Study on Fuzzy Mathematical Evaluation Method of the Machine Translation Quality Based on Network Survey

Yi-qun SUN\textsuperscript{1,2,a,*} and Min-kang ZHOU\textsuperscript{1,b}

\textsuperscript{1}Facultat de Traducció i d'Interpretació, Universitat Autònoma de Barcelona, Spain;
\textsuperscript{2}International Education College of Changchun University, Changchun University, China

\textsuperscript{a}Email:13969161305@163.com, \textsuperscript{b}Email: Minkang.Zhou@uab.cat

*Corresponding author

Keywords: Machine translation, Evaluation, Network survey, Fuzzy mathematics.

Abstract. In order to make the evaluation of the quality of translation more representative, literature lovers with different age, gender, occupation and cultural level are organized by establishing network reading groups and will be taken as appriciator and evaluators of text using machine translation. A number of evaluation indexes for translation text quality are set and a questionnaire is designed according to these indexes. The questionnaires are sent to literature lovers and data are obtained. 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. This paper provides a tool for developing and improving translation software.

Introduction

With the development of computer and software technology, machine translation is used more and more widely and has become one of the hotspots of research in the field of computer. The accuracy of evaluation method about the translation quality is the main basis of system development, and has become one of the main powers to promote the development of machine translation system.

A lot of translation evaluation focus on the qualitative assessment of the text itself at present, such as Yan Fu's Elegant Writing (1898), Fu Lei's Spirit Likeness (1951), Qian Zhongshu's Transferred (1964); Liu Chongde’s The Art of Translating (1993), etc. Both qualitative and quantitative assessment is now attracting the attention of people gradually, such as The equivalent evaluation method [1]; A Functional Linguistic Perspective of Translation Quality Assessment Model—With the Evaluation of English Version of Kong Yi-ji as Example[2]; The parameter model taking "decent" as supreme principle[3]; TQA in the Perspective of the Relevance Theory [4]; Most of the parameters used in evaluating translation are not accurate, such as fluency, apropos, misnomer, etc. These parameters are "fuzzy". Fuzzy mathematics is a kind of mathematics method which mathematically research and treat fuzzy phenomenon, so fuzzy mathematics in evaluating the quality attracts attention of public [5]. But in this research the judges must be experts with high cultural level and number of participants in the evaluation is limited, so the evaluation results are easily influenced by their own language and culture.

This article adopts the way of establishing network reading group, collecting literature lovers widely with different age, gender, occupation and cultural level as appriciator and evaluators of text using machine translation in order to make the evaluation about the quality of translation more representative. This paper will
evaluate the translation quality with a number of evaluation indexes and design questionnaire according to these indexes, surveying literature lovers in the network reading group to get the data. By using fuzzy mathematical tool for data analysis and processing, this research could give quantitative assessment of quality of the translation[6] which provide technical data for development and the improvement of the software using for machine translation.

**The Fuzzy Evaluation Method of Translation Quality**

For the translation to be evaluated, set m evaluation factors ($u_1, u_2, \ldots, u_m$), the U said set of factors about comprehensive evaluation, $U = (u_1, u_2, \ldots, u_m)$.

Evaluation results of each evaluation factor are divided into n levels (1, 2, ..., n). The number of people ranked level 1 in $u_1$ is marked as $R_{11}$. The number of people ranked level 2 in $u_1$ is marked as $R_{12}$. The number of people ranked level j in $u_i$ is marked as $R_{ij}(i=1,2,\ldots,m; j=1,2,\ldots,n)$, ..., and the number of people ranked level n in $u_1$ is marked as $R_{1n}$. Each evaluation factor is divided into n level, the evaluation results is showed with fuzzy matrix $RR(m \times n)$.

$$RR = \begin{bmatrix}
R_{11} & R_{12} & \cdots & R_{1j} & \cdots & R_{1n} \\
\vdots & \vdots & \cdots & \vdots & \cdots & \cdots \\
R_{m1} & R_{m2} & \cdots & R_{mj} & \cdots & R_{mn}
\end{bmatrix}$$

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

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

The computational formula of normalization is,

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

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

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

Get the fuzzy vector $X$ which stand for the weight of evaluation factors,

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

$P$ is the comprehensive evaluation results

$$P = X \cdot R$$
$P$ is the synthetic of weighted fuzzy vector $X$ and fuzzy evaluation matrix $R$, 
Where $P = (p_1, p_2, \ldots, p_m)$, \( p_i = V_{k=1}^{m} \left(x_k \land r_k \right) \)

The evaluation results $P$ is normalized by the following formula,

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

The comprehensive evaluation results $Y$ is obtained,

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

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;

The relationship of Input $X$, output $Y$ and fuzzy matrix $R$ is shown in Fig.1. Fuzzy matrix $R$ obtained by questionnaire survey is the fuzzy transducer. $X$ stand for the input, and $Y$ is the output. $X$ is the fuzzy vector of the weights of evaluation factors for the quality of the translation, and the number of weight values for age, gender, cultural level and other factors of readers. Output $Y$ can be calculated by weighted fuzzy vector $X$ and fuzzy variator $R$, thus we could get the audiences’ comprehensive evaluation of the translation.

