Building a List of Synonymous Words and Phrases of Japanese Compound Verbs

Kyoko Kanzaki and Hitoshi Isahara
Toyohashi University of Technology
1-1 Hibarigaoka Tenpaku-cho, Toyohashi, Aichi 441-8580
kanzaki@imc.tut.ac.jp, isahara@tut.jp

Abstract

We started to construct a database of synonymous expressions of Japanese "Verb + Verb" compounds semi-automatically. Japanese is known to be rich in compound verbs consisting of two verbs joined together. However, we did not have a comprehensive Japanese compound lexicon. Recently a Japanese compound verb lexicon was constructed by the National Institute for Japanese Language and Linguistics(NINJAL)(2013-15). Though it has meanings, example sentences, syntactic patterns and actual sentences from the corpus that they possess, it has no information on relationships with another words, such as synonymous words and phrases. We automatically extracted synonymous expressions of compound verbs from corpus which is "five hundred million Japanese texts gathered from the web" produced by Kawahara et.al. (2006) by using word2vec and cosine similarity and find suitable clusters which correspond to meanings of the compound verbs by using k-means++ and PCA. The automatic extraction from corpus helps humans find not only typical synonyms but also unexpected synonymous words and phrases. Then we manually compile the list of synonymous expressions of Japanese compound verbs by assessing the result and also link it to the "Compound Verb Lexicon" published by NINJAL.

Keywords: database, Japanese compound verbs, synonymous words and phrases

1. Introduction

In this work, we deal with Japanese verbs and in particular, those of the compound variety. Japanese "Verb+Verb" compounds frequently appear in daily communication which is related to human actions. In terms of morphology, Japanese compounds involving verbs and other predicates are productive and widespread in both syntactic and lexical domains. We have started to construct a database of synonymous expressions of Japanese "Verb + Verb" compounds semi-automatically.

Recently the Japanese compound verb lexicon was constructed by the National Institute for Japanese Language and Linguistics (NINJAL) (2013-15). It has meanings, example sentences, syntactic patterns and actual sentences from the corpus that they possess.

However, it has no relationship information with another words, such as synonymous words and phrases. We try to detect them automatically as much as possible in order to help humans find not only typical synonyms but also unexpected synonymous words and phrases and manually compile a lexicon of synonymous expressions of Japanese compound verbs. We conducted three actions, 1) extract synonymous expressions (words and phrases) from corpus by using word2vec and cosine similarity measure, 2) classify synonymous expressions into several clusters by using the clustering method k-means++ (Arthur and Vassilvitskii 2009) and 3) find a suitable number of clusters corresponding to the meanings of each compound verb by PCA and compile a list of synonymous expressions and their semantic classes by assessing them manually.

2. Compound verbs that we treat

Japanese compound verbs consist of two verbs joined together. The morphological form is a combination of the first verb in an adnominal form and the 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'.

These 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 the variety of second verbs, no restriction on the first verbs and so on. We exclude the syntactic compound verbs and treat only lexical compounds which tightly combine two verbs as one word and also not productive than syntactic compound verbs

3. Related Researches

So far, 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 thegeneration of compounding words(Sag et.al.2002; Katz and Giesbrecht2006; Hashimoto and Kawahara 2008 and so on). In Japanese, Uchiyama and Ishizaki (2003), and Uchiyama and Baldwin (2004) investigated the ambiguities of compound verbs and tried to discover the rules for generation, but the number of compound verbs that they treated was insufficient. As a research on predicative verbs, which was not limited to compound verbs, Izumi et.al (2013) proposed the recognition of semantically equivalent predicate phrases by using definitions in a dictionary and several thesauri as features of verbs. They showed a good result in their experiment, however, in our research, compound verbs that we deal with are not always registered on those lexicons. We have to explore possible methods to find similar expressions of words from corpus.

4. Data

We use "five hundred million Japanese texts gathered from the web" produced by Kawahara et.al. (2006) as corpus for extracting synonymous words and phrases. For compound verbs, we treated compound verbs registered in the "Compound Verb Lexicon (CVL)" created by the National Institute for Japanese Language and Linguistics (NINJAL). The total number of compound verbs in this lexicon is 2700, and each compound verb has meanings, syntactic patterns and example sentences. We also utilized it for an assessment.

