Establishment of Parallel Text Corpus of Equipment Manufacturing Industry Based on Data Mining Technology

Dongxia Liu¹, ², *, Jianguo Liu³, Xianghui Zhang², Manqian Chen²
¹Harbin Institute of Technology Harbin Heilongjiang 150001 China
²Harbin Vocational & Technical College Harbin Heilongjiang 150081 China
³Harbin Open University Harbin Heilongjiang 150001 China

*Corresponding author: Liudongxia0626@163.com

Abstract. In the era of language big data, traditional data analysis methods can't analyze semi-structured or unstructured data such as text, but all the contents in the equipment manufacturing corpus belong to text data. The equipment manufacturing corpus is a linguistic information base for legal activities and equipment manufacturing research, which aims to study equipment manufacturing and collect equipment manufacturing cases. At present, the construction of legal database in China is not perfect, and there are still many problems. In this paper, a method based on template transformation is proposed to automatically acquire parallel corpus on the Internet, and a method based on the number of transformation patterns and the retrieval and sorting of transformation patterns is adopted to verify bilingual parallel texts. This system can build a large-scale parallel corpus of equipment manufacturing industry by automatically acquiring a large number of parallel texts from the Internet.

Keywords: Data mining, Equipment manufacturing industry, Parallel text, Corpus

1. Introduction
Thesaurus research is not only an important part of language research, but also has important application value in the field of foreign language teaching. Relevant research shows that [1-3], not all words have the same importance. In order to realize highly adaptive personalized learning and create ecological teaching contents, teaching methods and teaching environment, multi-modal teaching concepts are used to trigger learners' multi-intelligence. Parallel corpus is a text corpus composed of texts with mutual translation relationship, which is an important resource in natural language processing fields such as machine translation, cross-language information retrieval and dictionary construction [4]. Since 1990s, corpus and corpus linguistics have developed rapidly, and various forms of corpus have played an important role in many aspects. Therefore, one way to solve this problem is to set up a parallel corpus of the whole equipment manufacturing language.

Text data is not numeric data. In the past, the common analysis method of this kind of data was to intercept and quote a part of the original text, and then attach the analyst's explanation and research. If a document meets a specific matching rule, it is judged to belong to this kind of statistical learning.
technology, so that computers can learn a large number of similar documents like human beings and sum up their own experience as the basis for classification [5]. At present, the known research methods mainly include parallel text mining based on intra-page patterns, parallel text mining based on URL and link patterns, parallel text mining based on anchor text, parallel text mining based on search engine snapshots and so on.

This study aims to build a parallel text corpus of equipment manufacturing industry as a research platform, and explore how to integrate the ontology research of popular science translation under the framework of translatology and linguistics on the basis of large-scale data analysis, aiming at further promoting the teaching and practice of popular science translation.

2. Definition and Classification of Corpus
Corpus stores a large number of real corpora by computer, and makes various processing annotations for research purposes, and uses the developed retrieval tools to search and classify the annotated corpora quickly, so as to discover and analyze the language phenomena that have not been noticed due to conditions [6]. For example, after reading speeches and speeches with naked eyes, a person can have a general impression of speeches or speeches. Equipment manufacturing corpus is collected for equipment manufacturing research and equipment manufacturing. It is composed of written language and spoken language, processed and stored by computer, and is a collection of a large number of real legal texts. A large number of linguistic facts provided by corpus can stimulate learners' cognitive ability and construct learners' knowledge schema. Obviously, the order of vocabulary acquisition should also follow the laws of natural language, that is, learn the vocabulary with higher frequency first, and then learn the vocabulary with lower frequency, so as to minimize the memory burden and improve the learning efficiency.

