Word Sense Disambiguation Based on Bayes Model and Information Gain*

Yu Zheng-tao[1, 2], Deng Bin[1], Hou Bo[3], Han Lu[1], and Guo Jian-yi[1, 2]
[1]*The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051;*
[2]*The Institute of Intelligent Information Processing, Computer Technology Application Key Laboratory of Yunnan Province, Kunming 650051;*
[3]*Computing Center of Kunming University of Science and Technology, Kunming 650051*
E-mail: ztyu@bit.edu.cn

Abstract

Word sense disambiguation has always been a key problem in Natural Language Processing. In the paper, we use the method of Information Gain to calculate the weight of different position's context, which affect to ambiguous words. And take this as the foundation. We select the ahead and back six position's context of ambiguous words to construct the feature vectors. The feature vectors are endued with different value of weight in Bayesian Model. Thus, the Bayesian Model is improved. We use the sense of the HowNet to describe the meaning of ambiguous words. The average accuracy rate of the experiments of 10 Chinese ambiguous words was 95.72% in close test and the average accuracy rate was 85.71% in open test. The results showed that the method was proposed in this paper were very effective.

Keywords: *Word sense disambiguation, Bayes Model, Information gain.*

1. Introduction

There are many ambiguous words in natural language, and there are about 30%~43% [1] ambiguous words in English information and 42% [2] ambiguous words in Chinese information. Word sense disambiguation has long been a hot problem in Natural Language Processing, which can be well applied in many Natural Language Processing systems, such as Text Categorization, Information Searches, Machine Translation, Text Mining, Speech Recognition, and so on. Some literature indicates that the accuracy of system can be improved from 29% to 34.2% [3] when Natural Language Processing systems added the Word Sense Disambiguation. Word sense disambiguation was recognized the most difficult problem at the lexical level of Natural Language Processing field.

There are many methods for the Word Sense Disambiguation, which can be divided into two main categories. One kind of method based on rules of grammar. In grammar-based regular approaches, the sense of ambiguous words is determined by rules of linguistic grammar. For

* This paper is supported by National Nature Science Foundation (60863011, 60663004), Yunnan Young and Middle-age Science and Technology Leaders Foundation (2007PY01-11), The Key Project of Yunnan Nature Science Foundation, Yunnan Educational Department Foundation (07Z11139), Kunming University of Science and Technology PH.D Foundation (2006-12).

example, Syntactic Relations, Tie-in of Semantics, Lexical Characteristics, POS Feature, and so on. Cowie(1992)[4]、Agirre and Rigau(1995)[5]use a method ,which based on dictionary's Machine-readable method to assign a sense of ambiguous words. The rules of grammar must be constructed in this approach, and the rules are difficult to set up and obtain. The other kind of methods based on statistics of corpus. In Corpus-based Statistical approaches, calculate the word's probability weighting in given the context of an ambiguous word. The sense with the greatest probability weighting is the optimal results. The method including Naïve-Bayesian Classifier, Hidden Markov Model, Maximum Entropy Model [6], vector space model, and so on. With development of Corpus linguistics, the Corpus-based Statistical approaches have occupied mainstream status in the Word Sense Disambiguation Field.

The feature vector space is all words of the sentence that include ambiguous words. The feature vector space will become huge when the sentence is long. And there is same influence between all feature vectors and ambiguous words. Therefore, the influence degree of the feature vector of the Naïve-Bayesian model for ambiguous words are improved by Information Gain, the scope of ambiguous words' context is limited. The feature vector is endued with different value of weight. Thus, this paper presents an improved scheme for Information Gain which based on Bayesian model in the Word Sense Disambiguation.

