Predicting the Semantic Related words based on Hidden Markov Model

Fuping Yang[1, a], Huafeng Gu[2, b]

[1,2] College of Computer Science and Technology

Chongqing University of Posts and Telecommunications Chongqing 400065, China;

[a] yangfp@cqupt.edu.cn, [b]guhuafeng1991@163.com

Keywords: Hidden Markov Model; Semantic Relatedness; Natural Language Processing

Abstract. This paper presents a method of predicting the words with semantic relation based on Hidden Markov Model (HMM). Two words are set as an observation sequence, combined with HMM and the corpus, which has taken some works in Natural Language Processing, to calculate the most probable sequence with semantic relation by the given observation sequence. By Reducing the impact of high frequency words on the traditional method of semantic prediction based on the Text-window Co-occurrence. The experiment results show that this method is effective.

1. Introduction

In natural language, lexical semantics can be represented by its associated context. We can make statistics on the corpus, using co-occurrence frequency to obtain the semantic relatedness between the two words [1], and then to predict the word that has semantic relation with two given words. The semantic relatedness is usually calculated by two methods: one is Text-window Co-occurrence; the other is based on the Syntactic Dependency.

Text-windows Co-occurrence is a statistical method to calculate the frequency of vocabulary in the X-window. The X- window is a vocabulary sequence which has X words before and after the content word [2, 3, 4]. Syntactic Dependency is the way to compute the semantic relatedness by using the rules of grammar [5]. Due to the complexity of natural language, summarizing the rules of grammar by human has been unable to meet the requirements. So we usually use the statistical methods to calculate the correlation degree of words or to predicate the words that have relation in semantics.

When anlysing the method of Text-window Co- occurrence, we found that the traditional method ignored the semantic relationship between sentences. What's more, the high frequency of the words in the full text will affect the prediction result. Therefore, we proposed a new method of predicting the words in semantic relation, considering the semantic relation between the two given words with others which occurred in the full text, reducing the impact of high frequency words on the results,

improving the prediction results.

2. Computation of the semantic relation degree based on Text-window co-occurrence

We can use the Formula (1) to computer the degree of the words semantic relatedness based on the Text-window Co-occurrence.

$$P(Wi,Wj,Wk)=\frac{Dn(\mathrm{Fre}(Wk))}{Count(Dn)} \tag{1}$$

In Formula (1), P（Wi, Wj, Wk）means to calculate the semantic relatedness between Wk. and the two given words (Wi, Wj). Dn is a Text-window which is determined by given words. Fre（WK）means the number of words Wk appeared in the Text-window Dn, Count（Dn）is to count the total number of words in the text-window Dn. The method of Prediction the words based on the Text-window Co-occurrence.is to find the word Wk that makes the maximum value of P（Wi, Wj, Wk）. But during the test, we found that the high frequency vocabulary in the full text will affect the prediction results. A specific example can be seen in Table 1.

Table 1. a test example in Text-window Co-occurrence

(China, capital)	Frequency in full test (1.9 million words)	Frequency in text window (1796 words in test 5-windows)
People	0.0062	0.044
Beijing	0.0018	0.015

On the Table 1, the prediction result of Text-window Co-occurrence is People. But in fact, the latter word Beijing has a higher degree of semantic relation. It shows that there is a false to calculate the semantic relation degree by the method of Text-window Co-occurrence. Because of the high frequency words will also appeared in the Text-Window, sometimes a strong correlation leads to a weak result, or a weak correlation leads to a strong result.

In order to solve the problem, we proposed a method of prediction the semantic relation words based on Hidden Markov model. Reducing the influence of high frequency words in the results.

3.Predicting the Semantic Related words based on Hidden Markov Model

3.1 Theory of Hidden Markov Model

The Markov chain has been applied in many fields such as Speech Recognition, and Biological Sequence Analysis. Hidden Markov Model (HMM) is a Markov chain. In a HMM, the state is not directly visible, but the output dependent on the state is visible. The HMM can be represented by five tuple(S, Q, π, A, B).

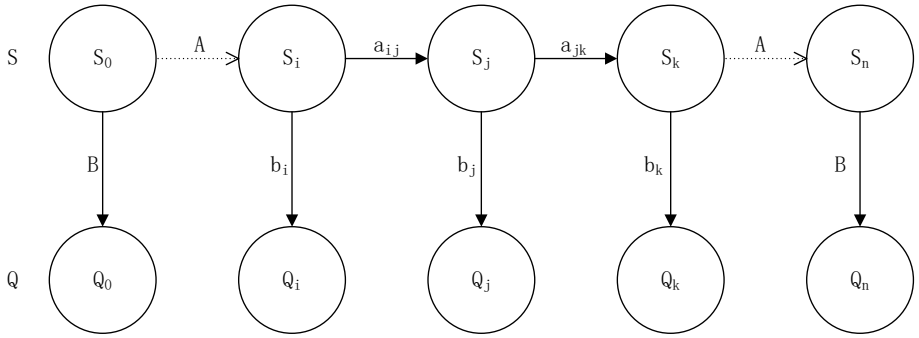

Figure.1.Hidden Markov Model

The parameter S is a hidden state sequence that cannot be observed directly. Q is the state that only can be directly observed after influenced by the hidden state,. π is a N-dimensional vector at the beginning, it must sum to 1. A is the probability of transition from state Si to state Sj. B is the emission parameter from state Si to state Qi.