At present, the mainstream translation technology in the world is machine translation based on statistics, and the language resources that statistical machine translation depends on are large-scale bilingual parallel corpora. In fact, the corpus users are mainly language teachers and language learners. Language teachers, students, textbook developers and language acquisition researchers use and develop their own corpora, and gradually become the mainstream users of corpora [7]. The frequency of words is not the only criterion for selecting words. Knowing which parameters will affect the position and order of words in the vocabulary is obviously helpful for curriculum setting, textbook compilation and dictionary compilation. Therefore, the application of data mining technology in the equipment manufacturing corpus is of great value. With the help of data mining technology, the equipment manufacturing corpus can be quantitatively analyzed and visualized.

In practical application, text data mining has three major elements: the first is to extract data; The second is data analysis; The third is the visualization of analysis results. Translation corpus is composed of texts in different languages with translation relations. The difference between translation corpus and parallel corpus lies in that it is only the correspondence between texts, and does not necessarily have other levels of correspondence, such as words, sentences and paragraphs. Therefore, these materials which have been accurately classified, namely corpus, are the important basis for classifier design and the classification standard.

3. Automatic Acquisition Process of Parallel Corpus
Firstly, this system obtains the candidate parallel texts of equipment manufacturing industry through cross-language information retrieval, and then uses the method based on template transformation mode to extract the parallel corpus of equipment manufacturing industry. The research of this system needs to use Web crawler tools to mine web pages and get a large number of web pages. Thesaurus designers should consider the scope, level and level of users from the actual needs of users, because different users have different expectations and needs for thesaurus. If copyright is not considered, individuals can download and process a large number of corpus in a short time by using relevant tools [8]. The Web Getter tool in Word Smith V4.0, a common tool software for corpus, also has the function of collecting a large number of corpora in a short time.
There are four steps to obtain parallel corpus of large-scale general equipment manufacturing industry from Web pages, which are web page preprocessing, candidate parallel text mining, parallel text verification and duplicate elimination. The acquisition process is shown in Figure 1.

![Parallel corpus acquisition process of equipment manufacturing industry](image)

The content of the corpus includes not only modern Chinese and ancient Chinese, but also social network language. With the integration of various corpora, the processing of corpora has been deepened and the refinement of corpora has been improved. The candidate parallel text mining module builds an index for the extracted English text, and retrieves the English text most similar to each Chinese text through the cross-language information retrieval system to obtain the candidate bilingual parallel text. Parallel corpus is to write sentences corresponding to each other in two different languages into the computer, and then compare and analyze them to find out the relationship between them, so as to find out how the same content is presented in two languages. By means of probability and statistics, this paper observes and describes the language features, development rules and translation strategies of popular science translation texts, and then makes theoretical analysis based on the results of quantitative analysis.

The construction of bilingual parallel corpora is an interdisciplinary technical work. People engaged in this work should not only have good foreign language and Chinese proficiency, but also master IT technology. They should not only skillfully use related software, but also learn script language and use script language and language knowledge to process corpus. Up to now, the mainstream of language research, the scale of corpus has changed from hundreds of thousands of words to over 100 million, and corpus has become the foundation of natural language processing and language research. However, there is no given definition of the size of corpus and the amount of corpus needed for research. At present, when it comes to the representativeness of small corpus, it only means that the number of corpus in small corpus is relatively small or the sampling ratio is too low, which is not comprehensive representative. Therefore, the research based on small corpus is not effective. However, a large number of experimental results show that the vocabulary has good universality, including many important academic vocabulary neglected in previous language teaching, and has important value.

4. Parallel Corpus Construction of Language Engineering and Equipment Manufacturing Industry

4.1 Determine the Statistical Unit of Word Frequency
Before discussing the statistical units of word frequency, we should first distinguish some related terms used in corpus linguistics. Find out the commonness and individuality of interlanguage phenomenon of different learners, combine it with Chinese ontology corpus, and use data mining technology to process a large amount of data provided by ontology corpus again, so as to make a plan that really meets the learning needs of learners. Users use a certain language to retrieve the required information from the data set expressed in another language or languages. Mark the corpus according to COLSEC format; Document annotation should mark ten kinds of specific information including category, name, author, type, collected person, time, place and method, so as to facilitate the subsequent use of corpus to understand the source and provide convenience for further research; Traditional translation tools mainly rely on bilingual dictionaries, while bilingual dictionary compilation is often confined to lag, subjectivity, few examples and out of context, which is difficult to meet the needs of popular science
translation tasks to some extent [10].

