Conversion of Standard Japanese into Young People's Words

Using word2vec

Eishi Takeshita^a, Hiromi Takahama^b, and Kenji Matsui^{a,*}

^a Graduate School of Robotics & Design Engineering, Osaka Institute of Technology, Japan

^b Graduate School of Engineering, Osaka Institute of Technology, Japan

*Corresponding Author: kenji.matsui@oit.ac.jp

Abstract

Billions of people use social media every day, and the number of users are growing rapidly. That trend has been changing not only our life style but the way of our business. Especially, young people are using social media heavily. Therefore, it is important to understand and use young people's words to make better communication. This paper describes our preliminary study of conversion method from standard Japanese into young people's words using latest AI technologies. We use word2vec in order to quickly analyze and grasp features of those words. The corpus we used for the training and testing is from Japanese microblogging, where we can obtain various types of data from wide range of generations. Based on the analysis result, we confirmed that we might be able to classify standard words and young people's words using SVM with around 70% accuracy. Finally, we made preliminary conversion test. The test result showed that our approach using word2vec and SVM seems promising.

Keywords: word2vec, young people's words, SVM, microblogging

1. Introduction

With the latest spread of communication using social media, sharing opinions each other has become quite common, especially among young generations. Because billions of people use social media every day, and the number of users are growing rapidly, that trend has been changing not only our life style but also the way of our business. Usually people use different type of colloquial expressions depend on the generations. Therefore, it is important to understand and use such expressions in terms

of business applications. Among various generations, young people's words are quite different from the standard expressions. To make better communication with those generations, it is effective to utilize young people's colloquial words. One of the promising web-based application is so-called "Chatbot". Such dialog system applications are becoming very important for industries because of the necessity of reducing human resources. However, at the same time industries require natural and attractive dialog platform. We aim to construct a basic framework for converting the response of the existing dialog system to a young people's words which is a crumbled representation from a standard word. This paper describes our preliminary study of conversion method from standard Japanese into young people's words using latest AI technologies. We use word2vec in order to generate a vector space, in our case around 200 dimensions. Word vectors are positioned in the vector space, and words which have common contexts are located in closely in the space. For example 「めっちゃ」, 「すごい」. These two words mean 「very」 in Japanese, and very close locations. The corpus we used for the training and testing is from Japanese microblogging, where we can obtain various types of data from wide range of generations. Based on the word2vec conversion result, we looked into the vector space and confirmed that the standard words and the equivalent young people's words can be extracted from the word2vec output. Our approach is to use word2vec to extract the words which have same meaning but different type of expressions. Then we use SVM as a classifier to discriminate standard words and colloquial words. Finally, we try to replace the standard expressions by the colloquial expressions.

2. Training of word2vec

2.1 Basic principle of Word2vec

Distributed representations of words in a vector space help developing algorithms to achieve better performance in natural language processing tasks by grouping similar words. To be able to analyze and grasp features of various types of words, we use word2vec⁽⁴⁾ because it is a kind of auto encoder and it can quickly convert words into vectors. word2vec is one of the powerful method of natural language processing. Those word vectors are placed into hundreds dimension vector space. Then, we can calculate distance and direction between words. That means we can research co-occurrence information of each words. In brief, Using vector space, similar words can be grouped and similarity value can be calculated.

2.2 Continuous Bag-of-Words(CBOW)

CBOW is one of the word2vec training method. CBOW uses both the n words before and after the target word $W(t)$ to predict it as shown in Figure 1. As a result of dimensional compression, vector space is generated.

Fig. 1. Continuous Bag-of-Words.

2.3 Development of distributed word representations

We acquire distributed representations by learning with word2vec and analyze those learned vectors to see various linguistic regularities and patterns. Since we are focusing on young people's words, a good place to start is to create young people's word dictionary. Based on the vector patterns of the dictionary, we can explore the characteristics of young people's words. The basic learning process shows in Figure 2.

Fig. 2. Flow of creating corpus and learning.

2.4 Create young people's word dictionary

There are a couple of web sites which have a set of young people's words and thesaurus. With refer to those information, we have created a small dictionary which has standard expressions versus young people's expressions. Table 1 shows an example of the dictionary.

Table 1. Part of bilingual dictionary.

