International Journal of Asian Language Processing

Semantic Similarity of Inverse Morpheme Words Based on Word Embedding

Jiaomei Zhou
Institute of Chinese Information Processing
Beijing Normal University
Beijing, China
zhoujiaomei@outlook.com

Zhiying Liu*
Institute of Chinese Information Processing
Beijing Normal University
Beijing, China
liuzhy@bnu.edu.cn

Received (30 July 2021)

Inverse morpheme words are compound words that have the same morphemes but are arranged in the opposite order. The majority of related works on the subject have focused on a narrow investigation of dictionary definitions, with few studies based on large-scale corpora. We used the People's Daily corpus (1946-2017) to add and delete words from a base list and obtained a word list of 668 pairs of inverse morpheme words. Furthermore, the cosine similarity is computed by using word embedding based on the distributed representation, and the Pearson correlation coefficient between it and the manually annotated value is 0.907, indicating that this method can measure the semantic similarity of inverse morpheme words very close to human judgment. We also discovered that 76 percent of inverse morpheme words have a cosine similarity of 0.4 or higher, and that word formation, part-of-speech, and frequency all have an impact on semantic similarity.

Keywords: Inverse morpheme words; Semantic similarity; Word embedding.

1. Introduction

In the history of the Chinese language, one of the clearest developmental changes has been a shift from monosyllabic to bisyllabic words[1], where one of the main reasons for the early production of bisyllabic words is the temporary combination of synonymous monosyllabic words. The ancient Chinese language was dominated by monosyllabic words, and it was common for these monosyllabic words to be used synonymously or antonymically. The order of words was relatively free, so there were compound words with the same morphemes but in the opposite order. In modern Chinese, such compound words are referred to as inverse morpheme words since their morphemes can be reversed[2,3]. Because of the reasons above, they often have the same or similar meanings. Consider the following example:

宋荣子之议，设不**斗争**，取不随，仇不羞，囹圄见侮不辱，世主以为宽而礼之。

（《韩非子·显学》）

处乡不节，憎爱无度，则**争斗**之爪角害之。嗜欲无限，动静不节，则痤疽之爪角害之。（《韩非子·解老》）

In this case, the words “争斗” and “斗争” come from Han Fei Zi. Their meanings are the same. This conclusion is based on my personal reading experience of ancient Chinese books. Is it possible to reach a conclusion using a statistical approach?

Many approaches to calculating semantic similarity between two words have been proposed, and in Section 1.2 and Section 2, we will briefly introduce one that is based on word embedding and explain why we chose it. An experiment will follow in Section 3, obtaining the cosine similarity of words relying on word embedding and calculating the Pearson correlation coefficient between it and the manually annotated value to see if this method is feasible. The study of the factors that influence semantic similarity is another contribution to this paper (Section 4).

1.1. *An Overview of previous work on inverse morpheme words*

The consensus on the definition of inverse morpheme words is that they are compound words that have the characteristics of two identical morphemes with the same pronunciation. Their semantics, however, remain a point of disagreement. Cao Wei[4] defined inverse morpheme words more strictly, believing that true inverse morpheme words should have the same phonological and written forms, opposite linear order, and the same meanings, such as “累积/积累，吞并/并吞，通畅/畅通，离别/离别，斗争/争斗”. While the word pairs “工人/人工，情敌/敌情” have the same morphemes but different meanings, the word pairs “孙子/子孙，结巴/巴结” which are pronounced differently, are pseudo inverse morpheme words. Based on this definition, he listed 51 pairs of completely synonymous inverse morpheme words, such as “逃窜/窜逃”, and 56 pairs of words with different meanings, such as “嘴快/快嘴”, etc. Bo[1] argued that not all homographs are semantically identical. He pointed out that only if the morphemes are all in parallel structure and their morphemes have exactly the same meaning, it is only possible that their meanings are exactly equal.

