Journal of Theoretical and Applied Information Technology
30th November 2012. Vol. 45 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENTROPY THEORY: AN EFFECTIVE TOOL OF IMPROVING THE COMPUTATIONAL TRANSLATION

JUANJUAN SUO, HUIQQING TIAN, YANCANG LI
Hebei University of Engineering, Handan 056038, Hebei, China

ABSTRACT

Elimination of the ambiguity is the key to the computational translation. In order to find an effective way to improve the accuracy of the computational translation, the cross entropy was introduced. After the analysis of the reasons of the low accuracy, the information entropy was introduced into the disambiguation. The practice of ambiguity elimination shows the method has high accuracy and this study provides an effective way to improve the computational translation. The study has significance in theory and practice for the development of computational translation.

Keywords: *Computational Translation, Ambiguity Elimination, Entropy*

1 INTRODUCTION

With the information explosion and economic globalization, international information exchange is surging tremendously. Machine translation is the effective way to solve the problem of the translation of mass information quickly and inexpensively. Machine translation technology has undergone a history of over 70 years since its debut in the United States for collecting intelligence in 1940s. Computational translation is more and more important for the computational linguistics. As one of the computational linguistics research field, the emergence of the machine translation drives the development of the information society.

In the past half century, lots of works have been done on it. And there are many machine translation systems available today. It has the advantages of speed, cost-efficiency, and the ability to deal with sheer volume of translation task [1-2]. However, there is one thing computer can not beat human being, at least at the present time and near future, which is the quality of ambiguity. How to improve the accuracy has significance in theory and practice for the development of the computational linguistics and the information society. Many Scholars have done a lot on the ambiguity elimination [3-7], but we still have a long way to go.

To find an effective method for the improvement of computational linguistics, the entropy theory was employed to the study of the ambiguity elimination. The example shows that the method is simple and effective, and is easy to computer adaptive realization. This research can improve the digital level of our society, and promote the construction of digital city and harmonious society.

The rest of the paper was organized as follows. Fist, the reasons of the low accuracy of the computational translation was analyzed. And then the basic knowledge of entropy was introduced and then it was employed to the ambiguity elimination. Finally, its efficiency was shown through an example.

2 REASONS OF LOW ACCURACY OF COMPUTATIONAL TRANSLATION

Computational linguistics has become not only a basic linguistic of the information society. Many experts have pointed out that, as the key and difficult point of computational linguistics, ambiguity is one of the biggest obstacles in the computer analysis and understanding. Many scholars devoted to the research away discrepancy. But we still need to explore new calculate method, to focus on the study of language level, especially need to put forward more effective, perfect ambiguity description and eliminate the theory and the method.

Ambiguity is the bottleneck problem of the natural language processing. The natural language processing decades of history is actually the history of ambiguity struggle. Ambiguity, according to sources, is divided into vocabulary ambiguity and structural ambiguity. Vocabulary ambiguity is one of parts of speech ambiguity to carry on the syntactic analysis. It easily leads to the extremely syntactic analysis errors. Meaning ambiguity directly leads to the wrong statement [8]. Structural ambiguity is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

generally caused by the same syntactic structure, and it should be eliminated through the text analysis of the subject and the analysis of sentences by other components. In 1993, Lancaster University Corpus Research Center developed automatic SEMTAG. Through automatic classification of the each word, phrase and sentence, the discourse of the semantic features of general appearance and distribution state, and the calculation formula of the original text can be obtained. This method can solve the exact nature of context translation.

Somers pointed out that the main difficulty of machine translation can be summed up in one word: ambiguity, although problems of style and interpretation should not be ignored [9]. Ambiguity is a pervasive phenomenon in human languages. It is very hard to find words that are not at least two way ambiguous, and sentences which are (out of context) several ways ambiguous are the rule, not the exception. The ambiguity can be easily distinguished by human translators in most cases, but can not be distinguished by machines, simply because machines can not understand the text [10]. Ambiguous categories can be classified into the semantic and syntactic ambiguity. Ambiguity, according to sources, is divided into vocabulary ambiguity and structural ambiguity. Vocabulary ambiguity is one of parts of speech ambiguity to carry on the syntactic analysis. It easily leads to the extremely syntactic analysis errors. Meaning ambiguity directly leads to the wrong statement [11]. Syntactic differences between languages make it impossible for word translation. When one single sentence can produce more than one interpretation, the structural ambiguity will appear. Ambiguities exist not only on lexical level but also in syntactical level. Distinctive syntactic difference between Chinese and English is the use of passive voice. Passive voice is not used so commonly in Chinese as in English. Large amount of English passive sentences need translating into active voice in Chinese to match Chinese grammar and syntactical rules. Otherwise, the translation will be rigid or distorted. Structural ambiguity can be categorized according to the‘range’‘depth’of ambiguity. Structural ambiguity is generally caused by the same syntactic structure, and it should be eliminated through the text analysis of the subject and the analysis of sentences by other components. W. John Hutchins and Harold L. Somers put three headings under ambiguity, respectively morphology problems, lexical ambiguity and structure ambiguity. The second category has three components, which are category ambiguity, homograph and polysemy, and transfer ambiguity. And the third one has two: real structural ambiguity and accidental structural ambiguity. When one word represents more then one speech, or has multiple meanings, or the combination of these two, the lexical ambiguity will appear. Where one word can be interpreted in more than one way is the meaning of lexical ambiguity. And one word would have different translations when being put into different sentences and contexts.

