Constructing a Corpus of Japanese Predicate Phrases for Synonym/Antonym Relations

Tomoko Izumi[†], Tomohide Shibata[‡], Hisako Asano[†], Yoshihiro Matsuo[†], and Sadao Kurohashi[‡]

[†]NTT Media Intelligence Laboratories, Nippon Telegraph and Telephone Corporation
1-1 Hikarinooka Yokosuka, Kanagawa, Japan

[‡]Graduate School of Informatics, Kyoto University
Yoshida-honmachi Sakyo-ku, Kyoto, Japan

E-mail: izumi.tomoko@lab.ntt.co.jp, shibata@i.kyoto-u.ac.jp, asano.hisako@lab.ntt.co.jp, matsuo.yoshihiro@lab.ntt.co.jp, kuro@i.kyoto-u.ac.jp

Abstract

We construct a large corpus of Japanese predicate phrases for synonym-antonym relations. The corpus consists of 7,278 pairs of predicates such as "*receive*-permission (ACC)" vs. "*obtain*-permission (ACC)", in which each predicate pair is accompanied by a noun phrase and case information. The relations are categorized as synonyms, entailment, antonyms, or unrelated. Antonyms are further categorized into three different classes depending on their aspect of oppositeness. Using the data as a training corpus, we conduct the supervised binary classification of synonymous predicates based on linguistically-motivated features. Combining features that are characteristic of synonymous predicates with those that are characteristic of antonymous predicates, we succeed in automatically identifying synonymous predicates at the high F-score of 0.92, a 0.4 improvement over the baseline method of using the Japanese WordNet. The results of an experiment confirm that the quality of the corpus is high enough to achieve automatic classification. To the best of our knowledge, this is the first and the largest publicly available corpus of Japanese predicate phrases for synonym-antonym relations.

Keywords: predicates, synonym, antonym

1. Introduction

Identifying synonymy and antonymy relations between words and phrases is one of the fundamental tasks in Natural Language Processing (NLP). Understanding these semantic relations is crucial for realizing NLP applications including QA systems, information retrieval, text mining etc.

Among various words and phrases, identifying the semantic relations between *predicates* is especially important because predicates convey the propositional meaning of a sentence. For example, identifying synonymous predicates such as *"can't repair X"* and *"unable to fix X"* is crucial for text mining systems. Recognizing semantic opposites (antonyms) and their *aspect* of oppositeness are also important for NLP tasks. For example, predicate pairs expressing opposite attributive meaning such as *"beautiful"* and *"ugly*" can be used to detect contradiction in texts while antonym pairs expressing perspective differences of an event such as *"sell*" and *"buy"* are important to identify paraphrasing ("sell" and "buy" become a paraphrase if they share the same participants.).

However, it is hard to obtain a rich language resource that can completely cover synonymy/antonymy relations of predicates. This is because the meaning of a predicate varies depending on the context. For example, "ignore" and "break" can express the same meaning if they are combined with the argument "rule" (break the rule vs. ignore the rule).

In this paper, we introduce a large human annotated set of predicates for synonym-antonym relations. Synonym relations in this paper denote a mutual entailment relation from Chierchia and McConnel-Ginet, (2000), which can also be defined as "content synonyms" and "paraphrasing". We also annotate an entailment relation, whose synonymy is unidirectional. Antonym relations include semantic opposites, in which the events expressed by these opposite predicates result in contradiction. Accompanied by a noun phrase and case information, our data consists of 7,278 pairs of predicates such as "*receive*-permission (ACC)" vs. "*obtain*-permission (ACC)"; the relations are categorized as synonyms (mutual entailment and entailment), antonyms, or unrelated. Antonyms are further categorized into three different classes depending on their aspect of oppositeness.

