Study on evaluation method of machine translation quality based on questionnaires and data analysis

Yiqun Sun[1, 2, a], Minkang Zhou[1, b]

[1] Facultat de Traducció i d'Interpretació, Universitat Autònoma de Barcelona, Spain;
[2]International Education College of ChangChun University, ChangChun University, China

[a] 13969161305@163.com, [b] Minkang.Zhou@uab.cat

Keywords: Evaluation, machine translation, questionnaires, Euclidean distance

Abstract. In order to do a global evaluation of translation software and compare their differences within specific indicators, the study applied Euclidean distance and cosine similarity to evaluate 4 commonly used translation software tools, based on the analysis of the data from the questionnaire. The results suggest that software A and B are good in general, especially at the vocabulary translation. However, there are still some shortcomings in the discourse translation. Therefore, they should accumulate more materials in the Chinese corpus. It is concluded that Euclidean distance covers the shortage of randomness and the limitations to the micro level of the present evaluation methods such as BLEU. Meanwhile, Euclidean distance has higher accuracy and more convenient calculation than the cosine similarity. It can be used to evaluate translation software effectively.

Introduction

In today's information era, the main carrier of the cultural exchanges is the language, and the translation among different languages is exactly the key to the cultural exchanges. The machine translation has been a major hot spot in the computer science and computing-related fields. An accurate evaluation method is the main basis for the system development process, and it is one of the main driving forces to promote the development of MT system.

After a few years' development, the study of automatic MT evaluation method in the world has achieved fruitful results. Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is"– this is the central idea behind BLEU【1】, the method of Yokoyama based on two-way MT[2], the method of Yasuda、Akiba and Papineni based on N-gram language models for sentence similarity computing【3,4】. 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, in other words, 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.

In this paper, we obtain the data through questionnaire and calculate directly the similarity between MT and original text, basing on the analysis of the data, without turning to the artificial translation. In this way, we overcome the disadvantages of the indeterminacy of the artificial translation. We hope carry out an objective and accurate evaluation on currently commonly used MT software. Our aim is to find out the possible problems and put forward feasible suggestions for improvement.

The questionnaire design To evaluate the quality of a translation, we have to consider three main factors: vocabulary, grammar and discourse[5]. The vocabulary contains four specific aspects: semantic collocation, rhetoric, terminology and use of dialect; the grammar: only one aspect, that is to investigate whether the grammar of the translation is correct or not; the discourse seven aspects: cohesion, coherence, intentionality, acceptability, informedness, context and intertextuality. In this paper, we design a questionnaire based on the above parameters, which contains three major categories of indicators and twelve minor categories of indicators. According to the concept of Likert scale, these twelve minor categories of indicators, each one contains a total of five levels: from the best to the worst.

Level 1: Complete transfer of the original text information; only minor revision needed to reach professional standard. Level 2: Almost complete transfer; there may be one o two insignificant inaccuracies; requires certain amount of revision to reach professional standard. Level 3: Transfer of the general ideas but with a number of lapses in accuracy; needs considerable revision to reach professional standard. Level 4: Transfer undermined by serious inaccuracies; through revision required to reach professional standard. Level 5: Totally inadequate transfer of the original text content; the translation is not worth revising. The translation is mostly incoherent.

To answer the Questionnaire, choose one of the 5 levels of each indicator.

The selection of the software and the text for the investigation In this paper, we select 4 commonly used translation software: Google, Bing, youdao and baidu translator, which are marked as A, B, C and D. The first one, developed by Google Company, can provide instant translation between 103 languages; the second one, developed by Microsoft Asia Research Institute, 40 languages; the third one, by Netease company, 52 languages and the last one, by Baidu company, 27 languages.

