Novel method for quality assessment of computational translation

Juanjuan Suo*, Huimin Dong

Hebei University of Engineering, Handan 056038, China

Received 1July 2014, www.tsi.lv

Abstract

To overcome the shortcomings that there are few feasible methods and models in the comprehensive assessment on the quality of the computational translation, a novel mathematical tool, the unascertained measure was introduced. After the introduction of the basic knowledge of the Unascertained Sets, the unascertained measure was defined and the comprehensive assessment model was set up. Then the method was introduced to the quality assessment of the machine translation. Engineering practices shows that the method can complete the assessment systematically and scientifically without any assumption.

Keywords: computational translation, quality assessment, unascertained measure, model

1 Introduction

As one of the computational linguistics research field, the emergence of the machine translation drives the development of the information society [1]. In the past years, lots of works have been done on it. And there are many machine translation systems available today [2]. They have the advantages of speed, cost-efficiency, and the ability to deal with sheer volume of translation task. However, there is one thing computer cannot beat human being, at least at the present time and near future, which is the quality of ambiguity. As the key and biggest difficulty of computational linguistics, the ambiguity is the chief bottleneck of computer analysis and understanding. So, the assessment of the computational translation has significance in theory and practice for the development of the computational linguistics and the information society. Many scholars devoted to the related research and have proposed many effective theories and methods [3-10]. But, we still have a long way to go.

According to this situation, a new method, the unascertained set was introduced to solve the unascertained problem of the assessment. The unascertained measure was introduced and the credible identification was set up for the reliability assessment. Application results showed that it could complete the reliability assessment systematically and scientifically.

The rest of the paper was organized as follows. In the introductory part, attention was paid to the basic concepts of the Unascertained Mathematics. In the following part, a reliability assessment model was set up. Then, its application in practice was introduced. Finally, the advantages of the method proposed here were pointed out.

2 Machine translation

Machine translation is also called computer translation and electronic translation. The research of machine translation in China was started in 1956 and the first test was carried out successfully in 1959. With the rapid development of network technique, machine translation becomes more prosperous when more challenges appear. And unsolved problems for many years still exist, many work have to be finished until machine translation technique mature.

2.1 PRINCIPLE AND FUNCTION OF MACHINE TRANSLATION

The process of Machine translation can be divided into five parts, includes original language input, original language analysis, transfer of the original language to target language, target language generation and target language output. Although the research work of machine translation has been underway for many years, its application has great limitations. Currently, the most advanced machine translation systems are only used to replace human translation in a limited range.

The fundamental principle of machine translation is to build machine dictionary, terminology database, data bank for translators by huge storage capacity and rapid retrieval ability of computer. Thus the retrieval time of translators can be saved greatly. Some machine translation system can store, revise and print translated text to help to improve work efficiency apart from providing retrieval function. Machine translation system allows people intervene during analysis, transformation and generation. The problems can be solved easily by people intervene under certain circumstance, such as language construction ambiguity. In order to improve translation efficiency and ensure the

*** *Corresponding author*** e-mail: suojj@163.com

consistency of translated text, machine translation system usually contains a series of tool components. And two major tool components are translation memory and terminology management.

Translation memory is equivalent database of original and translated text by machine building. Computer can store translated words needed translation in language database when translators are working. During the process of translation, computer will show translators matching sentences translation when the same and similar sentences appear. Most of translation memory software support fuzzy match, users can set minimum matching degree. The translators still can obtain a sentence of fuzzy match by fuzzy match, and then all that is needed is to translate or revise the different parts. Machine memory update constantly and store automatically new translation users providing. Along with constant rich of memory database, the work efficiency of translation will becomes much higher. The work principle of terminology management is to scan one by one original terminology by machine and check in its dictionary base. Terminology management ensures consistency and accuracy of translation terminology and the work efficiency of translators.