The Instance of Translation Quality’s Evaluation

Design of Questionnaire and Standards about Evaluation

To establish an evaluation standard which is relatively objective is the premise of accurately evaluating the quality accurately. In this paper, the evaluation standards mainly refer to a combination of EUROTRA with evaluation criteria putting forward by Japanese experts. The quality evaluation factors of the translation are: loyalty, the application of translation skills, syntax application and vocabulary application. Each factor is divided into five levels: level 1, level 2, level 3 and level 4 and level 5, corresponding to the outstanding, good, medium, low and the poor respectively.

The translation ranked level 1 faithfully reflects the content of the original and is translated smoothly, meaning clear and using accurate words. They almost do not need to be edited after translation. The translation ranked level 2 faithfully reflects the content of the original, translated smoothly and has clear meaning, but they have flaws in words or grammar and could be relatively smoothly translated without referring to the original except a little change; The translation ranked level 3 is essentially the faithful translation of the original, but the word order may be improper and words’ meaning is not accurate, what’s more, the tense relationship is improper and sometimes relationship between phrases and attributive as well as adverbial’s location is wrong, some sentences is difficult to understand, these translation need a careful editing after translation processing; The translation ranked level 4 is partly the faithful translation of the original, the original structure cannot be translated. Many
prepositions phrases, clauses structure and judgment of subordinate clause are wrong, and some contents is lost; The translation ranked level 5 can't reflect the content of the original at all and a lot of places are failed to be translated. Sometimes there may be a complete or a more completely translation while most of the translation can't form complete sentences.

According to the text analysis to the original, German literary theorist Jauss receives anticipation vision of the aesthetic as the theoretical basis and analyze the original text from three levels (the stylistic anticipation, image anticipation and meaning anticipation). Among these factors, the stylistic anticipation includes three elements as language, structure and plot, while image anticipation includes four elements(the image appearance and personality characteristics of minor characters and heroes).Meaning anticipation includes five elements( aesthetic meaning of works, the author's emotional state, life attitude, mindset and writing intention ) there are a total of 12 elements. From these twelve elements, we determine what needed to be passed in the original and design 12 questions to inspect the loyalty of the translation. To each question you could enter "1" or "0" respectively by selecting "yes" or "no". The number of Right answer 12, 8 to 11, 4-7, 2-3, 0 and 1, is corresponding to level 1, level 2, level 3, 4 and 5 respectively.

Input 1 in the skills of Translation, vocabulary and syntax on the level selected and input zero in others.

The Original and Software of Machine Translation

This paper selected a typical chapter of the winner of 1956 Nobel Prize, Spanish writer Jimenez’s famous masterpiece Platero y yo. Three software is chosen from the widely translation software, such as Google translation, Bing translation, Baidu translation, Youdao translation and General translation, Lingoes, etc and marked them as a, b, c, respectively. Obtain the evaluation data of translation quality by means of Internet survey data, using comprehensive fuzzy evaluation method to analyze the data.

The Processing of Questionnaire Data and Analysis of Result

Sent these 18 chapters selected from the translating a, b and c and the questionnaire to the Internet audience, investigating the audience of readers and statistics the effective questionnaires. Count up the number of “YES(1)” or “NO(0)” under the topic of the faithfulness of translation, and determine the level of the faithfulness of translation as well as the number of people in each level of loyalty. As for Translation skills, vocabulary and syntax application, add up the number of “YES(1)” or “NO(0)” typed in by readers in the same level of every factors in the questionnaire so as to get the number of readers who choose this level. The questionnaires whose questions are all answered formally by readers were deemed to be valid. For example, if readers chose two “1” and three “0” or five “0”, the questionnaire is deemed to be non-standard questionnaire. Only if readers chose one “1” and four “0”, the questionnaire is deemed to be valid questionnaire.