Semantic Relation	# of Pairs	Examples
Synonyms	3,188	denwa-o-tsukau vs. denwa-o-riyou "use a phone" "utilize a phone" kyoka-o-eru vs. kyoka-o-syutoku "receive permission" "obtain permission"
Entailment	1,557	eigo-ga-tannou vs. eigo-o-hanasu "fluent in English" "speak English"
Antonyms		
Incompatible Attribute/Event Relation	426	meirei-o-ukeru vs. meirei-o-kyohi "follow the order" "reject the order"
Sequential Event Relation	131	denwa-o-kakeru vs. denwa-o-kiru "make a phone call" "hang up a phone"
Counterpart Perspective Relation	159	sain-o-motomeru vs. sain-ni-oujiru "ask for an autograph" "give an autograph"
Unrelated	1,817	denwa-ga-hairu vs. denwa-de-tsutaeru "receive a phone call" "announce by phone"
TOTAL	7,278	

Table 1: Corpus of Japanese Predicate Phrases for Synonym-Antonym Relations.

Using the data, we further propose a supervised classification task of synonymous predicates based on linguistically-motivated features. We believe that this is the first and the largest publicly available corpus to specify detailed synonym-antonym relations between different predicates in Japanese. For the sake of simplicity, we use the term synonyms/antonyms to refer to semantic relations of predicate phrases throughout this paper.

2. Related Work

WordNet (Miller, 1995), one of the largest thesauruses, provides word-level semantic relations including synonyms (synsets), hyponyms, and antonyms. WordNet is also available in different languages including Japanese (Bond et al., 2009), but the Japanese WordNet only provides synonyms (synsets) and hyponyms.

Aizawa (2008) automatically constructs semantically similar word pairs (synonyms and hyponyms) using the pattern "C such as A and B (e.g., "*publications* such as *magazines* and *books*)". Here, A and B (i.e., *magazines* and *books*) can be synonyms while C (i.e., *publications*) can be a hyponym of A and B. This pattern, however, can only be used for extracting noun phrases; for synonymous predicates with an argument, we need a different pattern set.

Mitchell and Lapata (2010) construct adjective-noun, noun-noun, and verb-object phrases with human judged similarity ratings. However, the data only shows relatedness scores, so one cannot distinguish whether the relation of a pair is synonymous or antonymous.

As shown, existing resources cannot directly used to measure the semantic relations of predicates or predicate-argument structures, such as "*break* the rule" vs. "*ignore* the rule".

3. Constructing Predicate Phrases in Synonym-Antonym Relations in Japanese

To express the relations of synonyms and antonyms, we formed predicate pairs that are accompanied by a noun and a case marker. Synonymous predicates are further categorized into mutually synonymous or not (i.e., entailment)[1].

Antonyms are also subcategorized into three classes depending on their aspects of oppositeness. The following is an example.

(1)	ugly	vs.	beautiful
(2)	matriculate	vs.	graduate
(3)	buy	vs.	sell

(1) expresses the semantic opposite in quality, having ugly in one extreme and beautiful in the other. (2) expresses events which cannot coincide. Interestingly, these two events are also have a past-future event relation such that one enters school, and then graduates (sequential event). (3) also expresses an opposite event, but the oppositeness is rather related to the difference in perspective. That is, (3) involves the same event but focuses on the different roles of the two participants, the buyer and the seller.

We also included predicates whose meanings are unrelated because these pairs can be used by supervised methods as negative examples. Table 1 summarizes our data.

[1] In this paper, we treat both synonyms and entailment as synonymous predicates because synonyms can also be called mutual entailment.

Features		Description
Recognizing Synonyms	**Definition sentences in a dictionary**	-Binary features indicating whether a predicate appears in the definition sentences of the other predicate -Word overlap among definition sentences between predicate pairs
	Abstract predicate categories	-Predicate categories that the two predicates share - Ratio of overlap in predicate categories
	Distributional Similarities	-Distributional similarities between predicates -Distributional similarities between predicate-argument pairs
	Modality and Negation	-Modality and Negations that each predicate has -Ratio of overlap in Modality and Negations between two predicates
Recognizing Antonyms	**Compounding and the *tari* contrastive construction**	- The Frequency and Ngram scores of the compounding word of predicate pairs - The ngram score of the string in which two predicates are combined by the *tari* conjunct.
	Suffix combination	- The combination of the first character of antonym pair and its Ngram score and frequency.
POS		- Part-of-Speech of each predicate

Table 2: Summary of Features.