2. The improved model for Information Gain based on Bayesian model in Word Sense Disambiguation

2.1. The processing of Word Sense Disambiguation

The HowNet has been as resources for the Word Sense Disambiguation in this pager. The meaning of words in context was determined by the sense of the HowNet. At first, the training corpus are tagged and segmented by the using of the sense of the HowNet. The test corpus also must be segmented. Secondly, according to the information of context in the test corpus, the weight of context position and the effective range of context are calculated. Thirdly, according to the information of the test corpus, the necessary parameters are calculated for disambiguation models. The fourth, according to the improved model for Information Gain based on Bayesian model in the Word Sense Disambiguation, a sense of ambiguous words are calculated in the test corpus.

2.2. HowNet

HowNet is a knowledge library, which takes Chinese and English words as the description object and reveals the inter-concept semantic relations and the inter-attribute semantic relations. HowNet use the limited sememe to explain concept, these sememes are organized separately in the respective document, including entity, event, attribute and attribute value and so on. In the HowNet, according to certain rules and relate concepts, the sense is formed from the sememe. It use a serial number (NO.) expressed. W_C, G_C, E_C expressed Chinese words, the lexical category and the example respectively, DEF is the semantic expression. The words possibly have the different serial number (NO.) with the semantic expression (DEF). For easily count, we define a unique semantic identification code as each sense (NO_MARK). Like "材料" have three senses. The NO_MARK of the first sense is "9406", NO.is "009406", DEF is “attribute|属性,quality|质量, &human|人". The NO_MARK of the second sense is "9408", NO.is "009408", DEF is “information|信息". The NO_MARK of the third sense is "9409", NO. is "009409",DEF is " material|材料, generic|统称". During computation, processing according to

DEF will be carried out and a NO_MARK-WORD database table will be created simultaneously.

The identical word possibly has different NO_MARK in this database, and different words may also have the same NO_MARK. There are 17308 senses from the database. Therefore, by these senses we can express all concepts and tag the word's meaning

2.3. Information Gain

In Word Sense Disambiguation, the context of the ambiguous word is very important when ambiguous word is understood. There is specific relevance to the meaning in neighboring words. A certain scope's context can provide the sufficient language information for ambiguous words. The relations with Information Gain can be quantified. A certain scope's context may be further analyzed to determine how many information contents for each position of an ambiguous word's context can provide, including that how to determine the information content of the each context position and the context scope.

2.3.1. Words Context Information Matrix Determination: The certain scope's context of the core word can be defined as a context vector of the core word. After being segmented and tagged, context vector set of the core word is constructed, and the core word's context vector is defined as the Context Information Matrix. Thus, the number of core word's sense is the number of core word's Context Information Matrix. For further formalization of the sense of core word expression, the form of context information matrix provides a direct way to express. In this section, the paper discuss that how to determine the scope of the context information.

A method named information gain to obtain the effective range of the context is used. In order to obtain the statistical average result of the context position's weight which based on the corpus, a set that includes 1000 high frequency words which obtained from 1998 "People's Daily" were used as the object of study[7].

2.3.2. Determine the Information Content of the Context by Information Gain: The high frequency words and their contexts are formalized as a symbol information system. Source prior uncertainty (entropy) is statistical uncertainty of the high frequency words, the letter places posterior uncertainty is uncertainty of the context in a known position. The difference between them is the information gain of condition entropy in the known context position, and the information gain determines the information content in each position [8].

Assume the set of high frequency words is W, the set of the contextual words is cw, the formula of Information Gain is as follow:

$$IG_P = H(W) - H(W \mid V_P) \quad (1)$$

The formula (1) is the information content of in p position context (IG_P), Gain information is the reduction that between the entropy of overall system $H(W)$ and the entropy of conditions which known in the location p of the context $H(W \mid V_P)$.

The formula (1) each explanation is as follows:

$H(W)$ is the information entropy of the high frequency words source in the context's information matrix. It is defined as follows.

$$H(w) = \sum_{w \in W} P(w) \times \log_2 P(w) \quad (2)$$

And $P(w)$ is the frequency statistics probability of the high frequency words w, define for form (3):

$$P(w) = \frac{|\, fre(w) \,|}{|\sum_{w \in W} fre(w)\,|} \quad (3)$$

Description: $|\sum_{w \in W} fre(w)\,|$ is the total frequency of 1000 high frequency words which appears in the corpus; $|\, fre(w)|$ is the frequency of the word w which appears in the corpus.