5. Extraction of synonym candidates of each Compound Verb

We utilized word2vec (Mikolov, 2013), one of the deep learning methods, for the vectorization of words. The learning model of word2vec that we used is CBOW (Contiguous Bag of Words) and the range of window is five words. We estimate w(t), a word located in position "t" in a sentence, due to two words each before and after w(t) (that is, w(t-2), w(t-1), w(t+1), and w(t+2)).
In our experiment, we vectorized all of the 2700 compound verbs with word2vec by using the five hundred million Japanese web corpus and extracted synonyms of 2700 compound verbs based on this vectorization.

Step (1): preprocessing

In the first trial, we simply put the output of the morphological analyzer JUMAN into word2vec, however, the result was unsatisfactory. Consequently, we decided to utilize syntactic information for the input data. Syntactic information means the case relations between verbs and nouns extracted from web corpus by KNP parser. The sets of a noun, a verb and a case marker consist of the input data for word2vec.
We needed to treat phrasal expressions as "verbs", because some compound verbs can be paraphrased into phrases. We concatenated modification relations between verbs and adverbial words and made units which we treated as "verbs" (e.g. correctly / understand → "correctly understand").

Step (2): vectorization and similarity

We performed then vectorization of all verbs and nouns in the web corpus by using word2vec and 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.

Step (3): list creation

For each compound verb in the CVL, 2000 similar expressions were chosen in order of 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 polysemous meaning in this list.

6. Getting Clusters for a Compound verb

In order to identify each of the polysemous meanings we classify synonymous expressions for each compound verb by using the clustering method k-means++ and Principal Component Analysis (PCA) (Pearson 1901).
Our procedure is shown in detail in Figure 1 below.

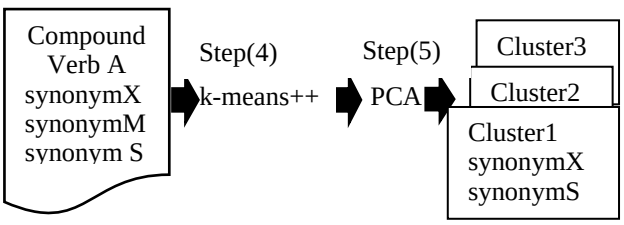

Figure1: the process of obtaining polysemous clusters

Step(4):choosing 100 synonymous expressions

In order to select a suitable clustering method, we compared the results obtained by hierarchical clustering and k-means++. For the hierarchical clustering, we adopted Ward's method. k-means++ is a partial optimization clustering of k-means developed by Arthur and Vassilvitskii (2007), and its initialization method is improved. Comparing the results, these methods both had advantages and disadvantages, and the plausibility of classification of synonymous expressions is similar, however, we chose k-means++ in this work because it found some unexpected expressions.
At the beginning, each compound verb has 2000 synonymous expressions extracted from the web corpus. We needed to narrow down the number of expressions in order to detect the plausible synonymous expressions. Our idea is to decrease the number of expressions in a step-by-step approach by iterating the k-means++. The actual procedure is described below.
1. Firstly, we set 64 clusters for the k-means++. In this stage, 2000 expressions are classified into 64 clusters.
2. For each cluster, we extract 10 expressions with the highest similarity values in the list. In this stage, we narrow down to 640 expressions (64 clusters * 10 expressions).
3. We iterate the same process as step2 for the 640 expressions. In this stage, we set 10 clusters. As a result, we obtained 100 synonym expressions classified into10 clusters (10 clusters * 10 expressions).
This data is then used as the input data for Principal Component Analysis (PCA).

Step (5): Clustering by PCA

We settled the 100 synonymous expressions 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 PCA. We implemented PCA with 100 expressions for each compound verb. We manually found clusters from the result by PCA. In the scatter diagram we extracted clusters clearly separated from the groups of unrelated expressions that were gathered together.