Foregoing are all the monolingual uncertainties. In order to make the English-Chinese machine translation more effectively, lots of works have been done [8-12]. Here we will summarize them up. F. Zheng et al. proposed an ambiguities technique in HENU automatic Chinese segmenting system. The method places emphasis on the discovery of segmentation ambiguities and the removal of ambiguous words and phrases. First, the longest word and the second longest word are formed by means of the major dictionary based matching strategy. Second, segmentation ambiguities are found by the use of leap test so as to judge whether the segmentation ambiguities are of intersection type or combination type. Then, on the basis of the different kinds of segmentation ambiguities, disambiguation is done. The disambiguation of intersect ion type segmentation ambiguities is done by using the rule based strategy and the statistics based strategy. The combination type of ambiguities are removed by the rule based strategy so that the ex act place for segmentation is found. S. Du described the structure of modified Hidden Markov Models (HMM) on condition that observation noise is not independent of the Markov chain, and proposed the Baum Welch algorithm of modified HMM and derived the update parametric estimation equations for modified HMM based on traditional HMM. Y. Liu combined Support Vector Machines (SVM) with rules and proposed a new algorithm (SR algorithm) to deal with the combinatorial ambiguous phrases in Chinese word segmentation [13]. The key idea of the SR algorithm is to solve combinatorial ambiguous phrases making use of the theory of SVM and rules of parts of speech. In a test of several kinds of Chinese corpus, it indicates that the accuracy of segmentation for combinatorial ambiguous phrases reach 83%. It provides a new method for solving Chinese word segmentation problems. W. Tan proposed a method based on the bayes and machine readable dictionary which could disambiguate by the training of a small scale corpus and the definition of semantic in machine dictionary. The experimental results show that it has a high accuracy rate of word sense ambiguation when the scale of markup corpus has been limited. X. Wang

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

proposed a new method of Chinese automatic segmentation that can check all overlapping ambiguity in sentence. This algorithm is based on the principle of Choose Longer Word. It solves the problem that the count of segmentation way is exponentially increasing with the sentence length, and provides a method to handle overlaying ambiguity and overlapping ambiguity separately. Duo to the word limit, we will not repeat the others. They all improve the overall quality of machine translation and promote the development of the computational linguistics.

The essence of the ambiguity is the shortage of the corresponding relation between the expression of the language form and its meaning. Ambiguity arises when there is a certain concept in language A but there is no such concept in Language B or a concept which is described by one single word in one language may have several words to express in another language. When words or sentences are translated into other languages, ambiguities may occur because of cultural, grammar or syntactic differences among languages. This is the inherent characteristics of the natural language and it is one of the characteristic of the difference between natural language and artificial language. Human translators can handle this kind of complexity by investigating the cultural differences and conducting research to produce correct translations. However, if translated by machine, it would be impossible. The studies to natural language processing system has guiding significance to researchers, but the complex of the ambiguity phenomenon needs to put forward more perfect and more suitable methods for the ambiguity description and eliminate. There are many factors contributing to the ambiguity of the machine translation translations other than in linguistic perspective, such as computational problems. The studies to natural language processing system has guiding significance to researchers, but the complex of the ambiguity phenomenon needs to put forward more perfect and more suitable methods for the ambiguity description and eliminate. This is the inherent characteristics of the natural language and it is one of the characteristic of the difference between natural language and artificial language.

3 AMBIGUITY ELIMINATION BY USING ENTROPY

3.1 Introduction Of Entropy Theory

First proposed in 1864 by R. Clausius, the entropy has 140 years of history. As an important concept, entropy is the best measure to "uncertainty". It is widely used in natural science and social science. Some scholars put forward that the 21st century is the century of entropy [14].