3.1. Extracting Predicate Phrases

We extracted predicates from Japanese Wikipedia. [2] Based on the frequencies of nouns in Wikipedia, we selected every fifth noun starting from the most frequent noun until the total number of nouns selected reached 100. We used these 100 nouns as an argument, and extracted the top 20 predicates based on mutual information between the noun and the predicate. The following is an example;

(4) Predicates with the argument *denwa* "phone"

- denwa-o-tsukau
  phone-ACC-use
  "use a phone"

- denwa-ni-deru
  phone-DAT-answer
  "answer the phone"

If a predicate appears with an auxiliary verb expressing negation, passive, causative, and potentials, we retain the negation and modality information by setting semantic labels, such as "negation" and "passive."

3.2. Annotation based on Linguistic Tests

Annotation was done by three annotators, all with a solid background in linguistics. We divided the data into three based on argument noun phrases, and each annotator annotated two to three predicate-argument pairs for each semantic category of an assigned noun phrase.

In order to make the data as consistent as possible, we also created several linguistic tests based on Chierchia and McConnel-Ginet (2000). For simplicity, we use the term Predicate A and Predicate B to refer to predicate pairs. (# indicates a semantically wrong sentence.)

- Synonym (Mutual Entailment)

Definition: Predicate A and Predicate B denote the same event. (If the event expressed by Predicate A is true, the event expressed by Predicate B is also true and vice versa.) (e.g., *repair* vs. *fix*)
Test: Negating only one of the predicates results in a contradictory fact (i.e., does not make sense).
Example: # I repaired my PC, but I didn't fix it.

- Entailment

Definition: If the event denoted by Predicate A is true, the event denoted by Predicate B is also true, but not vice versa. (e.g., *snore* vs. *sleep*)
Test: Negating only Predicate B (i.e., sleep) does not make sense. However, the opposite is possible.
Example: # I snored last night, but I didn't sleep.
I slept last night, but I didn't snore.

- Antonym

Definition: If the event denoted by Predicate A is true, the event denoted by Predicate B must be false. (e.g., *long* vs. *short*)
Test: Predicate A and Predicate B cannot be combined by the conjunction "but" in a sentence.
Example: # His legs are long, but they are short.

[2] http://ja.wikipedia.org/wiki

We further categorized antonyms into the following three categories based on their oppositeness.

- Incompatible Attribute/Event Relation

The two predicates in this relation *always* denote a contradictory fact. (e.g., *ugly* vs. *beautiful*)

- Sequential Event Relation

The events expressed by the predicates cannot coincide but can be in the relation of *past-future event*. (e.g., *matriculate* vs. *graduate*)

- Counterpart Perspective Relation

The two predicates express different perspectives of a single event (e.g., *buy* vs. *sell*).

We evaluated the quality of the corpus by randomly selecting predicate-argument pairs of 100 different noun arguments and asked an evaluator, not one of the annotators, whether the assigned semantic relation was correct or not. 93.4% of the predicate-argument pairs were evaluated as being assigned the correct semantic relation.

4. Automatic Recognition of Synonymous Predicates

Using the corpus as a training set, we conducted the supervised binary classification of synonymous predicates. The purpose of automatic classification is to examine the quality of the data as well as to investigate the possibility of automatically constructing a thesaurus of synonymous and antonymous predicates. Because the number of predicates in antonym relations is relatively small (only 10 % of the corpus), we conducted a binary classification of synonymous predicates, making predicates in synonym and entailment relations as positive examples and those in antonym relations and others as negative examples.

