Positioning Unknown Words in a Thesaurus by Using Information Extracted from a Corpus

Naohiko URAMOTO
IBM Research, Tokyo Research Laboratory
1623-14 Shimo-tsuruma, Yamato-shi, Kanagawa-ken 242 Japan
uramoto@trl.ibm.co.jp

Abstract

This paper describes a method for positioning unknown words in an existing thesaurus by using word-to-word relationships with relation (case) markers extracted from a large corpus. A suitable area of the thesaurus for an unknown word is estimated by integrating the human intuition buried in the thesaurus and statistical data extracted from the corpus. To overcome the problem of data sparseness, distinguishing features of each node, called "viewpoints" are extracted automatically and used to calculate the similarity between the unknown word and a word in the thesaurus. The results of an experiment confirm the contribution of viewpoints to the positioning task.

1 Introduction

Thesauruses are among the most useful knowledge resources for natural language processing. For example, English thesauruses such as Roget's Thesaurus and WordNet [4] are widely used for tasks in this area [5, 6, 3]. However, most existing thesauruses are compiled by hand, and consequently, the following three problems occur when they are used for NLP systems.

First, existing thesauruses have insufficient vocabularies, especially in languages other than English. In Japan, there are no free thesauruses that can be shared by researchers. Furthermore, general-domain thesauruses do not cover domain-specific terms.

Second, the human intuition used in constructing thesauruses is not explicit. Most existing thesauruses are hand-crafted by observing huge amounts of data on the usage of words. The data and human judgements used in constructing thesauruses would be very useful in NLP systems; unfortunately, however, this information is not represented in the thesauruses.

Third, the structure of thesauruses is subjective. The depth and density of nodes in (tree-like) thesauruses directly affect the calculated distances between words. For example, nodes for biological words have many levels, while abstract words are classified in relatively shallow levels. However, existing thesauruses only represent uniform relationships between words.

This paper describes a way of overcoming the problems, using a medium-size Japanese thesaurus and large corpus. The main goal of our work is to expand the thesaurus automatically, explicitly including distinguishing features (viewpoints), and to construct a domain-sensitive thesaurus system.

To expand the vocabulary of the thesaurus, it is important to position new words in it automatically. In this paper, words that are not contained in the thesaurus but that appeared in the corpus more than once are called *unknown words.*[1] The proper positions of the unknown words in the thesaurus are estimated by using word-to-word relationships extracted from a large-scale corpus. This task may be similar to word-sense disambiguation, which determines the correct sense of a word from several predefined candidates. However, in positioning a word whose sense is unknown, a suitable position must be selected from thousands of nodes (words) in the thesaurus, and therefore it is very difficult to position the word with pinpoint accuracy. Instead, in this paper, we give a method for determining the area in which the unknown words belongs. For example, suppose the word "SENTOUKI" (fighter)[2] is not contained in a thesaurus. Calculation of the similarity between the word and those in the thesaurus assigns it to the area [flying vehicle [air plane, helicopter]].

Viewpoints are features that distinguish a node from other nodes in the thesaurus, and are good clues for estimating the area to which an unknown word should be assigned. The area can be efficiently estimated by extracting viewpoints.

Several systems have used WordNet and statistical information from large corpora [3, 5, 6]. However, there are two common problems: noisy co-occurrence of words and data sparseness. In WordNet, since each node in the thesaurus is a set of words that have synonym relationships (SynSet), various methods for similarity calculation using the SynSet classes have been proposed. In this paper, ISAMAP [8], a hand-crafted Japanese thesaurus, is used as a core. To overcome the problems of noise

[1] That is, unknown words do not mean very low-frequency words.

[2] A Japanese word in ISAMAP is represented by a pair of capital Roman letters and the word's English translation.