The predictions of words that have semantic relatedness are influenced by the two given words. So we use the 2-gram model to extended the HMM and the parameters become (S,Q,π,A1,A2 , B1,B2).

The parameters S and Q are the same as the normal HMM. The detailed parameters can be expressed by the following Formulas. N is the number of states.

$$\pi = \{\pi_i\}, \pi_i = P(q_1 = s_i), \sum_{i=1}^{N} \pi_i = 1, \pi_i \geq 0 \tag{2}$$

$$A1 = \{a_{ij}\}, a_{ij} = P(\mathrm{q}_t = \mathrm{s}_j \mid \mathrm{q}_{t-1} = \mathrm{s}_i), \sum_{j=1}^{N} a_{ij} = 1, a_{ij} \geq 0, 1 \leq i \leq N \tag{3}$$

$$A2 = \{a_{ijk}\}, a_{ijk} = P(\mathrm{q}_{t+1} = \mathrm{s}_k \mid \mathrm{q}_t = \mathrm{s}_j, \mathrm{q}_{t-1} = \mathrm{s}_i), \sum_{k=1}^{N} a_{ijk} = 1, a_{ijk} \geq 0, 1 \leq i, j \leq N \tag{4}$$

$$B1 = \{bi\}, bi = P(ot = Qi \mid qt = Si), 1 \leq i \leq N \tag{5}$$

$$B_2 = \{b_{ij}\}, b_{ij} = \mathrm{P}(\mathrm{o}_t = Q_j \mid \mathrm{q}_t = \mathrm{S}_j, \mathrm{q}_{t-1} = \mathrm{S}_i), 1 \leq i, j \leq \mathrm{N} \tag{6}$$

There are three typical problems in HMM:

a) Assessment: Calculating the probability of the observation sequence using the given HMM parameters and the observation sequences.
b) Decoding: Using the given observation sequence and the parameters of HMM to find out the maximum probability sequences of the implicit state.
c) Learning: How to get the relevant parameters of the HMM.

3.2 Learning

The parameters of HMM are obtained by the method of maximum likelihood estimation. We take each word as a possible state value, so the probability of initial state is the frequency of each word in the corpus. Part of the initial matrix is shown in Figure.2:

$$\pi i = \begin{pmatrix} \text{People} & \text{China} & \text{Country} & \textit{production} & \cdots\cdots \\ 0.01355 & 0.01223 & 0.00739 & 0.00737 & \cdots\cdots \end{pmatrix}$$

Figure.2. Part of the initial matrix

According to the definition of aij and aijk, we statistics on the corpus to get the number of sequence which current and next states is i to j or the states from last to next is i, j, k. With the help of Formula (3)(4),we can calculated the parameters of aij and aijk. The calculation method of bi and bij is similar to aij and aijk. Part of the transition matrix is shown in Figure.3:

i / j	*peolpe*	*China*	*job*	*Country*	*production*	*development*
peolpe	*XXXX*	0.2128	0.0612	0.1194	0.0344	0.0898
China	0.1939	*XXXX*	0.0426	0.0956	0.0168	0.0999
job	0.0470	0.0359	*XXXX*	0.0585	0.0951	0.0728
Country	0.0729	0.0641	0.0465	*XXXX*	0.0453	0.0843
production	0.0199	0.0107	0.0715	0.0429	*XXXX*	0.1047
development	0.0524	0.0640	0.0552	0.0805	0.1057	*XXXX*

Figure.3. Part of the transition matrix

After the above steps, we obtained the specific parameters of HMM based on the corpus of the People's Daily. The experimental corpus is consists of 48205 articles from 1946 to 2005.

3.3. Decoding

We take the latent semantic relatedness as the hidden sequences. Using Viterbi algorithm to calculate the optimal sequences of the hidden states when the words Wi, Wj and the parameters of HMM are given.

Definition of the Viterbi Algorithm: At the time t, in all cases of the state i, the maximum value of the probability is δt（i）.

$$\delta_t(i) = \max_{i_1, i_2, i_{t-1}} P(i_t = i, i_{t-1}, \ldots, i_1, o_t, \ldots o_1 \mid (\pi, \mathrm{A}, \mathrm{B})) \tag{7}$$

The Formula (7) is the all possibilities when the current state is i, the sequence of hidden states is it-1… i1 and the observation sequence is ot….o1. When t=1, the initial value of the recurrence relation is δ1（i）=πibio1.With the change of the parameter t the recurrence relation becomes formula(8).λ=（π,A,B）.

$$\delta_{t+1}(i) = \max_{i_1,\ldots,i_{t-1}} P(i_{t+1} = i, i_t, .., i_1, o_{t+1}, , o_1 \mid \lambda) = \max_{i \le j \le N} \left(\delta_t(j)\ \ a_{ji}\right) b_{io_{t+1}} \tag{8}$$

Finally, we calculate the maximum value of the formula (8) and get the state i. The state i is the most semantically related words with the given words (Wi, Wj) in the experimental corpus.