As features for recognizing semantically similar predicates, we used two different kinds of linguistically-motivated features; one for recognizing synonyms and the other for recognizing antonyms. These features are summarized in Table 2.

4.1. Linguistic Features for Recognizing Synonyms

Definition sentences in a dictionary

If two predicates (e.g., buy vs. purchase) express the same meaning, one (especially one with broader meaning) tends to occur in the definition sentence of the other (e.g., "to buy something, especially big of expensive" is a definition of *purchase*). We call this feature as "complementarity in definition sentences" because one predicate complements the meaning of the other synonymous predicate. We use the binary feature of existence of complementarity in definition sentences.

We also observed that if two predicates are synonymous, their definition sentences are also similar. The following is an example of definition sentences of "high-priced" and "expensive".

(5) "high-priced": Costing a lot of money
(6) "expensive": Costing a lot of money

We also used this characteristic, and the following are features extracted from definition sentences. [3]

- Complemantarity in definition sentences
- Commonality in the content words of two definition sentences (Frequencies of overlapped content words are used.)

Abstract Predicate Categories

If two predicates are synonyms, their abstract semantic categories must be the same. Therefore, we used the semantic categories that the predicates share as features. For example, the following two synonymous predicates share the same predicate attribute in *Goi-Taikei* (Ikehara et al., 1999).

(7) Predicate Attributes of *kau* "buy" and *kounyuu-suru* "purchase"
- *Kau* "buy" : [*Transfer in possession*], [*Action*]
- *Kounyu- suru* "purchase": [*Transfer in possession*], [*Action*]

Both share the same predicate attributes of *Transfer in possessions* and *Action.*

We use *yougen zousei* "predicate attributes" in *Goi-Taikei* (Ikehara et al, 1999) as features. The predicate attributes in *Goi-Taikei* are hierarchically organized; the deeper the shared attribute is, the more similar the two predicates are, so we use a weighted overlap ratio in predicate attributes, in which the deeper attributes are weighted more heavily. The weights are decided heuristically. Level x indicates the level at *Goi-Taikei's* Predicate Attribute Hierarchy (the highest being 1 and the lowest 4). PAtr is for Predicate Attributes.

$$\text{Weighted ratio of overlap in PAtr} = \frac{\begin{array}{c}(|\text{PAtr for Pred1 at Level 1} \cap \text{PAtr for Pred2 at Level1}| * 1 \\ + (|\text{PAtr for Pred1 at Level2} \cap \text{PAtr for Pred2 at Level2}|) * 1.5 \\ + (|\text{PAtr for Pred1 at Level3} \cap \text{PAtr for Pred2 at Level3}|) * 2 \\ + (|\text{PAtr for Pred1 at Level4} \cap \text{PAtr for Pred2 at Level4}|) * 2.5\end{array}}{(|\text{PAtr for Pred1} \cup \text{PAtr for Pred2}|)}$$

- Predicate attributes that two predicates share
- Weighted ratio of overlap in predicate attributes

[3] We use *Gakken Japanese Dictionary* (Kindaichi and Ikeda, 1988).

Distributional Similarities

We used distributional similarities between predicates and predicate-argument pairs calculated from a vector model constructed from 6.9 billion sentences on the Web, the methods proposed in Shibata and Kurohashi (2010). They use words in the dependency relations with the predicate-argument or predicate as features for vector models. The following are distributional similarity based features we used.

- Distributional similarity between predicates (e.g., buy vs. purchase)
- Distributional similarity between predicate-argument structures (e.g., break-rule vs. ignore-rule)

Modality and Negation

Because we target predicates, we also used information of modality and negations such as "cannot" if a predicate has one. Because an auxiliary verb in a predicate phrase is transformed into a semantic label, we simply use the label as features.