具体物	(Physical Object)	現象	(Phenomenon)
有意志体	(Creature)	関係	(Relation)
抽象物	(Abstract Object)	時	(Time)
方法	(Method)	場所	(Location)
活動	(Action)	空間	(Space)
属性	(Attribute)	単位	(Unit)
状態	(State)	操作	(Operation)
力	(Force)		

Fig. 1: Top Categories of ISAMAP

and data sparseness, relationships of connected nodes in the thesaurus are used. Resnik proposed a class-based approach, in which sets of words are used instead of words [5]. In his approach, each Synset is used as a class. In our approach, on the other hand, an area that contains connected nodes in the thesaurus is used as a class. The nodes are connected by IS-A relationships as well as synonym relationships, and therefore large areas represent strong similarities to unknown words.

2 Knowledge Sources

This section describes the thesaurus and statistical data used in this paper. A Japanese noun thesaurus called ISAMAP is a set of IS-A relationships. It contains about 4,000 nouns with about ten levels. Each node of ISAMAP is a word or a word and its (one or two) synonyms. Figure 1 shows the top categories of ISAMAP. Some words are placed at multiple positions in the thesaurus. For example, SENSUIKAN (submarine)" is classified as "water vehicle" and "weapon."

To extract viewpoints for the existing structure of the thesaurus in order to position unknown words in it, a collection of pairs of words and their relation markers, together with their frequency, was extracted from a corpus. The source of the words was articles published in a Japanese newspaper (Nikkei Shinbun) in 1993. The articles were morphologically analyzed automatically, and then stored in the following form:

`oc(word1, rel, word2) = n`

This means that `word1` and `word2` occur `n` times with a relation marker `rel`. Relation markers consist of case markers such as "GA", "WO", and "NI", and adnominal forms of adjectives and adjective nouns. The statistical data for each relationship are shown in Figure 2.

We use restrictive relationships with the markers, rather than word 2-grams, for two reasons. In the unknown-word-sense disambiguation task, the number of possible candidate word-senses (positions in the thesaurus, in this paper) is very large, and thus it is important to reduce noises that prevent the output of a result. Second, these case relationships can be used to identify classification viewpoints for thesauruses. For example, suppose that

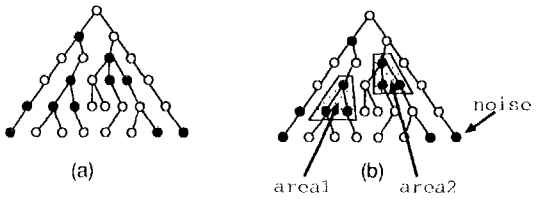

Fig. 3: Marked Nodes in the Thesaurus

the words *plane* and *ship* are located below *vehicle*. We can say that "planes fly" and "a plane in the sky," but not "ships fly" or "a ship in the sky." That is, (plane, SUBJ, fly) and (plane, in, sky) can be called viewpoints for the word "plane."

3 Positioning Unknown Words

This section describes the procedure for positioning of words in ISAMAP. In this task, the input is a word to be placed somewhere in ISAMAP. The goal is to determine the most suitable area for the word. The procedure consists of the following three steps:

Step 1: Extraction of viewpoints for each node in ISAMAP.
Step 2: Extraction of candidate areas for the input word.
Step 3: Evaluation of the candidates and selection of the most preferable area.

3.1 Basic Idea

The basic idea is very simple. For an unknown word, the word-to-word relationships that contain it are extracted. The similarity between the word and each node in ISAMAP is calculated. The nodes for which the similarity exceeds a predefined threshold are marked and connected in the thesaurus. The left tree in Figure 3 shows nodes in the thesaurus. The marked nodes are represented by black circles. For straightforward statistical similarity calculations, there are many similar words, including noisy words. In this paper, the following three hypotheses are used to resolve the problem. First, the marked words form certain areas (connected nodes) of words in the thesaurus. The areas that occupy a large space are preferred. The right tree in Figure 3 shows areas of words.

Second, specific words, that is to say, words at lower levels of trees are preferred. In Figure 3, `area1` is preferred to `area2`.

Third, each node in the thesaurus has *viewpoints* that distinguish it from other nodes. The viewpoints for each node are extracted by using case and modification relationships that contain statistical data extracted from the corpus. If an unknown word has the same viewpoints as a certain node, the similarity for the node is weighted. The next subsection describes how viewpoints are extracted.

