Scholar Publishing Group
International Journal of Multimedia Computing
https://doi.org/10.38007/IJMC.2022.030202
ISSN 2789-7168 Vol. 3, Issue 2: 8-15

Big Data in the Translation of Terminology for Foreign-Related Public Crisis Events

Yonghe Xiao*

School of Foreign Languages, Nanchang Institute of Technology, Jiangxi 330099, China

498154972@qq.com

**corresponding author*

Keywords: Big Data, Foreign-Related Public Crisis Events, Terminology Translation, Automatic Acquisition

Abstract: Today, as the world enters the age of informationization and digitalization, big data technology has become a key area of cooperation and development between China and foreign countries, and friction is inevitable when the language is not interchangeable, and foreign-related major public crisis events are major public crisis events involving foreign countries, which can easily lead to greater disasters if not handled properly, so it is very important to pay attention to the translation problems and propose countermeasures in this field. The purpose of this paper is to study the translation of terminology of big data technology in the handling of foreign-related public crisis events. This paper uses the research methods of description, induction and comparison to analyze the real corpus, mainly to further promote linguistic analysis of the terminology of foreign-related public crisis event handling in the context of big data. The paper analyzes and summarizes the patterns of terminology through individual cases, helping to identify accurate expressions and deepen the understanding of the field. The experimental results show that the accuracy rate of finding the terminology of foreign-related public crisis events through web pages and online dictionaries simultaneously reached 93%, indicating that the method and system of automatically obtaining terminology translation through web pages and online Chinese-English dictionaries under big data has good accuracy and the system is less time-consuming and more practical.

1. Introduction

The terminology in a specific domain is the customary convention in the domain, and its translation has a fundamental role in supporting machine translation, cross-lingual information retrieval, and ontology learning [1-2]. Foreign-related major public crisis events are much more complex than domestic-related major public crisis events, and they become even more complicated when it comes to terminology translation [3-4]. Foreign-related major public crises are not uncommon from ancient times to the present and are quite specific because they involve foreign countries, so a special study of them is absolutely necessary [5-6]. This can help us understand foreign ideas quickly and thus facilitate communication, of which terminology translation is the key and difficult point [7-8].

In the study of terminology translation of big data technology applied to foreign-related public crisis event processing, many scholars have conducted theoretical research and practical operation with good results [9]. For example, Noland R B proposed the method of combining syntactic and semantic features to identify entities, using statistical features for entity recognition, and the synthesis of both to achieve the recognition of Chinese entities. Gare G proposed the method of candidate translation suffix or prefix redundancy to remove noisy data [10].

This paper, on the basis of the main problems, difficulties and research status faced by big data-based translation acquisition, discusses the shortcomings of previous research, then puts forward the basic process and ideas of acquiring foreign-related public crisis event processing terminology translation from big data technology, and finds that three translations belong to the difficulties: first, terms appear sparse in non-professional corpus, and professional parallel corpus is not easily accessible, so statistical-based methods often encounter data sparsity problem. Secondly, specialized terms have flexible word constructions and special grammatical structures, which are also difficult to be handled by rule-based translation methods. Finally, the translation of terms is professional in nature, and for the fixed-use terms, the translation must be guaranteed to be strictly correct, and the paraphrase often does not meet the requirements.

2. Translation of Terminology of Big Data Technology in Foreign-related Public Crisis Events

2.1. The Main Problems of Terminology Translation Methods Based on Big Data Technology in Foreign-Related Public Crisis Events Handling

(1) Bilingual corpus co-occurrence web page retrieval

Retrieving the relevant web pages for terminology translation from a huge amount of web pages is the first problem that needs to be solved for terminology translation based on big data, and usually translation systems use search engines to search and obtain relevant information. Retrieving the most relevant results for foreign-related public crises from hundreds of millions of information is the key to the success of terminology translation system based on big data technology. Constructing accurate and efficient query terms into search engine queries has become the focus of researchers, and commonly used methods include query term construction methods using heuristics, semantic prediction or keywords.

