ATLANTIS PRESS
Advances in Engineering Research (AER), volume 61
International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2017)

Semantic Knowledge Acquisition based on Maximum Entropy

Maoyuan Zhang [a], Kai Xing [b] and Jianping Zhu [c]
School of Computer, Central China Normal University, Wuhan 430079, China
[a]zhangmy@mail.ccnu.edu.cn, [b]xingkai@ mails.ccnu.edu.cn, [c]zjp201@163.com

Keywords: Semantic Knowledge; Maximum Entropy; Semantic Distance.

Abstract. It's necessary to acquire semantic knowledge in Natural Language Processing research. In this paper, we present an approach for acquiring Chinese semantic knowledge based on maximum entropy model. Semantic knowledge units are composed of central word and a group of feature words. Because the maximum entropy to extract features, we utilize it to calculate the semantic distance between the central word and feature words in large-scale network corpus. In the experiment, tests on a number of manual defined data sets show that the Spearman correlation coefficient has been increased 6.2%-20.9%.

1. Introduction

Research on semantic knowledge between words is basic research of field about Natural Language Processing. When people understand the meaning of a word, we often associate the term with other terms, our mind establishment the relations between words to help us to understand the meaning [1]. According to this idea, this paper treats the word which we hope to acquire semantic knowledge as the central word, treat words that calculated the maximum degree of association with the central word as feature words. Feature words are helpful to the understanding of central words for computer. There are two sources of Chinese semantic knowledge[2], one is the artificial construction of knowledge base, such as Hownet, another is a large-scale real texts, including the massive text on the Internet, a variety of off-line text collections (such as various sizes corpus), various online encyclopedia (such as Wikipedia, Baidu Encyclopedia, Interactive Encyclopedia).

In this paper, the method of semantic knowledge acquisition is similar to the method of word sense similarity, so the method of this paper is compared with the method of calculating semantic similarity. Different from the similarity of word meaning, the related degree of the word reflects the other words that are associated with the word, the similarity of meaning about related words may be small, but related words can reflect the larger meaning of the word.

The remainder of this paper is organized as follows. Section 2 presents our method specifically. Section 3 describes the experiment. And finally this paper is concluded in Section 4

2. Acquisition of Semantic Knowledge

To acquire the semantic knowledge of a word, we need to get the feature words of the center word and the feature weights of each feature word, feature weights are used to describe the relationship between the feature words and the center word [3]. Therefore, first of all, we need to get the feature words to be used as the feature word group, and then the weights of the feature words are obtained to be used as the relationship between the feature words and the center word.

2.1 Acquisition of Feature Words

In this paper, we use the data collected from web news, which is downloaded from the network media monitoring center in 2015 and the number of news is 183747. We segment all of the text automatically by using the word segmentation software ICTCLAS, do preprocessing for all of the data, such as stopping word. The acquisition of feature words is based on the related words, which may appear in the context of the target words. In this paper, the context window size is 3; get the context words in the target word. Then calculate the PMI of the target word w_i and the context word c_j:

$$PMI\left(w_i, c_j\right)=\frac{p\left(w_i, c_j\right)}{p\left(w_i\right) * p\left(c_j\right)}=\frac{tf\left(w_i, c_j\right)T}{count\left(w_i\right) * count\left(c_j\right)} \quad (1)$$

Here, T is the total number of words in the corpus, count(x) is the frequency of x appearing in the corpus, Setting the threshold value is k, when PMI (w_i, c_j) < k, we filter out c_j. Get all the words in the context window of the center word, calculating the PMI of these words and the central word, filtering out the word which the PMI is less than the threshold value, the rest of the word as a feature word. Then, we need to train the feature words to get the weight of the feature words.

2.2 The Weight of Feature Words

The basic idea is to establish the maximum entropy method consistent with the known facts (training data) model, do not make any assumptions about the unknown factors - keep the distribution as evenly as possible. The maximum advantage of constructing the maximum entropy method is to select the most objective weight under the condition of limited information.Input model: according to feature words group which we obtain, take them as a training sample set D= $\{(x_1,y_1), (x_2,y_2)\ldots(x_n,y_n)\},\{(x_i,y_i)\}$, the context information is x_i when the y_i is represented in the corpus. The empirical probability distribution is obtained from the training examples:

