Recognition of False-friend and Machine Translation Technology

Xianggui Xie[a], Ying Jiang[b], Jinwei Li[c], Fan Wang[d], Siling Chen[e], Zijie Liang[f] and Jingyi Ling[g]

School of Management, Beijing Normal University, Zhuhai 519087, China.

[a]18998736215@163.com, [b]jpz6311whu@bnuz.edu.cn, [c]King_vey@163.com, [d]wfwfpaper@163.com, [e]echo_csl@163.com, [f]892875925@qq.com, [g]13631261705@163.com

Keywords: Chinglish, LanguageTool, machine translation, idiom.

Abstract. With popularity of learning English in China, many problems such as Chinglish have been exposed out. This paper is based on the existing technology of LanguageTool's rule-base. In this paper , we compile the rules which are suitable for English articles, analyze the errors of false-friend , put forward some suggestions on improving the machine translation and propose a small error correction program based on the rules of false-friend we compiled.

Introduction

Machine Translation (MT) indicates a wide prospect of development because of the rapid development of Internet and the more and more frequent global communication. The traditional manual work can't meet the rapid rising needs of translation. People often learn English by MT which can not only reduce the cost of learning English but also improve the ability of English quickly and effectively. But the disadvantages of MT become increasingly remarkable. First, MT is mostly based on literal translation so the accurate rate is not credible. Second, the difficulties for English learners are not only vocabulary, grammar and word-formation but also the linguistic differences between English and Chinese [1]. Practically, the deep-rooted Chinese thinking mode misleads the translation, which caused the English vocabulary and grammar with Chinese habit. We call it Chinglish. This phenomenon existed widely in MT. From what has been mentioned above, doing research on the Chinglish plays an important role in correcting Chinglish and promoting Cross-language communication.

Research Status of MT

Research Status at Home and abroad

Research Status at Home. In 2011, Haifeng Wang, vice chairman of the International Institute of Computational Linguistics (ACL), and the chief scientist of the Baidu, who also proposed " statistical machine translation technology in his research, which has the advantages of strong model learning ability, new language ability, good robustness, and so on. But the defects were also very prominent, mainly reflected in two aspects: long distance adjustment ability and poor translation logic" [2]. In 2014, Yuyu Wang in her research put forwards "Youdao and Kingsoft translation software have the mistranslation phenomenon of meaning of a word, preposition and conjunction in the vocabulary .And they also have the phenomenon of disordered words, improper collocation and tense mistranslation. In addition, there is also a mistranslation of literary language, place names and personal names caused by different cultural background." [3]. In 2015, Shuang Dong compared the translation of Google in the functional equal functional principle and proposed "Google translation only focused on the form of equivalence, but ignored the semantic equivalence" [4]. From these points of view, the domestic experts did not put forward the real meaning of the solution in the machine translation, just only listed some examples to prove machine translation has the qualitative problems, simply classified the error and did not put forward the effective improvement measures for translation machines.

Research Status at Abroad. In 2013, Chung-ling Shih proposed “Machine translation (MT) serves as a fast way of transmitting web information across countries and MT performance can be highly improved by adapting the source text using a controlled language (CL).” [5].In 2014, Yusuf Yaylaci, Arman Argynbayev proposed “False cognates, also known as false friends, are pairs of words which have a similar form and/or pronunciation but different meanings in two languages”. The aim of their study is to raise awareness in avoiding misunderstanding, which English-Russian false-friends cause in English-medium intercultural communication [6]. In the same year, M. Forsyth and B. Haggart's article investigated the difficulties in transplanting global legal norms into developing countries, specifically the problem of “false friends”. This is a linguistics concept describing the situation where there is a striking resemblance between two words in two different languages, leading speakers of each language to assume, incorrectly, that they understand the word's meaning in the other language [7]. Therefore, the English-Chinese false friends MT technology has not yet involved.

False-friend in machine translation. Domestic and foreign scholars can’t put forward an effective measure for the phenomenon of False-friend in the machine translation result. However, this phenomenon exists widely in many kinds of translation machines and hasn't been solved. To make the English learners in China have a better use of translation machines, we compile a False-friend rule-base and study on the phenomenon of the False-friend in the Baidu, Youdao, Google and Bing translation machines.