After a long time of development, the semantic relationship between a pair of inverse morpheme words has become more complicated, with various semantic relationships such as equal, similar, or different, etc. The meaning relationship between their constituent morphemes is also more complicated, and most of them are not identical, but one of the senses is the same. Therefore, in the current study, inverse morpheme words are defined as a pair of bisyllabic words with inverted morpheme orders and related morpheme meanings in modern Chinese.

In addition, some other researchers have extracted word lists of inverse morpheme words from dictionaries or corpus. Zhang[5] extracted 85 pairs from the recent Chinese corpus; Tang[6] extracted 136 pairs from both sides of the Taiwan Strait, among which there are a large number of Taiwanese words, such as “熊猫/猫熊，日昨/昨日” which are hardly

used in Mandarin. Huang[7] extracted 738 pairs from the Modern Chinese Dictionary (2005 edition) and the Applied Dictionary of Inverse Morphemes Words. The meaning relationships of these words were also classified according to the dictionary interpretation. In conclusion, there is still a lack of a word list based on a massive modern Chinese corpus.

1.2. *Word embedding and distributed representation*

It is necessary to digitally represent words that computers can understand and process for natural language processing. Word embedding (vectors), coined by Bengio et al.[8], is one of the most popular approaches. It can be understood as the creation of a word list in which each word corresponds to a vector in the word list, and the representation is performed by looking up the vector corresponding to each word in the word list[9].

There are two main approaches to word embedding: one-hot representation and distributed representation. One-hot representation represents each word as a high-dimensional binary vector with the length of the size of the corpus word list, and the position where the word appears is marked as 1 and the other positions are marked as 0, so that each word can be represented as a string of numbers consisting of 0 and 1. It is based on the mutual independence between words, which is a simple and effective encoding method. However, it still has two drawbacks: it tends to cause the curse of dimensionality and a loss of context. Because the encoding dimension of each word is the size of the whole vocabulary, the larger the number of words, the larger the dimension will be, so the encoding dimension is huge and sparse, making the computation more expensive. More importantly, one-hot representation assumes that words are independent of one another and cannot reflect the degree of relationship between words. For example, in the sentence “I am Chinese and I love China”, “Chinese” and “China” can be presented as [0,0,1,0,0,0] and [0,0,0,0,0,1], and their dot product is zero, while the dot product of “Chinese” and “and” is also zero. It means that “China” and “and” in this representation approach is the same as “Chinese” and “and”. There is no difference in the similarity between these word pairs. It shows that one-hot cannot represent the semantic relationship between words. Therefore, it is not suitable for the representation of inverse morpheme words in our research.

Distributed representation solves the problem of one-hot. The dimensionality of the vectors is not constrained by the size of the word list, and the text is represented as low-dimensional, dense continuous vectors. Each word in the word list is represented by a real vector, which is usually 50-dimensional or 100-dimensional. Each word is a point in the vector space, and the distance between points is proportional to the similarity between words. Word2Vec[10] is the most popular of the word embedding models, and a key benefit is that it can take additional context into account. It maps words to n-dimensional vectors. A word is represented by an n-dimensional vector, and a long text is represented by multiple short n-dimensional vectors. The closer the semantics of two words are, the closer their vectors are in the vector space. The most significant advantage of word embedding is that it can capture the semantic information of words, allowing semantically related or similar words to be close in vector distance.

2. Methods

2.1. *Pre-trained word embedding model*

In the current research, we used distributed representations of words obtained from Word2Vec to compute the semantic similarity of inverse morpheme words. The pre-trained Chinese word vector trained on the People's Daily corpus (1946-2017)[11]. The model contains 300-dimensional vectors for 356,053 words and phrases. We chose it instead of the Ancient Chinese Corpus mainly because we analyze the semantic features of modern Chinese inverse morpheme words from a synchronic perspective, rather than focusing on exploring the causes or historical changes of inverse morpheme words. The second reason is the language style. The People's Daily is China's largest newspaper. Its articles have qualities including accuracy, currentness, and clarity.

