The Integrated Evaluation Method of Machine Translation Quality Based on Z-numbers

Dong Qiu, Haihuan Jiang and Shuqiao Chen

Abstract—**The evaluation of machine translations is of great significance to machine translation research. The evaluations are based on the judgments of the evaluators, and it is necessary for us to ensure that the judgments are reliable. Since the Z-number has the potential to overcome the issue of information reliability, introduction of this concept brings an improvement to describe uncertain information in multicriteria decision making. Inspired by this development, we utilize the Z-number to formulate an integrated translation evaluation method. The concrete steps of the method are given, and two different machine translated target texts that are from the same original text are studied using the method. It is shown that the proposed method provides helpful guidance of machine translation quality evaluation.**

Index Terms—**Machine translation quality evaluation, Multicriteria decision making, Integrated evaluate method, Z-number.**

I. Introduction

MACHINE Translation is a subfield of natural language processing [1]–[4] that investigates the use of computers to convert one language into another. There are many different principles of machine translation systems. For the researchers of machine translation, the evaluation of the translation represents the weak point of machine translation systems and the direction of future machine translation research. Moreover, the users of machine translation systems can choose an optimal system based on the evaluation of machine translations. Therefore, the evaluation of machine translations plays a key role in machine translation research.

The development of machine translation systems has led to extensive interest in the evaluation of machine translation. Kishore Papineni et al. proposed a method for automatic evaluation of a machine translation [5]. The authors used statistical knowledge to measure the similarity degree between the machine translation result and the human translation result. They evaluated the machine translation system by the similarity degree. On the basis of [5], Chin-Yew Lin applied the recall rate to examine the sufficiency of the translation [6]. Satanjeev Banerjee et al. took advantage of the unigram recall and the weighted average harmonic number of the unigram precision to investigate the unigram matching between human translation and machine translation [7]. The results of the unigram matching served as the indicator of translation quality. Methods that utilize the similarity between the artificial translation and the machine translation to evaluate the translation quality are relatively convenient, simple and efficient. However, automatic evaluation of machine translations can only judge the sufficiency of the information in translated articles. These approaches are limited to the micro level, and they lack a macro grasp on the translated articles [8]. These methods cannot connect to the specific context and cannot understand the deep meaning of the articles. Moreover, the estimated contents of automatic evaluation systems do not involve the correctness of sentence structure, the clarity of language expression or the style of the translated text. Therefore, the artificial evaluation system is irreplaceable. The fuzziness of language makes it difficult for people to judge the quality of translation. Accordingly, the establishment of fuzzy set theory has solved this problem well. Christer Carlsson et al. showed that fuzzy set theory is helpful to address uncertain information [9]–[12]. According to [13], [14], Xiaojun Zhang introduced a fuzzy integrated evaluation method (FIEM) to evaluate translation quality [15]. Ageeva et al. introduced an open-source task management and gap-filling method to evaluate machine translation [16]. The gap-filling method combines the automatic evaluation method and human evaluation method to improve the evaluation. In [17], Ma et al. used human evaluation to guide the training process of automatic evaluation, and the approach improved the performance of the automatic evaluation.

Although the results of combined evaluation methods are more accurate, these methods still do not consider the reliability of the judgments made by the evaluators. Zedeh introduced the concept of the Z-number, which can be used to measure the reliability [18]. Subsequently, the Z-number has been widely applied in the realms of decision making. The Z-number was used as the tool of decision analysis in [19], [20]. Zhi-Quan Xiao exploited Z-numbers in multicriteria decision making and described each evaluation criterion with Z-numbers [21]. In this paper, we attempt to utilize the theory of Z-numbers to evaluate translation quality. The elements of evaluation are analyzed in the form of Z-numbers, and the evaluation matrix is built on the basis of Z-numbers. An evaluation method of machine translation quality is proposed.

The structure of the paper is as follows. In section 2, we review the theory of Z-numbers and introduce the process of converting Z-numbers to regular fuzzy numbers. An approach to evaluate translation quality based on Z-numbers is presented in section 3. We give a case study on the evaluation of machine translation to prove the effectiveness of the proposed method in section 4. The conclusions are discussed in the final section.

Manuscript received Jul. 1, 2019; revised Sept. 3, 2019. This work was supported by The National Natural Science Foundations of China (Grant no. 11671001).