Foundation of Rule-base Based on the Technology of LanguageTool

Chinese idiom (CI) is a fixed phrase with four Chinese characters. Idiom is the essence of language with strong national color and distinctive cultural connotation. In Jikuang Zhang’s study of the manifestation in CI, he classified the error of CI in six manifestations, included the words collocation, the misuse of words, the redundancy or loss of words, the improper sentence structure, rhetorical error and cultural background [8]. When the machine translation translated the CI and idioms, the error of the translation always contain these six manifestations. Here are some examples:

This example of idioms is mainly expressed the meaning of “take chair”. But the learner miswrited it into “take the chair”. The correct translation of “take the chair” in Chinese is “担任主席”, “主持会议” or “起步跳舞” and the correct translation of “take chair” in Chinese is “取椅子” or “拿椅子”.

This example of CIs is mainly expressed the meaning of “get cocky”. But the learner miswrited it into“tail up”. The correct translation of “tail up” in Chinese is “兴高采烈” or “兴趣盎然” and the correct translation of “get cocky” in Chinese is “翘尾巴”.

So we divided false-friend into idiom and Chinese idiom in this paper and collected the false-friend rules mainly by Internet and books, which ensure both the accuracy and the reliability. Finally, we've collected 377 CIs and 502 idioms. However, this collection way is not comprehensive so we will add more false-friend rules continuously.

The technology of rule-base written by us mainly was based on the rules of XML and Regular Expression (RE). By virtue of the Chinese Lexical Analysis System segmenting translation which is written in Chinese and Xinhua near-synonym finding the similar words, we successfully finished the compiling of rule-base. These two tools increase the practicality of rules.

Development of False-friend Translation Software

Idea of design. LanguageTool has the function of accessing rule base and identifying text error. We use these characters of LanguageTool to evaluate the accuracy of Baidu, Youdao, Google and Bing machine translation’s API so as to analyze the reason and propose solving scheme.

Matching Algorithm Between Rule-base and Machine Translation. Translation is a course of which translator expresses the meaning of the original words out by the way of acceptance, understanding, digest and then elaborates clearly in another language to make the readers gain the

same feeling or similar meaning to the original ones [9].Therefore, to judge whether the machine translation is correct, compared with the correct translation seems very necessary. Excellent matching algorithm can reduce the workload of the researcher and ensure probability of the error is small. The following is a compilation of the algorithm matching sample:

(1) Direct matching method

When compared with the translation of the machine translation and the correct translation, if these two translation is the same, the translation of machine translation is right. .Loop matching method.

(2) Loop matching method

The first loop matching is to use the first character of correct translation match with every character of the machine translation. If successfully matched, then go to the second loop matching. When the next character of machine translation compared with the next character of correct translation, if these two characters are not the same or the length of correct translation is longer than current length of machine translation, then exit the second loop matching. If the second loop matching can completely finished, the translation of machine is right. As shown in Figure 1:

```
For WordOfMT from FirstWordOfMT to LastWordOfMT step OneWord
  if WordOfMT = FirstWordOfMT
    TranslationIsRightOrNot←True
    for WordOfCorrectTranslation(CT) from FirstWordOfCT from to LastWordOfCT step
OneWord
      if WordOfCT != CorrespondingWordOfCT
        TranslationIsRightOrNot←False
      end if
    end for
    if TranslationIsRightOrNot = True
      return True
    end if
  end if
end for
```

Fig. 1 An example of loop matching method

(3) Word segmentation matching method

The word segmentation matching method is to segment the two translations with segmentation API before the operation of the loop matching, meanwhile the matching process is not a single character matching, but matching between segmented vocabulary and segmented part of speech, and finally combined with the next segmented vocabulary to promote the improvement of the cycle based algorithm.

(4) Near semantic matching method

In consideration of machine translation and correct translation may exist the situation of “The different words with the same meanings”, calling the synonym database API can greatly reduce this problem. After the segmentation of translation, each segmented vocabulary can query words that have the similar meaning in synonym database, and if this vocabulary is as same as the near-synonym of the correct translation, then matching with next segmented vocabulary so as to improve the segmentation algorithm.