3.2 Extraction of Viewpoints

A viewpoint is a set of distinguishing features for each node in a thesaurus. The viewpoint of a node

Marker	Distinct number	Total number	Relationship
GA	394,887	817,030	Subject (e.g. man go)
WO	483,400	1,210,581	Object (e.g. drink coffee)
HE	18,564	53,876	Goal (e.g. go to office)
NI	451,986	1,114,877	Goal, etc. (e.g. go to church)
DE	225,247	61,4619	Instrument, etc. (e.g. hit with hammer)
TO	176,738	570,475	Accompanier (e.g. man and woman)
NA	78,079	569,837	Adnominalization (e.g. basic word)
I	51,001	881,255	Adnominalization (e.g. large building)
Total	1,879,902	5,832,550	

Fig. 2: Number of Statistical Data

node is defined as a list *(node, marker, word)*. Though such features are implicitly used in the creation of most existing thesauruses according to human intuition, they are lost when the constructed thesauruses are used. An exception is the WordNet, in which the distinguishing features are manually listed. In this paper, the distinguishing features are extracted automatically, reflecting the characteristics of the corpus to be used.

For example, Figure 5 shows a part of ISAMAP. The viewpoint of a node in the thesaurus is estimated by using a certain procedure. Suppose we want to extract the viewpoint of the noun "HERIKOPUTAA" (helicopter). The word occurs 131 times in our corpus. Figure 3.2 shows examples of the relationships.

For each relationship, a search is made for nodes that have the same relationship. In the case of the pattern "TUKAU" (use), 385 nodes with the same relationship are extracted from areas, scattered throughout ISAMAP. On the other hand, the pattern "TOBU" (*fly*) shares only two nodes, *helicopter* and *airplane*. The nodes have direct IS-A relationships; in other words, the nodes are can be *connected in the hierarchy of nodes. Since the viewpoints of a node are inherited by its children in many cases, the existence of the connected nodes that include ISA relationships is strong evidence for the viewpoints. In this case, (fly, SUB) is a viewpoint for the node "airplane," which is the topmost of the connected nodes.*

Viewpoints are extracted by calculating the typicalness *of word-to-word relationships. Given a node* nd *and its candidate viewpoint (a pair of a relation marker* rel *and a word* w*), the typicalness of the viewpoint is calculated as*

$$typicalness(\mathrm{nd}, \mathrm{rel}, \mathrm{w}) = max\left(\frac{\sum_{c\in C} oc(c,rel,w)}{\sum_{n\in N} oc(n,rel,w)}, \frac{\sum_{c\in C} oc(w,rel,c)}{\sum_{n\in N} oc(w,rel,n)}\right),$$

where N is a set of nodes in ISAMAP, and C is a set of connected nodes that contain the word w*. Examples of the viewpoints (whose typicalness exceeds 0.5.) are as follows:*

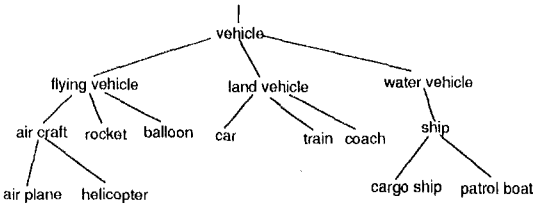

Fig. 5: Example of Viewpoints the Thesaurus

Node (word)	Viewpoints
airplane	(fly, SUBJ), (land, SUBJ), (take off, SUBJ)
rocket	(launch, OBJ)
ship	(come alongside the pier, SUBJ), (sink, SUBJ)
land vehicle	(transportation, by)

3.3 Example for Positioning Words in ISAMAP

Let us consider an example to see how algorithm works. Suppose the word "SENTOUKI" (fighter[3]) is to be placed in the thesaurus.

