Towards linking synonymous expressions of compound verbs to Japanese WordNet

Kyoko Kanzaki
Toyohashi University of Technology
Aichi Japan
kanzaki@imc.tut.ac.jp

Hitoshi Isahara
Toyohashi University of Technology
Aichi, Japan
Isahara@tut.jp

Abstract

This paper describes our project on Japanese compound verbs. Japanese "Verb (adnominal form) + Verb" compounds, which are treated as single verbs, frequently appear in daily communication. They are not sufficiently registered in Japanese dictionaries or thesauri. We are now compiling a list of the synonymous expressions of compound verbs in "compound verb lexicon" built by the National Institute of Japanese Language and Linguistics. We extracted synonymous words and phrases of compound verbs from five hundred million Japanese web corpora. As a result, synonymous expressions of 1800 compound verbs were obtained automatically among 2700 in the "compound verb lexicon". From our data, we observed that some compound verbs represent not only motion but also additional nuances such as an emotional one. In order to reflect the abundant meanings that compound verbs own, we will try to think of a link of synonymous expressions to Japanese wordnet. Concretely, in the case of synonymous phrases, we try to link adverbial expressions which are a part of phrases to the adverbial synset in Japanese wordnet.

1 Introduction

Japanese "Verb (adnominal form) + Verb" compounds, which are treated as single verbs, frequently appear in daily communication, however, they are not sufficiently registered in Japanese dictionaries or thesauri.

The Japanese "compound verb lexicon" was constructed by the National Institute for Japanese Language and Linguistics (NINJAL) (https://db4.ninjal.ac.jp/vvlexicon/). It has the meanings, example sentences, syntactic patterns and actual sentences from the corpus that they possess. However, it has no relation with another words, such as synonymous words and phrases.

We detect them automatically as much as possible in order to help humans find out synonymous expressions that they may fail to bring to mind and then manually compile a lexicon of synonymous expressions of Japanese compound verbs.

In this paper, firstly we explain how to build the list of compound verbs and their synonymous words and phrases, and then consider what should be considered for linking to the Japanese wordnet based on our obtained result.

2 Related researches

So far, in NLP domain researches on complexed verbal meaning have treated multi word expressions in order to distinguish a literal meaning with the metaphoric meaning, but their purposes are word sense disambiguation or generation of compounding (Sag et.al.2002; Hashimoto and Kawahara 2008 and so on). In Japanese, Uchiyama and Ishizaki (2003), and Uchiyama and Baldwin (2004) investigated ambiguities of compound verbs and tried to find the generation rules. As a resource on phrases, Tanabe et.al (2014) built Japanese Dictionary of Multi word Expressions.

However, works on the organization of words and phrases are few. Our goal is to compile a list of words and verbal phrases with linking similar relations by using both automatic and manual ways.

3 Japanese compound verbs

The morphological form of a compound verb is a combination of a first verb in an adnominal form and a second verb coming after it, as in *hikari* (adnominal form)-*kagayaku* (give.off.light

& shine) ‘shine like the sun’, *nage* (adnominal form)-*ireru* (throw & put.in) ‘throw in’.

Japanese compound verbs are divided into two types in terms of syntactic and morphological analysis; syntactic compound verbs and lexical compound verbs (Kageyama 1993).

Kageyama (1993) says that syntactic compound verbs are easily recognizable and interpretable due to some characteristics, that is, a limitation of a variety of the second verbs, no restriction on the first verbs and so on. For example, “*utai_hajimeru* (sing & start, ‘start singing’” “*hanashi_hajimeru* (speak & start, ‘start speaking’)” “*hashiri_hajimeru* (run & start, ‘start running’)” and so on. We can generate varieties of “Verb_*hajimeru* (Verb & start, ‘start V_ing’)”. The second verbs of syntactic compound verbs are mainly aspectual verbs and also are limited to 30 verbs which are classified into 9 categories; inception, continuation, completion, incompletion, excessive action, habitual, reciprocal action and potential.

We exclude the syntactic compound verbs and treat only lexical compounds which tightly combine two verbs as one word and not productive compared to syntactic compound verbs.