2.2. *Cosine similarity*

After the word embedding was obtained, the semantic similarity of inverse morpheme words could be obtained by calculating the embedding distance[12]. In this paper, we used cosine similarity to calculate the vector similarity of words. The cosine similarity is measured by measuring the cosine of the angle between the two vetors, and the value of the cosine of the angle is [-1, 1]. The larger the value, the smaller the angle between the two vectors, the higher the similarity. The formula is shown in equation (1).

$$\cos\theta = \frac{a - b}{||\mathrm{A}|| \cdot ||\mathrm{B}||} = \frac{\sum_{I=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}. \quad (1)$$

The cosine similarity of a pair of words was computed using the similarity function in the Gensim library (an open-source Python library), and the value of the cosine similarity was used to indicate their semantic similarity.

3. The experiment

3.1. *Word list extraction*

Several researchers have proposed methods for extracting inverse morpheme words as well as word lists. Tang's word list[6] included a large number of Taiwanese words, many of which were not applicable to Mandarin. With a total of 738 pairs of inverse morpheme words, Huang's word list[7] had the most words and basically covered Cao's word list. As a result, we used Huang's word list as a starting point and added and deleted words to create a new word list that was more applicable to modern Chinese. We discovered that Huang's

word list has some problems based on our observations. (1) There were many historical words that were no longer used in modern Chinese, with about 85 pairs having never appeared in the People's Daily corpus in the past 70 years. The words “熬煎” and “煎熬”, for example, “熬煎” is almost no longer used, whereas the latter is very common. (2) Some of the words, such as "习见/见习", "渊深/深渊”, are less commonly used frequently in Mandarin, resulting in insufficient semantic representation during word vector training, potentially leading to inaccurate final similarity calculation results. (3) Since it was extracted from a dictionary and was published in 2006, the inclusion of new words may be incomplete.

To solve the first problem, we removed these 85 pairs from the list that have never appeared in the People's Daily corpus in the past 70 years. To solve the third problem, we extracted all the inverse morpheme words from the top 10,000 words and phrases (a total of 356,053) in frequency in the word embedding file, yielding 46 pairs of words, 15 of which did not appear in Huang's word list, accounting for 32.6% of the total:

上海/海上　越南/南越　来到/到来　故事/事故　面前/前面　政党/党政
意愿/愿意　南海/海南　上网/网上　新高/高新　放开/开放　前年/年前
自来/来自　时有/有时　建党/党建

Obviously, “政党/党政”, “新高/高新”, “上网/网上”are more modern. Other words like “意愿/愿意”, “故事/事故” are well-known and have been used for a long time. Therefore, the previous hypothesis that Huang's word list was incomplete is confirmed. We added these 15 pairs of words to the word list to make it more complete. Finally, we provided a word list with 668 pairs of inverse morpheme words, some of which are shown in Table 1 as examples (the order of word 1 and word 2 does not make sense).

Table 1. Inverse morphemes word list.

W1	W2	W1	W2	W1	W2	W1	W2
爱抚	抚爱	伴侣	侣伴	编选	选编	藏躲	躲藏
爱心	心爱	膀臂	臂膀	爱情	情爱	草莽	莽草
鞍马	马鞍	包皮	皮包	谙熟	熟谙	侧翼	翼侧
拔海	海拔	保准	准保	白灰	灰白	查抄	抄查
板鼓	鼓板	报警	警报	摆钟	钟摆	查检	检查
办公	公办	本原	原本	半夜	夜半	查询	询查
扮装	装扮	笔画	画笔	傍依	依傍	产物	物产
邦联	联邦	闭关	关闭	宝珠	珠宝	畅通	通畅
保管	管保	边沿	沿边	暴风	风暴	潮红	红潮
报捷	捷报	爱恋	恋爱	倍加	加倍	尘烟	烟尘

3.2. *Data annotation*

We chose 20 Mandarin native speakers (14 females, 6 males) who were all university students with advanced language and literacy skills. Each subject was asked to rate the semantic similarity of 25 pairs on a scale of 0 to 1, with higher scores indicating greater similarity. Their average value would be used as the evaluation criterion for the cosine

similarity results. The results of manual annotation for 25 pairs of inverse morpheme words are shown in the table below.