There are 500 readers in this network reading group, and we only count up the valid questionnaires’ data. Fuzzy matrix $RR_a^1$, $RR_b^1$, $RR_c^1$ which stand for the evaluation results showed by number of readers 'chose of four evaluation factors and five ratings in the first chapter of Chinese version a, b and c.
Faithfulness: 431 26 23 10 10  
Translation skill: 462 31 4 3 0  
Syntax application: 480 18 1 1 0  
Vocabulary application: 390 59 32 14 5

\[ RR_a = \begin{pmatrix} 
254 & 170 & 46 & 13 & 7 \\
312 & 118 & 29 & 23 & 8 \\
189 & 204 & 63 & 19 & 15 \\
315 & 32 & 67 & 58 & 18 
\end{pmatrix}, 
RR_b = \begin{pmatrix} 
89 & 250 & 132 & 11 & 5 \\
59 & 273 & 114 & 29 & 12 \\
87 & 235 & 151 & 8 & 6 \\
290 & 35 & 69 & 72 & 21 
\end{pmatrix}, 
RR_c = \begin{pmatrix} 
7478 & 552 & 527 & 236 & 205 \\
8287 & 529 & 71 & 92 & 19 \\
8465 & 460 & 21 & 25 & 27 \\
6992 & 1068 & 577 & 258 & 103 
\end{pmatrix} \]

Fuzzy matrix \( R^a \), \( R^b \) and \( R^c \) which stand for the evaluation results showed by number of readers' chose of four evaluation factors and five ratings in this 18 chapter versions.

\[ RR^a = \sum_{k=1}^{64} RR^a_k = \begin{pmatrix} 
7478 & 552 & 527 & 236 & 205 \\
8287 & 529 & 71 & 92 & 19 \\
8465 & 460 & 21 & 25 & 27 \\
6992 & 1068 & 577 & 258 & 103 
\end{pmatrix}, 
RR^b = \sum_{k=1}^{64} RR^b_k = \begin{pmatrix} 
4500 & 3066 & 850 & 265 & 137 \\
5582 & 2140 & 533 & 418 & 145 \\
3118 & 3927 & 1125 & 393 & 255 \\
5633 & 586 & 1212 & 1057 & 330 
\end{pmatrix}, 
RR^c = \sum_{k=1}^{64} RR^c_k = \begin{pmatrix} 
1629 & 4509 & 2340 & 197 & 90 \\
1067 & 4879 & 2059 & 533 & 227 \\
1578 & 4199 & 2715 & 152 & 121 \\
5184 & 652 & 1203 & 1334 & 392 
\end{pmatrix} \]

Evaluation results of the five grades of each evaluation factor (fuzzy matrix of each row) is normalized, and then we can get the fuzzy evaluation matrix \( R^a \), \( R^b \), \( R^c \),

\[ R^a = \begin{pmatrix} 
0.8311 & 0.0613 & 0.0586 & 0.0262 & 0.0228 \\
0.9210 & 0.0588 & 0.0079 & 0.0102 & 0.0021 \\
0.9408 & 0.0511 & 0.0023 & 0.0028 & 0.0030 \\
0.7771 & 0.1187 & 0.0641 & 0.0287 & 0.0114 
\end{pmatrix}, 
R^b = \begin{pmatrix} 
0.5103 & 0.3477 & 0.0964 & 0.0301 & 0.0155 \\
0.6330 & 0.2427 & 0.0604 & 0.0474 & 0.0164 \\
0.3536 & 0.4453 & 0.1276 & 0.0446 & 0.0289 \\
0.6388 & 0.0665 & 0.1374 & 0.1199 & 0.0374 
\end{pmatrix} \]
It could be seen from the data that the evaluation of faithfulness, translation skills, vocabulary application and syntax application of these three translations are different. Take achieving excellent and good as the standard, and analyze the four evaluation factors of three translations.

Table 1 is the fuzzy evaluation results of translation quality of each evaluation factor, the L1 stands for the percentage of the number of people who determined level 1 in the total, the L2 stands for the percentage of the number of people who determined level 2 in the total.

It could be seen from table 1 that faithfulness of translation a is the best, and translation b is not as good as a, while translation c is the worst. The L1 for faithfulness of translation a is much higher than b translation, the L1 for translation c is only 18.59%, but L1+L2 of translation c reached 70.03%, therefore these three software is all good in terms of faithfulness.

In terms of translation skill, translation a is the best, and translation b is not as good as a, while translation c is the worst. The L1 for translation skills of translation a is as high as 92.10%, the L1 + L2 is 97.98%, so software a is very good in terms of translation skill, and software a is far better than b and c. L1 for translation c is only 12.17%, with L1+L2 is only 67.84%, so the software c need more improvement in terms of translation skills.

In terms of syntactic application, translation a is the best, translation b is not as good as a, translation c is the worst. The L1 for syntax application of translation a is 94.08%, the L1+L2 is as high as 99.19%, so the syntactic application of software a is excellent and it’s far better than that of b and c. The L1 of translation c is only 18%, so the software c needs more improvement in terms of syntactic application.