D. Qiu is with College of Mathematics and Physics, Chongqing University of Posts and Telecommunications, Nanan, Chongqing, 400065, P.R. China e-mail: dongqiumath@163.com; qiudong@cqupt.edu.cn

H. Jiang and S. Chen are with the same university as D. Qiu.

II. Basic Definitions

A. The definition of Z-number

Definition 1 [22] *A fuzzy number $\widetilde{A} = (a, b, c)$ is considered as a triangular fuzzy number if its membership is* *determined as follows:*

$$\mu_{\widetilde{A}}(x) = \begin{cases} \dfrac{x-a}{b-a}, & a \le x \le b, \\ \dfrac{c-x}{c-b}, & b \le x \le c, \\ 0, & otherwise. \end{cases} \quad (1)$$

Definition 2 [23] *A fuzzy number $\widetilde{A} = (a, b, c, d)$ is considered as a trapezoidal fuzzy number if its membership is determined as follows:*

$$\mu_{\widetilde{A}}(x) = \begin{cases} \dfrac{x-a}{b-a}, & a \le x \le b, \\ 1, & b \le x \le c, \\ \dfrac{x-d}{c-d}, & c \le x \le d, \\ 0, & otherwise. \end{cases} \quad (2)$$

Definition 3 [18], [24] *A Z-number $Z = (\widetilde{A}, \widetilde{B})$ is an ordered pair of fuzzy numbers. The first component $\widetilde{A}$ is a restriction on the real-valued variable X. The second component $\widetilde{B}$ is a measure of reliability of $\widetilde{A}$.*

B. Convert Z-number to a Regular Fuzzy Number

We use the method proposed in [25] to convert the Z-number to a regular fuzzy number. On the basis of [25], we let $Z = (\widetilde{A}, \widetilde{B})$ be a Z-number in which the first component $\widetilde{A}$ is defined as a trapezoidal fuzzy number and the second component $\widetilde{B}$ is defined as a triangular fuzzy number [26]. The details of the method are as follows:

step 1 Use equation (3) to convert $\widetilde{B}$ into a crisp number.

$$\alpha = \frac{\int x\mu_{\widetilde{B}}(x)dx}{\int \mu_{\widetilde{B}}(x)dx}, \quad (3)$$

where $\int$ is an algebraic integration.

step 2 Regard α as the weight of $\widetilde{A}$. That is, we utilize the reliability of the information to restrain $\widetilde{A}$. The Z-numbers after weighting can be expressed as:

$$\widetilde{Z}^{\alpha} = \{\langle x, \mu_{\widetilde{A}^{\alpha}}(x)\rangle \mid \mu_{\widetilde{A}^{\alpha}}(x) = \alpha\widetilde{A}(x),\ x \in [0,1]\}. \quad (4)$$

step 3 Let $\widetilde{Z}'$ be the transformed Z-number that is a regular fuzzy number. To ensure that the fuzzy expectation between the Z-number and the transformed Z-number is equal such that $E_{\widetilde{Z}'}(x) = E_{\widetilde{A}^{\alpha}}(x)$, $\widetilde{Z}'$ is defined as follows:

$$\widetilde{Z}' = \{\langle x, \mu_{\widetilde{Z}'}(x)\rangle \mid \mu_{\widetilde{Z}'}(x) = \mu_{\widetilde{A}}(\frac{x}{\sqrt{\alpha}}),\ x \in [0,1]\}. \quad (5)$$

Then, the process of converting the Z-number to a classical fuzzy number is complete.

III. Proposed Method

It is known that the core of translation is to find the same meaning among different languages. The standard of evaluation translation quality is whether the meaning of the translated text is expressed properly in another language. "Faithfulness, expressiveness and elegance" are regarded as the top pursuits of translation [27]. To obtain a high-quality translation, one should guarantee the semantics, syntactics, pragmatics and context appropriately in the translated text [15]. Consequently, we choose semantics, syntactics, pragmatics and context as the influence factors of the evaluation in our method.

Considering that the four influence factors are vague, we use the Z-number to measure their accuracy [28]. Let $U = \{u_1, u_2, u_3, u_4\}$ be the set of influence factors. Each element in U is denoted as $u_i, i = 1, ..., 4$, where u_1 is the evaluation of the translated text's semantics, u_2 is the evaluation of the translated text's syntactics, u_3 is the evaluation of the translated text's pragmatics and u_4 is the evaluation of the translated text's context. We use the analytic hierarchy process to construct a comparative judgment matrix and obtain the weight value. According to [29], the important degrees among the influence factors are those listed in Table 1 and Table 2.