On the basis of an in-depth explanation of the data source, corpus annotation, corpus data mining and presentation, and the relationship between data mining technology, visualization technology and equipment manufacturing corpus construction, this paper puts forward the basic process of equipment manufacturing corpus construction, as shown in Figure 2.

Fig.2 Tentative ideas on the construction of equipment manufacturing corpus system based on visual data mining

It can be seen from Figure 1 that the collaborative application of data mining technology and visualization technology avoids the unreasonable phenomenon caused by automatic machine analysis. The integration of manual intervention can make the results of data mining more personalized and humanized. The manual intervention of mining results is not only completed by teachers, but also integrates the interaction between teachers and learners. The establishment of subsidiary subspace enhances the dynamic nature of corpus and promotes learners' deep understanding of their own errors. The establishment of virtual space platform can improve learners' understanding of Chinese phenomena. It is of great significance in promoting the construction of a new generation of equipment manufacturing corpus and improving the visualization level of the learning process.

Text data mining method, which carries out digital analysis and research on text data, has no specific analysis method and process. The overall design and framework design of corpus is the first step of corpus research and creation, which determines the future research and application of corpus, and is closely related to the research needs and purposes. It is necessary to build an inverted index for the English text that has been translated into Chinese in the website, which is used to map the contents such as words or numbers in the text to the text number where they are located [11]. Since the appearance of computer corpus, most of the word lists have been built on the basis of corpus. The choice of corpus is closely related to the goal of thesaurus development, and the composition of corpus must meet the needs of thesaurus users. In addition, in the construction of equipment manufacturing corpus, it is necessary to make full use of visualization technology, establish corresponding subsidiary subsystems and virtual learning space, so that learners can have a deep understanding of their interlanguage phenomenon, and realize perceptual understanding and rational understanding of equipment manufacturing learning in the process of interacting with learners and teachers.

4.2 Relationship between Data Mining Technology and Construction of Equipment Manufacturing Corpus

The purpose of cluster analysis is to define data subsets with the same or similar characteristics, so as to enhance people's cognition and understanding of objective facts. In other words, in Chinese ontology corpus, the purpose of clustering analysis is to define Chinese corpus subsets with similar features, so as to enhance the understanding of teachers or learners on the application rules of a certain language entry, which is the value of clustering analysis. Researchers can choose existing corpora to create thesaurus, but most of the existing corpora are designed according to the needs of designers at that time, which may not fully meet the needs of existing research. Generally speaking, free composition is more in line with the normal language use of language learners, so the naturalness of data is strong; Although the
composition of test papers can reflect the current writing level of learners, the stress and anxiety in the examination room are that their compositions are not normally used in language, and the naturalness of their data is weak. Before establishing a corpus, it is necessary to formulate the corpus inclusion rules according to the specific application requirements, and stipulate the scope, content, standard, format, information labeling, proofreading and sampling methods.

In addition to the metadata of corpus records and the corpus itself, the corpus will be stored in different levels in the form of language unit segmentation, and in the form of processing and labeling at different levels, so as to serve the language knowledge mining and discovery.

The corpus warehouse can be formally defined as a multivariate group $CW = (C, P, V, D)$, where:

$C$ represents the non-empty and finite set of a batch of materials, with $n$ pieces of materials $C = (c_1, c_2, \cdots, c_n)$, namely.

$P$ refers to the attribute set of corpus, $P = A \cup T$, where $A$ is the meta attribute of corpus, non-empty and finite set, and is an ordered multivariate group $\{a_1, a_2, \cdots, a_i, \cdots, a_{|A|}\}$, which describes the basic attributes of corpus, $a_i$ represents the $i$-th meta attribute of corpus, $|A|$ represents the number of meta attributes of corpus; T represents the annotation attribute set of corpus. $A \cap T = \emptyset$.