Table 2. Results of manual annotation of inverse morpheme words.

W1	W2	manual annotation value
心中	中心	0.14
地基	基地	0.09
动机	机动	0.12
出发	发出	0.19
来历	历来	0.15
工人	人工	0.16
传言	言传	0.26
加强	强加	0.13
明文	文明	0.10
产物	物产	0.28
并吞	吞并	0.59
寻找	找寻	0.74
讲演	演讲	0.59
爱怜	怜爱	0.67
率直	直率	0.76
笔画	画笔	0.13
伴同	同伴	0.24
伴侣	侣伴	0.60
白花	花白	0.18
变形	形变	0.57
称号	号称	0.33
分工	工分	0.14
办公	公办	0.20
低压	压低	0.10
传言	言传	0.26

3.3. *Experimental results*

The semantic similarity of each pair of inverse morpheme words was calculated using Gensim's similarity function, with the minimum value being 0 and the maximum value being 1, and the larger the value, the more similar the group of words is. We divided the data into 10 groups to make it easier to observe, with values of similarity between 0 and 0.1 forming one group, and so on for subsequent counts. The findings are summarized in Figure 1.

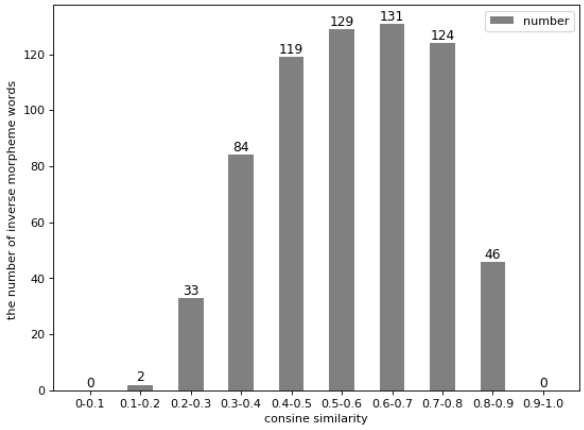

Fig. 1. Semantic similarity of inverse morpheme words with different order of words.

The data has the characteristic of being dense in the middle and sparse on both sides, as can be seen. The majority of word pairs have a similarity of 0.5 to 0.8, with 131 pairs having a similarity of 0.6 to 0.7, 129 pairs having a similarity of 0.5 to 0.6, and 124 pairs having a similarity of 0.7 to 0.8. There are 0 pairs of words with values below 0.1 and above 0.9 in the current word list, indicating that there are almost no completely unrelated pairs and no completely equivalent inverse morpheme words. This indicates that most inverse morpheme words have a high level of semantic similarity, i.e., most of them have similar meanings.

Figure 2 shows the linear relationship between the manually annotated value and the cosine similarity value. The Pearson correlation coefficient is about 0.907 ($p<0.05$), and their goodness-of-fit is 0.82. Therefore, cosine similarity based on word embedding is close to native speaker judgment and can be used as a reference standard for determining the meaning relationship of inverse morpheme words.

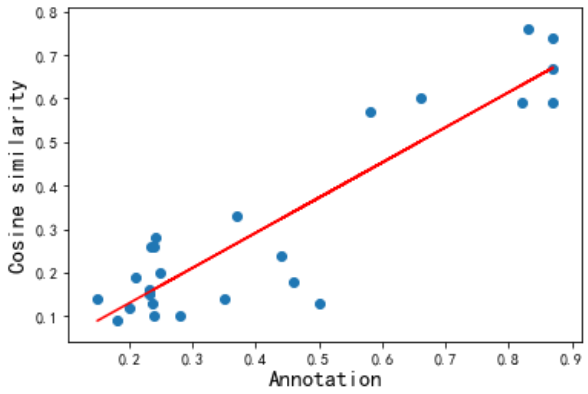

Fig. 2. The relationship between manual annotation and cosine similarity.

