# Study on the comprehensive evaluation method of machine translation quality

*Sun Yiqun* [1,*]

[1]Ocean University of China, Qingdao 266071, China

**Abstract.** The accurate evaluation of machine translation quality is the main basis of machine translation systems research and development. Based on the analysis of three translation quality assessment factors, i.e. vocabulary, grammar and discourse, this paper built twelve indicators, which classified into five levels of evaluation index system. Meanwhile questionnaire was designed as well. The people with different age, gender, occupation and cultural level were selected as respondents. Then, the data for machine translation evaluation was obtained from questionnaire. The analytic hierarchy process was applied to determine the weighted vector of machine translation evaluation factors. In addition, on this basis, fuzzy mathematics theory was used to establish the comprehensive evaluation vector of machine translation quality. As a result, the quantity comprehensive evaluation for the translation text quality was realized.

## 1 INTRODUCTION

Machine translation is exactly the key to the cultural exchange and information searching among different languages. The machine translation has been a major hot spot in the computer science and computing-related fields. An accurate evaluation of translation quality is the main basis for the system development. After a few years' development, study on the evaluation method of machine translation quality has achieved fruitful results. Evaluation of the translation's similarity between references and machine translation is one of the main methods[1].For instance, the automatic evaluation method developed by IBM named BLEU[2];the developed by Yokoyama based on bidirectional machine translation[3], the method of Yasuda, Akiba and Papineni based on N-gram language models for computing sentence similarity[4][5]. These evaluation methods have two problems: first of all, they need the help of a third party-an artificial translation as reference. Therefore, the results of the evaluation depend largely on the quality of the artificial translation, which often cannot be guaranteed. As a result, the accuracy of this evaluation method is random. Secondly, during text analysis and comparison, the current methods focus on comparing the similarities between language units at all levels, namely the similarities between words, phrases and sentences in compositions. But after all, the language is flexible, because of its lexical, grammatical, syntactic and contextual changes, the meaning will be very different. Accordingly, these evaluation methods are limited to a micro level and lack of analysis of the article from a macro point of view. Meanwhile, the machine can't perceive the specific context and understand the implication of the article. Therefore, the comprehensive evaluation method has drawn attention of the public and become one of the hot spots in research.

Based on the analysis above, from the point of narrowing the gap between subjective and objective evaluation as well as enhancing the reliability of the evaluation results, integrate fuzzy comprehensive evaluation method and analytic hierarchy process. Use the analytic hierarchy process to design the evaluation index system of quality of machine translation with the hierarchical structure, and determine the various levels and the weight of evaluation indexes. In this paper, we design the questionnaire, get the data of readers' evaluation of translation quality through questionnaire survey and give the data fuzzy a comprehensive evaluation by different level, and finally we got quantitative evaluation results.

Weight is the measure of the relative importance of index, and it is an important factor to affect the comprehensive evaluation result. All of the evaluation factors are compared in pairs based on the principle of analytic hierarchy process to construct the comparative judgment matrix, and we examined the consistency of the comparative judgment matrix, finally we got the weight of each evaluation index and each evaluation level.

## 2 Evaluation index weight value calculation and consistency check

### 2.1. Calculation of evaluation index weight value

After establishing the hierarchy index system, construct judgment matrix according to the subordinate relations between the upper and lower levels. Take an evaluation

level as the guidelines, based on its dominance relation to the next level factors, the relative importance of the next levels' factors to standards (evaluation levels) are compared in pairs and give a certain score. Comparison scale method is used to determine the score, and its standard is shown in Table 1[6].

**Table 1**. Score standard of comparison scale method

| Scale | Meanings |
|---|---|
| 1 | The two elements are equally important when compared |
| 3 | The former element is a bit more important than the later one when compared |
| 5 | The former element is obviously more important than the later one when compared |
| 7 | The former element is strongly more important than the later one when compared |
| 9 | The former element is extremely more important than the later one when compared |
| 2,4,6,8 | The intermediate value of the adjacent judgment above |
| Reciprocal | If the ratio of the importance of the i and j is aij, then the importance of element j and i is aji=1/aij. |

A few compared factors constitute $A$ judgment matrix when compared to a certain evaluation criterion. $A=(a_{ij})_{n\times n}$ ($a_{ij}$ is the scale of the ratio of the importance of the $B_i$ and $B_j$, and $a_{ii}=1$).

$$A=\left(a_{ij}\right)_{n\times n}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \quad (1)$$

After get comparative judgment matrix $A$, sum and product method is used to calculate sort of numerical of relative importance of the evaluation level to evaluation index according to the characteristics of the judgment matrix, named the weight value.