$H(W \mid V_P)$ is the entropy of conditions which known in the location V p of the context, define for formula (4):

$$H(W \mid V_p) = \sum_{cw \in V_p} P(cw) \times H(W \mid cw) \quad (4)$$

And V is the statistical probability of the contextual word cw in the contextual position p; $H(W \mid cw)$ is the entropy of conditions which the contextual words were known.

The result of the each weight of context position is as follow table:

Table 1.The Information Gain of Context Position

Context Position	The Value of the Information Gain	Context Position	The Value of the Information Gain
-1	2.27311089547	1	2.30594858001
-2	2.15387590152	2	2.13597409566
-3	1.94098164570	3	1.92702181067
-4	1.31981900741	4	1.11722149564
-5	1.15615596421	5	1.10859172459
-6	1.00150682983	6	1.00649586732
-7	0.19558745846	7	0.63259581121
-8	0.10980246139	8	0.30904760549

As a result, in this paper the best range of validity of the context is the first 6 and the latter 6 position of the high frequency words.

2.4. Improved the Bayesian Model of Word Sense Disambiguation based on Information Gain

Bayesian classifier use Bayesian decision-making rules of classification. Assume the ambiguous word $\vec{x}$ have two senses, must decide the word $\vec{x}$ belong to s_1 or s_2, first count the probability of $P(s_1 \mid \vec{x})$ and $P(s_2 \mid \vec{x})$, calculate separately $\vec{x}$ belong to the different category probability. If $P(s_1 \mid \vec{x}) > P(s_2 \mid \vec{x})$, the ambiguous word belongs to s_1, otherwise it belongs to s_2, using the Bayesian formula:

$$P(s \mid \vec{x}) = \frac{P(\vec{x} \mid s)}{P(\vec{x})} P(s) \quad (5)$$

According to Bayesian decision-making rules, may use follow formula to decision the type s' of the ambiguous word $\vec{x}$.

$$s' = \arg\max_{s} [\log P(\vec{x} \mid s) + \log P(s)] \quad (6)$$

In order to calculate conveniently, assume the vectors of feature in text are independent:

$$P(\vec{x} \mid s) = P(\{x_j \mid x_j\ in \vec{x}\} \mid s) = \prod_{x_j \text{ in } \vec{x}} P(x_j \mid s) \quad (7)$$

Because there is influence for the position of context by information gain, if ambiguous words have the sense set s, may change the formula (7) as follows.

$$s' = \arg\max_{s \in S} \prod_{p=-6}^{6} [IG_p P(x_p \mid s)] P(s) \quad (8)$$

Training parameters of the Bayesian model based on Information Gain using maximum likelihood estimation method, and IG_p obtain the each value according to Table1.1, the value of $P(x_p \mid s)$ and $P(s)$ be calculated separately based on the formula (9) and (10).

$$P(x_p \mid s) = \frac{C(x_p, s)}{C(s)} \quad (9)$$

$$P(s) = \frac{C(s)}{C(w)} \quad (10)$$

$C(x_p, s)$ is the number of times which the contextual word x of training corpus w appear together in position p with the ambiguous word, when the ambiguous word w is s. $C(s)$ is the number of times which the ambiguous word w appear in the training corpus, when the ambiguous word w is s. $C(w)$ is the number of times which the ambiguous word w appear in the training corpus.

2.5. Deal with Data smoothing

Because of Small-scale training corpus and the uneven distribution of parameters, some probability parameters possible are ″0″, namely existence data sparse matrix. In order to eliminate the influence which probability parameters are ″0″, we use the data smoothing technology to estimate these probability parameters.