4. Discussion

4.1. *Classification of inverse morpheme words*

According to the similarity results, we divided the inverse morpheme words into three categories: total synonyms with the same meaning are words with a score of 0.8 or higher, synonyms with similar meanings have a score of 0.4 to 0.8, and word pairs with different meanings have a score of less than 0.4. The percentages of these three groups are about 7%, 76%, and 17%. Table 3 shows some of the randomly selected words that were annotated after being randomly selected. We found that the average absolute error of 15 pairs of words was 0.16, with only one pair having an error of 0.3 or more, by looking at the absolute value of the error between the cosine similarity and the manual annotation. In terms of the average errors for different types of word pairs, the average errors for the first two synonyms are all 0.18, while the last one is only 0.1, which is much less than the previous two types. As a result, cosine similarity is more accurate for inverse morpheme words that appear to be more semantically different.

Table 3. Inverse morpheme words with three different meaning relationships.

semantic relationship	word pairs		cosine similarity	manual annotation	abs error
same meaning	并吞	吞并	0.87	0.59	0.28
	寻找	找寻	0.87	0.74	0.13
	讲演	演讲	0.82	0.59	0.23
	爱怜	怜爱	0.87	0.67	0.20
	率直	直率	0.83	0.76	0.07
similar meaning	笔画	画笔	0.5	0.13	0.37
	伴同	同伴	0.44	0.24	0.20
	伴侣	侣伴	0.66	0.60	0.06
	白花	花白	0.46	0.18	0.28
	变形	形变	0.58	0.57	0.01
different meaning	称号	号称	0.37	0.33	0.04
	分工	工分	0.35	0.14	0.21
	办公	公办	0.25	0.20	0.05
	低压	压低	0.28	0.10	0.19
	传言	言传	0.24	0.26	0.02

In conclusion, by comparing with the manual annotation values, cosine similarity is a viable method for distinguishing inverse morpheme words, and cosine similarity's referenceability is higher for inverse morpheme words with different meanings.

4.2. *Analysis of factors influencing the semantic similarity of inverse morpheme words*

After annotation, twenty native speakers were asked to vote on the factors that influenced their judgment. Four options were set: meaning, part of speech, frequency of use, and others. Each person cast two votes and the results were 20, 15, 4, and 1. As can be seen, all

participants believed that the meaning of the words influenced their decisions, and 75% believed that part of speech also influenced them. Only a few participants believed that frequency and other factors had an impact on the decision. In the following section, we will look at some of the factors in greater depth.

4.2.1. *Word formation*

Words are made up of one or more morphemes, and the way the morphemes are put together is called word formation[1]. The majority of inverse morpheme words are coordinative compound words[3]. We randomly selected three categories of words classified in Section 4.1 and labeled their word formation. Each category has 10 pairs of words, for a total of 60 inverse morpheme words. The following are the results based on the five different types of Chinese compound word formation: coordinative type, attributive type (modifier and word it modifies), complementing type, predicate-object type, and subject-predicate type are as follows (Table 4).

Table 4. Word formation of inverse morpheme words.

	coordinative	attributive	subject-predicate	complementing	predicate-object
same meaning	20	0	0	0	0
similar meaning	9	8	2	0	1
different meaning	5	12	1	1	1

It demonstrates that different word formations have different influences on semantic relations. The coordinative and the attributive types of words predominate, accounting for 90% of the total words, with the coordinative type being primarily total synonymous and the attributive type being primarily synonymous with similar meanings. However, the number of subject-predicate types, complementing types, and predicate-object types is very small, and they are primarily words with similar or different meanings. In the complementing type, for example, only the word "压低" appears. There are 10 pairs of total synonyms with the same morphemes among different semantic types, indicating that all of them are coordinative types, indicating that morpheme position reversal has no effect on the semantics of these words. We also counted the words whose formation changes as morpheme position changes and found that 5 and 6 pairs of words with similar and different meanings have different formations, respectively. The difference between these two groups is that there are fewer coordinative words and more attributive words as the meaning shifts from similar to different.