2.2 PROBLEM OF COMPUTATIONAL DISTINSTICS AMBIGUITY

The ambiguity problem is one of the main core problems in Computational Linguistics. In the beginning of machine translation, this problem didn't obtain enough attention. So the machine translations soon fell into an unprecedented crisis and directly lead to the appearance of ALPAC report. It made people further realize the importance of ambiguity problem. Language disambiguation is a challenge in machine translation. And the ambiguity phenomenon is the universal phenomenon in nature language. Ambiguity processing is key to improve the translation quality. 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. 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. 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. 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 Basic knowledge of unascertained mathematics

The Unascertained Mathematics, proposed by Guangyuan Wang in 1990 [11], is a tool to describe the subjective uncertainty quantitatively. It mainly deals with the unascertained information, which differs from the stochastic information, fuzzy information and grey information. The unascertained information refers to the decision-making-demanded information. The information itself has no uncertainty, but because of situation constrain, the decision-maker cannot grasp the total information of them. The decision-maker himself produces the uncertainty. Since 1990s, Kaidi Liu and other scholars have done a lot of work and the Unascertained Mathematics has been successfully used in many fields [11-13].

The definition of unascertained sets is introduced systematically in [11]. Here we will briefly introduce some key points of the unascertained sets.

3.1 MEMBERSHIP FUNCTION CONSTUCTION

The membership function of the unascertained set meets the three principles of measure and it is defined in the topology space (F,E). Yet, the membership function of the Fuzzy Set is a function of a single variable defined in the space of U. The key to the unascertained set is the membership function construction which demands the decision-maker's experience and knowledge background.

3.2 INDEX IDENTIFICTION WEIGHT

The index's weight used in the determining of the composed membership by the single index is and only is the identification weight of the index. In this case, the common methods are used to get the weight value of each assessment index, such as Delphi method, Brainstorming, Analytic Hierarchy Process (AHP) and so on, are helpless. Here, the information entropy is employed to determine the index's identification weight.

Entropy which used to be a thermodynamic concept, it was introduced into information theory in 1948 by C. E. Shannon who put forward the concept of information entropy to measure the level of system chaos or disorder. And Shannon information entropy, which is an objective and applicable method for the determination of weight value, was introduced into the comprehensive assessment. It can calculate weight value of each index more effectively in the comprehensive assessment of marine ecological environment. In the application of Shannon information entropy method, the greater entropy weight indicates greater variation extent of relevant index, much more information and has the greater effect. So, weight value of corresponding index also should be bigger. In contrast, for the smaller entropy weight which has little effect, its weight value should be the smaller [13].

For the discrete stochastic variables, their information entropy is:

$$S = -k\sum_{i=1}^{n} p_i \ln p_i \,, \tag{1}$$

where p_i is the probability and $p_i \ge 0, \sum_{i=1}^{n} p_i = 1$. As u_i $\left(0 \le u_i \le 1, \sum u_i = 1\right)$, suppose $H(\alpha) = -\sum_{k=1}^{K} u_{ijk} \log u_{ijk}$:

$$u_j^{(i)} = 1 + \frac{1}{\log k} H(\alpha)\,, \tag{2}$$

then $\omega^{(i)} = \left(\omega_1^{(i)}, \omega_2^{(i)}, \ldots, \omega_m^{(i)}\right)$ is the weight of the $I_1, I_2, \ldots, I_m$, where:

$$\omega_j^{(i)} = \frac{u_j^i}{\sum_{j=1}^{m} u_j^{(i)}}\,. \tag{3}$$

For the identification principle, if the ranks are orderly, the principle of maximum degree of membership is not applicable and the credible identification is often used.

4 Quality assessment model of machine translation

4.1 INDEX SYSTEM

The establishment of the assessment index system should use the system engineering theory and it should be closely associated with the practice. After consulting the specialists, seven independent factors constitute the main assessment indexes. The indexes include readability, formality, convey degree of the related implications, convey degree of the implicit implication, covertly erroneous error, overtly erroneous error and meaning function. The system of assessment indexes is shown in Figure 1.