4 Extracting synonymous expressions from corpus

4.1 Data

We use “five hundred million Japanese texts gathered from web” produced by Kawahara et.al. (2006) as corpus for extracting synonymous words and phrases. The data has been processed into morphologically analyzed data.

As for compound verbs for an extraction of synonymous expressions, we dealt with compound verbs registered in “compound verb lexicon” built by NINJAL. The total number of compound verbs in this lexicon is 2700, and each one has meanings, syntactic patterns and example sentences.

4.2 Procedure

For the first step, we extracted synonymous words and phrases of compound verbs from corpora.

Step1: Preprocessing

Some compound verbs can be paraphrased into phrases. Therefore we concatenated modification relations between verbs and adverbial words and made them into units which we treated as “verbs” (e.g. correctly / understand >>> “correctly understand”). Also compound verbs which are not registered in a dictionary of a morphological analyzer need to combine two verbs (verb in adnominal form + verb (*nage* ‘throw’/ *ireru ‘put in’* >>> *nageireru ‘throw in’*)).

For the first experiments, we had put all words segmented by Japanese morphological analyzer and calculated the similarity between compound verbs and another verbs by cosine similarity measure, but the result was not good. We obtained many unrelated words for each compound verb. Therefore, we decided to exclude the passive and causative form and so on which make an alternation of case markers.

After that, we generate the list of the sets of a noun, a verb and a case marker, which is an input data for vectorization.

Step2: Vectorization and cosine similarity

We performed vectorization of all verbs and nouns in the web corpus by using word2vec (Mikolov 2013), one of the deep learning methods. The learning model of word2vec that we used is CBOW (contiguous bag of words). Then we explored the semantic distance between verbs (including verbal phrases) by cosine similarity. For each compound verb, the verb and verbal phrases were arranged in descending order from the highest score.

Step3: Creating a list of candidates of synonymous expressions

For each compound verb, 2000 similar expressions were chosen in order from the highest score of cosine similarity. Here, the lists of synonymous expressions for each compound verb were created. However, in this list, the polysemy of compound verbs was not taken into account. That is, the synonymous expressions of compound verbs were stored together without distinction of their polysemous meaning in this list.

Step4: Shrinking synonymous expressions and getting clusters for each compound verb

A rough diagram of the process to get categories for each compound verb is shown below.

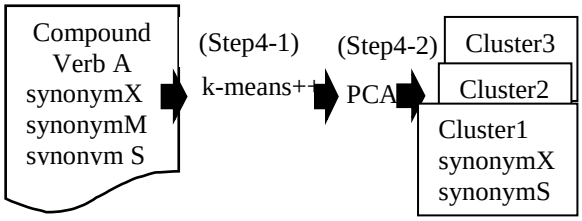

Figure1. The process of getting clusters

Step4-1) Decreasing candidates of synonymous expressions

At the beginning, each compound verb has candidates of 2000 synonymous expressions extracted from the web corpus. To easily determine plausible synonymous expressions, we decrease 2000 synonyms in a step-by-step approach by iterating the k-means++ (Arthur and Vassilvitskii 2007). As for input data for k-means++, we used vectors of synonymous expressions obtained by word2vec. Centroids were calculated by Euclidian distance. Our procedure is described below.

P1. Firstly, we set 64 clusters for the k-means++. In this stage, 2000 expressions are classified into 64 clusters.

P2. For each cluster, we extract 10 expressions with the highest cosine similarity values. In this stage, we narrow down to 640 expressions (64 clusters * 10 expressions). We considered that cosine similarity values between a compound verb and other expressions would be plausible to choose synonym candidates for a compound verb.

P3. We iterate the same process as the step P2 for 640 expressions. In this stage, we set 10 clusters. For each cluster, we extract 10 expressions with the highest similarity values in the list. As a result, we obtained 100 synonym expressions classified into 10 clusters (10 clusters * 10 expressions). This data is used as input data for Principal Component Analysis (PCA).

