The Extraction of Emotional and Attribute Words in Perspective Sentences

Naijun Cao¹, Zhengkui Lin¹, Hong Yang¹, Huancheng Zhang¹ and Zenghui Qiu¹

¹School of Computer Science and Technology, Dalian Maritime University

m18742521258_1@163.com

Abstract: The article focuses on the research on the classification criteria of emotional words and attribute words. Firstly, the semantic similarity and its representative method based on distance, the Euclidean distance calculation method and the cosine distance calculation method are introduced. Based on this, a comparison experiment of semantic similarity between emotional words and attribute words based on Euclidean distance and cosine distance was carried out. According to the experimental results, a method for determining the classification criteria of emotional words and attribute words based on the distribution rules of words and cosine distances was proposed.

1. Introduction
In recent years, online shopping reviews have become one of the most important sources of information for online consumer evaluation [1]. The consumer sentiment of the online shopping review information on the purchased goods has an important influence on the purchase decision of subsequent potential consumers, and the online shopping commentary in the online shopping commentary extracts the emotional words and attribute words in the online shopping review. The key to the premise and foundation of the sentiment analysis is to select the appropriate classification feature model, in other words, the classification criteria, to support the extraction of emotional words and attribute words. This dissertation focuses on the semantic similarity calculation method of emotional words and attribute words and the related research on the classification methods of emotional words and attribute words based on the distribution law of words and cosine distances.

2. Distance-based semantic similarity calculation method

2.1. Word distance and semantic similarity
Dekang Lin thinks that the similarity of any two words depends on their Commonality and Differences, and then from the perspective of information theory, it gives the definition formula of semantic similarity:

$$\text{Sim}(A, B) = \frac{\log p(\text{common}(A, B))}{\log p(\text{description}(A, B))}$$ (1)

Among them, the numerator represents the amount of information required to describe the commonality of A and B; the denominator represents the amount of information required to fully describe A and B.
At present, there are two main methods for calculating the semantic similarity of words: one is to use the statistical information method to solve the word context information, and the other is to calculate the word similarity based on distance. Simple correspondence, and then through the semantic dictionary, the concept of related words organized in a tree-like structure, by computing the distance between nodes in the tree to solve the semantic similarity between words.

\[
\text{Sim}(w_1, w_2) = \frac{k}{\text{Dis}(w_1, w_2)} \quad (2)
\]

Among them, \(\text{Dis}(W1, W2)\) is the distance of the tree represented by \(W1, W2\), and \(k\) is the proportional coefficient.

In distance-based word similarity calculation methods, common distance calculation methods mainly include Euclidean distance, cosine distance, Manhattan distance, and the like. This paper focuses on the performance comparison of semantic similarity between Euclidean distance and cosine distance calculations.

2.2. Euclidean distance
Euclidean distance, or euclidean metric, is a widely used definition of distance, which refers to the true distance between two points in \(m\)-dimensional space, or the natural length of a vector (that is, the point to the origin Distance) \(\textnormal{[2]}\), calculated as follows:

\[
\text{distance}(X, Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (3)
\]

Among them, \(X\) and \(Y\) are \(n\)-dimensional vectors, \(X_i\) is the \(i\)-th component of \(X\), and \(Y_i\) is the \(i\)-th component of \(Y\).

2.3. Cosine distance
Cosine distance is also called cosine similarity. The cosine distance is calculated as follows\([3]\)[4]

\[
\text{sim}(X, Y) = \frac{\sum_{i=1}^{n}(X_iY_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2} \sqrt{\sum_{i=1}^{n}(Y_i)^2}} \quad (4)
\]

Among them, \(X\) and \(Y\) are \(n\)-dimensional vectors, \(X_i\) is the \(i\)-th component of \(X\), and \(Y_i\) is the \(i\)-th component of \(Y\).

The cosine distance represents the angular difference between the two vectors and is applied to the word vector. Since the word vector has semantic meaning, the cosine distance between two word vectors can be used to represent the difference between the semantics expressed by the two word vectors.

3. Experiment Based on Distance Emotional Words and Attribute Words

3.1. Experimental Design
Both Euclidean distance and cosine distance can be used to calculate the semantic similarity between words.\([5]\) In order to obtain better classification of affective words and attribute words and classification criteria of emotional words and attribute words, this paper develops Euclidean distances and cosine distances. The similarity degree calculation of the emotional words and attribute words is compared with the experiment.