$V$ is the set of attribute values, $V = V_A \cup V_T$, where $V_A$ is the set of meta attribute values of corpus, $\forall a_i \in C$ is the set of meta attribute values of corpus, $V_A^{c_i} = \{V_A^{c_k}(K | K = 1, 2, \cdots, |A|) | V_A^{c_k} \in V_A \cup V_T \cup \cdots V_A^{|A|} \}$ is the set of annotated attribute values of corpus.

$D$ represents dimension attribute set, which is an ordered multivariate group, $D = \{d_1, d_2, \cdots, \cdots |D| \}$, where $d_i$ represents the $i$th dimension attribute, $|D|$ represents the number of dimension attributes.

In the development of thesaurus, the choice of statistical units of word frequency is closely related to the purpose of research. From the perspective of teaching with Chinese corpus, the dependent variable in the corpus is language phenomenon, while the independent variable is user, environment, place and time. After the index is established, the text similarity can be calculated. This process can be realized by various models, and we choose to use vector space model [12]. The parallel corpus which maintains "one-to-many" correspondence from source language to target language has high research value. This is conducive to the multi-dimensional and multi-level comparison of the translated texts; A large number of authentic corpora provide a source for compiling examples of dictionary entries. At the same time, the quantitative research method of corpus provides a large number of reliable and real data, which is convenient for studying the specific usage of terms in a specific period or genre.

A part of corpus data which is independent from the corpus warehouse, passes through a specific processing level and corresponds to a specific language unit can be called corpus sub-database. The corpus marked by word segmentation and part of speech is stored in the word material database; If it is syntactically marked, save the sentence material library (tree library); Those marked by text information are stored in the text sub-database. In this way, the corpus warehouse contains multi-level sub-corpus sets from raw corpus to text sub-database, which correspond to different language units and serve the needs of various special studies (Figure 3).
In addition, linking keywords with high degree of association into a line to form an association network can not only see the relationship between entries, but also understand the relationship between entries and times. The corpus of classified corpus should be a slowly changing text repository, so as to reflect the detailed evidence of language evolution, correctly reflect the new application features of language and show the hierarchical structure of categories in detail. These research methods are based on morpheme analysis. Morpheme is the smallest phonetic combination in language. Later, it is proposed to develop a text editor with a translation strategy assignment set annotation plug-in to annotate translation strategies for a certain number of hierarchical sampling versions. Therefore, the value of cluster analysis lies in the induction and definition of interlanguage corpus with similar or identical characteristics, so as to reduce the cognitive burden of teaching researchers or teachers. It provides an excellent platform and reference tool. By making full use of the corpus, we can better understand terms and expressions, save time and translate accurately, reduce errors and achieve higher efficiency.

4.3 Develop Vocabulary Selection Criteria

Word frequency is a very important index in word selection. The higher the frequency of occurrence, the higher the position of words in the word list. On the contrary, the lower the frequency of occurrence, the lower the position of words in the word list. The purpose of building a corpus of equipment manufacturing industry is to reveal the essential characteristics of legal language through a large number of real legal corpus, so as to facilitate the next language research and apply it to specific equipment manufacturing industry. On this basis, the potential text factors that may affect the translation strategy, such as phrase meanings, syntactic patterns, language discourse, metaphor cognition, etc., are manually marked; The manual completion of a corpus involves the ability of professionals, and whether the corpus selection is correct depends on their judgment ability. The system will automatically search for articles that meet this condition, and list the words that appear very frequently in a table. Some classical or important corpora should be selected. The habit of using language is often derived from some classic works.