TABLE I
The Meaning of The Important Degrees

The degree	The meaning of the degree
1	The importance of factor one and factor two are the same
3	Factor one is slightly more important than factor two
5	Factor one is significantly more important than factor two
7	Factor one is intensively more important than factor two
9	Factor one is extremely more important than factor two
2,4,6,8	The intermediate value

TABLE II
The Comparing Judgment Matrix Based on Z-number

	Semantics	Syntactics
Semantics	$[(1,1,1,1),(1,1,1)]$	$[(1/3,1/2,1/2,1)$
Syntactics	$[(1,2,2,3),VH]$	$[(1,1,1,1),(1,1,1)]$
Pragmatics	$[(2,3,3,5),H]$	$[(1,4/3,4/3,2),MH]$
Context	$[(1/3,2/5,2/5,2),H]$	$[(1,4/3,4/3,2),M]$
	Pragmatics	**Context**
Semantics	$[(1/5,1/3,1/3,1/2),H]$	$[(1/2,5/2,5/2,3),H]$
Syntactics	$[(1/2,3/4,3/4,1),MH]$	$[(1/3,3/4,3/4,1),M]$
Pragmatics	$[(1,1,1,1),(1,1,1)]$	$[(1/7,1/3,1/3,1),M]$
Context	$[(1,3,3,7),M]$	$[(1,1,1,1),(1,1,1)]$

Then, we can turn the comparing judgment matrix into a fuzzy number as shown in Table 3:

TABLE III
The Classical Fuzzy Number

	Semantics	Syntactics
Semantics	$(1,1,1,1)$	$(0.33,0.49,0.49,0.98)$
Syntactics	$(0.95,1.90,1.90,2.84)$	$(1,1,1,1)$
Pragmatics	$(1.90,2.85,2.85,4.74)$	$(0.84,1.12,1.12,1.67)$
Context	$(0.32,0.38,0.38,0.90)$	$(0.71,0.94,0.94,1.41)$
	Pragmatics	**Context**
Semantics	$(0.19,0.32,0.32,0.47)$	$(0.47,2.37,2.37,2.85)$
Syntactics	$(0.42,0.63,0.63,0.84)$	$(0.24,0.53,0.53,0.71)$
Pragmatics	$(1,1,1,1)$	$(0.10,0.24,0.24,0.71)$
Context	$(0.71,2.12,2.12,4.95)$	$(1,1,1,1)$

Let S_i be the fuzzy comprehensive degree of the factor u_i. From [30], we can obtain:

$$\begin{aligned} S_i \quad = \quad & (\sum_{j=1}^{m} a_{ij}, \sum_{j=1}^{m} b_{ij}, \sum_{j=1}^{m} c_{ij}, \sum_{j=1}^{m} d_{ij}) \\ / \quad & (\sum_{i=1}^{4}\sum_{j=1}^{m} a_{ij}, \sum_{i=1}^{4}\sum_{j=1}^{m} b_{ij}, \sum_{i=1}^{4}\sum_{j=1}^{m} c_{ij}, \sum_{i=1}^{4}\sum_{j=1}^{m} d_{ij}) \end{aligned}$$

and

$$S_1 = (0.0745, 0.2337, 0.2337, 0.4741),$$
$$S_2 = (0.0977, 0.2370, 0.2370, 0.4499),$$
$$S_3 = (0.1438, 0.2912, 0.2912, 0.7263),$$
$$S_4 = (0.1026, 0.2482, 0.2482, 0.7388).$$

From Figure 1 and Figure 2, we can observe that the triangular fuzzy number is the special trapezoidal fuzzy number. For the triangular fuzzy number $B(a, b, c, d)$, it will be a trapezoidal fuzzy number if $b = c$. Thus, we use formula (6) to compare the fuzzy degree between $S_i(a_i, b_i, c_i)$ and $S_j(a_j, b_j, c_j)$.

$$V(S_i \geq S_j) = \begin{cases} 1, & b_i \geq b_j \\ \dfrac{(a_j - c_i)}{(b_i - c_i) - (b_j - a_j)}, & other \\ 0, & a_j \geq c_i \end{cases} \quad (6)$$

Then,

$$V(S_1 \geq S_2) = 0.9913, V(S_1 \geq S_3) = 0.4014,$$
$$V(S_1 \geq S_4) = 0.9624, V(S_2 \geq S_1) = 1,$$
$$V(S_2 \geq S_3) = 0.8496, V(S_2 \geq S_4) = 0.9688,$$
$$V(S_3 \geq S_1) = 1, V(S_3 \geq S_2) = 1,$$
$$V(S_3 \geq S_4) = 1, V(S_4 \geq S_1) = 1,$$
$$V(S_4 \geq S_2) = 1, V(S_4 \geq S_3) = 0.9326.$$

We set the weight as formula (7):

$$w_i = min(V(S_i \geq S_j)). \quad (7)$$

Thus, we can obtain

$$W' = (0.4014, 0.8496, 1, 0.9326)^T. \quad (8)$$

Normalizing formula (8), we can obtain the weight

$$W = (0.1261, 0.2669, 0.3141, 0.2929)^T.$$

With the vague and subjective evaluation words such as "good" and "very good" to describe the influence factors, we use the Z-number to quantify the translation evaluation words. The concrete method is implemented by the following steps:

step 1 Let R' be the evaluation matrix that is related to the influence factor $U(u_1, u_2, u_3, u_4)$.