(2)Term translation extraction

Due to the unstructured nature of the web, how to extract term translation from web pages is another problem faced by term translation based on big data technology. Some researchers have used information such as brackets, colons, and keywords to process well-structured term translations and achieved high accuracy rates. However, due to the small number of such types of translations, the translation information in the remaining unstructured web pages present in the web is not extracted, which affects the recall rate of the system. How to extract the maximum number of translations with guaranteed accuracy is a problem that needs to be solved urgently.

(3) Noise data processing

The terminology translation system of traditional technology pays little attention to the processing of noisy data, however, the unstructured nature and complexity of the web lead to the extraction results containing irrelevant interference items. In the absence of parallel corpus, it is difficult to process the noisy data effectively. The poor accuracy of the results leads to a less practical and less reliable system, which requires manual re-calibration of the results. This becomes the third problem faced by terminology translation based on big data.

2.2. Terminology Translation Methods of Big Data Technology in Handling Foreign-Related Public Crisis Events

(1) Using minimal intersection word alignment

A simple string-to-serial alignment method is used to address the characteristics of foreign-related public crisis events in which the terms contain relatively few words and are syntactically difficult to analyze. The terminology bilingual alignment corpus construction based on the minimal intersection alignment effectively makes up for the shortage of bilingual dictionaries, and the obtained terminology level alignment contains forms such as present participle or past participle forms of English words, which can better meet the requirements of obtaining partial translations of terms.

(2) Using advanced web search

To find terminology translations, the search engine returns results in the form of a summary of the co-occurring parts of the terminology in the source web page and the online dictionary. In order to improve the efficiency of crawling information related to the handling of foreign public crisis events, the summary returned by the search engine is obtained directly without entering the specific web page. Also, in the case of multiple search terms of the same web page, an ellipsis is used to omit some irrelevant and overly long information. When the term and the translation in the source web page are in the same sentence or adjacent sentences, the search engine presents the original form in the web page, with the source term and the translation directly connected or separated only by punctuation, blanks or sign words.

Online dictionaries are word translation resources compiled by human beings, providing real-time translation functions of online Chinese and English words with high credibility. Online Chinese-English words have better effect on terminology translation for foreign-related public crisis events handling, so the module of obtaining terminology translation from online Chinese-English dictionaries is added to the system. The accuracy of obtaining translations from online Chinese-English dictionaries is high and less time-consuming.

However, due to the limited capacity of dictionaries, some specialized terms cannot be translated from online dictionaries. Therefore, it is necessary to obtain translations from online dictionaries first, and continue to obtain translations from web pages for terms for which translations cannot be obtained. Combining the two acquisition methods, the online dictionary and webpage resources are fully utilized to ensure the accuracy and recall of the results.

(3) Data cleaning

For the candidate translations extracted from the big data information, the candidate translations need to be normalized. Different variants of the same translation, such as singular-plural forms, different tenses, etc., are merged, and the form with the highest frequency is selected as the final form of different variants, and the frequencies of different variants are summed up after merging as the frequencies of the final candidate translation items.

For the cases that cannot be handled in the extraction mode, the English phrases that generate conflicts will be added to the translation conflict list first and left for final processing. After completing data cleaning, the generated candidate translation list is sorted from largest to smallest by the frequency of candidate translations extracted. For the term tuples that generate conflicts that cannot be judged, the English phrases that appear in the candidate translation list with the highest frequency are selected as candidate translations and added to the candidate translation list, and the rest are discarded.

2.3. Bilingual Word Alignment Techniques

Bilingual word alignment is an important problem in the field of natural language processing, aiming to determine the correspondence between words or phrases in bilingual translation pairs. Bilingual word alignment needs to be performed under the condition of aligned bilingual sentences. Bilingual word alignment is the cornerstone of statistical machine translation systems, and some parameters of statistical machine translation models often need to refer to word-aligned bilingual corpora. There are two main types of bilingual word alignment: lexicon based bilingual word alignment and statistical based bilingual word alignment.