Suppose there is a table with $R$ rows and $C$ columns, which we call contingency tables. Each column of contingency table represents a category, each row represents the value of random variable, the data in the table $A_{i,j}$ represents the sample frequency when the $i$ value of $j$ category appears, $\sum_{i}^{nR_{i}}$ represents the frequency sum of all categories with $R_{i}$ value, and $\sum_{j}^{nC_{j}}$ represents the frequency sum of
category $C_j$. $n$ is the sum of total frequencies. The formula of $\chi^2$ statistic is as follows:

$$\chi^2 = n\left(\sum_{i,j}^{x_{ij}} - 1\right)$$  \hspace{1cm} (1)

It can be easily seen that the special form of $\chi^2$ distribution in $CR$ contingency table is the special case of $C = R = 2$. The selected text data set is the same as the text data set from which feature words are extracted by the two-class $\chi^2$ statistics method.

In order to quantify the "possibility" of the extracted text data category marking errors, we define a text data error rate $ER$. The error rate of text data refers to the ratio of text data with wrong category marking to all text data. Let $C = \{x_1, x_2, \ldots, x_N\}$ be the extracted text data, $N$ be the data amount, $x_i$ be the $i$th text data, $y_i$ be the category mark of the $i$th text data, $y_{0i}$ be the real category of the $i$th text data, and $I(\cdot)$ be an illustrative function.

$$ER = 1 - \frac{\sum_{i=1}^{N} I(y_i = y_{0i})}{N}$$  \hspace{1cm} (2)

The larger the value of $ER$, the more text data with wrong category marks.

Bilingual corpus refers to a corpus that collects texts in two languages. Multilingual corpus refers to a corpus that collects texts in more than two languages. Pan-knowledge domain learning and sorting: general knowledge absorption pays attention to the understanding of principle and practice; professional knowledge absorption focuses on professional concepts, terminology discrimination and word understanding. Only by exploring, can we get the valuable laws and rules hidden behind Chinese interlanguage, so as to better understand learners' learning characteristics and laws. Ideally, this independent corpus has the same type, similar size and different content as the source corpus. We regard the string in which all Chinese characters in the Chinese part and all English words in the English part of each parallel text pair are connected as a record, ignoring the numbers and symbols in it, and storing only one copy for each record, thus eliminating the situation of duplicate records.

5. Experiment
In shorthand, we often encounter some words that directly or indirectly affect understanding, including keywords that convey the key points of language information and express practical meaning, such as nouns, verbs, adjectives, numerals and so on; What must be required is the corpus with the features mentioned in the previous section, and some web pages that are very relevant to the topic. For example, in the category of data structure, we should obtain web pages that are highly relevant to the course of data structure, not just pages containing the four words of data structure. However, some words are used less frequently and become obsolete as time goes by, but most of these changes are not reflected in the vocabulary. After finishing the warehousing work, the corpus should be proofread to ensure the quality of corpus in the database. There are certain differences in the accuracy of automatic alignment of corpus, so manual proofreading is necessary in paragraph alignment and sentence alignment.

In the process of parallel text mining, there are many pairs of parallel texts sharing the same conversion mode. We use bilingual text verification based on the number of conversion patterns. When the number of occurrences of the set conversion mode $t$ is greater than the threshold $m$, the text corresponding to $t$ is judged as bilingual aligned text, as shown in Table 1.
There is a problem that must be considered when downloading pages from various websites on the Internet to the local area through search engines, that is, page coding. Search engines roam web pages from websites in different regions and with different technologies. It is of great reference value and significance to innovate the teaching methods of popular science translation, provide authentic and timely translation resources and improve the quality of popular science translation teaching. The starting point of constructing a shorthand note symbol recognition corpus based on text data mining technology is to solve the problem of blocked note production in shorthand. This has an incomparable effect on language and translation studies, and it is also convenient for researchers to study the collocation characteristics of words by sorting the search results. By comparing and analyzing a series of search results, we can find and summarize the language rules and reveal the essence of language on the basis of full description.