The experiment is based on the previous viewpoints. Each word and its corresponding word vector in the viewpoint are used as the basic corpora;\([6]\) then the basic corpora are paired with the word pairs, and after removing the duplicates, formula 1.1 and formula 2 are used respectively. 1.2 Calculate the Euclidean distance and cosine distance between pairs of words to obtain experimental corpus; Finally, select several representative pairs of words from experimental corpus and determine distance-based classification criteria of emotional words and attribute words.\([7]\)[8][9]The properties of the manually labeled word pairs and their corresponding Euclidean distance and cosine distance calculation results are shown in Table 1 and Table 2 respectively.
Table 1. Examples of Euclidean distance

| Word pair            | Word pair nature                | Euclidean distance |
|----------------------|---------------------------------|-------------------|
| System - Bluetooth   | Attribute word - attribute word | 3.545893          |
| Function - Signal    | Attribute word - attribute word | 4.362586          |
| Cache - Satisfaction | Attribute word – non attribute  | 3.250933          |
| Good-big             | Emotional Words - Emotional     | 3.385302          |
|                      | Words                           |                   |
|                      | ......                           | ......             |
| Good- appearance     | Emotional Words – non           | 3.351089          |
|                      | Emotional Words                 |                   |

Table 2. Examples of Cosine distance

| Word pair              | Word pair nature                | Cosine distance  |
|------------------------|---------------------------------|------------------|
| Signal - Touch Screen  | Attribute word - attribute word | 0.3550653294881015 |
| Screen - Battery       | Attribute word - attribute word | -0.0213554490291685 |
| Touch screen - old man | Attribute word – non attribute  | 0.5220542183771003 |
| Beautiful - beautiful  | Emotional Words - Emotional     | 0.7212404951757746 |
|                       | Words                           |                  |
|                       | ......                           | ......            |
| Decision - Touch Screen| Emotional Words – non           | 0.6637506361054051 |
|                       | Emotional Words                 |                  |

3.2. Experimental analysis of distribution rules of emotional words and attribute words based on Euclidean distance

(1) Experimental process and experimental results

The maximum Euclidean distance of all word pairs in the experiment corpora obtained according to the experimental scheme is 16.096646, which is a study of attribute word-attribute word pair, attribute word-non-attribute word pair, emotional word-emotion word pair and emotional word-non-emotion word pair. The distribution of the law, first of all, the Euclidean distance in experimental corpora is divided into [0.0, 1.6), [1.6, 3.2), [3.2, 4.8), [4.8, 6.4), [6.4, 8.0), [8.0, 9.6], [9.6, 11.2), [11.2, 12.8), [12.8, 14.4], [14.4, 16.096646] 10 intervals, and then statistically count the number of attribute word-attribute word pairs and attribute words in 10 intervals. The number of non-attribute word pairs, the number of emotional word-emotion word pairs, the number of emotional word-non-emotion word pairs. For convenience of statistics, the symbol "a" represents the attribute word-attribute word pair, and the symbol "b" On behalf of attribute words - non-attribute word pairs, symbol "c" represents emotional word-emotion word pairs, symbol “d” represents emotional word-non-emotional word pairs, and finally, based on statistical results, the number of pairs of words plotted Summarize its distribution pattern. The algorithm described above is described in Table 3 below.

Table 3. Euclidean distance segmentation is used to calculate the attribute statistics

Process: For segmented statistics for word pairs that have been calculated for Euclidean distance, the formal representation of this algorithm takes one of the intervals as an example.

Symbol Description:

Dic_0_0_1_6_a: Euclidean distance in the range [0.0, 1.6) and the word pair property is the statistical result of the attribute word-attribute word
Dic_0_0_1_6_b: The Euclidean distance is in the range [0.0, 1.6) and the word pair property is the statistical result of attribute word-non-attribute word

Dic_0_0_1_6_c: Euclidean distance in the range [0.0, 1.6) and the word pair property is the statistical result of the attribute word-attribute word

Dic_0_0_1_6_d: Euclidean distance in the range [0.0, 1.6) and the word pair property is the statistical result of the attribute word-non-attribute word

a : The word pair property is attribute word-attribute word; b : The word pair property is attribute word-non-attribute word

c : The word pair nature is emotional word-emotion word; d : The word pair nature is emotional word-non-emotion word

Vector_dis : experimental corpora after artificially marking word pair properties and calculating Euclidean distance

Results : statistical results of the nature of various word pairs in 10 intervals

Process:
Input → vector_dis
FOR line IN vector_dis
    Get the word pair for each line and write it as word-word
    Get the word pair attribute in each row, and write it as words-mark
    Get the corresponding Euclidean distance for word pairs in each row, written as word-distance
    IF words-mark == a and word-distance ∈ [0.0, 1.6) THEN
        Word-word → dic_0_0_1_6_a
    ELSE IF words-mark == b and word-distance ∈ [0.0, 1.6) THEN
        Word-word → dic_0_0_1_6_b
    ELSE IF words-mark == c and word-distance ∈ [0.0, 1.6) THEN
        Word-word → dic_0_0_1_6_c
    ELSE IF words-mark == c and word-distance ∈ [0.0, 1.6) THEN
        Word-word → dic_0_0_1_6_d
    END IF
END FOR
Output → results