$$R' = \begin{pmatrix} Z_{11} & Z_{12} & Z_{13} & Z_{14} \\ \vdots & \vdots & \vdots & \vdots \\ Z_{k1} & Z_{k2} & Z_{k3} & Z_{k14} \end{pmatrix},$$

where $Z_{ij} = (A_i, B_j)$ is the evaluation of the expert j to influence factor i. A_i is a trapezoidal fuzzy number, and B_j is a triangular fuzzy number.

step 2 According to [31], [32], linguistic variables are considered as the quantitative indexes of evaluators' reliability. The fuzzy numbers for the linguistic variables are given in Table 4.

step 3 Use equations (3) and (5) to convert the Z-number to a regular fuzzy number. Then, the evaluation matrix R'

TABLE IV
Linguistic Variables for the Evaluators' Reliability

Linguistic Variables	Trapezoidal Fuzzy Number
Very Low	$(0, 0, 0.1)$
Low	$(0, 0.1, 0.25)$
Middle Low	$(0.15, 0.3, 0.45)$
Middle	$(0.35, 0.5, 0.65)$
Middle High	$(0.55, 0.7, 0.85)$
High	$(0.8, 0.9, 1)$
Very High	$(0.9, 1, 1)$
	Triangular Fuzzy Number
Very Low	$(0, 0, 0, 0.1)$
Low	$(0, 0.1, 0.1, 0.25)$
Middle Low	$(0.15, 0.3, 0.3, 0.45)$
Middle	$(0.35, 0.5, 0.5, 0.65)$
Middle High	$(0.55, 0.7, 0.7, 0.85)$
High	$(0.8, 0.9, 0.9, 1)$
Very High	$(0.9, 1, 1, 1)$

is changed into $R' = [\widetilde{Z}'_{ij}]$. The specific representation of matrix R' is as follows:

$$R' = \begin{pmatrix} Z'_{11} & Z'_{12} & Z'_{13} & Z'_{14} \\ \vdots & \vdots & \vdots & \vdots \\ Z'_{k1} & Z'_{k2} & Z'_{k3} & Z'_{k14} \end{pmatrix}$$

where

$$Z'_{ij} = \{(x, \mu_{Z'_{ij}}(x)) | \mu_{Z'}(x) = \mu_{Z'_{ij}}(\frac{x}{\sqrt{\alpha}})\},$$

and

$$x \in \sqrt{\alpha} X_A.$$

step 4 Using the defuzzification algorithm, we can obtain

$$R' = \begin{pmatrix} r'_{11} & r'_{12} & r'_{13} & r'_{14} \\ \vdots & \vdots & \vdots & \vdots \\ r'_{k1} & r'_{k2} & r'_{k3} & r'_{k14} \end{pmatrix},$$

step 5 We utilize formula (9) to obtain the score of each influence factor $R(r_1, r_2, r_3, r_4)$.

$$r_{ij} = \bigwedge_{j=1}^{k} r_{ik}, \quad (9)$$

step 6 Finally, the evaluation result is

$$b = \sum_{i=1}^{4} r_i w_i. \quad (10)$$

In the proposed method, the values of matrix Z are determined by questionnaire investigation [33]. Therefore, the values are dynamic.

IV. Case Study

A. Evaluation Software and Evaluation Text

We use Baidu Translate and Google Translate to test translation of the same text from English to Chinese separately. Moreover, we excerpt parts of the twentieth chapter of Pride and Prejudice as the original text for translation. The results of the two softwares' translation are as follows:

Original text [34]: Mr. Collins was not left long to the silent contemplation of his successful love; for Mrs. Bennet, having dawdled about in the vestibule to watch for the end of the conference, no sooner saw Elizabeth open the door and with quick step pass her towards the staircase, than she entered the breakfast room, and congratulated both him and herself in warm terms on the happy prospect of their nearer connection. Mr. Collins received and returned these felicitations with equal pleasure, and then proceeded to relate the particulars of their interview, with the result of which he trusted he had every reason to be satisfied, since the refusal which his cousin had steadfastly given him would naturally flow from her bashful modesty and the genuine delicacy of her character.

Baidu Translate: Mr. Collins并没有长久地沉浸在他成功的爱情的沉思中；Bennet夫人在前厅里闲逛，为了结束会议的结束，一看见伊丽莎白打开门，迅速地把她递给楼梯，她就走进了早餐室。奥姆，并祝贺他和她在温暖的条件下的幸福前景，他们更密切的联系。Collins先生收到并归还了这些祝贺，然后继续讲述他们采访的细节，结果他相信他有充分的理由感到满意，因为表兄坚定地给予他的拒绝自然会流露出来。她羞怯的谦逊和她性格中真正的精致。