- Overlapped semantic labels (Semantic labels that both predicates have)
- Asymmetric Occurrence of Negation, and Passive
- Overlap rate of Semantic labels

4.2. Linguistic Features for Recognizing Antonyms

Measures such as distributional similarities often mistakenly assign a high score to antonym pairs. We, therefore, add several linguistic features that are peculiar to antonyms.

Compounding and the *tari* contrastive construction

In Japanese, Antonymous phrases tend to make a compound such as *uri-kai* (buy and sell) in which the conjunctive form of *uru* "sell" is combined with the conjunctive form of *kau* "buy". Similarly, antonymous phrases have a tendency to appear in the *tari* construction, in which the contrast of two different events/actions is described.

(8) hon-o ut-tari kat-tari dekimasu.
book-ACC sell-*tari* buy-*tari* can
"You can sell and/or buy books here."

We used the likelihood of making a compound and of appearing in the *tari* contrastive construction as features for distinguishing semantically similar phrases from antonymous phrases. By automatically generating a compound and a word string in the *tari* contrastive construction (Predicate A-*tari*-Predicate B-*tari*) for each predicate pair, we use the following two features as compounding and *tari* contrastive features.

- Document frequency (df) of the compound calculated from the web
- Ngram score calculated based on Japanese *google ngram*.
- Ngram score of the string with the *tari*

The higher frequency/score of the two compounds is used.

Suffix combination

Additionally, we used the information of *Kanji* character in each predicate pair.

(9) 入院 vs. 退院
"enter the hospital" "leave the hospital"

The kanji "入" expresses the action of entering, while the kanji "退" expresses leaving, which themselves are antonyms. In order to represent these properties, the following prefix combination features are used. The prefix combination is constructed by combining the first character of each predicate. The higher ngram score/document frequency is selected.

- Prefix combination of predicate pairs
- Document frequency of prefix combination
- Ngram score of prefix combination
- Overlap Flag in prefix combination (indicating whether prefixes extracted from each predicate are the same)

5. Experiment and Discussion

5.1. Experiment and Result

We conducted an experiment of automatically classifying synonymous predicates using the features discussed in Section 4. For training, we used LIBSVM(Chang & Lin, 2011) and conducted a five-fold cross validation for evaluation.

As a baseline, we used the Japanese WordNet (Bond et al., 2009), one of the largest thesauruses. If the synonymous predicate pairs are listed in the synsets in WordNet, they are counted as correct. The results are evaluated based on Precision (Prec), Recall (Rec) and F-score (F).

$$\text{Prec} = \frac{\#\ of\ True\ Synonyms\ \cap \#\ of\ Preds\ Classified\ as\ Synonymous}{\#\ of\ Preds\ Classified\ as\ Synonymous}$$

$$\text{Rec} = \frac{\#\ of\ True\ Synonyms\ \cap \#\ of\ Preds\ Classified\ as\ Synonymous}{\#\ of\ True\ Synonymous\ Preds}$$

$$\text{F} - \text{score} = \frac{2 * Prec * Rec}{(Prec + Rec)}$$

	Precision	Recall	F-Score
Baseline-WordNet	0.977	0.349	0.514
Proposed	0.899	0.932	**0.915**
Only Using Features for Recognizing Synonyms	0.730	0.917	0.812
Only Using Features for Recognizing Antonyms	0.721	0.972	0.828

Table 3: Results of Experiment.

As shown in Table 3, using the data as a training set, our supervised classification of synonymous predicates achieved the high F-score of 0.915, compared to the baseline (0.514).

5.2. Discussion

Although the use of WordNet yielded the highest precision, it suffered from low recall. The following are examples that are not listed in synsets of WordNet but were correctly categorized as synonymous predicates by our method.

(10) meisho-o-annai vs. meisho-o-shoukai
landmark-ACC-guide landmark-ACC-introduce
"guide a landmark" "introduce a landmark"

(11) shien-ni-ataru vs. shien-o-zisshi
support-DAT-take part in support-ACC-carry out
"take part in supporting s/th" "carry out support for s/th"

Predicates such as "ataru (take part in/ do)" and "zisshi (carry out/ conduct)" become synonymous with the argument "shien (support)." WordNet tends not to include predicates in synsets that become synonyms in a certain context, which degrades recall.