According to the above algorithm, the number of word pairs of different word pair properties in 10 Euclidean distance intervals is counted. The results are shown in Table 4 below:
Table 4. Euclidean distance distribution data

| Euclidean distance range | Attribute Word - Attribute Word Pair Number | Attribute Word - Non-attribute Word Pair Number | Emotional Words - Emotional Word Pairs | Emotional Words - Non-Emotional Word Pairs |
|-------------------------|---------------------------------------------|-----------------------------------------------|--------------------------------------|------------------------------------------|
| [0.0 , 1.6)             | 2080                                        | 88                                            | 1174                                 | 840                                      |
| [1.6 , 3.2)             | 1230                                        | 943                                           | 902                                  | 160                                      |
| [3.2 , 4.8)             | 860                                         | 490                                           | 1730                                 | 172                                      |
| [4.8 , 6.4)             | 22                                          | 34                                            | 27                                   | 1530                                     |
| [6.4 , 8.0)             | 990                                         | 1860                                          | 780                                  | 2315                                     |
| [8.0 , 9.6)             | 103                                         | 2127                                          | 166                                  | 1025                                     |
| [9.6 , 11.2)            | 157                                         | 1000                                          | 157                                  | 34                                       |
| [11.2 , 12.8)           | 1200                                        | 161                                           | 132                                  | 193                                      |
| [12.8 , 14.4)           | 109                                         | 136                                           | 1900                                 | 500                                      |
| [14.4 , 16.0]           | 400                                         | 300                                           | 400                                  | 300                                      |

(2) Analysis of results

On the one hand, since the range of Euclidean distances is \([0, +\infty)\), the maximum Euclidean distance of experimental corpora obtained at present is only 16.096646, which will lead to incompleteness of the distribution rules. With the continuous expansion, although the maximum value of the Euclidean distance may change, the overall trend of the distribution may also change, but the analysis of the distribution of the nature of word pairs will not have much impact.

On the other hand, according to statistical data, for the attribute word-attribute word pair and emotional word-emotion word pair, although there is a gap in the overall distribution, the difference is not significant, that is, different pairs of words are different. Since the number of Euclidean distances is equivalent, it is not possible to use a range of Euclidean distances as a classification criterion (threshold) for distinguishing attribute words, non-attribute words, emotional words, and non-emotion words.

3.3. Experimental Analysis of the Distribution Rules of Emotional Words and Attribute Words Based on Cosine Distance

(1) Experimental process and experimental results

The experimental process is similar to 3.2. Firstly, the cosine distances in the experiment corpora are divided into \([-1.0, -0.8), [-0.8, -0.6), [-0.6, -0.4), [-0.4, -0.2), [-0.2, 0.0), [0.0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), [0.8, 1.0]\) 10 intervals; then statistics on 10 intervals Word-attribute word pair number, attribute word-non-attribute word pair number, emotional word-emotion word pair number and emotional word-non-emotion word pair number, for convenient statistics, also note the symbol “a” for “Proxy attribute word-attribute word pair, symbol “b” for attribute word-non-attribute word pair, symbol “c” for emotional word-emotion word pair, and symbol “d” for emotional word-non-emotion word pair. Finally, according to the statistical results, draw the distribution of word pairs and summarize their distribution rules. According to the above algorithm, the number of word pairs of different word pair properties in 10 cosine distance intervals is counted.

According to the algorithm, the number of word pairs of different word pair properties in 10 cosine distance intervals is counted.

(2) Analysis of experimental results

On one hand, according to the cosine characteristics, since the range of the cosine distance is \([-1.0, 1.0]\), different from the Euclidean distance, the distribution rule reflected by the experimental results will not only occur because of the change in the number of experimental corpora. The fundamental changes, on the contrary, with the continuous expansion of the scale of experimental corpus, the above-mentioned distribution characteristics will become stronger and stronger;
On the other hand, according to the statistical results in Table 4, the following rules can be found: The distribution law of cosine distances for word pairs of different natures is used as a classification basis for emotional words and attribute words.

4. Classification criteria of emotional words and attribute words based on word-to-cosine distance distribution

Based on the different comparison of emotional words and analysis of the words, the distance can be used to calculate the words to be classified. Knowing the semantic similarity of emotional words and attribute words, and comparing and analyzing the distribution characteristics of cosine distances of different nature words, the classification criteria of emotional words and attribute words are determined.

