English Translation Effect Evaluation System Based on Big Data

Wang Xudong*
School of Foreign Languages, Longnan Teachers College, Chengxian, Gansu, China
*Corresponding author: houty2019@stu.scu.edu.cn

Abstract. In recent years, ontology has been widely used to model world knowledge at the conceptual level. In recent years, machine translation has been widely used. This paper focuses on the evaluation of machine translation system. This paper reviews and compares the types and standards of machine translation, the content of system evaluation, and the main methods of system evaluation. The experimental results show that a high-quality machine translation system must fully integrate linguistic knowledge and language neutral world knowledge. This paper introduces an ontology based English Chinese machine translation model system, which organizes concepts into a hierarchical structure and establishes rich conceptual connections among concepts. The accuracy of machine translation is improved by mapping words in a language to concepts in ontology. The experimental results show that the accuracy of machine translation can be improved by 10.8%.

Keywords: Ontology, Knowledge Modeling, Machine Translation, Mapping

1. Introduction
Different from manual translation, the evaluation of machine translation not only evaluates the quality of translation, but also evaluates other performance of machine translation system. In recent years, more and more attention has been paid to the evaluation of machine translation. In the past few years, there have been a number of influential evaluation activities in the world, such as the message understanding Conference (MUC) to evaluate the problem of proper noun recognition. Text retrieval conference (TREC) evaluates the development of information retrieval, as well as many evaluation activities of machine translation and speech technology. All these activities have a great influence on the development of machine translation.

The applicable scope of machine translation is relatively small, which is usually a "relatively limited professional field". "It is not used for texts with strong literary or cultural flavor, but for popular science documents, financial and commercial transactions, administrative memorandum, legal documents, manuals, agricultural and medical materials, industrial patents, brochures, newspaper reports, etc." [1]. Nagao limits the application scope of machine translation to scientific and technological documents, article titles and general sentences, and excludes poems, literary works, legal documents, tender contracts, etc. [2-3]. Therefore, when people evaluate machine translation, they limit the types of texts evaluated. However, there are different standards in the evaluation of machine quality. It is not only limited to the pure translation quality, but also involves the operability of
The 2nd International Conference on Computing and Data Science (CONF-CDS 2021)  
Journal of Physics: Conference Series 1881 (2021) 042067  
doi:10.1088/1742-6596/1881/4/042067

People often need to synthesize multiple criteria in order to evaluate machine translation system comprehensively, objectively and fairly [4-5].

In this sense, it coincides with (2) (faithfulness standard: faithfulness standard, accuracy standard and dynamic equivalence standard) of general category basic standard system (abstract standard) and (8) (translation standard of scientific and technological works) in non class special basic standard system (specific standard) of Gu Zhengkun's translation pluralistic system. Arnold et al. Suggested that the following factors should be considered in the evaluation of machine translation [6-7]. Figure 1 shows the machine translation system.

![Figure 1. The machine translation system](image)

1. The commodity system of machine translation should be distinguished from the study of machine translation
2. Evaluate machine translation system from the perspective of users;
3. The quality of machine translation should be viewed from the perspective of adaptation of needs and results.

The index system of machine translation evaluation directly determines the R&D route of machine translation researchers and the development direction of machine translation. Evaluation criteria have an important impact on the development and research of machine translation, and the famous ALPAC report in the United States in 1960s has a great impact on MT evaluation [8]. Therefore, how to evaluate a machine translation system objectively and fairly constitutes an important research topic [9].

2. Types and evaluation criteria of machine translation

In 1995, C. Boitet first distinguished four types of machine translation [10].

1. For the watcher, it is called MT-W, which is designed to help readers access foreign language materials. In this case, they prefer to accept the "rough" translation (sometimes after a certain degree of post editing) rather than nothing.

2. For the reviser, known as MT-R, is designed to automatically generate "rough" translations, which are similar to human translated drafts. Therefore, it can save professional translators time-consuming and laborious work, and make them become revisers.

3. For the translator, it is called MT-T. It aims to assist translators in their work by providing online dictionaries, thesaurus, translation case base, etc;

4. For the author, known as MT-A, is provided to the author who wants the work to be translated into one or more languages. The author is willing to write under the control of the system or help the system eliminate the possible ambiguity in the translation.

