Translation of Network Popular Words

Luo Shihan[1]

[1]*Graduate School and Faculty of Information Science and Electrical Engineering, KyuShu University, Fukuoka, Japan*

Keywords: Network popular words, Word2vec, Sentiment analysis, Sentiment polarity

Abstract: In today's SNS, numerous network popular words have been creating. Due to most network popular words are only used in a short period of time, it is hardly to find a dictionary that includes network popular words. So, it is unpractical to translate network popular words by traditional machine translation method. Now, the research of extracting network popular words has obtained some outcomes to a certain extent. However, the existing research of paraphrasing network popular words, which uses distributed representation method to analyse network popular words, does not make much progress. The purpose of this paper is to improve the translation of network popular words by adding sentiment analysis into distributed representation method. By analysing network popular words in eight sentiment polarity (joy, sadness, trust, disgust, fear, anger, surprise, anticipation), the accuracy of translation of network popular words is obviously improved.

1 INTRODUCTION

As the Internet becomes more and more popular, a lot of new network words appear, which have greatly enriched the vocabulary. However, if these words have no explanation, only certain people can understand them. In an investigation, over 80% Chinese Internet citizens use network popular words in SNS, and there are more and more network popular words being created every year. But, most elder people and foreigners do not understand those network popular words.

Network popular words mainly have three characteristics:

(1) Short time existence:

the existing time of network popular words is shorter than normal words and phrases.

(2) Semantic deviation:

the meanings of network popular words are different with their original meaning.

(3) Self-creation:

there are more and more network popular words are being created every year.

Due to those three characteristics, most network popular words are unknown words, which means Dictionary Based Machine Translation is not suitable for network popular words.

Many researchers have done a lot of work in cause of occurrence, Part of Speech tagging and extraction of network popular words, and they have got an ideal result in the extraction of network popular words. Nowadays, some translation systems just copy the network popular words as their translation, e.g. the translation of ‘lmk when you can go’ in those systems is ‘大丈夫 lmk 行くことができるとき‘, so the translation of network popular words is necessary. To improve the accuracy of Machine Translation, this paper aims to find a way to translate network popular words.

2 RELATED WORK

The extraction of network popular words mainly has three methods: Rule-based method, Statistics-based method and Combine Rule-based and Statistics-based method. Zheng has got an ideal result in the extraction of network popular words: precision rate: 91.2%, recall rate: 95% (Zheng and Wen-Hua, 2002). But only a few researchers pay their attention to translate network popular words, and the accuracy in translation of network popular words is not sufficiently high. (Zhao Xinyi, 2015) reported that the accuracy of the translation is around 80%. Translation of network popular words still has room for improvement.

Because the existing researches mainly aim to extract network popular words, there only a few Japanese and English papers aim to translate

Shihan, L.
Translation of Network Popular Words.
DOI: 10.5220/0008097500380043
In *Proceedings of the International Conference on Advances in Computer Technology, Information Science and Communications (CTISC 2019)*, pages 38-43
ISBN: 978-989-758-357-5

unknown words, and all those papers do not focus on network popular words. Fortunately, a few of Chinese papers can be read as reference materials. Those existing researches about translation of network popular words mainly have three methods: Rule-based method, Word vector method and Combine Rule-based and Word vector-based method.

Network popular words have lots of characteristics and categories, such as homophonic network popular words (e.g. バイ バイ---- bye bye), simplified network popular words (e.g. イケメン----いける man), English homophonic network popular words (e.g. トリクルダウン ---- trickle-down theory) and pictographic network popular words (e.g. %>_<% ---- crying). Rule-based method can use those characteristics to translate some network popular words, but it is not able to translate certain network popular words which do not have significant characteristics. The accuracy of Rule-based method is from 20% to 80% (Shang Fenfen, 2015).