4.2 MODEL BASED ON UNASCERTAINED MEASURE

Suppose $x_1, x_2, \ldots, x_n$ are n translation results, $I_1, I_2, \ldots, I_m$ are indexes for the assessment of x_i, $I = \{I_1, I_2, \ldots, I_m\}$, x_{ij} is the observed value of x_i under index I_j, and c_k is the k^{th} comment $(1 \le k \le K)$.

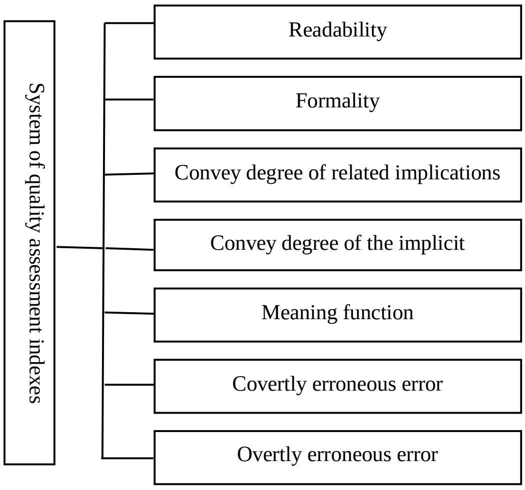

FIGURE 1 The index systems of quality assessment of translation

4.2.1 Single-index unascertained measure.

μ_{ijk} is the degree that the observed value x_{ij} of x_i belongs to the assessment rank c_k.

The Delphi method is employed to get the scores of every factor. The number of the specialists is k. Every specialist should rank the degree that $I_j\,(1 \le j \le m)$ belongs to $c_k\,(1 \le k \le K)$ by using 0-10. If the k^{th} specialist thinks that the degree I_j belongs to c_k is x_{ijk}, $\sum_{k=1}^{K} x_{ijk} = 10$, then $\mu_{ijk} = \frac{x_{ijk}}{10}$ is the unascertained measure. After obtaining the comprehensive indexes of assessment system, standardized processing of the data should be firstly finished. Suppose the research plan is x_{ij} $(i = 1,\ldots,n;\ j = 1,\ldots,m)$. It denotes that there are i samples and j indexes in the research plan. Based on the characteristics of assessment target, the indexes are divided into the positive type and negative type. The positive type indexes are the indexes whose values are the bigger the better. The negative type indexes refer to the indexes whose values are the smaller the better.

For the positive indexes, the normalization is as follows:

$$x_{ij} = \frac{x_{ij}^{'} - \min x_j}{\max x_j - \min x_j}. \tag{4}$$

For the negative indexes, the normalization is as follows:

$$x_{ij} = \frac{\max x_j - x_{ij}^{'}}{\max x_j - \min x_j}. \tag{5}$$

where, $\max x_j$ and $\min x_j$ represent the maximum and minimum value of x_j respectively.

After the standardized processing, the standardized matrix $x_{ij}^{'}$ be obtained.

The single-index measure assessment matrix of x_i is:

$$\left(\mu_{ijk}\right)_{m\times k} = \begin{bmatrix} \mu_{i11} & \mu_{i12} & \cdots & \mu_{i1k} \\ \mu_{i21} & \mu_{i22} & \cdots & \mu_{i2k} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{im1} & \mu_{im2} & \cdots & \mu_{imk} \end{bmatrix}, (i = 1,2,...,n). \tag{6}$$

4.2.2 Identification weight of the index.

$\omega_j^{(x)}$ is the identification weight and:

$$\omega_j^{(x)} = \frac{L_j^{(i)}}{\sum_{i=1}^{m} L_j^{(i)}}. \tag{7}$$

4.2.3 Comprehensive assessment system.

The common unascertained membership functions are as follows and four common measure functions, includes straight line distribution, parabola distribution, exponent distribution and sine distribution are shown in Figures 2-5.