At the same time, he proposed that the success or failure criteria of machine translation can be divided into five standards: conceptual standard, engineering standard, practical standard, commercial standard and communication standard

1. Concept standard: whether to put forward interesting new concept and demonstrate its feasibility and superiority with the help of simulation or experimental prototype, it is mainly related to
researchers.

(2) Engineering standard: whether technical innovation is completed or more advanced programming techniques are used to build prototypes or systems. It is mainly related to system developers.

(3) Practical standard: whether to run the prototype or complete system economically and reasonably under practical conditions and achieve satisfactory results. It mainly involves users.

(4) Commodity standards: judging from an economic point of view, not just the number of installations or sales.

(5) Communication standard: whether the decision maker or the company in the field has established a good image.

The above criteria are peripheral evaluation standards, or auxiliary standards. The translation quality of a system's output plays a key role in the evaluation of the system, so it is very important to establish the standard of translation quality. Figure 2 shows the structure of neural network. At present, Alpac report is widely used in the world to evaluate the quality of translation, which is divided into intelligibility and fidelity.

The scale of intelligibility was divided into 9 levels:

- **Level 9:** completely clear and easy to understand, just like reading ordinary text, with appropriate style.
- **Level 8:** completely or almost completely clear and easy to understand. However, there are small grammatical errors or style defects, and / or small improper usage, which are easy to "correct".
- **Level 7:** it's clear and easy to understand, but it's not as good as grade 8 in style, word selection, or syntactic arrangement.
- **Level 6:** the main idea can be understood at a glance. However, due to the poor choice of words, improper style, unaccustomed expression, untranslated words and incorrect grammar, the overall understanding is affected. Only by post editing can the translation be recognized.
- **Level 5:** only after careful consideration can you understand the main idea. However, despite the "noise" interference, the main idea can be detected.
- **Level 4:** In fact, the sentence is not so much intelligible as incomprehensible. However, the meaning can be vaguely understood, word selection, syntactic arrangement and / or expression are strange, and the key words are not translated.
- **Level 3:** Generally speaking, it is incomprehensible and does not read like gibberish. However, after careful consideration and consideration, the meaning of the sentence can be assumed.
- **Level 2:** it doesn't read like gibberish. After a lot of thinking and deliberation, it can not be understood.
- **Level 1:** can't understand at all, how to think and ponder can't read the sentence.

Unfortunately, the original report did not include the levels of fidelity. Intelligibility and
faithfulness can reflect the quality of the translation. Conceptually, these two indicators are independent of each other. One translation can be clear and easy to understand, but lacks fidelity. For example, Lin Shu's translation and Jean Francois ducis (1733-1816) [11] "beautiful but not faithful". But on the other hand, it can also be "very accurate and faithful, but poor comprehensibility", or "faithful but not beautiful", such as Lu Xun's "believe rather than Shun". In the latter case, ALPAC thinks that it is the poor intelligibility of the original text [ALPAC, 1966:67]. ALPAC also gives the index of information degree, which is divided into 10 levels. At that time, 18 native speakers of English and strong reading ability of scientific and technological Russian were evaluated. The specific method was to let the judges read the translation first, then the original Russian sentences, and then decide whether they got new information when reading the original Russian sentences. The maximum number of new information is 9, and when there is no new information, it is 1. The less new information is added, the more accurate the translation is, the higher the quality of translation is. Figure 3 shows the data processing.

The Japanese agency for science and technology's MT translation evaluation consists of two criteria: intelligibility (5 levels in total) and faithfulness (7 levels in total). Scale of intelligibility:

1. The meaning of the article is clear and there is no objection to it.
2. The meaning of the article is clear and understandable, but there are some problems in the use of words, grammar and style. However, these problems are easy to correct.
3. The significance of the article can be understood on the whole. However, due to the use of words and grammar. Some details of the understanding of doubt, the reader can not completely rely on their own correction, but want to ask people who understand the original text.
4. The quality of the translation is poor, and there are many problems in the use of words and grammar. After repeated and careful thinking, it is better to guess the meaning of the original text to a certain extent. It is better to have people revise it than to translate it manually.
5. The translation is totally incomprehensible and has to be retranslated.