First of all, each column of the comparative judgment matrix is summed up, and then they are normalized processed by the following formula:

$$\bar{a}_{ij}=\frac{a_{ij}}{\sum_{k=1}^{n} a_{kj}} \quad i,\ j=1,2,\ldots,n \quad (2)$$

$$A'=\left(\bar{a}_{ij}\right)_{n\times n}=\begin{bmatrix} \bar{a}_{11} & \bar{a}_{12} & \cdots & \bar{a}_{1n} \\ \bar{a}_{21} & \bar{a}_{22} & \cdots & \bar{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{a}_{n1} & \bar{a}_{n2} & \cdots & \bar{a}_{nn} \end{bmatrix} \quad (3)$$

Each line of $A'$ are summed up to $M$,

$$M=(\sum_{j=1}^{n}\bar{a}_{1j},\sum_{j=1}^{n}\bar{a}_{2j},\ldots,\sum_{j=1}^{n}\bar{a}_{nj},)^{T} \quad (4)$$

$M$ is normalization processed to get characteristic vector $W$ of $A$, component of $W$ is the relative importance of the arrangement of numerical evaluation indexes, namely the weight value.

$$W=(\frac{\sum_{j=1}^{n}\bar{a}_{1j}}{\sum_{i=1}^{n}(\sum_{j=1}^{n}\bar{a}_{ij})},\frac{\sum_{j=1}^{n}\bar{a}_{2j}}{\sum_{i=1}^{n}(\sum_{j=1}^{n}\bar{a}_{ij})},\ldots,\frac{\sum_{j=1}^{n}\bar{a}_{nj}}{\sum_{i=1}^{n}(\sum_{j=1}^{n}\bar{a}_{ij})})^{T} \quad (5)$$

### 2.2 Consistency check of Comparative judgment matrix

For matrix $A=(a_{ij})$ n×n is artificially assignment, the consistency check is necessary for the judgment to evaluate the reliability of the judgment matrix. Consistency ratio $CR$ is usually measured to determine the consistency.

$$CR=\frac{CI}{RI} \quad (6)$$

$$CI=\frac{\lambda_{max}-n}{n-1} \quad (7)$$

The biggest characteristic root for judgment matrix $A$ is $\lambda_{max}$. $\lambda_{max}=\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{n}a_{ij}\cdot w_i}{w_i}$. $n$ is the order for judgment matrix. Wi is the characteristics of the judgment matrix $A$ vector-valued (namely weights).

$RI$ is the average random consistency index. Value standard given by T.L.Saaty is commonly used, as showed in Table 2.

**Table 2**. Average random consistency index

| *n* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *RI* | 0 | 0 | 0.58 | 0.89 | 1.12 | 1.24 | 1.36 | 1.41 | 1.45 | 1.49 | 1.52 |

When $CR<0.1$, the consistency of judgment matrix A is acceptable.

## 3 The fuzzy evaluation method of translation quality

For the translation to be evaluated, set m evaluation factors ($u_1, u2\ldots u_m$), evaluation results of each evaluation factor are divided into $n$ levels (1, 2...$n$). The number of people ranked level $j$ in $u_i$ is marked as $R_{ij}$($i$=1,2……,m ; $j$=1,2,…,n), the evaluation results is showed in fuzzy matrix $RR_{(m^*\,n)}$.

$$RR=\begin{pmatrix} R_{11} & R_{12} & .. & R_{1j} & .. & R_{1n} \\ .. & .. & .. & .. & .. & .. \\ R_{i1} & R_{i2} & .. & R_{ij} & .. & R_{in} \\ .. & .. & .. & .. & .. & .. \\ R_{m1} & R_{m2} & .. & R_{mj} & .. & R_{mn} \end{pmatrix} \quad (8)$$