Word-vector-based method basically use the semantic relationships between words in the context to find the synonyms of network popular words. This method compares network popular word's word vector with other word's word vector to get the closest word as the synonym of the network popular words, then the synonym can be regarded as the network popular word, the accuracy of this method can reach 80% (Zhao Xinyi, 2015). Word2vec is the most popular tool, based on deep learning and released by Google in 2013, to train word vector now. This tool adopts two main model architectures, continuous bag-of-words (CBOW) model and continuous skip-gram mode. To learn the vector representations of words: The CBOW architecture predicts the current word based on the context, and the skip-gram use the current word to predict surrounding words (Jansen and Stefan, 2018).

The combining Rule-based and word-vector-based method has the best precision to translate network popular now. Because these two methods are independent, this method has achieved a higher accuracy, the precision rate is about 85% in certain context data.

Since network popular words are most used in SNS, not in formal paper or news. The very precise translation of network popular words is unnecessary. It is enough to just transmit the rough meaning and feeling of network popular words in the social internet. Therefore, this paper aims to get the meaning of network popular words by not only semantic analysis but also sentiment analysis.

However, the existing two methods Word-vector-based method and combining Rule-based and word-vector-based method are mainly using word's vector, it is important to obtain a very precise and complete data to train word vector. Due to the imperfect training data and training tool, the data of word vector cannot be perfect so that a few of chosen synonyms are not suitable. In this paper, to adjust those unsuitable synonyms and get a higher translation precision, we use sentiment polarity analysis to supplement the word-vector-based method.

Sentiment analysis can be improved by dividing sentiment into different types. As Plutchik's Wheel of Emotions shows that sentiment words not only can be divided into positive or negative polarity, but also can be divided into detailed emotional types (Plutchik and Robert, 1991), such as joy, anger, sadness and so on. This is also a way to get more precise result of sentiment analysis. And different sentiment words which belong to the same polarity usually have different sentiment intensity (Wang, 2015), e.g. "laugh" has a larger intensity than "smile."

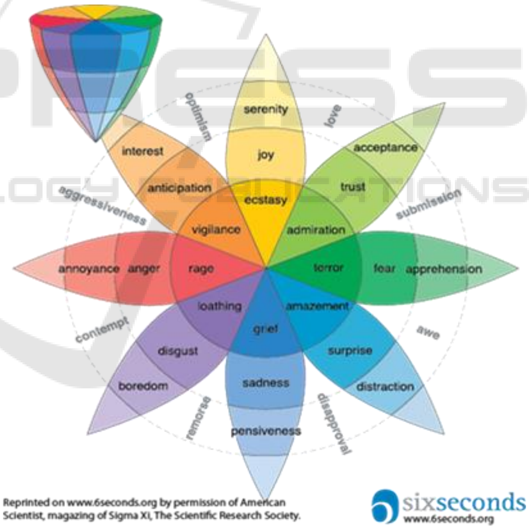


Figure 1: Plutchik's Wheel of Emotions.

There are mainly three types of methods in identifying polarity of Chinese sentiment words: Thesaurus-based method, Corpus-based method and Morpheme-based method.

Thesaurus-based method which computes similarity distance between reference words and the given sentiment word in thesaurus. It acquires sentiment words mainly by synonyms, antonyms, and hierarchies in thesaurus such as WordNet and HowNet (Kim and Soo-min, 2004). This method uses some seed sentiment words to bootstrap via synonym and antonym relation in a thesaurus.

Corpus-based method which computes similarity between reference words and the given sentiment words by statistic method in corpus. This approach has an implied hypothesis that sentiment words have the same polarity with the reference words of the greatest cooccurrence rate and the opposite polarity with the reference words of the least cooccurrence rate. The polarity of sentiment words is assigned by computing cooccurrence in corpus (Turney and Littman, 2003).

Take the characteristics of Chinese character into consideration (a single Morpheme also has meaning, e.g. 事 =affair, 情 =circumstances), some researchers proposed morpheme-based methods (Yuen and Raymond, 2004). Yuen et al. proposed a method by calculating similarity between reference morphemes and sentiment words in corpus to get the polarity.