Scale of fidelity:

1. The translation faithfully reflects the content of the original.
2. The translation faithfully reflects the content of the original text, and the meaning of the article is easy to understand, so only simple modification is needed;
3. Basically, it can faithfully reflect the content of the original text, but it needs to adjust word order and other similar modifications.
4. The content of the original text is basically faithfully translated. However, there are errors in the relationship between phrases, past tense, perfect tense, the difference between singular and plural, the position of adverbs, etc., which need to be adjusted in the structure after translation.
5. The content and structure of the original text are not well reflected, and some of them are omitted, and the collocation of phrases and clauses is wrong.
6. The structure and content of the original text are not well translated, and there are omissions in the translation of phrases and sentences, but it can still be seen as a sentence in general.
7. The translation can't reflect the content and structure of the original text at all, because the subject or predicate falls off, and it can't be a sentence.

EUROTRA of the European Community has adopted another set of evaluation criteria:
(1) Identification criteria: (a) legibility, (b) loyalty, (c) coherence, (d) usefulness, (e) reading speed, (f) acceptability;
(2) Economic standards: (a) input time, (b) edit and correct time, and (c) clear time.
(3) Language standards: (a) coherence of sentence structure and semantics, (b) vocabulary evaluation, and (c) translation errors.
(4) The difficulty of using the system.
IBM's Bleu (Bilingual evaluation under study) evaluation method considers that if the translation of a translation system is closer to manual translation, the higher its translation quality will be. Therefore, the key to the evaluation is how to define the similarity between the system translation and the reference translation. Bleu uses the method of comparing and counting the number of N-ary words co-occurrence that is, counting the number of N-ary words that appear in both the system translation and reference translation. Finally, the number of N-ary words matched is divided by the number of words in the system translation. The evaluation results show that Bleu method is simple and feasible, but does not consider the recall rate of translation (recall, which refers to the ratio of the number of unknown words identified and the total number of unknown words in the text). Figure 4 shows the data verification structure.

![Figure 4. Data verification structure](image)

The domestic research on MT evaluation is mainly conducted by the Institute of computational language, Peking University. After the 1990s, the Institute has researched and developed a machine translation evaluation system. The system uses the classification evaluation method and establishes the machine translation test outline. The National 863 program is also organizing expert evaluation from time to time. The evaluation results reflect the development level of machine translation in China at that time.

From the above-mentioned standards, the evaluation of machine translation system takes the quality of translation as the core of evaluation, and the criteria of translation evaluation also fall on fidelity and intelligibility.

3. Content of machine translation system evaluation
The translation quality of machine translation is the core problem of machine translation system research. However, there are many other important factors to be considered. Some factors are inherent characteristics of the system and cannot be measured by the current "black box". At present, it is generally accepted that there are seven main factors involved in the evaluation of machine translation system [12].

3.1 The quality of machine translation
Translation and machine translation are the most important indicators. However, the quality of translation is difficult to quantify, and its evaluation is still a very difficult task for people. Equivalence between two languages is a vague concept. In fact, the same source language can have unlimited number of different target language translations, not to mention the source language itself is full of ambiguity.

3.2 Application efficiency
Whether machine translation can improve work efficiency is another important indicator that users,
especially professional translation companies, pay attention to. However, machine translation should be placed in the whole process of language information processing, such as retrieval, recognition, input, pre editing, translation, post editing, output, type setting, printing, distance communication, etc. To test whether the machine instead of human translation saves a lot of time.

3.3 Practical environment
The application environment includes the hardware requirements of machine translation system, the dependence on other software, the requirements of input text, the quality and compatibility of user interface, etc. Human machine interface is one of the main factors that determine the machine translation system. In general, users prefer to choose a system with poor translation quality but easy translation modification. They are reluctant to choose the interface with better translation quality but not easy to modify.

There are two kinds of man-machine interface, which are user oriented and development and maintenance personnel (including linguists). The user oriented interface should have the functions of pre editing, post editing and dictionary expansion. The interface for development and maintenance personnel should be able to easily modify dictionaries, rules and language models. It can also provide statistical data such as word frequency, sentence pattern and error type.

