Corpus Translator's Style in the Era of Big Data Under Data Mining Algorithm

Bing Chen(✉)

Chengdu Polytechnic, Chengdu 610000, Sichuan, China
applechen123@126.com

Abstract. The research of corpus translation style in the era of big data under data mining algorithm is to use big data to analyze the characteristics and changes of translation style. This study aims to achieve the following goals: to analyze how translation styles have changed in different periods from ancient times to the present; Study how translation style changes according to various factors such as age, gender and major; Investigate whether there are any differences between the translations of professional translators and non professional translators.. In this article, we will introduce two problems in dealing with big data: (1) big data analysis and (2) translation quality evaluation. We will discuss how to solve these problems by using machine learning algorithms such as neural networks and deep learning models.

Keywords: Data mining · Corpus · Translation style

1 Introduction

"Corpus refers to the real corpus collected and stored in the computer on a large scale according to the specific purpose of language research using computer technology and certain linguistic principles. These corpora are marked to a certain extent, easy to retrieve, and can be applied to descriptive research and empirical research." The application of u corpus provides a lot of data support for the study of translator's style, which makes up for the shortcomings of traditional translation studies [1]. The research method of corpus has also become a rookie in translation studies. However, by searching the existing corpus based research on translator style, it is found that this kind of research mostly focuses on the Chinese translation of English works and the English translation of Chinese works, and there is little research on the translation of other languages.

Corpus not only brings new methods to traditional translation studies, but also provides new research ideas. With the extensive use of corpus, the study of translator's style with English text as the research object is emerging in endlessly, constantly adding a new research perspective to corpus translatology. However, it is rare to use corpus to analyze the translator's style of English texts. "Corpus translatology refers to the study of systematically analyzing the essence, process and phenomenon of translation based on corpus, with real bilingual corpus or translation corpus as the research object, data statistics and theoretical analysis as the research methods, and linguistic, literary and cultural theories

J. C. Hung et al. (Eds.): IC 2023, LNEE 1045, pp. 53–58, 2023.
https://doi.org/10.1007/978-981-99-2287-1_8

and translatology theories.“ Although corpus translatology did not appear early, it has developed rapidly and achieved a number of excellent research results, which can be observed from two aspects: methodology and case study. The characteristics of English texts, word form reduction and part of speech tagging of English texts involved in this study provide some methodological references for the study of corpus translatology, and also enrich the research results of translator style with English novels as the research object to a certain extent [2]. Based on this, our research is about the style of corpus translators in the era of big data under the data mining algorithm.

2 Related Work

2.1 Data Mining Text Classification

There are many ways to extract text features. The teacher of this course focuses on TF-IDF and chi square verification. Let’s first look at the calculation method of if-idf:

Term frequency (TF) refers to the frequency of a given word in the file.

Inverse document frequency (IDF). The IDF of a specific word can be obtained by dividing the total number of files by the number of files containing the word, and then taking the logarithm of the obtained quotient [3].

IDF is a measure of the general importance of a word. TF-IDF value is the product of TF value and IDF value.

TF-IDF comprehensively represents the importance of the word in the document and the document differentiation. However, it is not enough to use tf-df to judge whether a feature has discrimination in text classification. It does not consider the distribution of feature words among classifications. If a feature word is evenly distributed among various classes, such a word has little contribution to classification; However, if a feature word is concentrated in a certain class and hardly appears in other classes, such a word can well represent the characteristics of this class, and TF-IDF cannot distinguish between the two cases. The distribution of feature words in class internal documents is not considered. In documents within a class, if the feature words are evenly distributed among them, this feature word can well represent the characteristics of this class [4]. If it appears only in a few documents and does not appear in other documents of this class, it is obvious that such feature words cannot represent the characteristics of this class. Figure 1 below shows the data mining text extraction code.