$$\tilde{p}(x,y)=\frac{1}{N} * Count(x, y) \quad (2)$$

Here, Count(x,y) is the number of times appear in the corpus, N is the total number of words.Feature f refers to a specific relationship between x and y, which is expressed by two-valued function f(x,y).When PMI>I , I is the threshold value, f(x,y)=1.Feature empirical probability expectation:

$$\tilde{p}(f) = \sum_{x,y} \tilde{p}(x,y)f(x,y) \quad (3)$$

The expected probability of features is the true distribution of the features in the learning of random events:

$$p(f) = \sum_{x,y} \tilde{p}(x)p(x,y)f(x,y) \quad (4)$$

Here, $\tilde{p}(y|x)$ is the empirical probability of y, when x appears; $p(y|x)$ is the true probability of y, when x appears. The empirical probability of feature is consistent with the expected probability: $\tilde{p}(f) = p(f)$, it constitute a constraint equation.We assume that there are k features: $f_i(i = 1,2, \ldots, k)$, Multiple constraint equation expressed as a constraint set: $C=\{s \in S|p(f)=\tilde{p}(f)\}, \ i \in \{i=1,.., k\}$.

According to the conditions of distributed consensus model, we can get the best values of $p(y|x)$. At the same time, we can determine the best measure of the standard. Here we take optimal conditional entropy, record as:

$$M(s)= -\sum_{x,y} \tilde{p}(x)p(y|x)\log p(y|x) \quad (5)$$

The maximum entropy model is expressed as: we choose the maximum entropy in the constraint condition. This is a problem of constrained optimization: $S^*=\text{argmax}M(s)$ Here, S is a statistical model which satisfies the constraint set C. The argmax represents a parameter that finds the maximum score. We use the Lagrange multiplier method to solve this optimization problem. We introducing a parameter λ_i (Lagrange multipliers) for each feature f_i. The λ_i indicates the importance of the feature f_i for the model. In addition, because of the conditional probability $p(y|x)$, $\sum_y p(y \mid x)=1$, we introduce a parameter T(x) for each instance x. Maximum entropy of Lagrange's function expressed as:

$$\Lambda(s, \lambda)=M(s)+\sum_i \lambda_i\left(p\left(f_i\right)-\tilde{p}\left(f_i\right)\right)+\sum_x T(x)\left(\sum_y p(y|x)-1\right) \quad (6)$$

Then we get the derivation of it, we can get the $\sum_y p(y \mid x)$ when $\Lambda(s, \lambda)$ is maximum. The weight of feature f_i is expressed by the corresponding parameter λ_i. The p(y|x) with the maximum entropy condition is expressed as the exponential form:

$$\sum_y p(y \mid x)=V_\lambda(x) \exp\left(\sum \lambda_i f_i(x, y)\right) \quad (7)$$

Here, $V_\lambda(x) = 1 / \sum_y \exp \sum \lambda_i f_i(x, y)$ represents a normalization factor. In this way, we transform a constrained optimization problem into an unconstrained optimization problem. Using the maximum entropy model, we can get the feature word weight after training.

3. Experiment and Results

In this chapter, we need to determine the correctness of semantic knowledge acquisition through experiments. The criteria for evaluating semantic relationships are often manually labeled data sets [4], but there is no standard test set for computing the semantic correlation of Chinese words. In the experiment, on the one hand, we manually translated WordSimilarity-353 test set. The data set consists of 353 pairs of English words; the similarity value of these words is obtained by artificial judgment. It can train and test the method for semantic similarity. Because WordSimilarity-353 is a test set in English, we translate it into Chinese. On the other hand, this paper uses the test set Words-240 which is constructed by 613 Department of National University of Defense Technology. The test set includes 240 pairs of Chinese words, the creation of the test set is guided by Lev Finkelstein which the founder of WordSimilarity-353. The research on using this test set published related results in the Journal of Chinese Computer Systems. [5]

In this paper, we use WordSimilarity-353 and Words-240 to do the test. We compare to the results of manual annotation, and then we use Spearman and Pearson correlation coefficient as a measure of the results of the algorithm. We calculate three semantic similarity methods respectively. The three methods are computational method of lexical semantic similarity based on Hownet, and computational method of lexical semantic similarity based on Chinese Wikipedia, and the method in this paper. We select calculation results randomly. The similarity calculation results are in the range of 0-1, 0 representatives are not related, the more close to 1 point, the greater the degree of similarity.