Matching algorithm process

First, get the corresponding original text and the correct translation from the rule-base. Second, calling translation machine API to attain the translation, and then use the direct matching method, loop matching method, word segmentation matching method and near semantic matching method. If one of the methods is right, the translation of the machine is also correct. Otherwise is wrong. As shown in figure 2:

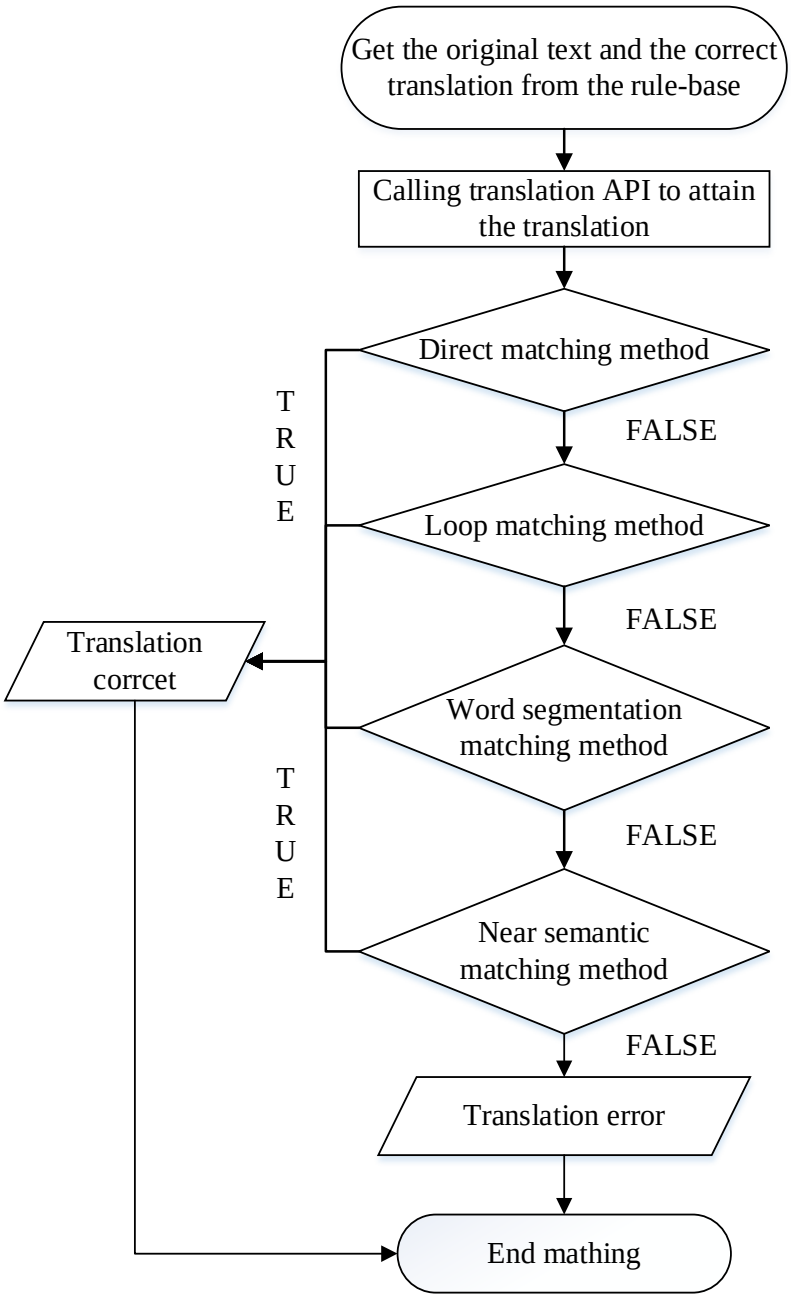

Fig. 2 Process of matching algorithm

Test of Translation Machine

Correct Rate of Matching

We take the translation machine's accuracy to evaluate the machine.

Translation Machine's Accuracy: numbers of correct translation divided by all numbers of translation.

Machine translation accuracy is not only counted by matching algorithm but also by artificial contrast.