Each evaluation results of evaluation factors, namely each row of the fuzzy matrix is normalized to get a fuzzy matrix $R$,

$$R=\begin{pmatrix} r_{11} & r_{12} & .. & r_{1j} & .. & r_{1n} \\ .. & .. & .. & .. & .. & .. \\ r_{i1} & r_{i2} & .. & r_{ij} & .. & r_{in} \\ .. & .. & .. & .. & .. & .. \\ r_{m1} & r_{m2} & .. & r_{mj} & .. & r_{mn} \end{pmatrix} \quad (9)$$

where $r_{ij}=\frac{R_{ij}}{\sum_{k=1}^{n}R_{ik}}$

These evaluation factors have different weights. The weight of factors $u_i$ is $q_i$; these weights for $q_1$ and $q_2$... $qi$... $q_n$ are normalized as follows,

$$X=(x_1, x_2,\ldots x_m) \quad (10)$$

where $x_i=\frac{q_i}{\sum_{k=1}^{n}q_k}$

$P$ is the comprehensive evaluation result

$$P=X\cdot R \quad (11)$$

$P$ is the synthetic of weighted fuzzy vector $X$ and fuzzy evaluation matrix $R$,

$P=(p_1, p_2 \ldots p_m)$ ， $p_i = \bigvee_{k=1}^{m}(x_k \wedge r_{ik})$

The evaluation results $P$ is normalized, the comprehensive evaluation results $Y$ is obtained,

$$Y=(y_1, y_2 \ldots y_m) \quad (12)$$

where $y_i = \frac{p_i}{\sum_{k=1}^{n} p_k}$

According to the comprehensive evaluation results $Y$, set threshold $\lambda_1$、$\lambda_2$、…、$\lambda_m$, and make the conclusion about translation evaluation.

If $(y_1 \geq \lambda_1)$ Then the quality evaluation for translation is level 1;

Else If $(y_1 + y_2 \geq \lambda_2)$Then the quality evaluation for translation is level 2;

……..

## 4 Quality evaluation of Machine translation

### 4.1 Evaluation index system and questionnaire design of machine translation quality

The quality of the translation needs to be judged by three main factors: "vocabulary", "syntax", and "discourse"[7], words are judged from the four aspects :word meaning collocation, rhetoric, technical terms, the use of dialect; Syntax is mainly inspects whether the grammar of the translation is correct; discourse Include textual cohesion and coherence, intentionality, acceptability, informational, context and inter-textuality. Therefore, the translation quality evaluation is decomposed into 12 different evaluation indexes in this article, and these evaluation indexes formed three evaluation levels according to the internal relationship and the subordinate relations between them, and then build up evaluation index system which take the quality of translation as the evaluation target, as shown in Table 3.

**Table 3.** Evaluation index system of Chinese translation quality

| Evaluation target | Evaluation level | Evaluation indexes |
|---|---|---|
| The Chinese translation quality Y | Vocabulary A1 | B1: Word meaning collocation |
| | | B2:Rhetoric |
| | | B3:Technical terms |
| | | B4:The use of dialect |
| | Syntax A2 | B5:Grammar |
| | Discourse A3 | B6:Textual cohesion |
| | | B7:Coherence |
| | | B8:Intentionality |
| | | B9:Acceptability |
| | | B10:Informational |
| | | B11:Context |
| | | B12:Inter-textuality. |

This paper designs a questionnaire based on the evaluation index system of Chinese translation quality, which contains three major categories of indicators and twelve minor categories of indicators. According to the concept of Likert scale [8], each minor categories of indicators contains a total of five levels: from the best to the worst. Level 1: They are complete transferred from the original text information; only minor revision needed to reach professional standard. Level 2: They are almost complete transferred; there may be one or two insignificant inaccuracies; requires certain amount of revision to reach professional standard. Level 3: They have general ideas but with a number of lapses in accuracy; needs considerable revision to reach professional standard. Level 4: Sentences can be well understood, a few of the content of the original should be speculated that can't fully express the original meaning .Level 5: They are totally inadequate transfer of the original text content; the translation is not worth revising. The translation is mostly incoherent.