In 1957, E. T. Jaynes put forward the famous idea of maximum entropy principle. The entropy optimization principle is mainly composed of maximum entropy principle and the cross entropy principle proposed by Kullback. The former comes from Shannon information entropy, the latter is derived from the concept of distance measure in probability. Both are the further development of information entropy.

From the concept of information entropy, we know that when probabilities of the random events are same, the entropy is the maximum. When the Probability space is $p_0 = (p_1, p_2, \cdots, p_m)$, the information entropy value is the largest. Because $p_0 = (p_1, p_2, \cdots, p_m)$ is the maximum probability distribution of uncertainty, we can define another uncertainty of the measurement, namely "relative measure".

For a distribution:

$$p_0 = (p_1, p_2, \cdots, p_m) \tag{1}$$

It has the shortest distance with

$$F(p,u) = \sum_{i=1}^{m} p_i \ln \frac{p_i}{u_i} \tag{2}$$

Then the corresponding uncertainty should also become the biggest. Based on some kind of prior factor, Kullback and Leibler proposed cross entropy

$$F(p,u) = \sum_{i=1}^{m} p_i \ln \frac{p_i}{u_i} \tag{3}$$

The cross entropy has following characters:

For discrete form of cross entropy, if $x_i, y_i \geq 0, i = 1,2,\cdots,n$ and $\sum_{i}^{n} x_i = 1 \geq \sum_{i}^{n} y_i$,

$$h(X,Y) = \sum x_i \log \frac{x_i}{y_i} \geq 0 \tag{4}$$

is the cross entropy of X to Y, and

$$X = (x_1, x_2, \cdots, x_n)^T$$

Journal of Theoretical and Applied Information Technology
30th November 2012. Vol. 45 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$Y = (y_1, y_2, \cdots, y_n)^T \quad (5)$$

And

(1) $x_i = y_i$, if and only if $x_i = y_i$;

(2) The bigger cross entropy, the greater distance of the distribution.

(3) $0 \log\left(\frac{0}{q}\right) = 0,\ p \log\left(\frac{p}{0}\right) = \infty$.

3.2 Application Of Cross Entropy In Disambiguation

Cross entropy (relative entropy) is a measure of the distance from the two relatively random physical quantities. This paper tried to use the system engineering methods, to the definition of a cross entropy for extended, and introduce it to the extinction of computational linguistics, to provide a more effective away for the ambiguity description and elimination. We have the following algorithm to eliminate ambiguity:

(1) Calculation of the probability value of the statements, as the prior information of training set. From the rules of A→λ, obtain the probability α,

$$\alpha = P(A \to \lambda) \prod_{i=1}^{n} P(x_i)$$

$$\omega_i, \omega_{i+1}, \cdots, \omega_j \quad (6)$$

(2) Calculate the probability of machine translation β, as the test set of posterior information.

$$\beta_i = P(A \to \lambda) \beta \prod_{k=1}^{i-1} \alpha_k \prod_{j=i+1}^{n} \alpha_j$$

$$i \in [1, n] \quad (7)$$

3) Calculation of the cross entropy.

$$H = -\frac{\sum_{s \in c} \log P_s}{\sum_{s \in c} |s|} \quad (8)$$

Where C is the training set, S is a sentence in the library, P_s is the probability of sentence, $|S|$ is the sentence length.

4) The minimum value cross entropy of information statement as a output of the machine semantic. That is, if $H(T_2 P_1) \prec H(T_2 P_2)$, take P_1 as a statement of machine translation.

5) Take the cross entropy comprehensive differences $\overline{D} = \left[(D_1 - D_2)/2 \right]^{\frac{1}{2}}$ as stop standards, $\overline{D} \le 0.01$ end program. Otherwise, back to step 1). The process of the algorithm is shown in Figure 1.

The algorithm takes the latent semantic indexing method to the calculation of the similarity.