7. Evaluation for 40 compound verbs

In order to predict how many suitable synonyms and clusters 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, four examinees evaluated the suitability of the 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.
Furthermore, 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 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.
Our method could not extract proper clusters for two verbs among 40 verbs, i.e. "*toi-awaseru* (make an inquiry)" and "*sashi-dasu* (holdout)". Most of the candidates of synonymous expressions we extracted for these compound verbs were unsuitable.

8. Manually Making a List of Synomymous Expressons from the Result of PCA

From the result of PCA, we manually classify synonymous expressions into clusters, i.e. drawing circles in Fig. 2.

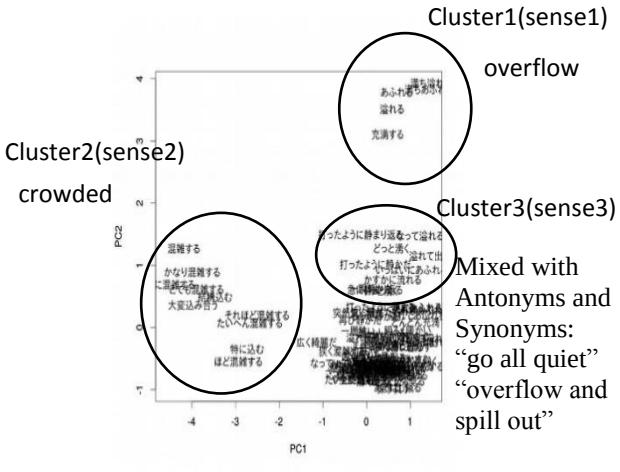

Fig.2. A distribution of synonymou expressions of あふれかえる (*afure-kaeru*) derived from PCA

The list of clusters extracted for あふれかえる (*afure-kaeru*) is shown as follows.

[Cluster 1]

Meaning : overflow
Synonymous expressions :
あふれる (*afureru*, overflow), 満ち溢れる (*michi-afureru*, bubble), 充満する(*juumansuru*, be full of) ...

[Cluster2]

Meaning : crowded
Synonymous expressions :
混雑する(*konzatsusuru*, crowded), かなり込み合う (*kanari komi-au*, very crowded), 特に込む(*tokuni komu*, extreamly congested), ...

[Cluster3]

Meaning : go all quiet
Synonymous expressions :
水を打ったように静かになる (*mizu-wo uttayouni shizukani naru*, become a complete silence as if someone can hear a pin drop.),
かすかに流れる *(kasukani nagareru*, flow faintly)
いっぱいにあふれる (*ippaini afureru*, full of and overflow),
あふれて出る (*afurete deru*, overflow and spill out) ..

Expressions in Cluster3 are a mixture of anotonymous and synonymous expressions. There are antonymous expressions like 水を打ったように静かになる (*mizu-wo uttayouni shizukaninaru*, hush fell over) and かすかに流れる (*kasukani nagareru*,, flow faintly) and also synonymous expressions like いっぱいにあふれる (*ippaini afureru*, full of and overflow) and あふれて出る (*afurete deru*, overflow and spill out). That is, the expressions in Cluster 3 are not separated clearly. When we finally make a list of synonymous expressions, even if the synonymous expressions are classified into an unsuitable cluster, i.e. Cluster 3 in this case, we do not ignore such expressions but move them to a suitable cluster. In this case, we pick up the synonymous expressions いっぱいにあふれる (*ippaini afureru*, full of and overflow) and あふれて出る (*afurete deru*, overflow and spill out) from Cluster3 and move them to a feasible cluster, in this case, Cluster1, manually.
The list for あふれかえる (*afurekaeru*) that we finally obtain is :

[Cluster 1]
Meaning : overflow
Synonymous expressions :
あふれる (*afureru*, overflow), 満ち溢れる (*michi-afureru*, bubble), 充満する (*juumansuru*, be full of), いっぱいにあふれる (*ippaini afureru*, full of and overflow), あふれて出る(*afurete deru*, overflow and spill out)

[Cluster2]
Meaning : crowded
Synonymous expressions :
混雑する(*konzatsusuru*, crowded), かなり込み合う (*kanari komi-au*,very crowded),特に込む (*tokuni komu*, extremely congested), ...