Meanwhile, the text for the investigation is an article of February 12, 2016 from Asahi shim bun, *Presidential election season brings reality of U.S. democracy into spotlight*

Data processing and analysis

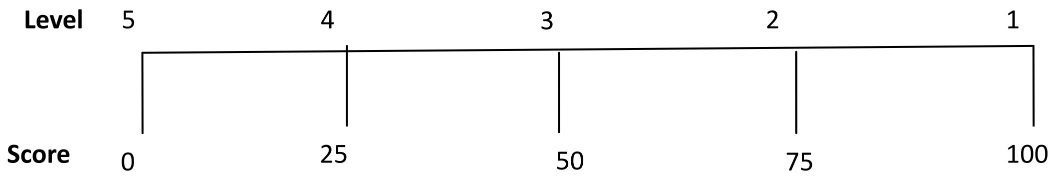

Fig.1 The correspondence between levels and scores for translation quality

In this paper, we surveyed 100 readers, which were asked to fill in the questionnaire. After the investigation, we calculated the number of each choice and make the data into tables. The statistical results of the Software A are showed en Table 1. In order to carry out quantitative analyses, we assigned respectively the 5 levels. The correspondence of the 5 levels and scores is shown in Fig.1. According to the Fig.1 and the number of each choice of the Software A, we processed the data, calculated scores of each indicator of the Software A and make the data into the right column of Table 1. In the same way, the scores of each indicator of the Software B, C and D are showed in the Table 2.

Table 1 The statistical results of the software A

Major Categories Indicators	Minor Categories Indicators	Number of People					Score
		Level1	Level2	Level3	Level4	Level5	
Vocabulary	Semantic Collocation	95	3	2	0	0	0.9825
	Rhetoric	94	2	3	1	0	0.9725
	Terminology	96	3	1	0	0	0.9875
	Use of Dialect	100	0	0	0	0	1
Grammar	Grammar	95	1	1	2	1	0.9675
Discourse	Cohesion	80	6	5	7	2	0.8875
	Coherence	78	5	8	4	5	0.8675
	Intentionality	82	9	3	3	3	0.91
	Acceptability	85	2	9	3	1	0.9175
	Informedness	84	7	7	1	1	0.93
	Context	86	6	5	2	1	0.935
	Intertextuality	90	5	2	1	2	0.95

In this paper, the quality of the translation is represented by a one-dimensional vector. Through the correlation analysis of the vectors, we carry out a quantitative research on the quality of the translations of the four software and compare them in the various indicators, to help the reader to choose a good software and point out the defects of the software for their developers.

As shown in the Table 2, the one-dimensional vector of quality of the software A‘s translation, AQ is:

AQ= （semantic collocation, rhetoric, terminology, use of dialect, grammar, cohesion, coherence, intentionality, acceptability, informedness, context and intertextuality）=（0.9825 0.9725 0.9875 1 0.9675 0.8875 0.8675 0.91 0.9175 0.93 0.935 0.95）

In the same way, the one-dimensional vector of quality of the software B‘s, software C‘s and software D‘s translations, BQ, CQ and DQ are:

BQ= （0.9675 0.975 0.98 1 0.8975 0.83 0.8375 0.8825 0.98 0.9725 0.9 0.9525）

CQ= （0.8 0.8125 0.84 0.88 0.8975 0.9475 0.995 0.985 0.9525 0.97 0.96 0.9575）

DQ= （0.76 0.7225 0.795 0.8225 0.905 0.9575 0.9725 0.955 0.96 0.965 0.955 0.9375）

From the right column of the Table 2, we can see that all the indicators for the best translation take the highest score 1, as a standard for the comparison between the quality of the translations of the software A, B, C and D. The one-dimensional vector of the quality of the standard translation, OQ is:

OQ= （1 1 1 1 1 1 1 1 1 1 1 1）

Take some of the minor categories of indicators and make them into a vector. Through the correlation analysis of the vector, we study the pros and cons of the software in some aspects, for example, the vector of the software A that reflects the quality of the vocabulary translation, AC is:

AC=（semantic collocation, rhetoric, terminology, use of dialect）=（0.9825 0.9725 0.9875 1）

In the same way, we can also give the vectors of the software B, C and D that reflect the quality of the vocabulary translation, BC, CC and DC and the vector of the standard translation OC.

The vector of the software A that reflects the quality of the discourse translation, AY is:

AY= （cohesion, coherence, intentionality, acceptability, informedness, context, intertextuality）= （0.8875 0.8675 0.91 0.9175 0.93 0.935 0.95）

In the same way, the vectors of the software B, C and D that reflect the quality of the discourse translation, BY, CY and DY and the vector of the standard translation OY.

We often use Euclidean distance and cosine similarity to describe the correlation of two vectors. This article is in this context, and does some works about the vectors that reflect the quality of the translation.