The high quality of translation can make the reader obtain the same understanding and feeling as the original article. As for algorithm can't judge whether the translation is correct or not in semantic and context, it is necessary to carry out artificial statistics. Artificial statistics can also test the accuracy of matching algorithm. In the end, the error number of the matching algorithm is 317, the accuracy rate can reach 90.98%. Therefore the artificial statistics can be replaced by the matching algorithm. As shown in Figure 3:

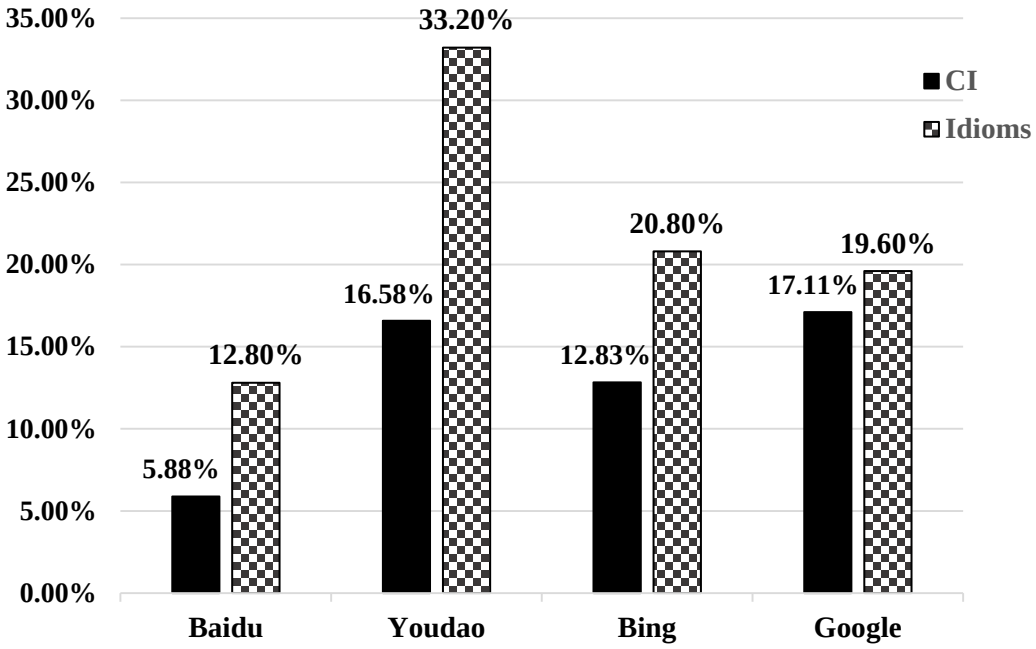

Fig. 3 Correct rate of translation machine

Whether it is CI or idioms, the correct rate of these four translation machines have the following features:

① The correct rate is low, there are only 33.2% correct rate and 17.11% correct rate on CI and idioms;

② The correct rate of idioms is higher than CI;

③ Youdao translation machine's correct rate is higher than others;

④ Baidu translation machine's correct rate is the lowest.

Thus it can be seen: it is necessary and significant to improve the machine translation.

Evaluated Result

Through the study of the correct rate of Chinglish in the translation machine of Baidu, Youdao, Bing and Google, we found that there was great difference among the correct rates of these four translation machines. In view of the difference, we made a deeper research on it.

After reading a lot of related authoritative literatures [2][10][11][12], we know that Baidu, Youdao, Bing and Google these four translation machines translation of Google are using statistical machine translation technology, the basic idea is to build a statistical translation model, and then use this model to translate. All the process need a large number of parallel corpus as a support. [12]Therefore, the different correct rate of each machine translation is due to the difference of the corpus.

Among these companies, Youdao company has developed its own dictionary called Youdao Dictionary. We also have done some researches on it. As shown in Table 1:

Table 1 Correct rate of Youdao dictionary translation

Research types	Total numbers	Translation numbers	Correct numbers	Correct rate
CI	377	179	147	82%
Idioms	502	360	331	92%

The correct rate of Youdao dictionary translation has the following features:

① The correct rate is high: the correct rate of idiom translation can be as high as 92%, and the correct rate of idiom translation can also be as high as 82%;

② The dictionary is lack of adequacy: There are 377 total numbers of CI and 502 total numbers of Idioms. As we can see from the Table 1, Only 179 numbers of CI can be translated and 142 numbers of Idioms cannot be translated.

Due to the existence of some translation can't be translated by Youdao Dictionary, the reliability of the final statistics's accurancy is not high. In order to statistic the real accuracy of

the Youdao dictionary, we need to count up the situations that those are correctly translated by Youdao dictionary meanwhile they are wrong interpreted by Yaodao Machine:

Table 2 Correct rate of Youdao dictionary translation

Research types	D(T), M(F) numbers	D(F), M(T) numbers	Others numbers
CI	104	3	207
Idioms	195	2	305

As shown in Table 2:

① This kind of D(F)M(T) situation happened rarely, and the probability of this kind error is less than 1%. Therefore, it may have a side effect, but very little.