The bilingual dictionaries contain higher quality lexical inter-translation information, which makes the lexicon-based word alignment methods highly accurate and easy to implement, and can achieve better results for data sparsity and small-scale data. However, the lexicon-based approach is more restricted by the lexicon capacity and domain, while the recall rate is often low due to the serious ambiguity alignment phenomenon in bilingual alignment, and the replacement between synonyms is difficult to handle. In order to improve the coverage of dictionaries, the interrelationship between bilingual words can be measured by calculating the Dice coefficient through bilingual dictionaries with the following formula:

$$\mathrm{Dice}(S1, S2) = 2|S1 \cap S2|/(|S1| + |S2|) \tag{1}$$

Given the Chinese word C and the English word E, the similarity of words C and E can be estimated by Dice (Sc, Se) assuming that the sets of sentences in which C and E appear in the corpus of sentence alignment are Sc and Se, respectively.

Statistical-based bilingual word alignment is mainly based on the noisy source channel model and Hidden Markov (HMM) model, etc. Brown et al. of IBM first proposed the noisy channel model to study bilingual word alignment, a process that can be described as a decoding process, where the translation process is understood as a source channel, and the source language word string T is sought for the target language string S. From the Bayesian formula, we can obtain:

$$\mathrm{P(T|S))} = \frac{\mathrm{P(t)\times P(s|T)}}{\mathrm{P(S)}} \tag{2}$$

The maximum value of T can be obtained from the above equation as:

$$\mathrm{P(T|S)} = \mathrm{argmaxP(T)} \times \mathrm{p(S|T)} \tag{3}$$

Among them, $\mathrm{P(T)}$ is the language model, which describes the fluency of the language, and $\mathrm{P(T|S)}$ is the translation model, which describes the accuracy of the translation. The problem of obtaining translations is transformed into a word alignment problem by means of a noisy channel model. the IBM family of models are both parametric models of word alignment, and they differ in the number and type of model parameters.

3. Experimental Study on Big Data Technology in the Translation of Terminology for Foreign-related Public Crisis Events

3.1. Research Subjects

The experimental data include 400 Chinese and English professional terms of foreign-related public crisis events, each term contains 2-10 Chinese characters or English words.

3.2. Experimental Design

More than 300,000 Chinese and English translation pairs of foreign-related public crisis events were obtained by manual or dictionary collation, etc., which were used to obtain partial translations of Chinese and English terms and used as bilingual translation resources to verify the accuracy of candidate translations when the candidate translations were verified.

3.3. Comparison Test System

In order to verify the effectiveness of the translation method and to correctly recognize the effect of the translation system implemented in this paper, it is necessary to use other translation systems for horizontal comparison. In this paper, Pharaoh, a phrase-based column search decoder, can use multiple alignment models and language models and combine them with different weight coefficients, but it is still based on the translation principle of the noise channel, except that multiple alignment models and language model. The decoding speed of Pharaoh is fast and easy to configure, and there is no special requirement on the input format of the source language, but it does not provide the source code, only the compiled system, and it is impossible to explore the internal implementation process. Pharaoh is a system compiled in a 32-bit environment, and it also uses the traditional memory way to manage and organize the alignment models.

4. Analysis of Big Data Technology in the Translation of Terminology for Foreign-related Public Crisis Events

4.1. Analysis of Using Big Data Technology to Obtain Translations of Professional Terms from Web Pages and Online Dictionaries

In the experiment, 388 terms can be obtained from the online dictionary for translation, and 368 terms can be obtained from the web page for translation. From the experimental results, it can be seen that the accuracy rate of web pages and online dictionaries in finding professional terms of foreign-related public crisis events at the same time reaches 93%, and the quality of translation pairs is satisfactory, as shown in Table 1.

Table 1. Comparison of web pages and online dictionaries to obtain translations of technical terms

Percentage	Web page	Online dictionary	Web page + Online dictionary
Accuracy rate	92%	97%	93%
Recall rate	83%	75%	90%

According to Figure 1, the accuracy rate is also 92% when only choosing to obtain translations from web pages. When the system uses only online dictionaries to obtain translations, the result reaches 97%, but the recall rate is only 75% due to the limited number of online dictionary translations.