Google Translate: 科林斯先生并没有长时间沉默地思考他成功的爱情; 对于班纳特太太来说，在前厅看见会议结束时，伊丽莎白刚刚打开门，快步上了楼梯，然后她进入早餐室，并祝贺他和她自己。温暖的条件是他们更近的联系的愉快前景。柯林斯先生平等地接受并回复了这些祝贺，然后继续将他们的采访细节联系起来，结果他相信他完全有理由感到满意，因为他的表弟坚决拒绝给他的拒绝自然从她害羞的谦虚和她性格的真正美味中流露出来。

Human Translation: 柯林斯先生独自一个人默默地幻想着美满的姻缘，可是并没有想上多久，因为班纳特太太一直待在走廊里混时间，等着听他们俩商谈的结果，现在看见伊丽莎白开了了门，匆匆忙忙走上楼去，她便马上走进饭厅，热烈地祝贺柯林斯先生，祝贺她自己，说是他们今后大有亲上加亲的希望了。柯林斯先生同样快乐地接受了她的祝贺，同时又祝贺了她一番，接着就把他跟伊丽莎白刚才的那场谈话，一五一十地讲了出来，说他有充分的理由相信，谈话的结果很令人满意，因为他的表妹虽然再三拒绝，可是那种拒绝，自然是她那羞怯淑静和娇柔细致的天性的流露。

B. Analysis of the Translation Results

When the translation system of Baidu translated the names of the characters in the book, it still expressed the names in English. At this point, Google Translate performed better than Baidu Translate. From the translated text,“not left long to the silent contemplation” was translated into “ 并没有长久地沉浸...的沉思” in the first machine translation, while the sentence was translated into “并没有长时间沉默地思考” in the second machine translation. At the same time, both of the target texts have “successful love” turned into “成功的爱情”, which is not appropriate. In the first translated text, the Baidu translation system translated “quick step pass her towards the staircase” to “迅速地把她递给楼梯”. In the original text, the author wants to express that Elizabeth steps up the stairs, but the first target text obviously does not have the same meaning. What is more, the syntactics in the sentence is wrong. “Staircase” should be the object in the translated text, and “her” should be the subject in the translated text.

In short, there are many problems about semantics, pragmatics, grammar and context in the target texts translated by both the Baidu translation system and Google translation system. However, the text translated by the Google translation system is relatively better than that of the Baidu translation system.

C. Questionnaire Data

The questionnaire was administered to 40 English major postgraduates, 40 non-English major postgraduates and 20 college English teachers. Though 100 people took part in the questionnaire survey, they comprised three kinds of people: non-English major postgraduate, English major postgraduates and college English teachers. The non-English major postgraduates all have passed the CET-6 test. The English level of non-English major postgraduates is middle. The English major postgraduates all have passed TEM8. The English level of English major postgraduates is middle high. The college English teachers all have some experience in teaching English, and they are all English major doctoral students. The English level of college English teachers is middle high. Thus, we assume that the reliability of non-English major postgraduates' evaluation is middle, the reliability of English major postgraduates' evaluation is middle high, and the reliability of college English teachers' evaluation is high. For the convenience of calculation, we regard them as 3 evaluators. The reliability set $B(B_1, B_2, B_3) = (M, MH, H)$. Table 5 presents the statistical data with respect to the translation quality questionnaire of the Baidu translation software. Table 6 shows the statistical data with respect to the translation quality questionnaire of the Google translation software.

TABLE V
Evaluation of the Baidu Translation System

	Syntactics	Pragmatics	Context	Semantics
B_1	M	M	ML	ML
B_2	MH	ML	M	M
B_3	M	L	M	ML

TABLE VI
Evaluation of the Google Translation System

	Syntactics	Pragmatics	Context	Semantics
B_1	ML	M	MH	ML
B_2	M	ML	MH	M
B_3	M	M	M	L

D. Evaluation Process

From Table 4 and Table 6, we can obtain the evaluation matrix of the Google translation system.

$$R_1 = \begin{pmatrix} (ML, M) & (M, M) & (MH, M) & (ML, M) \\ (M, MH) & (ML, MH) & (MH, MH) & (M, MH) \\ (M, H) & (M, H) & (M, H) & (L, H) \end{pmatrix}$$

According to formula (3), we can obtain

$$\alpha_M = \frac{\int x\mu_M(x)dx}{\int \mu_M(x)dx} = 0.5,$$

$$\alpha_{MH} = \frac{\int x\mu_{MH}(x)dx}{\int \mu_{MH}(x)dx} = 0.75,$$

$$\alpha_H = \frac{\int x\mu_H(x)dx}{\int \mu_H(x)dx} = 0.9.$$

So

$$\begin{aligned}(M, M) &= ((0.35, 0.5, 0.5, 0.65), (0.35, 0.5, 0.65)) \\ &\approx (0.25, 0.35, 0.35, 0.46),\end{aligned}$$