Step4-2) Finding the number of senses for each compound verb by PCA

The 100 synonymous expressions are now classified into 10 clusters for each compound verb. However, the number of senses of a compound verb differs from each other. We tried to detect the appropriate numbers of senses of each compound verb by using principal component analysis (PCA). We implemented PCA with 100 expressions for each compound verb.

5 Result

As a result, synonymous expressions for 1800 compound verbs were obtained automatically.

The following is a result of "持ち込む (*mochi_komu*)" derived from PCA. The synonymous expressions that we decided are surrounded with circles.

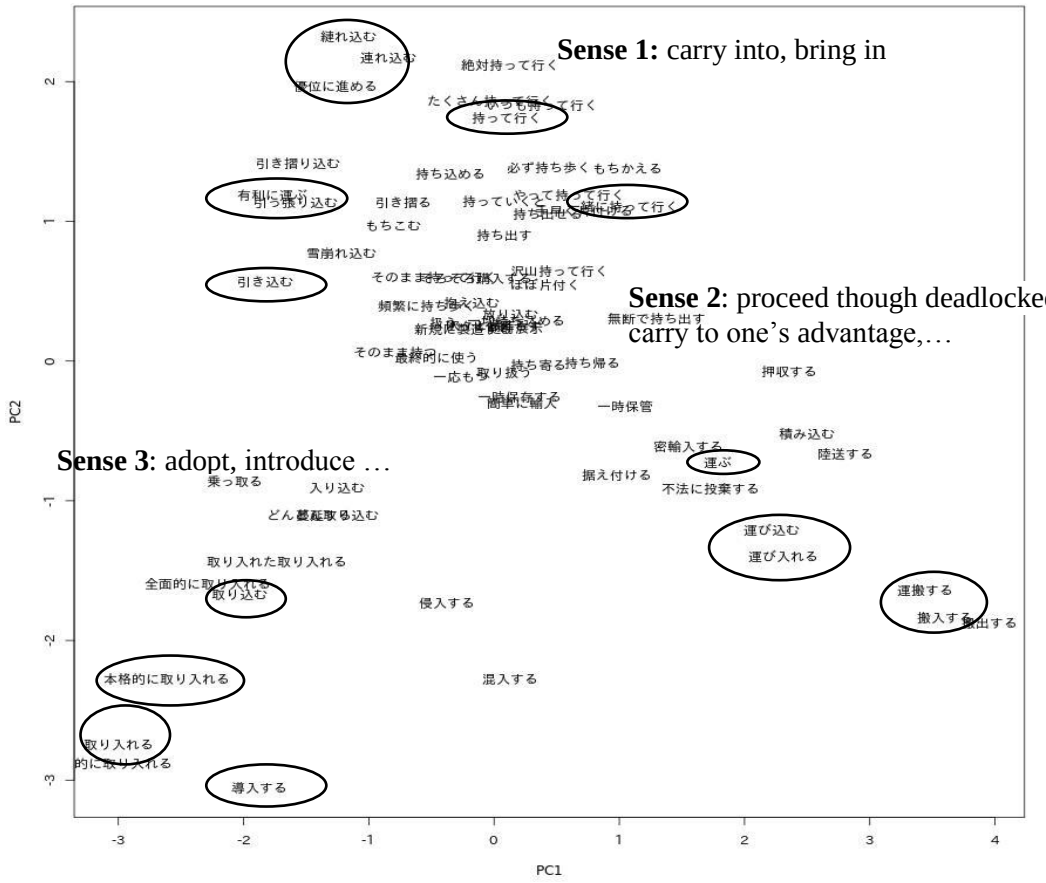

Figure 2. A distribution of synonymous expressions of "持ち込む (*mochikomu*) " derived from PCA

The list of synonymous expressions of “持ち込む (*mochikomu*)” is as follows.