These are many kind methods in the data smoothing technology. The ″Add One″ and the ″Good-Turing″ are common methods. In this paper we used ″Add One″ method [9]. This method is simple and effective. main idea of the ″Add One″ is that the words which are not counted appeared one time, simultaneously to guarantee that sum of all probability is ″1″, when calculate the probability the denominator must add the total which all words possibly appears. So the formula of the model parameter $P(x_p \mid s)$ and $P(s)$ like formula (11) and the formula (12) as follows:

$$P(x_p \mid s) = \frac{C(x_p, s) + 1}{C(s) + N} \quad (11)$$

$$P(s) = \frac{C(s) + 1}{C(w) + m} \quad (12)$$

And N is the total of words in the HowNet. But m is the number of word w all senses. If the result of the parameters which the massive parameters are ″0″ was calculated as ″1″ time in the smooth method, probability density distribution tend toward the words which have not counted. The result is not valuable in the statistics. Therefore, the words which have not counted is not ″1″ time, but records for ″λ″ time, Let ″λ″ in (0, 1) the scope, ″λ″ may make the corresponding adjustment according to the training corpus size. So $P(x_p \mid s)$ and $P(s)$ like formula (13) and the formula (14) as follows:

$$P(x_p \mid s) = \frac{C(x_p, s) + \lambda}{C(s) + \lambda N} \quad (13)$$

$$P(s) = \frac{C(s) + \lambda}{C(w) + \lambda m} \quad (14)$$

The parameter of Smooth algorithm in this paper $P(x_p \mid s)$ take λ=0.5, and $P(s)$ take λ=0.01.

3. Experiments and Results Analysis

The entire experiment divides into the training stage and the test phase. All experimental data from the 1998 "People's Daily" corpus, and the sense of ambiguous words come from the HowNet which explain ambiguous words. Experimental data has a total of 1,000 sentences, including 66,500 words. Tests were conducted close and open test, the close test data still come form 1998 "People's Daily", have 150 sentences, a total of 13,400 words. And Chinese information corpus of the open test come from the Internet, each sentence average long is 30 words; open test data have 200 sentences, including a total of 6,500 words.

3.1. Experimental Process

The training stage is as follows:

① Select the core words of certain coverage from the training corpus, simultaneously select the context of certain scope in the core words (this paper define around core words before and after 15 words) to form the contextual information matrix.

② calculate the context information matrix using the formula (1), obtain various information content IGp of context position to core word.

③ Count N in the database, calculate m in the training corpus.

The test phase is as follows:

① Segment and tag the each sentence of the test corpus.

② Count the total senses of each ambiguous word in the test sentence, calculate the probability of each sense in the training corpus, compute $P(s)$ and $P(x_p \mid s)$ of each ambiguous words′ sense in a sentence.

③ calculate correct word meaning s' for each ambiguous word with the formula (8).

④ repeat step ② and step ③ until all ambiguous words in the test corpus were completed.

3.2. Results Analysis

Using Improved the Bayesian Model(IBM) of Word Sense Disambiguation based on Information Gain to supervise disambiguating for the ambiguous words, train and test 10 ambiguous words. At the same time, we use the Naive Bayesian Model(NBM) to make the same training and the test for these 10 ambiguous words, as a contrast. The results are as follows:

AM strands for Ambiguous Words;

S stands for Senses;

CT stands for Amount of close test;

ACT stands for Accuracy of close test;

OT stands for Amount of open test;

AOT stands for Accuracy of open test.

Table 2. Results of experimentation

AM	S	CT	ACT(%)		OT	AOT（%）	
			IBM	NBM		IBM	NBM
材料	3	79	98.12	92.03	59	91.45	90.33
领导	3	46	96.23	91.57	45	88.36	86.47
打	28	56	95.05	90.34	35	81.65	80.90
拍	2	22	95.41	91.55	41	90.78	91.01
和	6	36	97.23	91.02	36	84.49	81.79
组织	3	55	93.02	89.46	22	81.09	80.72
活动	5	31	92.05	87.12	32	87.63	87.02
好	8	13	94.84	90.47	18	81.30	80.24
中心	3	26	95.25	90.31	10	80.09	78.77
指挥	2	19	100	96.29	13	90.27	90.16
average			95.72	91.02		85.71	84.74

Some conclusions can be found according to table's experimental result:

（1） select 6 contextual words which before and after ambiguous words as the feature vector, under these information's influence, the accuracy of close test is quite high. But the scope of context in Naive Bayesian Model is quite wide, therefore, the noise seriously impact on the ambiguous words, and the speed is quite slow.

