judgement was done in three leves; correct, partially correct and wrong. A partially correct case was counted as 0.5, which took 5 % of call cases.

Chen and Chang used the *Longman Dictionary of Contemporary English* (LDOCE) and *Longman Lexicon of Contemporary English* (LLOCE).

That link between D and C which gives the maximum $Sim(D, C)$ is selected as the sense mapping. When Sim is below a certain threshold (0 in our experiments), no link is established.

Applying the Linksense algorithm to our language resources, we were able to link 38,045 word senses to semantic classes, representing coverage of 68.1% and precision of 79.5%.

Table 2 summarizes the performace of Linksense and our proposed methods. We can see that Linksense is superior to ours in terms of coverage, but that we have the edge in terms of precision. We suggest that this is due to the robustness afforded by the information retrieval technique used in Linksense.

Table 2: Comparison with Linksense

	# word senses	Coverage	Precision
Our method	34,339	61.4%	85.5%
(sup-word)	27,853	49.8%	84.5%
(verb)	6,486	11.6%	90.0%
Linksense	38,045	68.1%	79.5%

We evaluated the overlap of word senses successfully linked by both methods, and present the results in Table 3. A total of 27,199 word senses were mapped onto semantic classes by the respective methods, of which 24,457 were mapped by both methods and 2,607 by only one of the two methods. For these word senses, there was no significant difference in precision between Linksense and our method.

7,139 word senses were linked by our proposed method but not by Linksense. The reverse case was observed for 10,846 word senses. Again, Linksense achieves higher coverage, but significantly lower precision (57% vs 83%).

We further qualitatively investigated the results of these two methods and discovered the following characteristic.

- Linksense tends to link word senses to related but not fully appropriate semantic classes. For example, Linksense linked " 愛書 (*to love books*)" to the two semantic classes "出版物 (*published matter*)" and "本 (内容) (*book (contents)*)". This occurs because Linksense uses an information retrieval technique.

- Our proposed method tends to link word senses to one semantic class, whereas Linksense tends to link them to more than one class. This tendency reflects the difference in distance (similarity) measure. Linksense similarity is based on overlap of keywords. Since the number of keywords in the dictionary definition is few, this measure often returns a tied score.

- In both methods, almost 80% of failed cases correspond to the instance of a link being established where there is no corresponding semantic class in the thesaurus.

6 Concluding remarks

This paper has proposed a method to find the relation between word senses in a dictionary and semantic classes in a thesaurus. We first analyze the dictionary definition of each word sense, and identify its superordinate word using a pattern matching technique. By calculating the distance between the entry word and its superordinate word, the corresponding semantic class is identified. Through experiments with the Iwanami dictionary and *Nihongo Goi Taikei* (NTT Japanese thesaurus), the proposed method was able to establish links for 49.8% of word senses in the dictionary, at 84.5% precision.

In order to improve coverage, we introduced another method based on common verbs appearing in dictionary definitions. This method improved the coverage to 61.4% and the precision to 85.5%.

We went on to compare our method with the Linksense algorithm, which is based on an information retrieval technique. We found

Table 3: Overlap with Linksense

Linked by	# word senses	Coverage	Precision	
			Our method	Linksense
Both methods	27,199	48.7%	86.1%	86.4%
Overlap	24,457	43.8%	87.0%	87.0%
No overlap	2,607	4.7%	82.0%	85.0%
Only our method	7,139	12.8%	83.0%	—
Only Linksense	10,846	19.4%	—	57.0%

Linksense to be slightly better in coverage, but our method to be superior in precision.

We obtained promising results in precision, but coverage needs further improvement. For this pourpose, further analysis of dictionary definitions is necessary. In particular, with the current method, only the first sentence of the word sense definition is analyzed in identifying the superordinate word. However, there are cases where related but slightly different meanings are defined in succeeding sentences.

Another issue to resolve is precise criteria to judge if a corresponding semantic class actually exists in the thesaurus. Both the proposed method and Linksense tend to establish wrong links due to the lack of a corresponding semantic class.

As the comparison of the proposed method with the Linksense algorithm suggested, the different approaches display common results to a certain extent, but do not overlap perfectly. Integrating the different approaches might improve overall performance. Investigation in this direction is necessary.

Another direction of future work includes integrating other types of linguistic resources. In this paper, we focused on nouns. When considering verbs, integration of semantic knowledge and syntactic knowledge such as subcategorization frames would be possible.

References

J. N. Chen and J. S. Chang. 1998. Topical clustering of MRD sense based on information retrieval technique. *Computational Linguistics*, 24(1):61–93.

C. Fellbaum. 1998. *WordNet, An Electronic Lexical Database.* MIT Press.

K. Hasida, H. Isahara, T. Tokunaga, M. Hashimoto, S. Ogino, W. Kashino, J. Toyoura, and H. Takahashi. 1998. The RWC text databases. In *Proceedings of The First International Conference on Language Resource and Evaluation*, pages 457–461.

S. Ikehara, M. Miyazaki, A. Yokoo, S. Shirai, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. 1997. *Nihongo Goi Taikei – A Japanese Lexicon.* Iwanami Syoten. 5 volumes. (In Japanese).

J. Nakamura and M. Nagao. 1988. Extraction of semantic information from an ordinary English dictionary and its evaluation. In *Proceedings of the 12nd Internationa Conference on Computational Linguistics*, pages 459–464.

M. Nisio, E. Iwabuti, and S. Mizutani, editors. 1994. *Iwanami Okugo Ziten.* Iwanami Syoten, 5 edition.

F. Pereira, N. Tishby, and L. Lee. 1993. Distributional clustering of English words. In *Proceedings of the 32nd Annual Meeting of the ACL*, pages 183–190.

T. Tokunaga, M. Iwayama, and H. Tanaka. 1995. Automatic thesaurus construction based on grammatical relations. In *Proceedings of IJCAI '95*, pages 1308–1313.

T. Tokunaga, A. Fujii, N. Sakurai, H. Tanaka, and M. Iwayama. 1997. Extending a thesaurus by classifying words. In *Proceedings of the Workshop Sponsored by the Association for Computational Linguistics "Automatic Information Extraction and Building of Lexical Semantic Resources"*, pages 16–21.

H. Tsurumaru, T. Hitaka, and S. Yoshida. 1986. An attempt to automatic thesaurus construction from an ordinary Japanese language dictionary. In *Proceedings of the 11th International Conference on Computational Linguistics*, pages 445–447.

N. Uramoto. 1996. Corpus-based thesaurus– positioning words in exsisting thesaurus using statistical information from a corpus. *Transaction of Information Processing Society of Japan*, pages 2182–2189. in Japanese.

T. Utsuro, T. Miyata, and Y. Matsumoto. 1998. General-to-specific model selection for subcategorization prefereence. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the ACL*, pages 1314–1320.