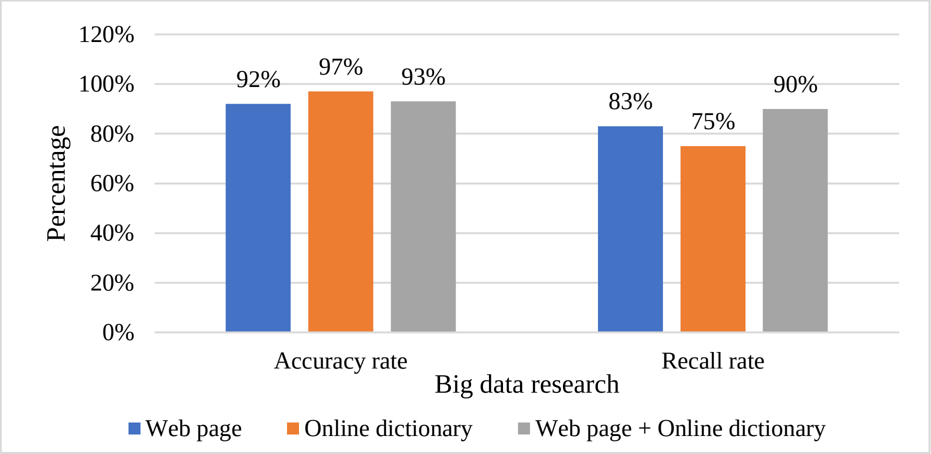

Figure 1. Comparison chart for obtaining terminology translations in web and online dictionaries

4.2 Analysis of the Speed of Acquiring Specialized Terms from Web Pages and Online Dictionaries by Using Big Data Technology

Based on the relevance feedback to dynamically control the amount of web page downloads, an effective translation decreasing rate is defined to portray the decreasing rate of the number of candidate translations obtained per page in the returned results of a query item relative to the previous page to dynamically terminate the query item. The average time consumption for each term in the experiment is stably between 4 and 7 seconds, as shown in Table 2.

Table 2. Terminology average time consumption data

Average term translation time/second	Web page	Online dictionary	Web page + Online dictionary
Experiment 1	4.3	3.7	5.9
Experiment 2	7	4.2	6.1
Experiment 3	5	3.5	5.8
Experiment 4	5.7	4.8	6.3

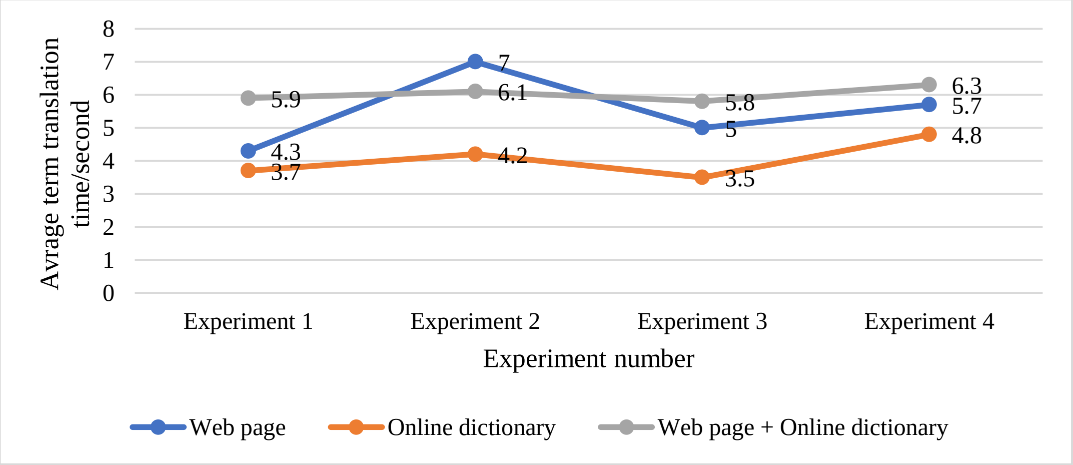

Figure 2. Average time spent on terminology

As shown in Figure 2, the convenient of searching the online dictionary further reduces the average terminology translation time consumption, which is basically within 5 seconds. The time taken for terminology search on the web page is slightly higher than that of the online dictionary, but it fluctuates relatively more and is not as stable as that of the online dictionary. If both web pages and online dictionaries are combined, the time spent is more stable, fluctuating around 6 seconds.