Table 2 The statistical results of the software A、B、C and D

Major Categories Indicators	Minor Categories Indicators	Software A	Software B	Software C	Software D	Score for the Best Translation
Vocabulary	Semantic Collocation	0.9825	0.9675	0.8	0.76	1
	Rhetoric	0.9725	0.975	0.8125	0.7225	1
	Terminology	0.9875	0.98	0.84	0.795	1
	Use of Dialect	1	1	0.88	0.8225	1
Grammar	Grammar	0.9675	0.8975	0.8975	0.905	1
Discourse	Cohesion	0.8875	0.83	0.9475	0.9575	1
	Coherence	0.8675	0.8375	0.995	0.9725	1
	Intentionality	0.91	0.8825	0.985	0.955	1
	Acceptability	0.9175	0.98	0.9525	0.96	1
	Informedness	0.93	0.9725	0.97	0.965	1
	Context	0.935	0.9	0.96	0.955	1
	Intertextuality	0.95	0.9525	0.9575	0.9375	1

The Euclidean distance R between two n-dimensional vectors a(x1,x2,…,xn) and b(y1,y2,…,yn) is,

$$R = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad （1）$$

The smaller the value of the Euclidean distance is, the smaller the difference between the two vectors is, that is to say, this software is better; on the contrary, the bigger the value of the Euclidean distance is, the bigger the difference between the two vectors is, in other words, this software is worse.

We can calculate the cosine of the angle between the two n-dimensional vectors a(x1, x2,…,xn) and b(y1,y2,…,yn) to measure the similarity between them, that is to say, we can figure out the cosine similarity of the two n-dimensional vectors. The cosine of the angle is,

$$\cos(\theta) = \frac{a \cdot b}{|a| \cdot |b|} \qquad （2）$$

That is：$\cos(\theta) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$ （3）

The range of values for the cosine similarity cos(θ) is [-1,1]. The bigger cos(θ) is, the smaller the angle between the vectors is, the more similar the two vectors are; on the contrary, the smaller cos(θ) is, the bigger the angle between the vectors is, the less similar the two vectors are. When the directions of the two vectors coincide，the cosine of the angle get the maximum: 1; when the directions of the two vectors are entirely opposite, the cosine of the angle gets the minimum: -1.

According to the above calculation method, we quantify the quality of the translations of the four software, which are shown in the Table 3. The data of the second line and the third line in Table 3 are respectively the Euclidean distance between AQ、BQ、CQ、DQ and OQ and cos(θ)of AQ、BQ、CQ、DQ and OQ, which are calculated respectively by the formula (1) and the formula (3). The Fig.2(a) shows the Euclidean distance between the vector of the four software's translations and the vector of the standard translation. Meanwhile, the Fig.2(b) shows the cos(θ) between the vector of the four software's translations and the vector of the standard translation. The smaller the Euclidean distance between the two vectors is, the more similar the two vectors are, the translation of the software is more similar to the standard translation. From the Table 3 and the Fig.2(a) we can see that, according to the quality of the translations, from the best to the worst, the four software can be ordered as A、B、C、D；The bigger the cos(θ) between the two vectors is, the smaller the angle between the two vectors is, the more similar the two vectors are, the translation of the software is more similar to the standard translation. From the Table 3 and the Fig.2(b) we can see that, according to the quality of the translations, from the best to the worst, the four software also can be ordered as A、B、C、D. However, from the results of the method of cos(θ) we can see little difference between the data: the difference appear on the third place after the decimal point. The distinguish ability is not very good. It is easy to cause deviations. As a result, we suggest that use the Euclidean distance to evaluate the quality of MT, rather than the method of cos(θ).