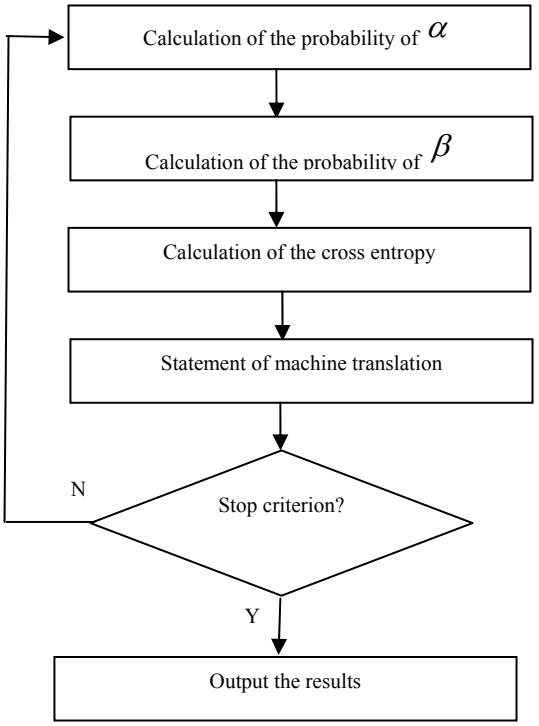

Figure 1: Framework Of The Algorithm

3. 3 Test Results

Using the B700 library of American Chrysler Corporation Corpus, we verified the method. The results of the disambiguation method proposed here are shown in Table I.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table I: Results Of Disambiguation

N0. of Corpus		B700
Sentence number		746
Word number		9270
Artificial rules disambiguation	Average ambiguity	1286
	Recall ratio	99.17
	precision	77.11%
Cross entropy method disambiguation	Average ambiguity	996
	Recall ratio	98.64
	precision	90.73

4 CONCLUSION

How to eliminate the ambiguity is the key to the computational translation. To find an effective way to describe and eliminate the ambiguity, we employed cross entropy to measure the distance from the difference the translations and the original meaning, and proposed the method based on cross entropy. The example shows that the method is simple and effective, and is easy to computer adaptive realization. But it is only a beneficial exploration, and many problems still need to be studied further.

ACKNOWLEDGEMENTS

The work was supported by the Youth Fund Project of Humanities and social science research of Hebei Province Department of Education (SQ125001). And the Fund of Social Science Development Program of Hebei Province (201103205).

REFRENCES:

[1] J. Yao and D.Zhao, "Disambiguation method in Chinese word segmentation based on phrase match". Journal of Jilin University (Science Edition), Vol. 48, No. 3, 2009, pp. 427-432.

[2] Hausser Roland, "Foundations of computational linguistics, human-computer communication in natural language", 2nd Edition, Berlin, New York, Springer, 2001.

[3] D.Hindle and M. Rooth, "Structural ambiguity and lexical relations", Computational Linguistics, Vol. 19, No. 1, 1993, pp. 103-120.

[4] Sil,i Wang and Bin Wang, "A Chinese overlapping ambiguity resolution method based on coupling degree o f double characters", Journal of Chinese Information Processing, Vol. 21, No. 14, 2007, pp. 14-17.

[5] D. McCarthy and R. Navigli, "The English lexical substitution task", Language Resources and Evaluation. Vol. 43, No. 2, 2009, pp. 139-159.

[6] Tomoharu Iwata, Daichi Mochihashi and Hiroshi Sawada, "Learning common grammar from multilingual corpus", Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), Uppsala University Uppsala, Sweden, 11-16 July, 2010. pp. 184-188.

[7] Roberto Navigli and Mirella Lapata, "An experimental study on graph connectivity for unsupervised word sense disambiguation", IEEE Transactions on Pattern Anaylsis and Machine Intelligence, Vol. 32, No. 4, , 2010, pp. 678–692.

[8] Muyun Yang, Shuqi Sun, Junguo Zhu, Sheng Li and Tiejun Zhao, "Improvement of machine translation evaluation by simple linguistically motivated features", Journal of Computer Science and Technology, Vol. 26, No. 1, 2011, pp. 57-67.

[9] R. Dale, H. Moisl and H. Somers, "A handbook of natural language processing", New York, Marcel Dekker, 2000.

[10] D. Arnold, L.Balkan, R. L.Humphreys, S. Meijer and L.Sadler, "Machine translation: an introductory guide", Oxford: Blackwell, 1994.

[11] Li Zhang, Liyong Zhang and Xiaomiao Zhang, "Research on Ambiguous words segmentation algorithm based on improved BP Neural Network", Journal of Dalian University of Technology, Vol. 47, No. 1, 2007, pp. 131-135.

[12] Mahsa Chitsaz and Chaw Seng Woo, "Software agent with reinforcement learning approach for medical image segmentation", Journal of Computer Science and Technology, Vol. 26, No. 2, 2011, pp. 247-255.

[13] Y. Liu and Z. He, "Algorithm based on SVM and rules for the disambiguation of combinatorial ambiguous phrases", Journal of Chongqing University, Vol. 28, No. 10, 2005, pp. 53-56.

[14] W. H. Qiu. "Management decision-making and entropy theory", Beijing, China Mechanical industry press, 2002.