Table 1. Relatedness of some word pairs in dataset WordSimilarity-353

English word pairs	standard value	Hownet	Wikipedia	Maximum Entropy
money-cash	0.908	1	0.633	0.842
psychology- Freud	0.821	0	0.86	0.63
cup-coffee	0.658	0.165	0.371	0.594
cup- article	0.24	0.087	0.13	0.314

Table 2. Relatedness of some word pairs in dataset Words-240

Chinese word pairs	standard value	Hownet	Wikipedia	Maximum Entropy
李白(lipo)-诗(poetry)	0.92	0	0.763	0.81
计算机(computer)-软件(software)	0.771	0.44	0.524	0.873
电影(movie)-爆米花(popcorn)	0.765	0	0.891	0.644
老虎(tiger)-猫(cat)	0.33	0.44	0.517	0.16

By analysis, there are some unknown words about similarity calculation method based on Hownet. For some new words and some of the less commonly used words can't calculate the similarity. There are some results which cannot reflect the semantic relations between words based on Chinese Wikipedia. Next, we compare the result of three ways with the results of manual marking on Spearman correlation coefficient and Pearson correlation coefficient.

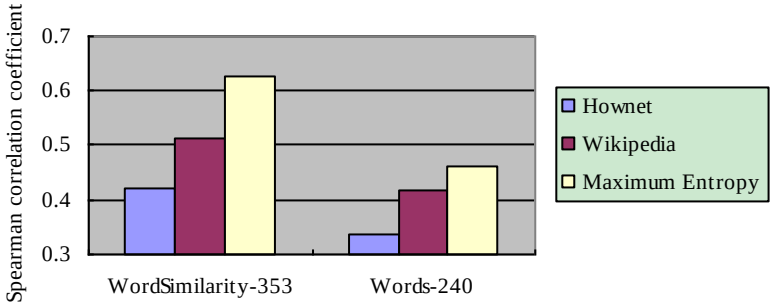

Fig. 1 Comparison of Spearman correlation coefficient

In our experiment, the Spearman correlation coefficient of Hownet, Wikipedia and Maximum Entropy is 0.402, 0.5126 and 0.6281 for test set WordSimilarity-353; for test set Words-240, the value of three methods is 0.3351, 0.416 and 0.4617.Compared with other two methods, our method improves precision of other methods by about 6.2%-20.9% for Spearman correlation coefficient. The Pearson correlation coefficient of Hownet, Wikipedia and Maximum Entropy is 0.5224, 0.5931 and 0.6317 for test set WordSimilarity-353; for test set Words-240, the value of three methods is 0.4194, 0.4832 and 0.513.Compared with other two methods, our method improves precision of other methods by about 6.17%-22.3% for Pearson correlation coefficient. The figure above demonstrates that the method we proposed shows a better result.

4. Conclusion

This paper proposes a method of acquiring semantic knowledge based on maximum entropy model. Firstly, we extract the center of the word in the context of the words in a massive corpus, and then we compute the PMI of the central word. Central words and characteristic words constitute the semantic units that we need. In order to test whether the result is correct, we make a comparison with the other methods of semantic similarity. The result is obviously better than the traditional method. For future work, we apply the semantic knowledge to the event tracking, and find the relation of events, and then forecast the development of the event.

Acknowledgements

This work was supported by Humanity and Social Science Youth Foundation of Ministry of Education of China (No. 15YJC870029), the self-determined research funds of CCNU from the colleges' basic research and operation of MOE (No. CCNU16A02049, No. CCNU16A06039)

References

[1]. Liu Yang, Tingting He, Xinhui Tu. The Chinese semantic knowledge acquisition based on Web Encyclopedia [C]// The Fifth National Conference on Computational Linguistics for young people (YWCL 2010). 2010.
[2]. Linhong Wang, Qiang Nv, Rui Xu. Research on Chinese semantic correlation computing model [J]. Computer Engineering and Applications, 2009, 45(7):167-170.
[3]. Jing Shi, Fangyun Wu, Kunli Qiu. A method for calculating semantic similarity of Chinese words based on large scale corpus [J]. Journal of Chinese Information Processing, 2013, 27(1):1-6.
[4]. Zhijian Zhan, Nali Liang, Pinxiao Yang. Word similarity calculation based on Baidu Encyclopedia [J]. Computer Science, 2013, 40(6):199-202.
[5]. Tian Xia. Research on semantic similarity computation of Chinese words [J]. Computer Engineering, 2007, 33(6):191-194.