Combining the features for recognizing synonyms and those for recognizing antonyms was very effective as the overall F-score drastically increased (Table 3).

An error analysis revealed that the proposed method failed to classify synonymous predicates when their meanings are idiomatic.

(12) ki-ni-yamu vs. ki-ga-yowai
eart-DAT-suffer heart-NOM-weak
worry" "be anxious"
"(lit., my heart suffers)" "(lit., my heart is weak)"

These idiomatic expressions need more sophisticated rules of inference. One possible solution would be to use how these expressions are translated into a foreign language because these idiomatic expressions might be translated into the same phrase as direct word-to-word translation is avoided for idiomatic expressions. The analysis of idiomatic expressions and their translations is for future study.

6. Conclusion

In conclusion, we constructed a large corpus of Japanese predicate phrases for synonym-antonym relations. The antonym relation was further categorized into detailed subclasses depending on their aspect of oppositeness. The corpus consists of a wide variety of expressions including idiomatic expressions. Using the data as training set, we proposed the supervised classification of synonymous predicates and achieved a promising result, indicating that the quality of the corpus is high enough to achieve automatic classification.

To the best of our knowledge, this is the first and the largest publicly available corpus of Japanese predicate phrases for synonym-antonym relations.[4] We hope that the corpus will accelerate research into the automatic acquisition of language resources.

7. References

Aizawa, A. (2008).On calculating word similarity using large text corpora, *Journal of Information Processing (IPSJ), vol. 49, 3,* 1426-1436.

Bond F, Isahara H, Fujita S, Uchimoto K, Kuribayashi T and Kanzaki K (2009). Enhancing the Japanese WordNet, *The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009,* Singapore.

Chierchia, G., and McConnell-Ginet, S. (2000). *Meaning and grammar: An introduction to semantics (2nd ed.).* Cambridge, MA: The MIT press.

Cruse, D. A., (1986). *Lexical Semantics.* New York: Cambridge University Press.

Chang, C-C., and Lin, C-J. (2011). "LIBSVM: A Library for Support Vector Machines" *ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), No.27.*

Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura K., Ooyama, Y., and Hayashi Y. (1999). *Goi-Taikei*, Tokyko:Iwanami.

Kindaichi, H., and Ikeda, Y. (1988). *Gakken Japanese Dictionary (2nd Ed.),* Tokyo: Gakusyuu Kenkyuusha.

Lin, D., Zhao, S., Qin, L., and Zhou, M. (2003). Identifying synonyms among distributionally similar words. *Proceedings of the 18th International Joint conference on Artificial Intelligence (IJCAI-03),*

[4] The corpus can be downloaded from http://nlp.ist.i.kyoto-u.ac.jp/index.php?PredicateEvalSet

1492-1493.
Mitchell J., and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science, Vol. 34, 8,* 1388-1429.
Miller, G. A. (1995). WordNet: A lexical database for English, *Communications of the ACM, Vol. 38, 11,* 39-41.
Mohammad, S., Bonnie, D., and Hirst, G. (2008). Computing word-pair antonymy. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2008),* 982-991.
Murphy, L. (2006). Antonym as lexical constructions; or, why paradigmatic construction is not an oxymoron. *Constructions, Special Volume 1,* 1-37.
Shibata, T., and Kurohashi, S. (2010). Bunmyaku-ni izonshita jutsugono dougikankei kakutoku. [Context-dependent synonymous predicate acquisition]. *Information processing society of Japan, Special Interest Group of Natural Language Processing (IPSJ-SIGNL) Technical Report,* 1-6.
Yih, W., Zweig, G., and Platt, J. (2012). Polarity Inducing Latent Semantic Analysis, *Proceedings of Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL),* 1212-1222.