$$\begin{cases} \mu_i(x) = \begin{cases} \frac{-x}{a_{i+1} - a_i} + \frac{a_{i+1}}{a_{i+1} - a_i} & a_i < x \le a_{i+1} \\ 0 & x > a_{i+1} \end{cases} \\ \mu_{i+1}(x) = \begin{cases} 0 & x \le a_i \\ \frac{x}{a_{i+1} - a_i} + \frac{a_i}{a_{i+1} - a_i} & a_i < x \le a_{i+1} \end{cases} \end{cases}, \tag{8}$$

$$\begin{cases} \mu_i(x) = \begin{cases} 1 - \left(\frac{x - a_i}{a_{i+1} - a_i}\right)^2 & a_i < x \le a_{i+1} \\ 0 & x > a_{i+1} \end{cases} \\ \mu_{i+1}(x) = \begin{cases} 0 & x \le a_i \\ \left(\frac{x - a_i}{a_{i+1} - a_i}\right)^2 & a_i < x \le a_{i+1} \end{cases} \end{cases}, \tag{9}$$

$$\begin{cases} \mu_i(x) = \begin{cases} 1 - \frac{1 - e^{x - a_i}}{1 - e^{a_{i+1} - a_i}} & a_i < x \le a_{i+1} \\ 0 & x > a_{i+1} \end{cases} \\ \mu_{i+1}(x) = \begin{cases} 0 & x < a_i \\ \frac{1 - e^{x - a_i}}{1 - e^{a_{i+1} - a_i}} & a_i < x \le a_{i+1} \end{cases} \end{cases}, \tag{10}$$

$$\begin{cases} \mu_i(x) = \begin{cases} 1 - \frac{1 - e^{x - a_i}}{1 - e^{a_{i+1} - a_i}} & a_i < x \le a_{i+1} \\ 0 & x > a_{i+1} \end{cases} \\ \mu_{i+1}(x) = \begin{cases} 0 & x < a_i \\ \frac{1 - e^{x - a_i}}{1 - e^{a_{i+1} - a_i}} & a_i < x \le a_{i+1} \end{cases} \end{cases}. \tag{11}$$

After the single index matrix and the identification weight are derived, the comprehensive assessment vector μ^i can be derived:

$$\mu^i = \left(\mu_{i1}, \mu_{i2}, \cdots, \mu_{in}\right) = \omega^i \cdot \left(\mu_{ijk}\right)_{m\times K}, \tag{12}$$

where μ^i is the unascertained classification, in order to obtain the certainty classification, the identification is needed.