Sense1 :
運び入れる (*hakobiireru*, ‘carry in’)
搬入する (*han’nyu_suru*, ‘carry to’)
運び込む (*hakobikomu*, ‘carry into’)
運ぶ (*hakobu*, ‘carry’)
持って行く (*motte_iku*, ‘take’)
一緒に持っていく (*isshoni motteiku*, ‘bring in’, ‘take … with [person]’.)
Sense2:
もつれこむ (*motsurekomu*, ‘to proceed though deadlocked’)
優位に進める (*yuui_ni* (adj in adverbial form) *susumeru*, ‘advance [the match]’)
有利に運ぶ (*yuuri_ni*(adj in adverbial form) *hakobu*, ‘carry to one’s advantage’)
引っ張り込む (*hipparikomu*, ‘pull’)
引き込む (*hikikomu*, ‘pull’)
Sense3:
取り込む (*torikomu*, ‘incorporate’)
取り入れる (*toriireru*, ‘incorporate’),
導入する (*dounyuusuru*, ‘introduce’)

4.1. Evaluation for 40 compound verbs

In order to predict how many suitable synonyms and clusters we’ve semi-automatically obtained by our method, we evaluate our results manually. For 40 compound verbs which are the most frequent compound verbs in our corpus, 4 examinees evaluated the suitability for synonymous expressions classified in each cluster. We evaluated the expressions for each cluster by comparing them to sense descriptions of the compound verb in CVL. As a result, 59% of extracted words are evaluated as synonyms. And we evaluated the suitability of clusters created by our method. We compared the clusters to sense descriptions of the compound verb in CVL. As a result, 65% of extracted clusters are evaluated as representing the proper meaning of the compound verb. For example, “*Furikaeru* (*Furu-+kaeru*)” has a single meaning like “look behind with twisting body” in CVL. Our method could extract another meaning, i.e. “think back on the previous episode.”

In terms of a recall, the total number of meanings of 40 compound verbs registered in CVL is 64. Among them 14 meanings could not be obtained by our methods (22%). These 14 meanings are included in 13 compound verbs.

5.1 Add more synonymous expression to the list

We selected these synonymous expressions from 100 synonyms candidates whose similarity score is 10 from the top in 10 clusters obtained from k-means++ (referred to step4-1). They are distributed on the PCA (step4-2). By performing this process, we could easily find senses for each compound verb from the distribution generated by PCA. On the other hand, when we observed candidates with a similarity score lower than Top 10, we found some examples which seem to be appropriate. Because of replenishing more synonymous expressions, we decided to check all the candidates in the 10 clusters for each compound verb.

For example, we added some examples to the list of “持ち込む (*mochikomu*)”. They appear lower than Top 10.

Sence1: 持参する (*jisan*_suru, ‘bring … with [person]’)

Sense2: 何とか制す (*nantoka* (adverb) *seisu*, ‘manage to get through’), 粘り勝つ (*nebari_katsu*, ‘compete tenaciously with each other, and finally win’)

5.2 Consideration

From our result, Japanese lexical compound verbs are found to be deeply related with adverbs and adverbial expressions. One of the reasons for this is that a compound verb represents a verbal meaning with the speaker’s emotional expressions. For example, “持ち込む (*mochikomu*)” in Sense2 implies that it’s not easy to realize a good result. Even “持ち込む (*mochikomu*)” in Sense1 has sometimes a meaning of an emphasis.

Also compound verbs are sometimes paraphrased into not only words but also phrases. Japanese compound verbs stand on a border of words and phrases.

6 Japanese compound verbs and Japanese wordnet

We try to incorporate a synonym list that we compiled into Japanese wordnet.

Currently, in Japanese wordnet, 2584 compound verbs are registered. In our experiment, we obtained 1800 compound verbs. If those compound verbs and their synonymous expressions are registered, it would be useful for not

only natural language processing like information retrieval, but also linguistic researches and language learning.
Our plan is:

(1) As for compound verbs registered in Japanese wordnet, we add or modify synonymous expressions of compound verbs and reconsider senses based on our result and Japanese dictionaries.
(2) As for compound verbs which are not in Japanese wordnet, we register synonymous expressions and then link them to the corresponding synsets in Japanese wordnet.
(3) We consider that the emotional and sensory meanings which compound verbs have are interesting and important information. The adverbial expressions included in synonymous expressions would be deeply related to compound verbs. We would like to try to put the adverbial expressions included in synonymous expressions into the Japanese wordnet. This means linking phrasal meanings to wordnet.