### 4.2 The calculation of weight value of machine translation quality evaluation index

According to the weight of the evaluation index system in section 2.1, the consistency test of the comparison judgment matrix is conducted according to the method of section 2.2, and the results are shown in Table 4 and Table 5.

**Table 4**. The weight value *W* of index system of the machine translation quality evaluation

| Evaluation target | Evaluation level | Weight *W* | Evaluation index | *W* related to evaluation level |
|---|---|---|---|---|
| The Chinese translation quality Y | vocabulary A1 | 0.4071 | B1 | 0.2361 |
| | | | B2 | 0.2361 |
| | | | B3 | 0.1806 |
| | | | B4 | 0.3472 |
| | Syntax A2 | 0.2643 | B5 | 1.0000 |
| | Discourse A3 | 0.3286 | B6 | 0.1256 |
| | | | B7 | 0.1256 |
| | | | B8 | 0.1078 |
| | | | B9 | 0.2064 |
| | | | B10 | 0.0994 |
| | | | B11 | 0.1702 |
| | | | B12 | 0.1651 |

**Table 5.** The result of the consistency check of comparative judgment matrix

| Comparative judgment matrix | $\lambda_{max}$ | *CI* | *CR* | Result |
|---|---|---|---|---|
| A1 | 4.2500 | 0.0833 | 0.0937 | Acceptable |
| A2 | Only one evaluation index which don't need to be inspected | | | |
| A3 | 7.3339 | 0.0557 | 0.04216 | Acceptable |
| Y | 0537 | 0.0557 | 0.04216 | Acceptable |

### 4.3 Evaluation results and analysis

Four software is chosen from the translation software which is widely used , such as Google translation, Bing translation, Baidu translation, Youdao translation and General translation, Lingoes, etc. and marked them as A, B, C, D, respectively. Obtain the evaluation data of translation quality by means of Internet survey data. O

Henry's short stories “The Gift of The Magi” was chosen as the test text.

Choose 100 respondents with different age, gender, occupation and cultural level. Questionnaire survey was conducted to determine the number of choices of each level. The statistical results of software A were shown in Table 6.

**Table 6.** Statistical results of the quality of the translation software A

| level | Index | The answer | | | | |
|---|---|---|---|---|---|---|
| | | L1 | L2 | L3 | L4 | L5 |
| A1 | B1 | 96 | 3 | 1 | | |
| | B2 | 94 | 4 | 2 | | |
| | B3 | 99 | 1 | | | |
| | B4 | 100 | | | | |
| A2 | B5 | 89 | 6 | 1 | 3 | 1 |
| A3 | B6 | 80 | 6 | 5 | 7 | 2 |
| | B7 | 78 | 5 | 8 | 4 | 5 |
| | B8 | 82 | 9 | 3 | 3 | 3 |
| | B9 | 99 | 1 | | | |
| | B10 | 92 | 7 | | | 1 |
| | B11 | 86 | 6 | 5 | 2 | 1 |
| | B12 | 92 | 5 | | 1 | 2 |

According to the statistics Table 6, get translation quality of fuzzy evaluation matrix $R_{YA}$ of software A:

$$R_{YA} = \begin{vmatrix} 0.9725 & 0.0200 & 0.0075 & 0.0000 & 0.0000 \\ 0.8900 & 0.0600 & 0.0100 & 0.0300 & 0.0100 \\ 0.8700 & 0.0557 & 0.0300 & 0.0243 & 0.0200 \end{vmatrix}$$

Have the weight vector $W$ and evaluation fuzzy matrix $R_{YA}$ on synthesis arithmetic, get the comprehensive evaluation vector of translation quality software A

$$Y_A = W \cdot R_{YA}$$

$$= (0.7780,0.0952,0.0476,0.0476,0.0317)$$

Translation quality evaluation target and evaluation level by the comprehensive fuzzy evaluation method of software A, B, C, D are processed as the methods above.

Table 7 and Figure 1 are the evaluation target for the quality of the evaluation results, L1 is assessment for level 1 in figure 2, the percentage of the total number of L2 for assessment for level 2 percentage. It can be seen that the L1 of software A, B, C, D were 0.7780, 0.6988, 0.6305 and 0.5684, the L1+L2 were 0.8732, 0.8128, 0.7333 and 0.6727, as a result, the order of translation software quality from good to bad order is A, B, C, D. L1 and L1+L2 of software A are respectively 1.37 times and 1.30 times of software D, so software D needs to be improved.