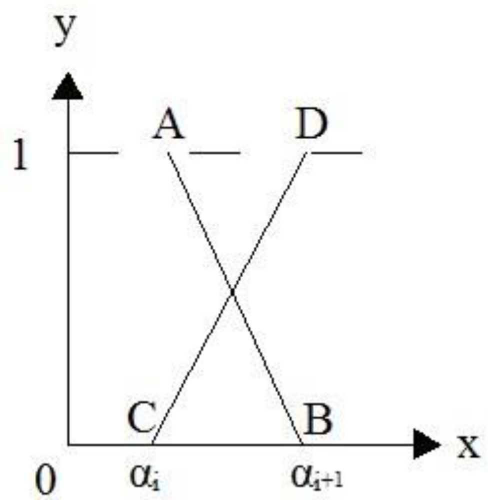

FIGURE 2 Straight line distribution

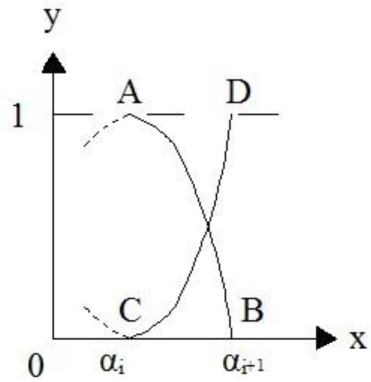

FIGURE 3 Parabola distribution

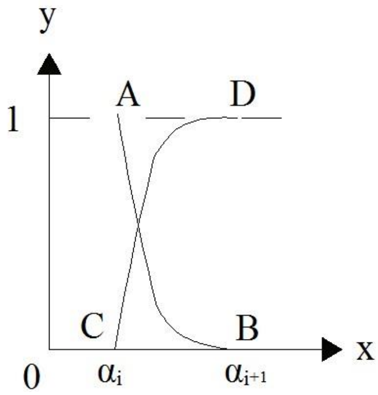

FIGURE 4 Exponent distribution

5 Practice application

Using the method mentioned above, we finished the reliability assessment of a translation result. Five

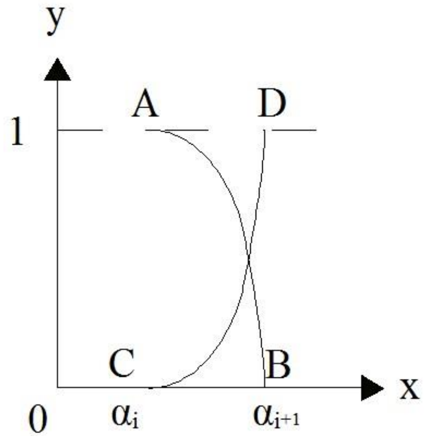

FIGURE 5 Sine distribution

4.2.4 Principle of identification.

Because the classification of the comment ranks is orderly, e.g. c_k is "better" than c_{k+1}, the identification principle of "maximum measure" is not available. The credible identification principle is needed. Let the credible identification be λ, it is always 0.6 or 0.7. If:

$$k_0 = \min\left[\left(\sum_{l=0}^{k} \mu_{jl}\right) \geq \lambda, k = 0,1,...,K-1\right], \tag{13}$$

then x_i belongs to the rank c_{k0}.

specialists are invited to give the values of indexes, which are listed in Table 1.

TABLE 1 Scores given by the specialists

Index \ Specialist	1	2	3	4	5
Readability	3.7	4.1	3.5	3.7	3.2
Formality	3.5	3.9	3.4	3.6	3.2
Convey degree of related implications	2.9	3.0	2.9	2.8	3.1
Convey degree of the implicit implication	3.1	3.2	3.0	2.9	3.0
Meaning function	3.0	2.8	3.1	3.0	3.1
Covertly erroneous error	3.9	3.8	3.4	4.0	3.8
Overtly erroneous error	3.5	3.4	3.3	3.8	3.5

Thus the policy-making matrix can be obtained:

$$x'_{7\times5} = \begin{bmatrix} 3.7 & 4.1 & 3.5 & 3.7 & 3.2 \\ 3.5 & 3.9 & 3.4 & 3.6 & 3.2 \\ 2.9 & 3.0 & 2.9 & 2.8 & 3.1 \\ 3.1 & 3.2 & 3.0 & 2.9 & 3.0 \\ 3.0 & 2.8 & 3.1 & 3.0 & 3.1 \\ 3.9 & 3.8 & 3.4 & 4.0 & 3.8 \\ 3.5 & 3.4 & 3.3 & 3.8 & 3.5 \end{bmatrix}.$$

Then, the single-index unascertained measure is obtained:

$$\left(\mu_{ijk}\right)_{m\times k} = \begin{bmatrix} 0.20 & 0.23 & 0.19 & 0.20 & 0.18 \\ 0.20 & 0.22 & 0.19 & 0.20 & 0.19 \\ 0.20 & 0.20 & 0.20 & 0.19 & 0.21 \\ 0.20 & 0.21 & 0.20 & 0.19 & 0.20 \\ 0.20 & 0.19 & 0.21 & 0.20 & 0.20 \\ 0.21 & 0.20 & 0.18 & 0.21 & 0.20 \\ 0.20 & 0.19 & 0.19 & 0.22 & 0.20 \end{bmatrix}.$$

The weight can be derived and shown in Table 2.