In terms of vocabulary application, translation a is the best, translation b is not as good as a, translation c is the worst. The vocabulary application of a and b are close, while c is the worst. However, the L1 for translation c also reaches 59.14% and performed just so-so, and it needs just a little improvement to be as good as b and c.

Table 1. The fuzzy evaluation results of translation quality about evaluation factors.

<table>
<thead>
<tr>
<th>Evaluation Factors</th>
<th>Translation software</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Total of level 1 and level 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faithfulness</td>
<td>a</td>
<td>0.8311</td>
<td>0.0613</td>
<td>0.8924</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>0.5103</td>
<td>0.3477</td>
<td>0.8580</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>0.1859</td>
<td>0.5144</td>
<td>0.7003</td>
</tr>
<tr>
<td>Translation skill</td>
<td>a</td>
<td>0.9210</td>
<td>0.0588</td>
<td>0.9798</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>0.6330</td>
<td>0.2427</td>
<td>0.8757</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>0.1217</td>
<td>0.5566</td>
<td>0.6784</td>
</tr>
<tr>
<td>Syntax application</td>
<td>a</td>
<td>0.9408</td>
<td>0.0511</td>
<td>0.9919</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>0.3536</td>
<td>0.4453</td>
<td>0.7989</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>0.1800</td>
<td>0.4791</td>
<td>0.6591</td>
</tr>
<tr>
<td>Vocabulary application</td>
<td>a</td>
<td>0.7771</td>
<td>0.1187</td>
<td>0.8958</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>0.6388</td>
<td>0.0665</td>
<td>0.7053</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>0.5914</td>
<td>0.0744</td>
<td>0.6658</td>
</tr>
</tbody>
</table>

Weights of these four evaluation factors are determined by questionnaire survey. According to data obtained by survey, the average number of weights of faithfulness,
translation skills, vocabulary application and syntax application are 0.2541, 0.3212, 
0.2180, and 0.2600, respectively. Weighted fuzzy vector \( X \) of evaluation factors 
are obtained through normalization processing.

\[
X = (0.2412, 0.3049, 0.2070, 0.2468)
\]

The \( P^a, P^b, P^c \) for comprehensive evaluation results,

\[
P^c = X \cdot R^a = (0.3049, 0.1187, 0.0641, 0.0287, 0.0228)
\]

\[
P^b = X \cdot R^b = (0.3049, 0.2427, 0.1374, 0.1199, 0.0374)
\]

\[
P^c = X \cdot R^c = (0.2468, 0.3049, 0.2412, 0.1522, 0.0447)
\]

The comprehensive evaluation result \( Y^a, Y^b \) and \( Y^c \) is obtained by normalizing \( P^a, P^b \) and \( P^c \).

\[
Y^a = (0.5565, 0.2201, 0.1189, 0.0532, 0.0423)
\]

\[
Y^b = (0.3620, 0.2881, 0.1631, 0.1423, 0.0444)
\]

\[
Y^c = (0.2493, 0.3080, 0.2437, 0.1538, 0.0452)
\]

The comprehensive evaluation results of translation quality are shown in Fig.2. In 
terms of \( Y^a, Y^b, Y^c \) and Fig.2, translation \( a \) is the best from the point of L1 and L1+L2, 
translation \( b \) is not as good as \( a \), translation \( c \) is the worst. The L1, L1+L2 of 
translation \( a \) are respectively 2.23 times and 1.39 times of translation \( c \), therefore, 
quality of translation \( a \) is much better than \( c \). The translation software \( c \) remains to be 
improved.

Set a threshold value of \( \lambda_1=0.25, \lambda_2=0.55, \lambda_2=0.65, \lambda_3=0.75, \lambda_4=0.85, \lambda_5=0.90, \) 
and it could be seen that \( y^a_i > \lambda_1, y^b_i > \lambda_1, \) so the level of 
translation \( a \) and \( b \) are deemed to be level 1 (excellent). 
\( y^c_i = 0.2493 < \lambda_3, y^c_i + y^c_i + y^c_i = 0.5573 < \lambda_4, y^c_i + y^c_i + y^c_i = 0.8011 > \lambda_5 \) , 
the level of 
translation \( c \) could be judged as level 3(medium).

Figure 2. The fuzzy evaluation results of translation quality.
Conclusions

(1) With the method of establishing network reading group, literature lovers with different age, gender, occupation and cultural level are organized and are taken as appreciator and evaluators of machine translation. Evaluation results which are more representative can be obtained.

(2) 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.

(3) 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 so as to improve the software design and the quality of software.

Reference