② The situations which those are correctly translated by Youdao dictionary meanwhile they are wrong interpreted by Yaodao Machine reflects that dictionary can promote the accurancy of machine translation. And the Table 2 shows us idioms increase the correct rate of 38.84%, CI increase the correct rate of 27.59%.

To this point, we put forward three suggestions on machine translation:

① Expand the machine translation corpus: statistical machine translation techniques rely on corpus. In theory, the more data added into the corpus, the more accurate machine would achieve. If it is possible to expand the corpus continuously, the phenomenon of Chinglish can be reduced.

② Develop a dictionary machine translation interface: As the research shows, the correct rate of dictionary translation is high. If it is able to build a interface between dictionary and machine translation, machine translation can greatly reduce the times of Chinglish. Theoretically, it can improve the accurancy of CI and idioms about 27.95% and 38.84%. But the dictionary has its own unresolved defects. First, false-friend rules included in dictionary is not sufficient; Second, only Youdao company has its own dictionary, other company in this dictionary field is blank.

③ Increase the false-friend rules: in the machine translation of the corpus, we targeted to increase the false-friend rules continuously. At the same time, the advantages of this proposed method is more simple, and do not need new techniques. Although the language error can not be completely eliminated, the correct rate can be infinitely close to 100% if added continuously.Therefore, the author suggests that this method can improve the machine translation.

According to these three suggestions, we make a small error-correcting program that can judge whether the translation obtained from translation machine has the phenomenon of Chinglish.

Summary

In this paper, the research on the recognition of Chinglish and the machine translation technology is based on the CI and idioms. The results show that: these four translation machines gives a lower right rate in Chinglish. To some extent, it proves the necessity of improving the defects of Chinglish. What's more, it will affect use of learning among the English scholars. We hope that this study of CI and idioms in Chinglish can give other researchers to provide a reference role, meanwhile we hope that our suggestions can improve these four major translation machines. We are moving forward all the time.

Acknowledgements

This work is supported by a project granted by the National Social Science Foundation of China (Project No. 14CTQ041), a grant from Science and Technology Plan Project of Guangdong Province (Project No. 2014A080804001), and a grant from the Soft Science Research Project of Guangdong Province(Project No.2014A030304013). The corresponding author of this paper is Ying Jiang (jpz6311whu@bnuz.edu.cn).

References

[1]Ran Li. An Analysis of the Causes and Characteristics of Chinese English [J]. Journal of Chinese. 2007(03)

[2]Haifeng Wang. Internet Machine Translation [J]. Chinese Journal of information, 2011,(25): 72-80

[3]Yuyu Wang. The Application of English Chinese Translation Software in College Students' English Learning [J]. Modern Chinese (language research edition), 2014,(10):147-150

[4]Shuang Dong. Research on the translation of scientific and technological translation based on Functional Equivalence Theory -- a compared study of Google translation and human translation [J]. Business Culture, 2015, (2):128

[5] Chung-ling Shih. Adaptations in Controlled Cultural Writing for Effective Machine Translation: A Register-specific Probe [J].Theory and Practice in Language Studies, 2013, Vol.3 (7), pp.1093-1102

[6]Yusuf Yaylaci, Arman Argynbayev. English-Russian False Friends in ELT Classes with Intercultural Communicative Perspectives [J].Procedia - Social and Behavioral Sciences, 2014, Vol.122

[7]M. Forsyth, B.Haggart. The False Friends Problem for Foreign Norm Transplantation in Developing Countries [J].Hague Journal on the Rule of Law, 2014, Vol.6 (2), pp.202-229

[8]Jikuang Zhang. Study on the Manifestation of Chinese English [J]. Journal of Neijiang Teachers College, 2006(03)

[9]Shuqin Li. Context -- the Basis of Correct Translation -- an Analysis of English Translation of Chinese English Translation of Eight Grade Examination (2000) [J]. Chinese Translation, 2001, 01:42-46.

[10] Information on http://shared.youdao.com/www/about.html

[11] Information on http://help.bing.microsoft.com/#apex/18/en-us/n9999/0

[12] Information on http://translate.google.com/about/intl/en_ALL/