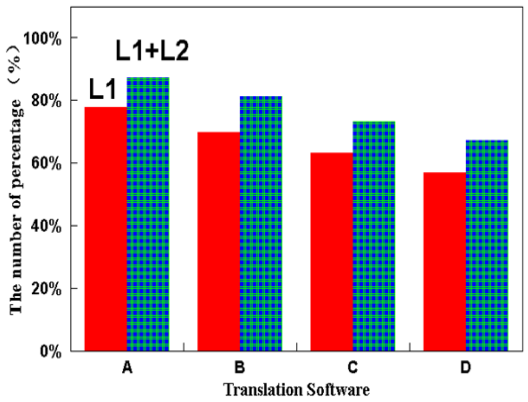


**Fig. 1.** The evaluation results of evaluation target of translation quality on level 1 and level 2

**Table 7.** The comprehensive fuzzy evaluation results of evaluation target of translation quality

| Software | L1 | L2 | L3 | L4 | L5 |
|---|---|---|---|---|---|
| A | 0.7780 | 0.0952 | 0.0476 | 0.0476 | 0.0317 |
| B | 0.6988 | 0.1140 | 0.0855 | 0.0529 | 0.0489 |
| C | 0.6305 | 0.1028 | 0.0964 | 0.0900 | 0.0803 |
| D | 0.5684 | 0.1043 | 0.0956 | 0.1130 | 0.1188 |

Table 8 and Figure 2 are the comprehensive fuzzy evaluation results for quality assessment of the level of vocabulary. It can be seen that the dates of level 1 of software A, B, C, D were 85.27%, 81.27%, 51.27% and 63.24%, the sum of level 1 and level 2 data were 95.09%、90.63%、63.98% and 81.62%, as a result, the order of translation software quality on the level of vocabulary from good to bad order is A, B, C, D Software A is as good as software Bon the level of vocabulary.L1 and L1 + L2 of software A are respectively 1.66 times and 1.49 times of software C, so software C needs to be improved on the level of vocabulary. Although software D is the worst in the field of comprehensive evaluation.

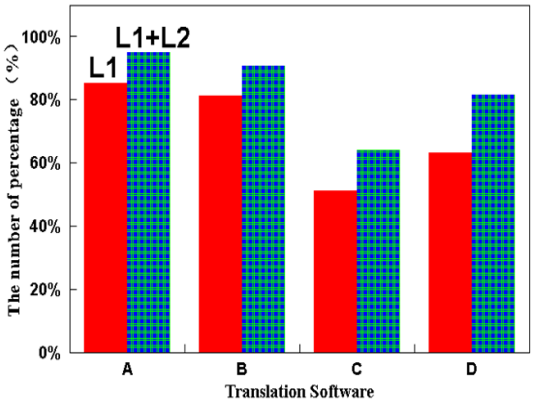


**Fig.2.** The evaluation results of the vocabulary of translation quality on the level 1 and level 2

**Table 8.** The comprehensive fuzzy evaluation results of translation quality on the level of vocabulary

| Software | L1 | L2 | L3 | L4 | L5 |
|---|---|---|---|---|---|
| A | 0.8527 | 0.0982 | 0.0491 | 0.0000 | 0.0000 |
| B | 0.8127 | 0.0936 | 0.0468 | 0.0234 | 0.0234 |
| C | 0.5127 | 0.1181 | 0.1181 | 0.1329 | 0.1181 |
| D | 0.6324 | 0.1838 | 0.1225 | 0.0306 | 0.0306 |

Table 9 and Figure 3 are comprehensive fuzzy evaluation results of quality assessment of the level of discourse, we can see that discourse of L1 software in A, B, C, D were 41.58%, 39.97%, 63.24% and 56.33% respectively; L1 + L2 for software A, B, C, D were 59.71%, 55.46%, 81.62% and 80.89% respectively. Therefore, discourse translation of C is the best, followed by D, came in third place was A, B is the worst. Therefore, although the overall evaluation of A is best, but it still has room to improve in terms of discourse, and overall evaluation of software B ranked second, worst in terms of discourse, the discourse presses for improvement in order to enhance software translation quality.