TABLE 2 Weight value of each assessment index

Assessment Index	Weight Value	Assessment Index	Weight Value
Readability	0.16	Meaning function	0.13
Formality	0.15	Covertly erroneous error	0.16
Convey degree of related implications	0.12	Overtly erroneous error	0.15
Convey degree of the implicit implication	0.13		

The assessment result can be obtained:

$$\mu^i = (0.20 \quad 0.21 \quad 0.19 \quad 0.20 \quad 0.20).$$

Let $\lambda = 0.7$, the final assessment results can be obtained: the translation belongs to the third rank, which means "normal".

Using the fuzzy comprehensive assessment [14], it belongs to the second rank, "better". Practice demonstrates that the result obtained by using the unascertained measure is more rational. The reasons are the unascertained measure pays more attention to the order of the assessment space and gives the rational rank and credible identification principles. All of those are not possessed by fuzzy comprehensive assessment.

6 Conclusion

The quality assessment of the computational translation can eliminate the possibility of failure and it is the key to ensure the quality of translation. In order to overcome the defects of subjectivity of common methods and evaluate effectively computational translation quality, here the unascertained measure model was established and employed in practice application. The application results show that it can easily realize the assessment without any assumption. And this study has great significance in improvement of machine translation and other fields.

Acknowledgments

The work was supported by the Program of Selection and Cultivating of Disciplinary Talents of Colleges and Universities in Hebei Province (BR2-206).

References

[1] Hutchins W J, Somers H L 1992 *An Introduction to Machine Translation* Academic Press: London
[2] Yuyang M, Shuqi S, Junguo Z, Sheng L, Tiejun Z, Xiaoning Z 2011 *Journal of Computer Science and Technology* **26**(1) 57-67
[3] King M, Popescu-Belis A, Hovy E 2003 *Proceedings of MT Summit IX, New Orleans, LA. Sept. 2003* 224-31
[4] Church K, Hovy E 1993 *Machine Translation* **8** 239-258
[5] Lavie A, Sagae K, Jayaraman S 2004 *Proceedings of AMTA 2004* Washington DC September 2004
[6] Papineni K, Roukos S, Ward T, Zhu W 2002 *ACL-2002: the 40th Annual meeting of the Association for Computational Linguistics* 311-8
[7] Turian J, Shen L, Melamed I D 2003 *Proceedings of the MT Summit IX* New Orleans USA 386-93
[8] Han A L F, Wong D F, Chao L S 2012 *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)* Mumbai India 441-50
[9] Han A L F, Wong D F, Chao L S, Lu Y, He L, Wang Y, Zhou J 2013 *Proceedings of the Eighth Workshop on Statistical Machine Translation ACL-WMT13* Sofia Bulgaria 414-21
[10] Wang G 1990 *Journal of Civil Engineering Institute of Haerbin* **23**(4) 1-9
[11] Wu H and Wu H 2000 *Journal of Quantitative & Technical Economics* **2** 21-4
[12] Liu K, Cao Q, Pang Y 2004 *ACTA Automatic Sinica* **30**(5) 747-56
[13] Qiu W 2002 Management Decision Making and Applied Entropy *China Machine Press: Beijing*
[14] Chen S 1998 Theory of Project Fuzzy Sets and Its Application *National Defence Industry Press: Beijing*

Authors

Juanjuan Suo, born in December, 1977, Hebei Province, P.R. China

Current position, grades: Teacher of Hebei University of Engineering.
University studies: M.D. at Hebei University in 2006.
Scientific interest: computational linguistics.
Publications: more than 10.

Huimin Dong, born in September, 1991, Shandong Province, P.R. China

Current position, grades: Postgraduate at Hebei University of Engineering.
University studies: B.E. at Hebei University in 2012.
Scientific interest: computational linguistics.