Study on the comprehensive evaluation method of machine translation quality

Sun Yiqun1,*

1Ocean 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

<table>
<thead>
<tr>
<th>Scale</th>
<th>Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The two elements are equally important when compared</td>
</tr>
<tr>
<td>3</td>
<td>The former element is a bit more important than the later one when compared</td>
</tr>
<tr>
<td>5</td>
<td>The former element is obviously more important than the later one when compared</td>
</tr>
<tr>
<td>7</td>
<td>The former element is strongly more important than the later one when compared</td>
</tr>
<tr>
<td>2, 4, 6, 8</td>
<td>The intermediate value of the adjacent judgment above</td>
</tr>
</tbody>
</table>

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 = (aij) where aij is the scale of the ratio of the importance of the B and Bj, and aii=1).

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

(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:

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

(2)

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

(3)

Each line of A are summed up to M,

\[ M = (\sum_{i=1}^{n} \bar{a}_{ij}, \sum_{i=1}^{n} \bar{a}_{2j}, \ldots, \sum_{i=1}^{n} \bar{a}_{nj})^{T} \]  

(4)

\[ M' = \left( \frac{\sum_{i=1}^{n} \bar{a}_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} \bar{a}_{ij}}, \frac{\sum_{i=1}^{n} \bar{a}_{2j}}{\sum_{i=1}^{n} \sum_{j=1}^{n} \bar{a}_{ij}}, \ldots, \frac{\sum_{i=1}^{n} \bar{a}_{nj}}{\sum_{i=1}^{n} \sum_{j=1}^{n} \bar{a}_{ij}} \right)^{T} \]  

(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.

\[ CR = \frac{CI}{RI} \]  

(6)

\[ CI = \frac{\lambda_{max} - n}{n-1} \]  

(7)

The biggest characteristic root for judgment matrix A is \( \lambda_{max} \). \( \lambda_{max} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i} w_{j}}{w_{i} w_{j}} \) 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

<table>
<thead>
<tr>
<th>n</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR</td>
<td>0</td>
<td>0</td>
<td>0.58</td>
<td>0.90</td>
<td>1.22</td>
<td>1.64</td>
<td>1.96</td>
<td>1.48</td>
<td>1.49</td>
<td>1.52</td>
<td></td>
</tr>
</tbody>
</table>

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 (u1, u2,...,um), evaluation results of each evaluation factor are divided into n levels (1, 2,...,n). The number of people ranked level j in ui is marked as \( R_{ij}(i=1,2,...,m; j=1,2,...,n) \), the evaluation results is showed in fuzzy matrix \( RR_{m \times n} \).

\[ RR = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1j} & \cdots & R_{1n} \\ R_{21} & R_{22} & \cdots & R_{2j} & \cdots & R_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ R_{m1} & R_{m2} & \cdots & R_{mj} & \cdots & R_{mn} \end{bmatrix} \]  

(8)

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

\[ R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1j} & \cdots & r_{1n} \\ r_{11} & r_{12} & \cdots & r_{1j} & \cdots & r_{1n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mj} & \cdots & r_{mn} \end{bmatrix} \]  

(9)

where \( r_{ij} = \frac{R_{ij}}{\sum_{k=1}^{m} R_{kj}} \)

These evaluation factors have different weights. The weight of factors \( u_i \) is \( q_i \); these weights for \( q_1 \) and \( q_2 \) q_i... q_n are normalized as follows, \( X=(x_1, x_2, \ldots, x_m) \)

\[ x_i = \frac{q_i}{\sum_{k=1}^{n} q_k} \]

(10)

P is the comprehensive evaluation result

\[ P = X \cdot R \]  

(11)

P is the synthetic of weighted fuzzy vector X and fuzzy evaluation matrix R.
The evaluation results $P$ is normalized, the comprehensive evaluation results $Y$ is obtained,

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

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

According to the comprehensive evaluation results $Y$, set threshold $\lambda_1, \lambda_2, \ldots, \lambda_m$, 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;

$$
\cdots \cdots
$$

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"[8], 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>
<thead>
<tr>
<th>Evaluation target</th>
<th>Evaluation level</th>
<th>Evaluation indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary A1</td>
<td></td>
<td>B1: Word meaning collocation</td>
</tr>
<tr>
<td>Syntax A2</td>
<td></td>
<td>B2: Rhetoric</td>
</tr>
<tr>
<td>Discourse A3</td>
<td></td>
<td>B3: Technical terms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B4: The use of dialect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B5: Grammar</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B6: Textual cohesion</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B7: Coherence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B8: Intentionality</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B9: Acceptability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B10: Informational</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B11: Context</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B12: Inter-textuality</td>
</tr>
</tbody>
</table>

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 completely 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>
<thead>
<tr>
<th>Evaluatio n target</th>
<th>Evaluation level</th>
<th>Weight $W$</th>
<th>Evaluation index</th>
<th>$W$ related to evaluation level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary A1</td>
<td></td>
<td>0.4071</td>
<td></td>
<td>B1 0.2361</td>
</tr>
<tr>
<td>Syntax A2</td>
<td></td>
<td>0.2643</td>
<td></td>
<td>B5 1.0000</td>
</tr>
<tr>
<td>Discourse A3</td>
<td></td>
<td>0.3286</td>
<td></td>
<td>B9 0.2064</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparative judgment matrix</th>
<th>$\lambda_{mn}$</th>
<th>CI</th>
<th>CR</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>4.2500</td>
<td>0.0833</td>
<td>0.0937</td>
<td>Acceptable</td>
</tr>
<tr>
<td>A2</td>
<td>Only one evaluation index which don’t need to be inspected</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>7.3339</td>
<td>0.0557</td>
<td>0.04216</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Y</td>
<td>0.0537</td>
<td>0.0557</td>
<td>0.04216</td>
<td>Acceptable</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>level</th>
<th>Index</th>
<th>The answer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>L1</td>
</tr>
<tr>
<td>A1</td>
<td>B1</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>B3</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>B4</td>
<td>100</td>
</tr>
<tr>
<td>A2</td>
<td>B5</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>B6</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>B7</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>B8</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>B9</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>B10</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>B11</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>B12</td>
<td>92</td>
</tr>
</tbody>
</table>

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

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

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 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 B on 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.

Figs. 1 and 2 show the evaluation results of the test text. It is to be noted that the percentage of the total number of L2 for assessment for level 2 and level 1 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.
Table 8. The comprehensive fuzzy evaluation results of translation quality on the level of vocabulary

<table>
<thead>
<tr>
<th>Software</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.8527</td>
<td>0.0982</td>
<td>0.0491</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>B</td>
<td>0.8127</td>
<td>0.0936</td>
<td>0.0468</td>
<td>0.0234</td>
<td>0.0234</td>
</tr>
<tr>
<td>C</td>
<td>0.5127</td>
<td>0.1181</td>
<td>0.1181</td>
<td>0.1329</td>
<td>0.1181</td>
</tr>
<tr>
<td>D</td>
<td>0.6324</td>
<td>0.1838</td>
<td>0.1225</td>
<td>0.0306</td>
<td>0.0306</td>
</tr>
</tbody>
</table>

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.

ACKNOWLEDGMENTS

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

References