In this paper, there are not too many strict requirements for words. This data mining experiment requires taking nouns, and all stop words are non nouns, so the operation of taking nouns only also removes all stop words. The problem of stop words mentioned below has been solved here. Similar operations are carried out on several corpora, turning each corpus into a noun [5]. If all nouns form a set without repeating elements (W1, W2, W3.. WN), this set is a dictionary. For each corpus, the values are 0 and 1 respectively according to whether the words in the corpus exist in the dictionary. So far, the work of converting corpus into word vector has been completed.

```
def getUrls(url):
    req = requests.get(url).text
    if (req == None):
        return
    bf = BeautifulSoup(req, 'html.parser')
    div_bf = bf.find('div', attrs={'class': 'content_list'})
    div_a = div_bf.find_all('div', attrs={'class': 'dd_bt'})
    urltxt = open(b'F:\data\url.txt', 'a', encoding='UTF-8')
    for div in div_a:
        link = div.find('a').get("href")
        urltxt.write(link+'\n')
    urltxt.close()
```

Fig. 1. Data mining text extraction code

2.2 Analysis of English Noun Structure

An English noun phrase can be composed of three parts: head, premodifier and postmodifier, and its structural sequence can be expressed in the following four forms:

1) Headword (noun phrase containing only the headword, such as noun phrase "China")
2) Prepositional modifier + headword (a noun phrase consisting of a headword and its prepositional modifier, such as a noun

   Phrase "the right procedure")

3) Headword + Post modifier (a noun phrase consisting of a headword and its post modifier, such as a noun

   Phrase "corruption aplenty")

4) Prepositional modifier + head word + Post modifier (composed of head word and its prepositional modifier and post modifier

   Noun phrases formed by words, such as the noun phrase "the talkative man in the center of theroom")

The head word is an indispensable part of any noun phrase and the core of the noun phrase; In front of it All modifiers of (or on the left) are collectively referred to as pre modifiers, and all modifiers after them are collectively referred to as post modifiers; in English grammar rules, the head word of a noun phrase can be acted as by a noun, pronoun or a nominalized word. Because nouns often represent new information in the language, and pronouns are a review of the previous old information, noun phrases with the head word as nouns are shorter than nouns with the head word as pronouns At the same time, relevant studies have proved that in English texts, the number of nouns is the largest and far greater than the number of pronouns, which may mean that the number of noun phrases with nouns as the center word is more than that with pronouns as the center word [6]. Therefore, in this study, we mainly identify noun phrases with nouns as the center word.

3 Research on Corpus Translator's Style in the Era of Big Data Under Data Mining Algorithm

In recent years, in addition to the traditional translation study, which takes the translated language as the main research object, the study of translator's style has gradually moved from invisibility to dominance under the framework of traditional translation theory. The translation and the translator have gradually got rid of the shackles of the source text and become one of the focuses of translation research. The study of translator's style is also developing towards the direction of independent and individual research. From the perspective of grammatical rules, the head word is the core of noun phrases and an indispensable part of any noun; In addition, there are relatively few parts of speech of words that can act as the head word, especially in this study, the part of speech of phrases whose head word is a noun is more specific. Therefore, we use the method of identifying the head word first and then its modifier for noun phrase recognition.

"Translation language features can be roughly divided into two categories: the generality of translation language and the language features of specific language on the translated text. The former refers to the universality and regular characteristics of the translated text, such as simplification, manifesting, normalization, implicit, etc. these characteristics are not limited by the source language and target language, and are closely related to the translation process itself [7]. The latter refers to the language features formed by the differences between specific source language and target language, which are mainly manifested in It refers to the characteristics of the translated text at the lexical and syntactic levels. "

After the recognition module completes the recognition of noun phrases of a sentence, the software will submit the sentence, the recognized noun phrases and their head words to the verification and judgment module, which will verify and judge the reliability and integrity of these noun phrases based on the corpus. The reliability judgment here mainly refers to determining the recognition of a noun phrase according to the frequency of the identified noun phrase and its colligation in the corpus as the main parameter: the integrity judgment refers to combining the two adjacent noun phrases with high reliability and the words between them into a new noun phrase in the order from left to right, Then by judging the reliability of this new noun phrase, we can determine the integrity of the original two adjacent noun phrases. If the reliability of the newly merged noun phrase is low, it indicates that the integrity of the original two noun phrases is high, and all the information of the newly merged noun phrase is discarded [8]; If the reliability of the newly merged noun phrase is high, it indicates that the integrity of the original two noun phrases is low and should be discarded, and the newly merged noun phrase should be used to replace the original two noun phrases; Press this cycle until the integrity of all noun phrases in the sentence is judged.