$$\begin{aligned}(MH, MH) &= ((0.55, 0.7, 0.7, 0.85), (0.55, 0.7, 0.85)) \\ &\approx (0.48, 0.61, 0.61, 0.74),\end{aligned}$$

$$\begin{aligned}(M, H) &= ((0.35, 0.5, 0.5, 0.65), (0.8, 0.9, 1)) \\ &\approx (0.33, 0.47, 0.47, 0.61),\end{aligned}$$

$$\begin{aligned}(M, MH) &= ((0.35, 0.5, 0.5, 0.65), (0.55, 0.7, 0.85)) \\ &\approx (0.30, 0.43, 0.43, 0.56),\end{aligned}$$

$$\begin{aligned}(MH, M) &= ((0.55, 0.7, 0.7, 0.85), (0.35, 0.5, 0.65)) \\ &\approx (0.39, 0.50, 0.50, 0.60),\end{aligned}$$

$$\begin{aligned}(L, H) &= ((0, 0.1, 0.1, 0.25), (0, 8, 0.9, 1)) \\ &\approx (0, 0.09, 0.09, 0.24),\end{aligned}$$

$$\begin{aligned}(ML, MH) &= ((0.15, 0.3, 0.3, 0.45), (0.55, 0.7, 0.85)) \\ &\approx (0.13, 0.25, 0.25, 0.38),\end{aligned}$$

$$\begin{aligned}(ML, M) &= ((0.15, 0.3, , 0.30.45), (0.35, 0.5, 0.65)) \\ &\approx (0.12, 0.21, 0.21, 0.32).\end{aligned}$$

Then, Table 6 can be expressed as Table 7.

TABLE VII
Evaluation Data of the Google Translation System

	Syntactics	Pragmatics
B_1	$(0.12, 0.21, 0.21, 0.32)$	$(0.25, 0.35, 0.35, 0.46)$
B_2	$(0.3, 0.43, 0.43, 0.56)$	$(0.13, 0.25, 0.25, 0.38)$
B_3	$(0.33, 0.47, 0.47, 0.61)$	$(0.33, 0.47, 0.47, 0.61)$
	Context	**Semantics**
B_1	$(0.39, 0.5, 0.5, 0.6)$	$(0.12, 0.21, 0.21, 0.32)$
B_2	$(0.48, 0.61, 0.61, 0.74)$	$(0.3, 0.43, 0.43, 0.56)$
B_3	$(0.33, 0.47, 0.47, 0.61)$	$(0, 0.09, 0.09, 0.24)$

After deblurring, we obtain Table 8.

TABLE VIII
Deblurring Data of the Google Translation System

	Syntactics	Pragmatics	Context	Semantics
B_1	0.3267	0.4118	0.5967	0.2267
B_2	0.4600	0.2033	0.7197	0.4600
B_3	0.4951	0.4951	0.4951	0.1023

From formula (9), we obtain the score of the Google translation system.

TABLE IX
Score of the Google Translation System

Syntactics	Pragmatics	Context	Semantics
0.3267	0.2033	0.4951	0.1023

In Table 10, we can observe that the syntactics of the Google translation system is 0.3267, the pragmatics of the Google translation system is 0.2033, the context of the Google translation system is 0.4951, and the semantics of the Google translation system is 0.1023. The results show that the Google translation system has done a relatively good job in ensuring the information integrity of the translated text. However, the Google translation system should make improvements in the areas of semantics, syntactics and pragmatics. The final evaluation score of the Google translation system is

$$\begin{aligned}b &= \sum_{i=1}^{4} r_i w_i \\ &= 0.1261 \times 0.3267 + 0.2669 \times 0.2033 \\ &\quad +0.3141 \times 0.4951 + 0.1023 \times 0.2929 \\ &= 0.2809.\end{aligned}$$

In the same way, we can obtain the evaluation results of the Baidu translation system, which are shown in Table 11.

TABLE X
Results of the Translation Systems

	Syntactics	Pragmatics	Context	Semantics	Finally
Google translation	0.3267	0.2033	0.4951	0.1023	0.2809
Baidu translation	0.4871	0.1802	0.3107	0.1955	0.2644

From Table 11, we can see that the pragmatics and context of the Google translation system are better than the factors of the Baidu translation system. Considering the weight of the influence factors, we find that the final evaluation score of the Baidu translation system is 0.2644. The evaluation result of the Google translation system is better than that of the Baidu translation system. On the basis of those facts, we can conclude that the target text of Google translation system is more satisfactory than the target text of the Baidu translation system. The conclusion is the same as the analysis that we made before.

E. Contrast Experiment

The machine translation evaluation is generally divided into manual evaluation and automatic evaluation [35]. The proposed method is compared with commonly used automatic evaluation methods.