4.2.2. *Part of speech*

Part of speech (POS) is a grammatical classification of words in a language that affects the syntactic function of words. As shown in Table 4, we list the 10 groups of inverse morpheme words with the lowest semantic similarity and label their POS (using the lexical markers from the Contemporary Chinese Dictionary (7th edition)). Although the meanings of these words are still somewhat related, they are no longer very similar in the sense of native speakers (see the results of manual annotation in Table 2), and the POS is no longer consistent after the morpheme is reversed. Only 4 out of 10 pairs of words have identical POS.

Table 5. Ten pairs of words with the lowest semantic similarity in inverse morpheme words.

Word1	POS	Word2	POS	cosine similarity
心中	adj	中心	n	0.155
地基	n	基地	n	0.184
动机	n	机动	adj	0.205
出发	v	发出	v	0.215
来历	n	历来	d	0.233
工人	n	人工	adj; n	0.234
传言	n; v	言传	v	0.235
加强	v	强加	v	0.236
明文	n	文明	n; a	0.240
产物	n	物产	n	0.242

We used the CpsWParser[13] to mark the POS of all words in the list and divided them into two groups: those with identical POS in a pair and those with different POS in a pair, to see if the effect of POS on similarity is significant (if a word has multiple classes and another word has only one of them, it is also counted as a different POS). The results show that there are 459 pairs of inverse morpheme words with identical lexicality, with a mean value of 0.593, and 209 pairs with different POS, with a mean value of 0.517. The ANOVA results reveal that there is a significant difference between the two sets of data, with $p=1.0439E-08$ ($p<0.01$). It demonstrates that POS has a significant impact on the semantic similarity of inverse morpheme words and that POS changes can cause a pair of inverse morpheme words to become more dissimilar, which is also consistent with the rules of language use, in which words with different POS may have different positions and syntax functions in the sentence. Consider the following example:

这是你大姑的扇子，那是你三姑的花鞋……都有了**来历** 。（萧红《呼兰河传》）

历来不立字据，全凭口头协议（霍达《穆斯林的葬礼》）

In the group of words “来历/历来”, “来历” is a noun that is primarily used as the subject and object in the sentence, and is the object of the verb “有” (have) in the above example; “历来” is an adverb that is primarily used as the gerund in the sentence. The definition of “来历” in the dictionary is “the history or background of a famous person or thing,” while the definition of “历来” is “historically” or “always”. Although they are related in meaning to some extent, they cannot be considered synonymous.

4.2.3. *Word frequency*

Word frequency may also influence the semantic similarity of inverse morpheme words, according to four native speakers who believe that word frequency influences their judgment. We calculated the proportion of different similarity values for all words in the list and used the same method for the 46 pairs of high-frequency (HF) pairs extracted in Section 3.1, as shown in Figure 3.

The two groups show different characteristics. The HF group has a peak between 0.3 and 0.4, whereas the full group has the most values between 0.5 and 0.6. Additionally, the HF group has a higher proportion of words in the low similarity interval, while the full group has a much higher proportion of words in the high similarity interval. Therefore, we believe that word frequency has an impact on semantic similarity. The more commonly a pair of inverse morpheme words are used, the more likely people are to separate them, and the further apart they are in vector space, the less similar they will be.