### 6. Conclusion

Corpus contains abundant linguistic phenomena, which is the basis of research methods based on statistical rules. This study puts forward the idea of building a parallel corpus for equipment manufacturing industry, and makes a preliminary demonstration and analysis on the research content based on the popular science parallel corpus. In this paper, a method of automatic acquisition of parallel corpus in large-scale general equipment manufacturing industry based on template transformation is proposed. This method does not depend on the existence form of specific parallel texts, and can obtain parallel corpus of equipment manufacturing industry from the Web. The purpose and goal of thesaurus creation is the foundation, and the selection of corpus, the standard of word selection and the determination of word frequency units must be closely combined with the research purpose and restricted by it; With the rapid development of corpus and corpus linguistics, the construction of equipment manufacturing database will be smoother and further reduce the difficulty of work. Therefore, only by using data mining technology can we analyze the language big data, find out the relevant content that meets the needs of equipment manufacturing industry, and provide a brand-new research idea for equipment manufacturing industry in the era of big data.

### Acknowledgements

This work was supported by 1) the special project of Heilongjiang Provincial Office of Philosophy and Social Sciences (No. 19YYD224), 2) This work was supported by the 2019 key topic of 13th Five-Year Plan from Heilongjiang Education Science Planning Leading Group (No. GZB1319015), 3) 2019-2020 vocational education "going Global" research topic of Nonferrous Metal Industry Talent Center (No. 2019YSRC07), and 4) Key research topic of Harbin Vocational & Technical College (No. HZXY2019Y018). 5) 2019-2020 vocational education "going Global" research topic of Nonferrous Metal Industry Talent Center (No.2019YSRCo3)

### References

[1] Liu Rui. Construction of parallel text corpus of film and television dialogue. Henan science and technology, 2017, 000no. 021, pp. 31-33.
[2] Zhang Ju. An empirical study on the English versions of English websites of British and American universities and Chinese universities from the perspective of parallel texts. Jiangsu Foreign Language Teaching Research, 2019, 000no. 004, pp. P.81-84.
[3] Wang Xiao, Cao Ya, Zhao Hongmei. Research on collaborative innovation of equipment
manufacturing industry based on C4.5 decision tree—from the perspective of improving quality management in production process. Internal Combustion Engine and Accessories, 2019, 282no. 06, pp. 192-194.

[4] Li Daipeng. Construction of Chinese Teaching Corpus Based on Data Mining Technology. Overseas Chinese Education, vol. 102, no. 01, pp. 115-124, 2019.

[5] Li Yi, Shu Chaohui, Li Manluo. Research on Integration and Development of Big Data Industry and Equipment Manufacturing Industry. Journal of Shenyang University of Technology (Social Science Edition), vol. 011, no. 006, pp. 519-524, 2018.

[6] Gu Xiao, Feng Xiaojie, Li Lei. Research on discrete manufacturing quality evaluation based on data mining. Digital World, vol. 000, no. 010, pp. 73-74, 2019.

[7] Gao yun, ma baiju. Translation of Chinese medicine publicity from the perspective of parallel texts. Journal of Chinese medicine management, vol. 27, no. 12, pp. 32-34, 2019.

[8] Liu Rui. Construction of parallel text corpus of film and television dialogue. Henan Science and Technology, vol. 000, no. 021, pp. 31-33, 2017.

[9] Huang Yanli, Ge Wenqian. Research on data source framework for aerospace manufacturing data mining. Informationization Research, vol. 44, no. 05, pp. 10-14, 2018.

[10] Wu Xuefeng, Malu. Data mining technology and its application in manufacturing. computer applications and software, vol. 034, no. 010, pp. 71-77, 2017.

[11] Zhang ju. Comparative empirical study of English versions of British and American university websites and Chinese university websites from the perspective of parallel texts. Jiangsu foreign language teaching research, vol. 000, no. 004, pp. 81-84.

[12] Wang Zuhua, Wu Ling, Liu Nianye. Practical stylistic translation model based on parallel texts-inspiration and theoretical interpretation. Journal of Ankang University, no. 4, pp. 38-43, 2020.