（2）improved the Bayesian Model of Word Sense Disambiguation based on Information Gain in a supervised closed-learning method can be achieved better disambiguation effect, the average closed test accuracy is higher than the Naive Bayesian Model 4.7 percent.

（3） the average closed test accuracy of Improved the Bayesian Model of Word Sense Disambiguation based on Information Gain is higher than the Naive Bayesian Model 0.97 percent in the open test.

While makes the effective disambiguation progress, also has the following some problems:

（1） still exist data sparse matrix problems. When the word did not appear in the training corpus, the system will not be able to accurately access the correct answer.

（2） the training corpus are restricted in People's Daily merely, the scope of the information is limited.

（3） the determination of context position is obtained by counting the 1000 high frequency words. So there is a little error in the information content IG of context.

4. Conclusions

In this paper, an improved Bayesian Model of Word Sense Disambiguation based on Information Gain is introduced to enhance the weight of context position. The Word Sense Disambiguation which improves the feature vector of the Bayesian Model for ambiguous words provided an effective reference. We finished an actual software system (http://222.221.6.149:7001/LIIPnew/xq.jsp). Experimental results show that the method of Information Gain can improve Bayesian Model effectively. This method broad prospect for development and it is very advantageous to Natural Language Processing for further development. Looking forward, in order to improve the accuracy of disambiguation, we need to overcome the issues caused by sparse matrix and more accurate information content of the context.

Acknowledgements

Thanks for information retrieval laboratory, Harbin Institute of Technology providing the LTP: Language Technology Platform API with us to word segmentation, POS tagging.

References

[1] Hwee Tou Ng, John Zelle, “Corpus-based approaches to semantic interpretation in natural language processing”, AI Magazine, 1997, 18(4): 45~64.

[2] Song Lu, Shuo Bai, etal, “Supervised word sense disambiguation based on Vector Space Model[J]”,Journal of Computer Research & Development, 2001, 38(6): 662-667.

[3] Hinrich Schutze, Pedersen J, “Information retrieval based on word senses”, In: Proc of the 4th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, 1995. 161~175.

[4] Cowie, J. and Guthrie, L., “Lexical disambiguation using simulated annealing[C]”, In Proceedings of the Fifth International Conference on Computational Linguistics COLING-92, 1992, 157－161.

[5] Agirre, E. and Rigau, G., “A proposal for word sense disambiguation using conceptual distance”, In Proceedings of the first International Language Proceeding, Velingrad, 1995.

[6] Li,J.Z., “An improved maximum language and its application[J]”, Journal of software, 1999, 3:257-263.

[7] LU Song, BAI Shuo, HUANG Xiong, and ZHANG Jian, “SUPERVISED WORD SENSE DISAMBIGUATION BASED ON VECTOR SPACE MODEL [J]”, Journal of Computer Research and Development,2001.06 Vol.38, No.6: 662-667.

[8] LU Song,BAI Shuo, “Quantitative Analysis of Con text Field in Natural Language Processing [J]”, Chinese Journal of Computer, 2001.7 vol24 No 7: 742-747.

[9] Lu Zhimao, Liu Ting,Zhang Gang, Li Sheng, “ Word Sense Disambiguation Based on Dependency Relationship Analysis and Bayes Model[J]”, High Technology Letters,2003.05. Vol13.

Authors

Yu Zheng-Tao, born in 1970, Ph.D., professor. His main research interes include natural language process, Chinese question answering system and Information retrieval