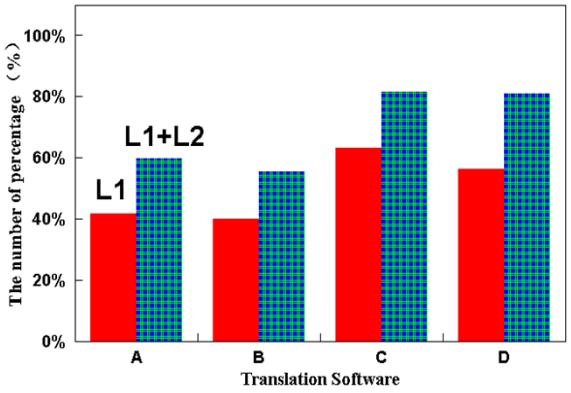


**Fig.3**. The evaluation results of the syntax of translation quality on the level 1 and level 2

**Table 9.** The comprehensive fuzzy evaluation results of translation quality on the level of syntax

| Software | L1 | L2 | L3 | L4 | L5 |
|---|---|---|---|---|---|
| A | 0.4158 | 0.1813 | 0.1612 | 0.1410 | 0.1007 |
| B | 0.3997 | 0.1549 | 0.1549 | 0.1549 | 0.1356 |
| C | 0.6324 | 0.1838 | 0.1225 | 0.0306 | 0.0306 |
| D | 0.5633 | 0.2456 | 0.0819 | 0.0546 | 0.0546 |

## 5 Conclusions

(1) Established the translation quality evaluation mathematical model based on analytic hierarchy process and fuzzy mathematics theory, which laid a foundation for quantitative evaluation of machine translation quality.

(2) Using analytic hierarchy process, as well as quantifying the weight of evaluation of machine translation related to the weight of evaluation target and evaluation level by constructing comparative judgment matrix in pairs, overcome the disadvantages of determining weight value by artificial.

(3) Based on the data of evaluation of translation quality, which is obtained through questionnaire survey, Fuzzy matrix is established to evaluate the quality of machine translation using the fuzzy mathematics method. A comprehensive evaluation vector is obtained and the quantity evaluation of the quality of the translation text is realized.

(4) The fuzzy comprehensive evaluation method can be used to analyze the differences of various indicators of quality of translation software, help readers to screen software with good translation quality and provide some aspects of the software defects for software developers as to improve the software design and the quality of software.

## ACKNOWLEDGMENTS

This article has been funded by the national social science foundation of 2016 (project No.:16BYY044).

## References

1. B.Wang: Research on Technologies of Evaluation and Diagnosis of Machine Translation Dissertation for the Doctoral Degree in Engineering, 2010 (In Chinese)
2. K.Papineni,S.Roukos,T.Ward, W.Zhu.BLEU: a method for automatic evaluation of MT. IBM research division, T J Watson Research Centre, Research Report: Computer Science RC22176 (W0109-022), 2001.
3. S. Yokoyama, H. Kashioka, etc. An automatic evaluation method for machine translation using two-way MT. MT summit conference. Santiago de Compostela, 2001: 568~573
4. K. Yasuda, F. Sugaya, etc. An automatic evaluation method of translation quality using translation answer candidates queried from a parallel corpus. MT summit conference. Santiago de Compostela, 2001: 373~378
5. Y. Akiba, K. Imamura, E. Sumita. Using multiple edit distances to automatically rank machine translation output. MT summit conference. Santiago de Compostela, 2001: 15~20
6. Thomas L.Saaty, Luis G. Vargas: Models, Methods, Concepts & Applications of the Analytic Hierarchy Process, Springer Science+Business Media New York, 2012 pp.23–40, 100–102, 149–158, 161–165, 203–247.
7. H. L Zhang, Parametric Analysis on Evaluation of Translation Quality, Yilin, 2011. No8 p70-76
8. Wuensch, Karl L. What is a Likert Scale? And How Do You Pronounce 'Likert?'. East Carolina University. October 4, 2005.