1) Automatic Translation Evaluation Method: The automatic translation method usually measures the quality of machine translation by the similarity between machine translation and human translation. Different automatic translation evaluation methods calculate the similarity from different angles. Bilingual evaluation understudy (BLEU) is based on n-gram co-occurrence between machine translation and human translation [5]. This method lets c be the length of the machine translation and s be the reference human translation corpus length, and the brevity penalty BP is as follows [5]:

$$BP = \begin{cases} 1, & \text{if } c > s, \\ e^{1-s/c}, & \text{if } c \leq s. \end{cases} \quad (11)$$

Then,

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log p_n), \quad (12)$$

where w_n is the weight of the n-gram, p_n is the penalty factor [36]

$$p_n = \frac{\sum_i^{K_n} \sum_i^{k_n} min(C_i, max(c_i))}{\sum_i^{k_n} c_i}.$$

The penalty factor can be used if the number of words in the machine translation is shorter than in the manual translation.

Meteor is based on the harmonic average of accuracy rate and recall rate [7]. The accuracy rate and the recall rate are between the machine translation and human translation. The score produced by the proposed method is compared with the meteor method and BLEU method, as shown Table 11.

TABLE XI
Comparison with Automatic Evaluation Methods

	BLEU	Meteor	Proposed Method
Baidu Translate	0.046896	0.091418	0.2644
Google Translate	0.103864	0.133523	0.2809
Time	0.073s	0.066s	11.43min

From Table 11, we find that automatic evaluation methods can obtain similar conclusions with the proposed method in evaluating which machine translation system is better. Moreover, the automatic evaluation methods use less time. However, the scores of automatic evaluation methods are obviously low. Furthermore, the automatic evaluation methods can only determine which machine translation system is best. They cannot analyze the concrete details of the difference. For example, the proposed method can express that the Baidu machine translation system is weak to address semantics, but automatic evaluation methods cannot do this. That is, automatic evaluation methods cannot interpret the deep meaning of the articles.

V. Conclusion

In this paper, we have proposed an integrated evaluation method of machine translation quality based on Z-numbers. By taking advantage of Z-numbers, we measure the reliability of evaluators, and we try to judge which translation is more accurate. To prove the effectiveness of the method, we translate two excerpts from the same original text by two different machine translation systems. Then, we collect evaluation data through a questionnaire survey, and we use the real data to test the method. The results of the automatic evaluation methods and the proposed method are compared. The facts confirmed that the results of the study correspond to the theoretical analysis. The proposed method can evaluate the machine translation quality adequately and intuitively. In the future, we will try to combine the theory of the proposed method with the theory of the automatic evaluation system to make the evaluation process more convenient.

References

[1] Z.Tu, Y.Liu, L.Shang, X.Liu, H.Li. "Neural machine translation with reconstruction," In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017:3097-3103.

[2] J.Mandravickaite, T.Krilavicius, K.L.Man, "A Combined Approach for Automatic Identification of Multi-Word Expressions for Latvian and Lithuanian," *IAENG International Journal of Computer Science*, 2017, 44(4):598-606.

[3] S.Elbagir, J.Yang. "Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment," *Proceedings of the International MultiConference of Engineers and Computer Scientists 2019*, Hong Kong, 2019:122-16.

[4] Y.Wang, A.Maeda. "Twitter User's Interest Detection by Using Followee Information Based on LDA Topic Model," *Proceedings of The International MultiConference of Engineers and Computer Scientists 2019* Hong Kong, 2019:40-44.

[5] K.Papineni, S.Roukos, T.Ward, W.Zhu. "BlEU: a method for automatic evaluation of machine translation," *40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, 2002:311-318.

[6] C-Y Lin. "ROUGE: A Package for Automatic Evaluation of summaries," In *ACL Workshop on Text Summarization Branches Out*, 2004:74 – 81.

[7] S.Banerjee, A.Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005: 65-72.

[8] H.Isozaki, T.Hirao, K.Duh, et al. "Automatic evaluation of translation quality for distant language pairs," *Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, 2010:944-952.

[9] C.Carlsson, R.Fullér. "Fuzzy multiple criteria decision making: Recent developments," *Fuzzy Sets and Systems*, 1996, 78(2):139-153.

[10] S.Wang. "A Manufacturer Stackelberg Game in Price Competition Supply Chain under a Fuzzy Decision Environment," *International Journal of Applied Mathematics*, 2017, 47(1):49-55.

[11] J.Liu, Q.Zhang, Z.Luo. "Dynamical analysis of fuzzy cellular neural networks with periodic coefficients and time-varying delays," *IAENG International Journal of Applied Mathematics*, 2016, 46(3):298-304.

[12] L.L.Wang, L.M.Wang. "Global Exponential Stabilization for Some Impulsive TS Fuzzy Systems with Uncertainties," *IAENG International Journal of Applied Mathematics*, 2017, 47(4):425-430.