4. Experiment

4.1 Preparation of Experiment

As mentioned in the previous, we used People's Daily as the corpus ,which consists of 48205 articles from 1946 to 2005.We do some operations of Natural Language Processing (NLP),such as filtering out the non-Chinese words or the stop words, word segmentation and so on.

Through the above method, there are total 1.9 million words and 55032 words are different. We take out the 1000 highest frequency vocabulary (45.9 % of the total number .), combining any two words as the given words (Wi, Wj). But there are a large number of invalid collocations in the given words. The invalid collocations mean the words Wi and Wj are never co-occurrence in the corpus or co-occurrence in few times by accident. Therefore, we filter out the pairs of words co-occurrence in the corpus less than 2 times, and get more than 80 thousand pairs of words.

4.2 Result

In order to verify the effectiveness of the algorithm, we randomly selected 2000 pairs of words as the experimental subjects. The results are evaluated by the following performance indicators which are commonly used in the statistical field.

$$\text{Recall}: \quad R = \frac{the\ \ number\ \ of\ \ the\ \ correct\ \ pairs\ \ in\ \ search}{the\ \ number\ \ of\ \ the\ \ correct\ \ pairs\ \ in\ \ test} * 100\%$$

$$\text{Precision}: \quad P = \frac{the\ \ number\ \ of\ \ the\ \ correct\ \ pairs\ \ in\ \ search}{the\ \ number\ \ of\ \ the\ \ pairs\ \ in\ \ search} * 100\%$$

$$F1: \quad F1 = \frac{2 * P * R}{P + R}$$

We set the same 2000 pairs of words as the experimental subjects to predict the word which has the closest semantic relation with the given words. It was done both in the HMM and Text-window Co-occurrence.

The evaluation index of the HMM and Text-windows Co- occurrence is shown in Table 2 to 4

Table 2. Recall of the HMM and Text-window co- occurrence

	1	2	3	4	5	6	7	8
HMM	0.863	0.886	0.896	0.907	0.917	0.925	0.936	0.940
Text-Windows	0.758	0.818	0.837	0.869	0.879	0.887	0.906	0.918

Table 3. Precision of the HMM and Text-window co- occurrence

	1	2	3	4	5	6	7	8
HMM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Text-Windows	0.984	0.982	0.978	0.975	0.975	0.968	0.964	0.960

Table 4. F1 of the HMM and Text-window co- occurrence

	1	2	3	4	5	6	7	8
HMM	0.926	0.939	0.944	0.950	0.955	0.960	0.967	0.966
Text-Windows	0.862	0.900	0.911	0.930	0.936	0.940	0.950	0.957

In Table 2 to 4, the horizontal coordinate indicates the number of the words we obtained as the result of the prediction. The observation of the Table 2 to 4 shows that the prediction based on HMM has a greater performance than Text-window Co- occurrence.

In order to verify the effect of the specific semantic prediction, we chosen 10 pairs of sequence from the corpus, the sequence consists of three words and have semantically related in human cognition. We select two words of them as the given words to predict the third word, the evaluation index shown in the Table 5.

Table 5. three evaluation indexes in specific prediction

	Precision	Recall	F1
HMM	0.857	0.700	0.771
Text-windows	0.500	0.600	0.545

After observing the Table 2 to 5, we can make a conclusion that predicting the words with semantic relation based on Hidden Markov Model is slightly better than the method of Text-window Co- occurrence both in universal and specific.

5. Conclusion

This paper introduces the Hidden Markov Model which is maturely applied in other fields and present a method of predicting the words with Semantic relatedness based on the Hidden Markov Model. This method uses the probability of state transition to reflect the semantic relatedness between words and extended the target words from the text window to the whole corpus, so as to reduce the impact of high frequency words on the traditional method of semantic prediction based on Text-window Co- occurrence. Finally, the experimental results not only show that the method based on Hidden Markov model is effective, but also improve the accuracy of prediction.

6. Acknowledgments:

This work is Supported by Scientific and Technological Research Program of Chongqing Municipal Education Commission（Grant No. KJ130532）.

7. References

[1] Firth J R. A synopsis of linguistic theory 1930-55[J]. Studies in Linguistic Analysis Oxford the Philological Society, 1957, 41(4):1-32.

[2] Agirre E, Alfonseca E, Hall K, et al. A Study on Similarity and Relatedness Using Distributional and WorldNet-based Approaches[C]// Naacl 09 Proceedings of Human Language Technologies:, the Annual Conference of the North America. 2013:19-27.

[3] Lee L. Measures of distributional similarity[C]// Meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, 1999:25-32.

[4] Dinu G, Lapata M. Measuring distributional similarity in context[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MitStata Center, Massachusetts, Usa, A Meeting of Sigdat, A Special Interest Group of the ACL. 2010:1162-1172.

[5] Hindle D. Noun Classification From Predicate. Argument Structures[J]. Acl Proceedings of Annual Meeting on Association for Computational Linguistics, 1990:268-275.