As this: the process of determining the classification criteria for emotional words and attribute words can be described as follows:

1. Establish basic attribute word lexicons, emotional word lexicons, and other word lexicons (that is words that are neither attributed nor sentimental);
2. For words of unknown nature, semantic similarity is calculated with the words in the dictionary of attribute words. If the value of semantic similarity is within the threshold of the attribute word-attribute word pair ([-1.0, -0.4]), When the attribute word dictionary is 80% of the number of words, it is determined that the keyword belongs to the attribute word, otherwise the semantic similarity is calculated with the selected emotional word, if the value of the semantic similarity is in the threshold of the emotional word-emotion word pair ([0.6, 1.0]), it is determined that the word is an emotional word, otherwise it is other words;
3. After the selection of emotional words, attribute words, and other words, the basic emotional word dictionary, basic attribute word dictionary, and basic other word dictionary are expanded to continuously improve the accuracy of the classification criteria. In terms of the extraction of attribute words and sentiment words, this paper proposes a semantic similarity calculation based on experimental corpora, and empirically researches the classification rules of semantic similarity based on Euclidean distance, cosine distance, emotional words, and attribute words. The emotional word-attribute word extraction method is based on corpus rules and matches emotional words and attribute words. For the emotional words that do not match the attribute words, the design algorithm extracts the implicit attributes of the words and compares them with the manually labeled results. The rate reached about 75%. However, because the subject involves a wide range of knowledge, coupled with my limited time and energy, and the following aspects can be further studied and carried out: cross-domain word contribution rate. For the value of the contribution of words to sentence semantics, the training corpus of this article is completed by manual annotation. However, the subjectivity of this part of people is relatively large. At present, there are no certain norms and standards for this aspect, resulting in possible final experimental results. Cannot be applied to all fields of language, for the word to sentence semantics contribution rate across the domain to achieve is the next major research content

5. Summary and Prospect

In terms of the extraction of attribute words and sentiment words, this paper proposes a semantic similarity calculation based on experimental corpora, and empirically researches the classification rules of semantic similarity based on Euclidean distance, cosine distance, emotional words, and attribute words. The emotional word-attribute word extraction method is based on corpus rules and matches emotional words and attribute words. For the emotional words that do not match the attribute words, the design algorithm extracts the implicit attributes of the words and compares them with the manually labeled results. The rate reached about 75%.

However, because the subject involves a wide range of knowledge, coupled with my limited time and energy, and the following aspects can be further studied and carried out: cross-domain word contribution rate. For the value of the contribution of words to sentence semantics, the training corpus
of this article is completed by manual annotation. However, the subjectivity of this part of people is relatively large. At present, there are no certain norms and standards for this aspect, resulting in possible final experimental results. Cannot be applied to all fields of language, for the word to sentence semantics contribution rate across the domain to achieve is the next major research content.

References
[1] https://baike.baidu.com/item/%E6%AC%A7%E5%87%A0%E9%87%8C%E5%BE%97%E5%BA%A6%E9%87%8F/1274107?fr=aladdin&fromid=2809635&fromtitle=%E6%AC%A7%E5%BC%8F%E8%B7%9D%E7%A6%BB.
[2] https://baike.baidu.com/item/%E4%BD%99%E5%BC%A6%E7%9B%B8%E4%BC%BC/17509249?fr=aladdin.
[3] Xiaobo Tang, Guangchao Liang. Summary of Research on Fine-grained Sentiment Analysis [J]. Joining of CI Books. 2017, 61(5):132-140.
[4] Turney P D, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems (TOIS). 2003, 21(4):315 − 346.
[5] Zhichun W, Juanzi Li, Zhigang W, et al. Cross-lingual knowledge linking across wiki knowledge bases [C]. In Proceedings of the 21st International Conference on World Wide Web, 2012: 459-469.
[6] Ku L W, Wu T H, Lee L Y, et al. Construction of an evaluation corpus for opinion extraction[C]. Ntcir. 2005: 513-520.
[7] Pang B, Lee L, Vaithyanathan S. Thumbs up: Sentiment classification using machine learning techniques[C]. In: Proceedings of Association for Computational Linguistics Conference on Empirical Methods in Natural Language Processing. ACL’02, 2002: 79 − 86.
[8] Sajib S, Ng V. Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification[C]. In: Meeting of the Association for Computational Linguistics. 2009: 701-709.
[9] Huancheng Zhang. Online shopping review information Research on Fine-grained Sentiment Analysis Method [D]. Dalian Maritime University, 2016.