First, for each node in the ISAMAP, the similarity between the word and the node is calculated. The similarity is calculated according to the following formula:

$$sim(w_1, w_2) = max(sim_1, sim_2)$$

$$sim_1 = \sum_{p\in P}\left(\frac{oc(w_1,r,p)}{oc(\_,r,p)} + \frac{oc(w_2,r,p)}{oc(\_,r,p)}\right)$$

$$sim_2 = \sum_{p\in P}\left(\frac{oc(p,r,w_1)}{oc(p,r,\_)} + \frac{oc(p,r,w_2)}{oc(p,r,\_)}\right)$$

P is set of words that co-occurs with $w_1 or w_2$, and the argument "_" can be any words. If the similarity value exceeds a pre-defined threshold, the node is marked. Figure 6 shows marked nodes that have high similarity.

[3]In English, a fighter means both a plane and a person; however, the original Japanese word SENTOUKI means only a plane.

Word	Marker	Freq	Word	Marker	Freq
TUKAU (to use)	wo	10	TYAATA (to charter)	WO	2
SIMA (on an island)	DE	5	HUKUMU (to contain)	WO	2
KYUUJYO (to save)	NI	3	KANPAN (on deck)	DE	2
TOBU (to fly)	WO	2	SYUTUDO (to take the field)	WO	2
TURIAGERU (hang by)	DE	2	JYOUKUU (in sky)	DE	2

Fig. 4: Example of Relationships for "helicopter."

	Word	Node-id	Relationships
1	HITO (human)	1.0	bad, protect
2	KABU (stock)	0.0.1.6	purchase, have, buy
3	SEIHIN (manufacture)	0.0.0.0	purchase, have, buy
4	MONO (object)	0.0	purchase, have, buy
5	KIN (gold)	0.0.6.3.0.6.0	purchase, have, rob
6	JYUUTAKU (house)	0.0.0.1.0.1.0	purchase, have, buy
7	GIJYUTUSYA (engineer)	1.0.67	have, send
8	KIGYOU (company)	1.2.0.0.2.3	have, buy, protect
9	BUHIN (parts)	0.0.0.0.1	purchase, buy, export
10	SETUBI (facilities)	0.0.0.1.3	purchase, buy, have
11	HON (book)	0.0.0.0.0.4.2	purchase, have, buy
12	TAI (party)	1.2.0.0.14	send, danger, collision
13	KOUKUUKI (air plane)	0.0.0.0.0.0.1.0.2.0	purchase, fly, buy
14	HEIKI (weapon)	0.0.0.0.0.2	purchase, have, buy

Fig. 6: Marked nodes with matched relationships

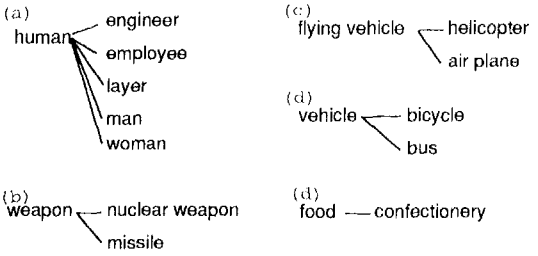

Fig. 7: Candidate connections for "fighter"

Areas that contain marked nodes are calculated. The results are given in Figure 7. The most suitable area for the word "fighter" must be selected from multiple candidate sets of connections.

The final phase is the evaluation of the candidates. Each candidate is evaluated according to the following four criteria.

Criterion 1: The size of the candidate. Given an input word w (in this case, "fighter"), and a node that is contained in the candidate C, $C1 = \sum_{node \in C} sim(w, node)$.

Criterion 2: The height of the candidate. C2 is the number of levels in the candidate. For example, in the candidate (a) in Figure 7, C2 = 2.

Criterion 3: The average depth of the nodes. For example, the depth of the node "airplane", whose node-id is 0.0.0.0.0.0.1.0.2.0, is 10.

Criterion 4: The number of viewpoints. For example, candidate (a) (whose top node is "human") has the largest number of nodes. However, as shown in Figure 6, the matched relationships ("bad human/fighter" and "protect human/fighter") are not typical expressions for the word "fighter"; that is, the relationships are not viewpoints. On the other hand, "airplane" in candidate (c) shares the "fighter (airplane) fly", which is the viewpoint of the node "airplane." C4 is the number of matched relationships that are considered as viewpoints of the node in the candidate.