[13] G.J.Klir, B.Yuan B. "Fuzzy sets and fuzzy logic - theory and applications," Prentice-Hall, Inc. 1994:283 – 287.

[14] L.Jin, G.Yin, D.Yang. "Fuzzy Integrated Evaluation for Measuring Quality of Feature Space-Based Component Model," *International Conference on Internet Computing in Science and Engineering. IEEE*, 2008:349-354.

[15] X.J.Zhang. "The fuzzy integrated evaluate method of translation quality," *Seventh International Conference on Fuzzy Systems and Knowledge Discovery. IEEE*, 2010:502-506.

[16] E.Ageeva, M.L.Forcada, F.M.Tyers, J.A.Pérez-Ortiz. (2015). "Evaluating machine translation for assimilation via a gap-filling task," In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*.

[17] Q.Ma, Y.Graham, S.Wang, et al. "Blend: a Novel Combined MT Metric Based on Direct Assessment—CASICT-DCU submission to WMT17 Metrics Task," *Proceedings of the Second Conference on Machine Translation*. 2017: 598-603.

[18] L.A.Zadeh, "A Note on Z-numbers," *Information Sciences*, 2011, 181(14):2923-2932.

[19] R.Aliev, "Decision Making under Z-Information, Control and Systems Engineering," Springer International Publishing, 2015:233-252.

[20] W.Jiang, B.Wei , X.Liu , et al. "Intuitionistic fuzzy power aggregation operator based on entropy and its application in decision making," *International Journal of Intelligent Systems*, 2018, 33(1): 49-67.

[21] Z.Q.Xiao, "Application of Z-numbers in multi-criteria decision making," *International Conference on Informative and Cybernetics for Computational Social Systems. IEE*E, 2014:91-95.

[22] Z.S.Xu. "A Method for Priorities of Triangular Fuzzy Number Complementary Judgement Matrices," *Fuzzy Systems and Mathematics*, 2002, 16(1):47-50.

[23] A.M.Yaakob, A.Gegov. "Interactive TOPSIS Based Group Decision Making Methodology Using Z-Numbers," *International Journal of Computational Intelligence Systems*, 2016, 9(2):311-324.

[24] R.R.Yager. "On a view of Zadeh's Z-numbers," In: Greco S, Bouchon-Meunier B, Coletti G, Fedrizzi M, Matarazzo B, Yager RR, editors. Advances in computational intelligence,*Proceedings of 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (IPMU 2012), Catania, Italy: Springer, Berlin; July 9 – 13, 2012. Part III,90 – 101.

[25] B.Y.Kang, D.J.Wei, Y.Li, D.Y. "A method of converting Z-number to classical fuzzy number," *Journal of Information and Computational Science*, 2012, 9(3):703-709.

[26] D.Qiu, Y.Xing, R.Dong. "On Ranking of Continuous Z-Numbers with Generalized Centroids and Optimization Problems Based on Z-Numbers," *International Journal of Intelligent Systems*, 2018, 33(1): 3-14.

[27] Z.Yao. "Translation Criticism in the Perspective of Cognitive Translation Studies," *Foreign Languages and Their Teaching*, 2014:15-19.

[28] G.Bordogna, G.Pasi, D.Lucarella. "A fuzzy object-oriented data model for managing vague and uncertain information," *International Journal of Intelligent Systems*, 2015, 14(7):623-651.

[29] Sun Y Q, Zhou M K. "Study on Fuzzy Mathematical Evaluation Method of the Machine Translation Quality Based on Network Survey," In *Preceeding of the 2016 International Academic Conference on Human Society and Culture*. Guangzhou: Atlantis press, 2016:298-303.

[30] Chang D Y. "Applications of the extent analysis method on fuzzy ahp," *European journal of operational research*, 1996, 95(3):649 – 655.

[31] R.C.Wang, Y.Zhang. "An Improved PROMETHEE Method Based on Z-numbers Theory," *Journal of Engineering Management*, 2017, 1(31):118-122.
[32] R.R.Yager. "On a semantics for neural networks based on fuzzy quantifiers," *International Journal of Intelligent System*s, 2010, 7(8):765-786.
[33] Sun Y Q, Zhou M K. "Study on Fuzzy Mathematical Evaluation Method of the Machine Translation Quality Based on Network Survey," 2016 *International Academic Conference on Human Society and Culture*.
[34] J.Austen. "Pride and prejudice," Wordsworth Editions Limited, 1992.
[35] H.Y.He, J.T.Yu. "LORANGE: an Automatic Meta-evaluation Method of Machine Translation Evaluation," *Compute Knowledge and Technology*, 2017, 13(5):138-142.
[36] M.Brkić, S.Seljan, T.Vičić . "Automatic and Human Evaluation on English-Croatian Legislative Test Set," *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, 2013, 7817(1):311-317.