4 Simulation Analysis

It refers to the establishment of corpora based on real translation corpora according to specific research objectives (including monolingual comparable corpuses, bilingual / multilingual parallel corpuses, translation corpuses, etc.). Based on the electronic text

of the corpus and computer statistics, it describes all kinds of translation phenomena in a large or specific range, analyzes and explains translation phenomena on the basis of full description, or verifies various hypotheses about translation [9]. In essence, corpus translation is an interdisciplinary product of the combination of descriptive translation studies and corpus linguistics. The research on corpus translator style in the era of big data under data mining algorithm is the effective application of natural language processing technology to web documents. This method first linearizes the source code of the reconstructed web page, and preliminarily filters the noise of the web page, then filters and clusters the text blocks by using the methods of classification and clustering to get the text paragraphs of the web page, and finally absorbs the pseudo noise paragraphs to get the text of the web page [10]. This method does not need to build a tree for web pages, and has the characteristics of fast and accurate. However, due to the use of text classification, clustering, data mining and so on, there is a certain complexity. As shown in Fig. 2 below, the character code of the captured text is as follows.

```
head = requests.head(url)
req = requests.get(url)
req.encoding = 'GB2312'
bf = BeautifulSoup(req.text, 'html.parser')
div = bf.find('div', attrs={'class': 'content'})
h1 = div.find('h1')
head = re.sub(r'\s+', '', h1.get_text())
out = open(filepath + '.txt', 'w', encoding='GB2312', errors='ignore')
out.write(head+'\n')
timediv = div.find('div', attrs={'class': 'left-t'})
time = timediv.get_text().replace(" ", "")[0:16]
out.write(time)
p = div.find('div', attrs={'class': 'left_zw'}).find_all('p', text=True)
for ptext in p:
    out.write('\n'+ptext.text);
out.close()
```

Fig. 2. Grab character feature code

5 Conclusion

The research on the translation style of corpus in the era of big data under data mining algorithm is a research aimed at analyzing and studying the translation quality of corpus. The main purpose of this study is to find out the difference between the translation quality of human translators and machine translators, which will help to improve the machine translation system. This research has another goal: to improve our understanding of the characteristics, characteristics and problems of machine translation.

Acknowledgments. 1.Research on the Construction and Development of "Cross-Border E-Commerce English" Loose-Leaf Textbook Based on Corpus (WYJZW-2021-2006).

2."Big Data Application Research Center for Higher Vocational Foreign Language Education" (19kypt06), a scientific research platform project of Chengdu Polytechnic.

References

1. Han, H., Jiang, Y., Yuan, X.: Corpus-based study of translator's style in the big data era. Foreign Lang. Educ. (2019)
2. Wang, Q.: Readjustment of translator's status in the era of big data. J. Panzhihua Univ. (2018)
3. A Brief Analysis of the Translator's Subjectivity in the Process of Com puter Aided Translation. **23**, 2 (2019)
4. Zhen, C., Jiang, C.: Overview of data mining in the era of big data. Int. Core J. Eng. **5**(10), 136–139 (2019)
5. Xinmin, Z., Qiang, N.: A comparative study of the translator's style——a corpus-based case study of lianghuiwang. **2**, 10 (2018)
6. Singh, K.K., Kushwaha, V.: Smart Wireless Network Algorithm in the Era of Big Data (2021)
7. Xiao, Z., Wang, Y.: The model of translator's information literacy in the new era. In: Proceedings of the 2019 4th International Conference on Modern Management, Education Technology and Social Science (MMETSS 2019) (2019)
8. Bazylev, V.N.: G.E. von Spilcker is the translator of the "Satires" of Antioh Cantemir (translation experience in linguistic and cultural context of the era) (2019)
9. Yang, Y.: On the embodiment of the literary and translator's subjectivity in the translation of literary works——taking the three translation versions of jane eyre as examples. J. Heihe Univ. (2019)
10. Wang, S.: The evolving paratexts in the C-E translation of Tang poems in the era of fragmentation reading. J. Xi'an Int. Stud. Univ. (2018)