The total preference P(word) is $p_1C1 + p_2C2 + p_3C3 + p_4C4$, where p_1, p_2, p_3, and p_4 are weights for each criterion. Intuitively, and according to a preliminary experiment, the contribution of C3 should carry more weight than the other criteria (in our experiment, $p_1 = 1$, $p_2 = 1$, $p_3 = 0.4$, and $p_4 = 3$). The most preferable candidate for the word "fighter" is (c); that is, "fighter" is placed in the area whose top node is "flying vehicle."

4 Experiment and Discussion

This section describes some experiments for positioning words in ISAMAP. Figure 4 shows part of the results. In the experiment, 2,000 nodes with the root "physical object" in ISAMAP were used.

Word	Position
court	(organization (union, meeting, party, class))
president	(human (man, woman, lawyer, family employee, etc))
Australia	(nation (Japan, China, Russia))
present	(object (food, hat, parts, etc))
the House of Representatives	(organization (union, meeting, party, team, etc))
author	(human (man, woman, lawyer, family, employee, etc))
seminar	(equipment (school, public equipment, parking, etc))
museum	(equipment (school, public equipment, parking, etc))
wife	(human (man, woman, lawyer, family employee, etc))
heavy oil	(object (material (fuel (gas, petroleum))))

Fig. 8: Result

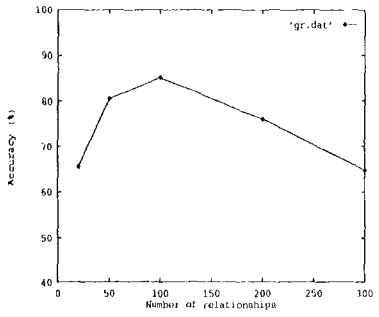

Fig. 9: Relationship between the number of relationships and the accuracy of positioning

The experiment yielded several observations. Viewpoints are strong clues for determining the suitable positions in the thesaurus for unknown words. In co-occurrence-based similarity calculation, words with strong similarities but whose relationships seem strange to human intuition reduce the accuracy of the proposed method. However, in many cases, these strong similarities are caused by less typical co-occurrences. In Figure 6, the words "buy," "purchase," and "have" convey less informative relationships than viewpoint relationships.

If there are many relationships for an unknown word, the possibility of the existence of viewpoints will increase. However, some relationships may be noisy. Figure 9 shows the relationships between the number of relationships and the accuracy of positioning. In this case, the accuracy means the percentage of words for which the most preferable area estimated by the proposed method contained the node that the word really belonged to. As shown in Figure 9, 50–100 relationships are needed to estimate the nodes. On the other hand, too many relationships prevent the extraction of useful viewpoints.

It is very difficult to position a word with pinpoint accuracy. Experiment showed that the following heuristic is useful. If an unknown word has conjunctive relationships with a node (word) in a particular area, it can be positioned as a sibling of the node. For example, the likely area of "heavy oil" is "(object (material (fuel (gas, petroleum))))", whose top node is very abstract. However, the relationship "heavy oil and gas[4]" suggests the position of "heavy oil."

By using the proposed method, the existing thesaurus was expanded to cover a large quantity of text. Though ISAMAP was designed for general purposes, the method allows it to reflect a specific domain through the use of a domain-dependent corpus. One of our goals is to develop a corpus-based thesaurus, consisting of a core thesaurus such as ISAMAP and a corpus that reflects domain knowledge. When a thesaurus is used for NLP applications, such as an information retrieval and disambiguation system, there is no need for it to have a well-defined tree-like structure. The system can use the thesaurus as a black box via certain functions. For example, the following functions are needed for a corpus-based thesaurus system:

position(w): returns the position (or path) of the word w.
superordinate(w): returns the superordinate words of the word w.
subordinate(w): returns the subordinate words of the word w.
similar(w): returns the words similar to w.
distance(w1, w2): returns the distance between w1 and w2.

It is important that the return values of the functions should be depend on the corpus and the local context of words.[5] The proposed method can be used to realize these functions. Viewpoints make it possible to realize a dynamic interpretation of distance.

5 Related Work

The method proposed here is related to two topics in the literature: automatic construction of a thesaurus and word-sense disambiguation.