Japanese	Japanese Slang		English
可愛い	きゃわたん	ぎゃんかわ	cute
かわいい	きゃわたん	ぎゃんかわ	cute
とても	ガチ	やばい	very
すごく	ガチ	やばい	very
いっぱい	めっちゃ		full
たくさん	めっちゃ		a lot

2.5 Collection of real-world test corpus

For the training and testing, we looked at some of Japanese microblog sites. We used 230 keywords extracted from the dictionary to search appropriate texts in the gathered data between November and January 15th 2016. Figure 3 shows an example of our corpus.

```
@kn_ee_brmsnya はい...!
とても十数年前のアニメとは思えないクオリティーですね RT @kanayan524: 今
は天使
グラデーションがとても良き。
撮影のいいてちょっと緊張してたけど楽しかった
なつみさん毎回ありがとうございます https://t.co/DXC4f6oxPntantan さん体調悪い中
とてもありがとうございました m(_ _)m RT @Tpat_m4E: 最近とても幸せなり
(*'-'*)@satokyun_xxx
```

Fig. 3. An example of collected twitter data.

2.6 Morphological analysis

Automatic morphological analysis is widely used to be able to make annotated corpora. We used Mecab⁽⁵⁾ to make segmentation of an input string into a sequence of units and assignment of part of speech (POS) tags. Figure 4 shows an example of the Mecab output. Because the IPA dictionary for Mecab has almost no young people's words, we added our young people's word dictionary and some more words into the user dictionary for the Mecab operation.

```
ぎゃんかわなポーチ見つけた
ぎゃんかわな 名詞,一般,****,
ポーチ 名詞,一般,****,ポーチ,ポーチ,ポーチ
見つけ 動詞,自立,*,一段,連用形,見つける,ミツケ,ミツケ
た 助動詞,*,*,特殊・タ基本形,た,タ,タ
! 記号,一般,****,!,!,!
EOS
このポーチぎゃんかわなんだけど
この 連体詞,****,この,コノ,コノ
ポーチ 名詞,一般,****,ポーチ,ポーチ,ポーチ
ぎゃんかわなんだけど 名詞,一般,****,
EOS
```

Fig. 4. Example of morphological analysis.

2.7 Word2vec Training

Word2vec is a particularly computationally-efficient predictive model for learning word embeddings from raw text. In order to learn the corpus, we conducted the following options. Table 2 shows the actual parameters of word2vec.

Table 2. Option of word2vec.

Size	Window	Min-count
100	5	30

“Size” is the number of dimensions. “Window” is the setting the number of words before and after the word to be estimated as described in the 2.2. We used 5 before and after the initial setting this time. Min-count exclude words whose number of occurrences of words is 30 or less. Learning is conducted using the corpus created in the preceding clause.

3. Feature analysis of young people's words

3.1 Analysis method

From the word2vec output you can acquire 100 dimensional word vector data. Fig.5 plots the axes of the primary and secondary components converted by t-sne method in terms of adverb group related to "とても", very in Japanese, and adjective group related to "可愛い", cute in Japanese. The t-sne method shows a probability distribution over pairs of low-dimensional objects. Next, we focused only on adjective group "可愛い", "かわいい", and "ぎゃんかわ". We picked up those words from the young people's word dictionary and visualize the positional relationship of the words using the t-sne method. (Fig.6) Then, clustering analysis was tested using both ward method and k-means method. From these result, it is difficult to see clear separation.

Fig. 5. The t-sne method.

Fig. 6. Cute relationship of word2vec.

Figure 7, Figure 8, and Figure 9 shows the similarity distance in terms of the each word "可愛い, かわいい, ぎゃんかわ".

Word: 可愛い Position in vocabulary: 245	Word	Cosine distance
	かわいい	0.816852
	かつこいい	0.789035
	似合う	0.679798
	カワイイ	0.671800
	カッコイイ	0.658275
	美人	0.624147
	お似合い	0.592735
	萌える	0.573489
	ずるい	0.565429
	羨ましい	0.565103

Fig. 7. Similarity distance regarding "可愛い".

Word: かわいい Position in vocabulary: 192	Word	Cosine distance
	可愛い	0.816852
	かつこいい	0.762375
	似合う	0.623624
	カワイイ	0.614310
	かわいかつ	0.604776
	ぎゃんかわ	0.600858
	可愛かつ	0.597997
	お似合い	0.576268
	カッコイイ	0.569803
	可愛く	0.568540

Fig. 8. Similarity distance regarding "かわいい".