Table 3 The Euclidean distance and the cos(θ) between the vector of the four software's translations and the vector of the standard translation

software	A	B	C	D
Euclidean distance	0.2429	0.3084	0.3677	0.4801
cos(θ)	0.9991	0.9982	0.9975	0.9953

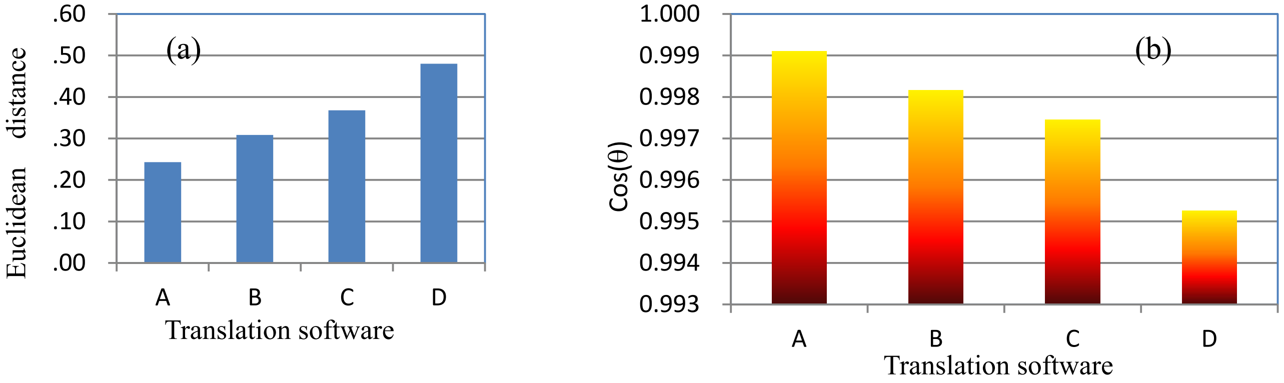

Fig.2 The Euclidean distance and cos(θ) between the vector of the four software's translations and the vector of the standard translation

Take some of the minor categories of indicators and make them into a vector. Through the correlation analysis of the vector, we study the pros and cons of the software in some aspects. For example, the vectors of the four software that reflect the quality of the vocabulary and discourse translation are AC、BC、CC、DC、OC and AY、BY、CY、DY、OY. The Euclidean distances between AC、BC、CC、DC and OC and the Euclidean distances between AY、BY、CY、DY and OY are shown in the Table 4.

From the Table 4 we can see that, the software A is the best for vocabulary translation, then the software B, C and D. For discourse translation, from the best to the worst, the order is D, C, A, B. In short, A is equivalent to B and C and D are at a level, in other words, google and bing are good at vocabulary translation, but not so good at discourse translation; baidu and youdao are just on the contrary.

Table 4 The Euclidean distance between the vector of the four software's translations and the vector of the standard translation

software	A	B	C	D
Vocabulary	0.0349	0.0456	0.3393	0.4562
Discourse	0.2382	0.2873	0.0978	0.1155

In summary, the translation quality of the software A and B is good in general, especially good at the vocabulary translation. However, there are still some shortcomings in the discourse translation. That is because the software A and B are both developed by foreign companies. With a limited Chinese corpus, for the translation of long sentences and complex sentences, the results will sometimes turn out to be not smooth and difficult to understand. Therefore, we suggest that accumulate more materials in the Chinese corpus.

It should also be noted that the text chosen for this study is a informative text, which has certain limitation. It cannot represent the situations of the expression text and the vocative text.

Conclusion

(1) Through questionnaire, we can collect data and analyze them effectively to determine the quality of the translation software.

(2) In this paper, the quality of the translation is represented by a one-dimensional vector. Through the correlation analysis of the vectors, we carry out a quantitative research on the quality of the translations of the four software and compare them in the various indicators, to help the reader to choose a good software and point out the defects of the software for their developers.

(3) We suggest that use the Euclidean distance to evaluate the quality of MT, because of its high accuracy and convenient calculation. On the contrary, the distinguish ability of the cosine similarity is not very good. It is easy to cause deviations.

(4) The results of the paper show that, among the four software: google, bing, youdao and baidu, the first two are good in general, especially good at the vocabulary translation. However, there are still some shortcomings in the discourse translation. On the contrary, the last two have advantages in this part.

References

[1] 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.

[2] 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

[3] 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]

[4] 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

[5] H. L Zhang, Parametric Analysis on Evaluation of Translation Quality，Yilin，2011.No8 p70-76

[6] Wuensch, Karl L. What is a Likert Scale? and How Do You Pronounce 'Likert?'. East Carolina University. October 4, 2005.

[7]P. Newmark, Approaches to Translation, Prentice Hall, 1988-1-1