We show “持ち込む (*mochikomu*)” as an example in Figure3. Sense1 and Sense3 are registered in Japanese wordnet. Sense2 is not registered. As for Sense1, for the sake of a precise meaning of “持ち込む(*mochikomu*)”, not only verb but also preposition “in”, a kind of modification of a verb, is added and linked to the Japanese wordnet.

On the other hand, Sense2 is a new meaning that we obtained from the data. The following expressions with underlines and those in boldfaces mean adverbial expressions which represent emotional nuance.

もつれこむ (***motsure****komu*, ‘to proceed **though deadlocked**’)
優位に進める (***yuui_ni***(adj in adverbial form) *susumeru*, ‘advance [the match]’)
有利に運ぶ (***yuuri_ni***(adj_in adverbial form) *hakobu*, ‘carry to **one’s advantage**’)
引っ張り込む (*hipparikomu*, ‘pull in’)
引き込む (*hikikomu*, ‘pull in’)
何とか制す (***nantoka*** (adverb) *seisu*, ‘**manage to get through**’)
粘り勝つ (***nebari****katsu*, ‘compete **tenaciously** with each other, and finally win’)

We will link not only verbs like “proceed” “carry” “get through” and “win” but also a kind of modification for verbs like “through deadlocked” “advantage” “manage to” “tenaciously” to the Japanese or English wordnet because they’re important to understand a nuance of the compound verb “持ち込む (*mochikomu*)”. As an example, one of synonymous expressions “粘り勝つ (*nebarikatsu*), ‘compete tenaciously with each other and finally win’ ” are shown in Figure 4.

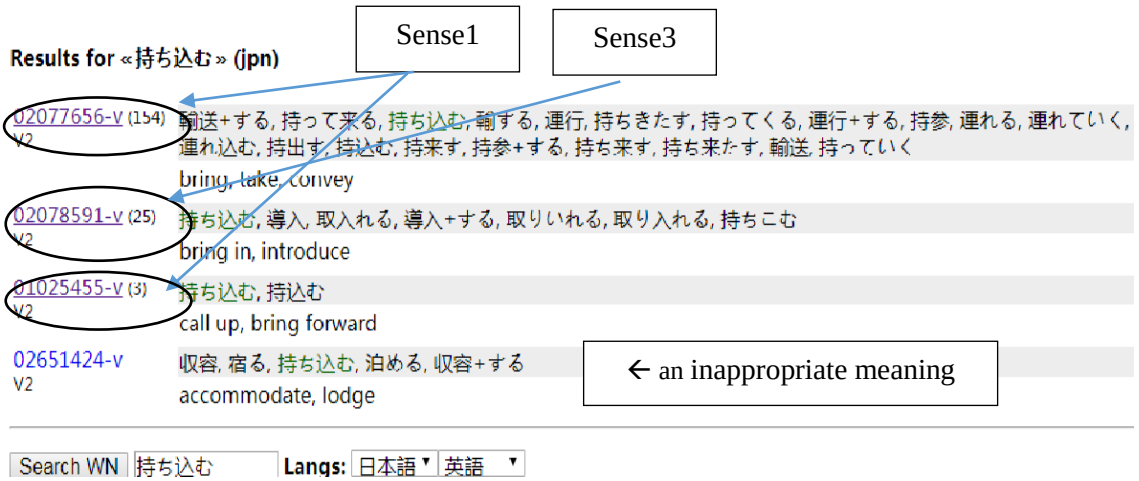

Figure 3. Comparison between senses that we obtained and synset of “持ち込む (*mochikomu*)” registered in Japanese wordnet