4. Main methods of machine translation system evaluation

4.1 Operational evaluation
Operational evaluation is also called practical evaluation, sometimes called economic evaluation. This evaluation is concerned with the economic value of machine translation system, considering the cost of machine translation, the price of each word and the cost of post editing, focusing on the cost of machine translation and human translation and the time ratio. In addition, the operational evaluation is too expensive and time-consuming for a user, and it is difficult to make the evaluation environment completely simulate the actual situation, which is ignored by most users.

4.2 Explanatory evaluation
Declarative Evaluation is also called Qualitative Evaluation, and this evaluation is also called Output Evaluation, which is also called a standard evaluation method. This method does not stick to the special situation of individual users. Instead, it tries to give a more general and extensive evaluation to the machine translation system, which tends to focus on the translation quality, focusing on evaluating the performance of the machine translation system by evaluating the translation quality.

Therefore, when using this method, there are at least three aspects to consider:
(1) Corpus used for testing: The material selection should be careful and fair. Different corpora have great influence on the evaluation of the same machine translation system.
(2) Scoring in experimental design: The basis of scoring has subjective color factors. Moreover, it is often influenced by environment, time and labor intensity. Therefore, more evaluators are needed to score more samples to obtain statistical significance.
(3) Result analysis: The results only show the quality of the translation, and the data obtained are difficult to be further processed.

5. Conclusion
To sum up, it is not difficult to see that the improvement of the quality of machine translation and the improvement of the evaluation system of machine translation system are both complex and long-term processes. If we say that the results of machine translation are still unsatisfactory, the way of computer-aided translation needs to be revolutionized. So what should we look for a breakthrough. To make computer translation more intelligent? The measurement and speed of information exchange in today's language requires a new technology that adapts to the global communication speed. To construct a multilingual digital interactive platform which can adapt to the rapid and large capacity
language information exchange. Maybe we can find the light of tomorrow in a new way of code switching.

**Acknowledgement**

Research of Application about College English Teaching on Problems Based Learning

- Item number: GS[2020]GHB4788
- Item Number:(GSSKG20-11), Educational Reform Project of Gansu Social Science and United Higher Education Press

**References**

[1] Cheng Yanan. Cross Cultural Communication and the Modernization of Chinese Libraries. Library and Information Services, 1997 (01): 11-14

[2] Chen Xianyuan. Motivation, Trend and Form of Intercultural Communication. Journal of Shanghai Jiaotong University: Philosophy and Social Sciences, 2000

[3] Dai Xun. Chinese Image and Its Migration in Cross-cultural Communication. Social Science Front, 2004 (1): 97-103

[4] Han Weihong. Cultural Differences and Intercultural Communication. Journal of Hehai University: Philosophy and Social Sciences, 1999 (04): 21-24

[5] Sun Weiping. on Cultural Diversity and Intercultural Communication. Shandong Social Sciences, 2011, 11: 5-9

[6] Xie Hua. Rationality and Inevitability of Cultural Misreading in Intercultural Communication. Jiangxi Social Sciences, 2006, 1: 186-189

[7] Zhang Yuan. Cross Cultural Communication and Communication of Chinese Dance in Southeast Asian Countries. Journal of Beijing Dance Academy, 2018, No.128 (03): 49-56

[8] Gu Mingdong. on Ultimate Equality in Cross-cultural Communication. Journal of Sun Yat Sen University, Social Science Edition, 2015, 55 (5): 1-12

[9] Peng Jixiang. Conflict and Fusion of Traditional Culture and Modern Consciousness. Chinese Film in Intercultural Communication. Film Art, 2001, 1: 4-9

[10] Li Mingjie. Characteristic analysis and operation control of large scale model high voltage AC / DC hybrid power grid. Power Grid Technology, 2016, v.40; No.389 (04): 10-16

[11] Fu Fuying, Zhang Xiaolu. Analysis of Cultural Memes in Intercultural Communication and Communication. Jiangxi Social Sciences, 2012, 000 (011): 242-245