[4]The marker TO indicates the conjunction.

[5]For this purpose, the functions can be expanded to contain the local context of the word as augmentations of the functions (e.g. position(w, context(w))).

There have been several studies of the automatic construction of thesauruses or sets of IS-A relationships [7, 1]. In these studies, the constructed relationships sometimes do not match human intuition. IS-A relationships do not appear in the corpora explicitly, and it is therefore difficult to extract them without including noisy relationships. In our approach, a core thesaurus is used to integrate human intuition with corpus-based co-occurrence information.

Yarowsky proposed a method for word disambiguation using Roget's Thesaurus [9]. In his approach, a word whose senses are known (a word may have several senses) is disambiguated by using "salient words" for each word-sense. A set of salient words is a list of words with no relationships. In our approach, word-to-word relationships with markers are used, in order to reduce noises and to extract viewpoints. Some other methods of word-sense disambiguation using WordNet have been proposed [2, 6, 3]. Their approaches are similar to ours, with, the difference that the sense of a word to be placed in the thesaurus is unknown. Thousands of nodes in the thesaurus are candidates, and therefore, more subtle knowledge is needed. Use of a core thesaurus and viewpoints — that is, word-to-word relationships with relation markers — makes it possible to estimate a suitable area for an unknown word.

6 Conclusion

This paper has described a method for positioning unknown words in an existing thesaurus by using word-to-word relationships with relation markers extracted from a large corpus. Suitable areas in the thesaurus for unknown words were estimated by integrating human intuition buried in the thesaurus with statistical data extracted from the corpus. Experiments showed that assigning "viewpoints" for each node gives important information can be used to estimate suitable positions in the thesaurus for unknown words. The following topics should be investigated in future work:

- When an unknown word has several word-senses, derivative meanings of it tend to lie buried among the candidates. If we take the example of the word "fighter" used in this paper, "weapon" is recognized as a candidate area, but is not given a strong similarity. One reason for the problem is the lack of viewpoints. More local contexts of the word are needed to specify such meanings.
- The similarity value and viewpoints can be used to refine the structure of the thesaurus. They make it possible to change dynamically the relationships between words in the thesaurus according to domain-sensitive corpus. We are now developing the general functions described in the previous section to realize a large-scale thesaurus for NLP systems.
- If the number of occurrences of an unknown words is low, the proposed method tends to output larger areas as positions. Other constraints such as use of local context are required.

Acknowledgement

We would like to thank Prof. Hozumi Tanaka, Tokyo Institute of Technology for allowing us to use ISAMAP.

References

[1] M. A. Hearst. "Automatic Acquisition of Hyponyms from Large Text Corpora". In *Proceedings of COLING-92*, pages 539–545, 1992.

[2] M. A. Hearst and G. Grefenstette. "Refining Automatically-Discovered Lexical Relationships: Combining Weak Techniques for Stronger Results". In *Proceedings of the AAAI Workshop on Statistically-based NLP Techniques*, pages 64–72, 1992.

[3] X. Li. "A WordNet-based Algorithm for Word Sense Disambiguation". In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, pages 1368–1374, 1995.

[4] A. Miller, R. Beckwith, C. Fellbaum, D. Gros, K. Miller, and R. Tengi. "Five Papers on WordNet". Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton University, 1990.

[5] P. Resnik. "WordNet and Distributed Analysis: A Class-based Approach to Lexical Discovery'. In *Proceedings of the AAAI Workshop on Statistically-based NLP Techniques*, pages 48–56, 1992.

[6] P. Resnik. "Disambiguating Noun Grouping with Respect to WordNet Senses'. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, pages 54–68, 1995.

[7] T. Strzalkowski. "Building A Lexical Domain Map from Text Corpora". In *Proceedings of COLING-94*, pages 604–616, 1994.

[8] H. Tanaka. "Construction of a Thesaurus Based on Superordinate/Subordinate Concept" (in Japanese). *IPSJ, SIG-NL*, 64(4):25–44, 1987.

[9] D. Yarowsky. "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora". In *Proceedings of COLING-92*, pages 454–460, 1992.