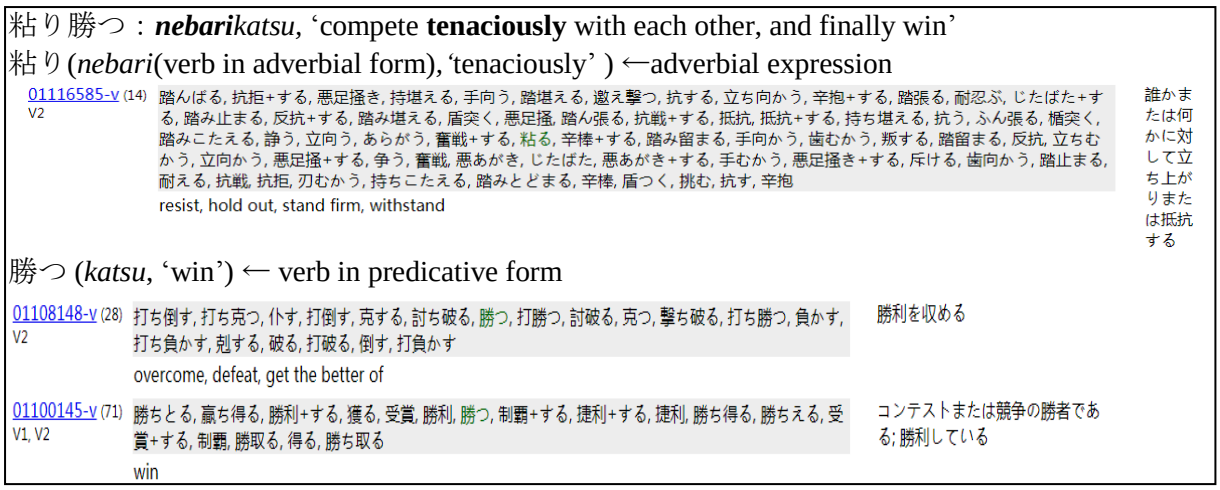

Figure 4. Link of one of synonymous expressions of “持ち込む (*mochikomu*)”: “粘り勝つ (***nebari**katsu*), ‘compete **tenaciously** with each other and finally win’ ”

7 Conclusion

In our work, first, we compiled a list of synonymous expressions of compound verbs by extracting from corpora semi-automatically and then try to link them to Japanese wordent. Japanese compound verbs have characteristics between words and phrases. We would like to consider how to combine phrasal expressions to wordnet. In addition, some compound verbs are deeply related with sense modalities. Therefore, it would be important to treat the adverbial meaning which it implies. If we registered a link of not only words but also phrasal expressions, Japanese wordnet would be useful for cross lingual works like linguistic researches, education and also, information retrieval.

References

David Arthur, Sergei Vassilvitskii. 2007. k-means++: The Advantages of Careful Seeding. In Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms, 1027-1035.

Chikara Hashimoto and Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on WSD in corporating idiom-specific features. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing(EMNLP 2008). 992-1001.

Taro Kageyama(1993), *Bunpō to Gokeisei* [Grammar and Word Formation], Tokyo: Hituzi Syobo

Daisuke Kawahara and Sadao Kurohashi. 2006. A Fully-lexicalized Probabilistic Model for Japanese syntactic and Case Structure Analysys. In Proceedings of Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL 2006), NY, USA, 176-183.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In proceedings of 27th Annual Conference on Neural Information Processing Systems,3111–3119.

Ivan A.Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickenger. 2002. Multiword Expressions: A pain in the Neck for NLP. In CICLing '02 Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, 1-15.

Toshifumi Tanabe, Masahito Takahashi and Kimiaki Shudo. 2014. A lexicon of multiword expressions for syntactically precise, wide coverage natural language processing, Computer Speech and Language, vol.28. No.6, 1317-1339, Elsevier.

Kiyoko Uchiyama and Shun Ishizaki. 2003.The Method on the Semantic Analysis for disambiguation of compound verbs. In proceedings of the 9th annual conference of Natural Language Processing,163-166.

Kiyoko Uchiyama, Timothy Baldwin,.2004. Automatic Disambiguation of Compound Verbs by Machine Learning. In proceedings of the 10th annual conference of Natural Language Processing, 741-744.

Language Resource References

National Institute for Japanese Language and Linguistics. Compound Verb Lexicon.(2013-2015) http://vvlexicon.ninjal.ac.jp/en/

Acknowledgements

This work was supported by JSPS KAKENHI （Grant-in-Aid for Scientific Research (C)）Grant Number JP 16K02727.