5. Conclusion

In the translation practice of foreign-related public crisis events, translators are often troubled by the translation of specialized terms. Since the relevant language is extremely strict, the major problem about how to accurately translate professional terms in translation cannot be ignored. In recent years, big data technology has made a great breakthrough, which has greatly improved the translation quality and provided a powerful means for the translation of professional terms. This paper analyzes the information of foreign-related public crisis events handling with broad content, clear classification, uniform format, standardized form, strict language and other characteristics suitable for translation by big data technology, uses bilingual parallel corpus based on minimum intersection word alignment to construct certain English-Chinese word translation pairs, and searches operators to merge terms and partial translations of terms. This paper proposes that using web pages and online dictionaries in big data technology can effectively improve the speed of professional terminology retrieval and ensure the accuracy and efficiency of translation.

Funding

Research project of Humanities and Social Sciences in Jiangxi Province (On the Present Situation and Countermeasures of Publicity Translation on Public Crisis Management of Chinese Government in Internet Era No. YY202009) & (On the Publicity Translation on Intangible Heritage of Jiangxi from the Perspective of Medio-Translatology No. YY19217).

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Introna, L. D. The Enframing of Code: Agency, originality and the plagiarist. Theory Culture & Society, 2011, 28(6):113-141. https://doi.org/10.1177/0263276411418131

[2] Shahhoseiny H. A Study of Errors in the Paragraph Writing of EFL Learners: A Case Study of First Year Translation Students at University of Applied Science and Technology in Bushehr, Iran. Embo Reports, 2015, 5(6):1307. https://doi.org/10.17507/tpls.0506.26

[3] Figar S , Aliperti V , Taliercio V , et al. P1-423Assessment of influenza outbreaks using a private healthcare information system: an analysis of the 2009 H1N1 epidemic in Buenos Aires.

Journal of Epidemiology & Community Health, 2011, 2(1):75-85. https://doi.org/10.1136/jech.2011.142976g.13

[4] Brenner L A, Homaifar B Y, Adler L E, et al. Suicidality and veterans with a history of traumatic brain injury: precipitants events, protective factors, and prevention strategies.. Rehabilitation Psychology, 2009, 54(4):390. https://doi.org/10.1037/a0017802

[5] Ernest, Sternberg. Planning for Resilience in Hospital Internal Disaster. Prehospital & Disaster Medicine, 2003, 18(4):291-299. https://doi.org/10.1017/S1049023X00001230

[6] Liu J, Kim J , Colabianchi N , et al. Co-varying Patterns of Physical Activity and Sedentary Behaviors and Their Long-Term Maintenance Among Adolescents. Journal of Physical Activity & Health, 2010, 7(4):465. https://doi.org/10.1123/jpah.7.4.465

[7] Szakács, Alexandru, Pécskay, Zoltán, Silye, Lóránd, et al. On the age of the Dej Tuff, Transylvanian Basin (Romania). Geologica Carpathica, 2012, 63(2):139-148. https://doi.org/10.2478/v10096-012-0011-9

[8] Fink E M. Post-Realism, or the Jurisprudential Logic of Late Capitalism: A Socio-Legal Analysis of the Rise and Diffusion of Law and Economics. Social Science Electronic Publishing, 2007, 55(4):931.

[9] Noland R B, Quddus M A, Ochieng W Y. The effect of the London congestion charge on road casualties: an intervention analysis. Transportation, 2008, 35(1):73-91. https://doi.org/10.1007/s11116-007-9133-9

[10] Gare G. A history of project management models: From pre-models to the standard models. International Journal of Project Management, 2013, 31(5):663-669. https://doi.org/10.1016/j.ijproman.2012.12.011