[Cluster3]
Meaning : go all queit
Synonymous expressions :
水を打ったように静かになる (*mizu-wo uttayouni shizukani naru*, become a complete silence as if someone can hear a pin drop), かすかに流れる (*kasukani nagareru*, flow faintly)

9. Compound Verb Lexicon compiled by NINJAL

The list that we are now compiling will be linked to CVL in order to cover meanings that we fail to find and also extend the CVL.
Compound Verb Lexicon (CVL) was constructed by the National Institute for Japanese Language and Linguistics (NINJAL) (2013-15). It says that comprising over 2,700 verb-verb compound verbs commonly used in contemporary Japanese, the Compound Verb Lexicon is designed to provide both researchers in linguistics and foreign learners of Japanese with useful information on their grammatical, semantic, and other linguistic features. In addition to Japanese representations, it offers English, Chinese, and Korean translations for the semantic definitions and example sentences.
As an example, the description of あふれかえる (*afurekaeru*) in CVL is as follows.

Meaning:

Japanese: 場所の収容力以上に、いっぱいである.
English: (Of a place) To be full beyond capacity.
Chinese : 远超过场所可容纳的量。［意译：爆满］
遠超過場所可容納的量。［意譯：爆滿］

Korea : 장소의 수용력 이상으로 가득차 있다.

Example :

駅前にはタクシー待ちの人たちがあふれ返っていた.
Ekimae-ni-wa takushiimachi-no hitotachi-ga afurekaette ita.
English : There were huge crowds of people waiting for taxis in front of the station.
Chinese :车站前挤满了等出租车的人
車站前擠滿了等計程車的人。
Korea : 역앞에는 택시를 기다리는 사람들이 넘쳐났다.

Syntactic pattern :

N1-*ni* N1-*ga afurekaeru*
N1-*ga* N2-*de afurekaeru*

10. Link our List of Synonymous Expressions to CVL

In our list, each cluster (cluster1, cluster2,..) corresponds to one of the meanings of each compound verb, therefore, in our list あふれかえる (*afurekaeru*) has two meanings (cluster1 and 2) and one antonymous meaning (cluster3). In CVL, this compound verb has one meaning, "(Of a place) to be full beyond capacity". CVL would be bundling cluster1 and cluster2 in our list by the same core meaning. However, in our result, synonym expressions in cluster 1 and those in cluster2 are clearly divided because they are used in a different context by figurative meaning. Actually, "people are overflowing at the station" is acceptable in Japanese, however, "water was crowded" is not acceptable. In our list, we keep two meanings, cluster1 (overflow) and cluster2 (crowded) and link cluster1 to the meaning "(Of a place) to be full beyond capacity" of あふれかえる (*afurekaeru*) in CVL and, on the other hand, add Cluster2 as a new meaning.

11. Future work

We try to compile a list of synonymous expressions for about 2700 compound verbs registered in CVL and link to CVL so that we can find what kind of paraphrases compound verbs have.

If we try to adopt our method to the construction of a large scale lexicon, it will consume significant time and human effort because the final step described in section 8 is manually conducted. In fact, it takes two people about one month to fully evaluate the result. For this task, the number of Japanese " Verb + Verb" compounds that we call "lexical compound verbs", not "syntactic compound verbs", is limited and they are not productive. Japanese VV lexical compound verbs are said to be about 3000 words, therefore our method can adopt to our task. For future work, although our result is interesting and is not bad, we intend to try another automatic method for the final clustering step instead of PCA.

Japanese compound verbs are often tough obstacles for beginning learners of Japanese to work through. By linking synonym expressions to CVL, we would like to contribute to linguistic, NLP, and Japanese language education.

12. Bibliographical 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.

Katz Graham, and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Association for Computational Linguistics, 12-19.

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.

Tomoko, Izumi., Tomohide Shibata, Kuniko Saito, Yoshihiro Matsuo and Sadao Kurohashi. 2013. *Recognizing Semantically Equivalent Predicate Phrases based on Several Linguist Clues*. Journal of Natural Language Processing, 539-561.

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.

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.

13. 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.