3 PROPOSED METHOD

Most sentiment polarity analysis methods only analyse words in positive and negative polarity. But, if we want to translate an unknown word, it is not sufficient. To improve sentiment analysis methods, a good way is to add different emotion types like the Plutchik's Wheel of Emotions shows into Corpus-based method and Morpheme-based method by not using only 'positive and negative' aspect but using 'joy and sadness', 'trust and disgust', 'fear and anger' and 'surprise and anticipation' 4 aspects as commendatory and derogatory. The following graph shows the outline of the way to improve Corpus-based method and Morpheme-based method by using detail benchmarks.

Since the word vector can be got by using word2vec, and by comparing words' word vector we can get the closest word as the synonym of network popular word, then the synonyms can be regarded as the network popular words. But due to the defect of word vector, the accuracy is not very high. In this research, because the extraction of network popular words already has an ideal result, this paper mainly focuses on the translation of network popular words rather the extracting of network popular words. By adding sentiment polarity analysis into word-vector-based method to get more suitable synonyms of network popular words, so that the accuracy will be improved. The following outline graph Figure 3 is the way about how to add sentiment polarity analysis into word-vector-based method. First, use Word-vector-based method to find top N close candidates. Second, use sentiment polarity analysis to get the sentiment polarity of network popular words, then compare the sentiment polarity with sentiment polarities of top N close candidates. Finally, choose the candidate which has closest sentiment polarity as synonym and translation.

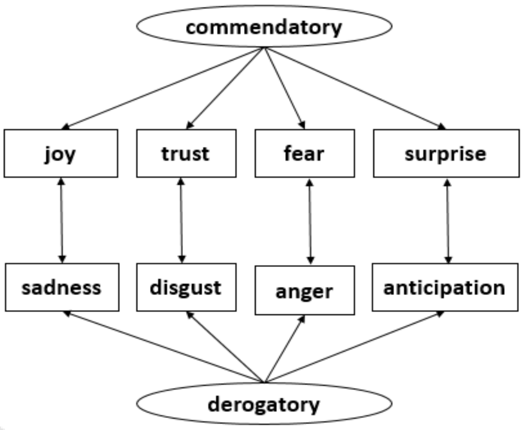


Figure 2: Benchmarks.

4 EXPERIMENTS

In this paper, we use 10G online novels as training data of word vector. As the proposed method, first, use cosine similarity to find top 10 close words as candidates of network popular words. In the example Figure 4, a network popular words 逆袭 which means 'counterattack' is spread widely in China. By single using Word-vector-based method, the closest candidate 王者归来(return of the king) is not a synonym of 逆袭(counterattack), it makes a mistake to get the synonym of the network popular word.

The meanings of top 5 close candidates of 逆袭 are 王者归来 (return of king), 绝地反击 (counterattack), 横空出世(come out suddenly), 开挂(cheat) and 咸鱼翻身(turnaround).

To adjust the inaccurate outcome of word-vector-based method, sentiment analysis of network popular words and candidate words is needed. In this paper, by means of improving standard Corpus-based method, we also use cosine similarity to calculate the distance between words and eight emotion words shown by Plutchik's Wheel so that it can get the emotion distribution of words in eight emotion directions.