Word: ぎゃんかわ Position in vocabulary: 84	Word	Cosine distance
	きやわたん	0.674648
	かわいい	0.600858
	HBirthday	0.587638
?	?	0.566828
?	?	0.561255
	推せる	0.559969
	ちゃん	0.551095
	めぐ	0.540671
	??	0.534735
	ダイスキ	0.532230

Fig. 9. Similarity distance regarding "ぎゃんかわ".

Figure 7 ~ 9 shows the ten most closest word lists to the word either "可愛い", "かわいい" or "ぎゃんかわ". It seems these three words are very close, however, when we tested k-means clustering with total cluster number is either 6 or 20. Figure 10 and Figure 11 shows which cluster those words belong to. The young people's word "ぎゃんかわ" is not in the same cluster as "可愛い", "かわいい". From those results, we tried to find the boundary between young people's words and others. Also, you can find quite frequent co-occurrence between young people's words and special emotional symbols such as "!!!!!!", "ah ah ah" or "wwwwwww". This is a powerful tip to search for the young people's words, and even we can use these symbols

to convert standard expression into more casual ones.

0	ぎゃんかわ
3	可愛い
3	かわいい

Fig. 10. Cluster number 6.

9	ぎゃんかわ
2	可愛い
2	かわいい

Fig. 11. Cluster number 20.

3.2 SVM

From our preliminary experiment results, we can see that the similar words have close in the vector space and young people's words seems to have different features from the standard expressions. To be able to investigate such hypothesis more clearly, we performed another experiment using SVM to find the boundary. We labeled 234 words which are both standard expressions and young people's words. It was 74.7% of accuracy as a result of guessing word types. That result was encouraging, and we tried to make a simple word conversion test.

3.3 Conversion test

We performed simple classification test using word2vec and SVM. The basic steps are as follows.

- (1) Detect an adverb from an input sentence which is converted into young people's expression.
- (2) Find top 60 words similar to that adverb in terms of the cosine similarity distance using word2vec.
- (3) Classify those 60 words using SVM, and discard the words which were classified as standard expressions.
- (4) Also discard emotional symbol words using SVM.
- (5) Replace that adverb by the word which is the closest word in the remaining words.

3.4 Conversion result

A result of the conversion test with 6 test sentences are as follows;

とてもかわいいぬいぐるみ	⇒	めちゃくちゃかわいいぬいぐるみ
隣の客はよく柿を食う客だ	⇒	隣の客はなんか柿を食う客だ
すぐに学校に行きます	⇒	速攻に学校に行きます
ときどき遊びに行く	⇒	固遊びに行く
かなり大変な仕事	⇒	けっこう大変な仕事
どうしたの？	⇒	どーしたの？

Fig. 12. Example of conversion.

We were able to convert these test sentences into young people's expression with reasonably good quality. However, there are some inadequate connections. Because the test corpus size was relatively small, it was difficult to find enough amount of candidates.

4. Conclusion

We investigated the possibility of conversion between standard Japanese expression and young people's expression. We utilized word2vec in order to quickly analyze and grasp features of those words. The corpus we used for the training and testing is from Japanese microblogging, where we can obtain various types of data from wide range of generations. Based on the analysis result, we confirmed that we might be able to classify standard words and young people's words using SVM with around 70% accuracy. Then, we made a preliminary conversion test. The test result showed that our approach using word2vec and SVM seems promising. However, large number of test sentences need to be tested with large corpus to be able to tune the conversion accuracy. In order to improve the performance, the following problems need to be solved.

- 1) We need to put more large number of young people's words into the corpus and dictionary to make both morphological analysis and conversion more accurately.
- 2) We should use not only word-level data, but also phrase-level information so that the conversion algorithm can handle more natural sounding expressions.

- 3) Depending on the part-of-speech, we need to figure out more detailed conversion rules or exceptional rules.

We will work on those issues as our future work.

References

- 1) Journal of Ministry of Internal Affairs and Communications : "Heisei era 28-year information communication white paper", Chapter.4, pp. 235, 2016
- 2) J. Weizenbaum. "ELIZA-a computer program for the study of natural language communication between man and machine", Communications of the ACM, Vol. 9, No. 1, pp. 36-45, 1966.
- 3) ['F8', 2016](https://developers.facebook.com/videos/)
- 4) <https://code.google.com/p/word2vec/>, "word2vec", 2016
- 5) <http://taku910.github.io/mecab/>, "Mecab", 2016