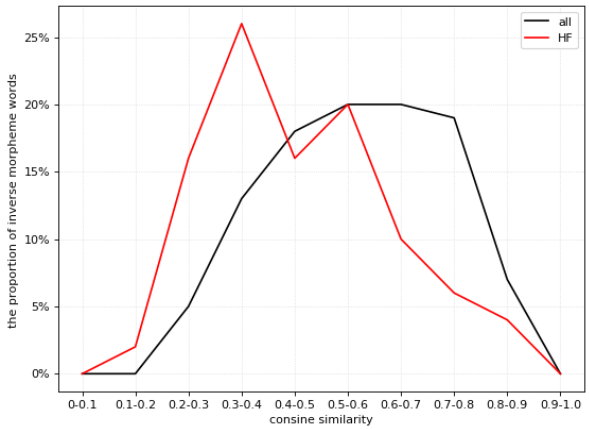

Fig. 3: Effect of frequency on cosine similarity of inverse morpheme words.

5. Conclusion

This study set out to investigate whether a word embedding-based approach to calculating word similarity is feasible and whether the results can be judged on par with native speakers. We concluded that this method is valid after analyzing the experimental data and calculating Pearson coefficients. The second goal of this study was to see which factors have a significant impact on similarity as determined by the semantic similarity calculation. As a result, we found that word formation, part-of-speech, and frequency are all important.

There is probably a lot more that can be done to improve the word embedding model and the method for calculating semantic similarity. The People's Daily corpus was used to train the model. More research should be done to compare models trained on various corpus. The cosine similarity is not the only way to calculate the distance between two vectors. As

the abs error shows, it is not very good at calculating words with the same meaning. As a result, investigation and experimentation should be carried out to determine which method is the most effective. Furthermore, the amount of manually annotated data in this paper is limited, and we will need to add more in the future.

Acknowledgements

This work is supported by the Ministry of Education Humanities and Social Science Planning Project (No. 18YJAZH112) and the Research on the corpus construction and intelligent application of graded reading for Chinese International Education (No. ZDI135-41). Zhiying Liu is the corresponding author.

References

1. Jerome L. Packard, *New Approaches to Chinese Word Formation: Morphology, Phonology and the Lexicon in Modern and Ancient Chinese* (De Gruyter, Berlin, 2011), p. 35.
2. J. Bo, *Study on Inverse Morpheme Words. Journal of Tianjin Normal University (Social Science Edition)* **6** (1996) 70-73.
3. L. Hong. *A Study of Inverse Morphemes Words in Ancient Chinese. Journal of Shenyang Normal College (Social Science Edition)* **21** (1997) 46-48.
4. W. Cao, *Studies in Modern Chinese Vocabulary (Revised Edition)* (Jinan University Press, Guangzhou, 2010).
5. Y. Zhang, *The Two-syllable Words in Modern Chinese with the Opposite Character Order. Studies of the Chinese Language* **3** (1980) 177-183.
6. Z. Tang, *The synchronic state of contemporary Chinese words and their transmutation* (Fudan University Press, Shanghai, 2001).
7. L. Huang, *A Study of Disyllabic Words of the Same Morpheme and Formation of Reverted Order in Modern Chinese* (Huazhong University of Science and Technology, Wuhan, 2006).
8. Y. Bengio, R. Ducharme, P. Vincent, & C. Janvin, *A Neural Probabilistic Language Model. Journal of Machine Learning Research* **3** (2003) 1137–1155.
9. M. Chen, *Introduction to Cognitive Computing* (Huazhong University of Science and Technology Press, Wuhan, 2017).
10. T. Mikolov, G. Corrado, K. Chen, & J. Dean, *Efficient estimation of word representations in vector space*, in *Proc. Workshop at ICLR*, Scottsdale, 2013, pp. 1-12.
11. S. Li, Z. Zhao, R. Hu, W. Li, T. Liu, X. Du, *Analogical reasoning on Chinese morphological and semantic relations*, in *Proc. The 56th Annual Meeting of the Association for Computational Linguistics*, Vol 2, Melbourne, 2018, pp. 138-143.
12. P.-N. Tan, M. *Steinbach & V. Kumar, Introduction to Data Mining* (Posts & Telecom Press, Beijing, 2011).
13. H. Xiao, *Study on word annotation of corpus* (2016). http://corpus.zhonghuayuwen.org/CpsWParser.aspx.