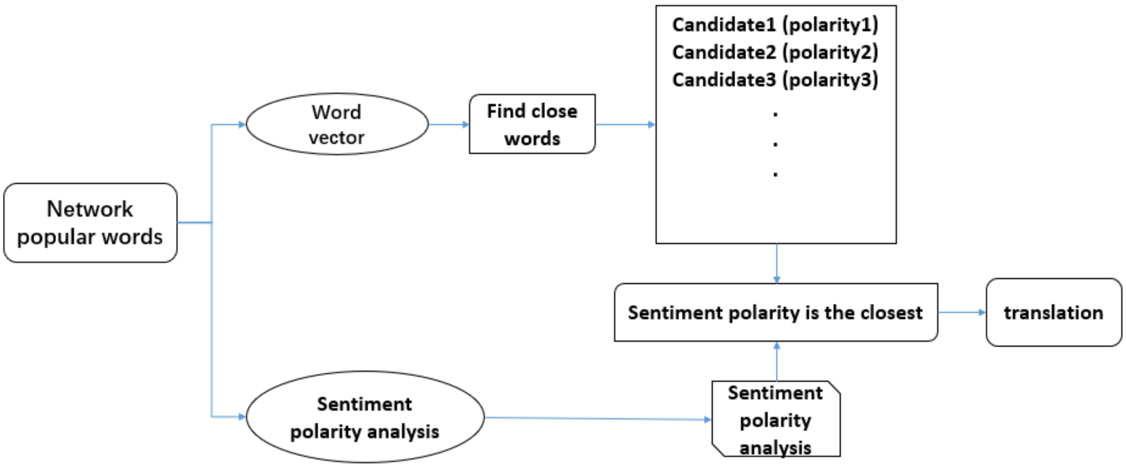
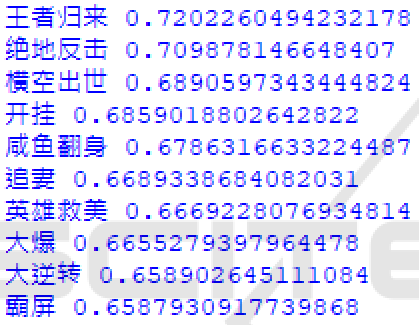


Figure 3: Proposed method to improve translation precision.



```
王者归来 0.7202260494232178
绝地反击 0.709878146648407
横空出世 0.6890597343444824
开挂 0.6859018802642822
咸鱼翻身 0.6786316633224487
追妻 0.6689338684082031
英雄救美 0.6669228076934814
大爆 0.6655279397964478
大逆转 0.658902645111084
霸屏 0.6587930917739868
```

Figure 4: Close words of 逆袭.

The emotion distribution of 逆袭 and some its candidates is as Table 2. Then we use KL divergence to compare the emotion distribution between network popular word and its candidates. KL divergence’s formula is as follow:

$$\mathrm{D(P||Q)} = \sum \left(P(x) \log \left(\frac{P(x)}{Q(x)} \right) \right) \qquad (1)$$

P and Q is probability distribution of data.

By using KL divergence to analyse the emotion distributions of 逆袭 and some of its candidates, we get the results as follow:

Table 1: Results of KL.

Candidates	王者归来	绝地反击	横空出世	咸鱼翻身
KL	0.0092	0.0005	0.0012	0.0015

Because 开挂(cheat) also is a network popular words, it is an unknown word, we delete it from the candidates. From the result of KL divergence, the emotion distribution of 绝地反击(counterattack) is the closest word to 逆袭(counterattack), their meaning is the same. Obviously, by supplementing sentiment analysis into word-vector-based method, the inaccurate candidate has been adjusted, and the right candidate has been found.

It necessary to figure out whether network popular words’ candidates contain sentiment polarity or not, and to prove it is the reason of the improvement. Calculating the variance of similarity of words is a way to figure out it. As the Table 3, the variances between network popular words’ candidates and sentiment benchmarks are significantly bigger than the variances between network popular words’ candidates and non-sentiment benchmarks {时间(time),星期(week),水分(moisture),位置(position),书籍(book),新闻(news),大陆(continent),月份(month)}. On the other hand, the candidates of normal words do not have specific tendency, so that network popular words indeed contain sentiment polarity, we can use it to improve the accuracy of translation.

In addition, we randomly chose 100 network popular words, and invite two volunteers to judge whether the translations are true or not. Through using Kappa-Statistic analysis, only 48 network popular words have been translated accurately by simply using Word-vector-based method, but our method was useful for adjusting 14 wrong translation so that 62 network popular words have been translated precisely.

Table 2: Emotion distribution of 逆袭 and its candidates.

Words Benchmarks	逆袭 （counter back）	王者归 （return of king）	绝地反击 （counter back）	横空出世 (break out)	咸鱼翻身 （turnaround）
喜悦（joy）	0.1897	0.2709	0.2400	0.0902	0.1628
悲伤（sadness）	0.1158	0.1924	0.1236	-0.0225	0.0060
信任（trust ）	0.1153	0.0943	0.1429	0.1325	0.1217
厌恶（disgust）	0.0635	-0.0041	0.0296	0.0250	-0.0078
害怕（fear）	0.0347	-0.0052	0.0866	-0.1181	0.0263
愤怒（anger）	0.1790	0.1717	0.2936	0.0787	0.0411
惊奇（surprise）	-0.0029	0.0222	0.0492	0.0654	0.0542
预期(anticipation)	0.1871	0.1919	0.2584	0.0980	0.1339

Table 3: Variance of words’ similarity.

Variance of candidates’ similarity	Network popular words				Normal words			
	他喵的 Damn it	鸡冻 excited	忧桑 sad	妈蛋 damn	风光 landscape	地图 map	社会 society	茶杯 cup
预期,惊喜,喜悦,悲伤,信任,厌恶,愤怒,害怕（average）	0.0063	0.0148	0.0120	0.0065	0.0125	0.0079	0.0123	0.0034
时间,星期,书籍,新闻,惊奇,预期,大陆,月份（average）	0.0034	0.0060	0.0073	0.0038	0.0114	0.0092	0.0158	0.0043

Table 4: Kappa-Statistic (Word vector).

V1 V2	Accurate	False	Total
Accurate	48	8	56
False	0	44	44
Total	48	52	100

Table 5: Kappa-Statistic (our method).

V1 V2	Accurate	False	Total
Accurate	60	6	66
False	2	32	34
Total	62	38	100

The K value of Table 4 and Table 5 is 0.84 and 0.83. So, the credibility of translation is sufficient credible.

Table 6: Outcome comparison.

Method	Word vector	Our method
Total number	100	100
Accurate number	48	62

5 CONCLUSIONS

In translation of network popular words, by supplementing sentiment polarity analysis into word-vector-based method, the accurate rate gets a 14% promotion, clearly, the result has been improved. It indeed makes certain progress.

And here are two types of network popular words that sentiment polarity method does not work: 1. The candidates given by word-vector-based method are already perfect (this is not bad); 2. The existed words have new meaning.

6 NEXT WORKS

Besides of online novels, it is necessary to try more types of training data. And 100 network popular words are also insufficient, it is necessary to experiment more words. Trying some different benchmark words to figure out whether benchmarks will influence the outcome. And the network popular words that our method does not work also is a good point to explore.

REFERENCES

Zheng J H, Wen-Hua L I, 2002. A Study on Automatic Identification for Internet New Words According to Word-Building Rule. *Journal of Shanxi University*.

Zhao Xinyi, 2015. Automatic Extraction and Translation of Popular Words on the Internet. *Central China Normal University*.

Shang Fenfen, 2015. Research on the Sense Guessing of Chinese Unknown Words. *Nanjin Normal University*.

Jansen, Stefan, 2018. Word and Phrase Translation with word2vec.

Plutchik, Robert, 1991. The emotions. *University Press of America*.

Wang, Bingkun, et al, 1991. A fuzzy computing model for identifying polarity of Chinese sentiment words. *Computational intelligence and neuroscience*.

Kim, Soo-Min, and Eduard Hovy, 2004. Determining the sentiment of opinions. *Proceedings of the 20th international conference on Computational Linguistics*. *Association for Computational Linguistics*.

Turney, Peter D, and Michael L. Littman, 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*.

Yuen, Raymond WM, et al, 2004. Morpheme-based derivation of bipolar semantic orientation of Chinese words. *Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics*.