A Lexical Analysis Algorithm for the Translation System of Germany and China Under Information Technology Education

Haopin Luo(✉)

School of International Education, Wuhan Business University, Wuhan 430056, China

Abstract. This paper proposes a rule-based lexical analysis algorithm for German Chinese machine translation. The algorithm can not only effectively restore the original morphemes of various deformed words, but also provide useful part of speech and various grammatical features for the subsequent parsing mechanism in the system, Through the part of speech information of specific morphological changes, we can only extract the part of speech information in the original dictionary definition of the deformed word and its corresponding dictionary entry definition, so as to facilitate the analysis and processing of the word.

Keyword: Lexical analysis of machine translation

1 Introduction

If only the original words are stored in the dictionary of the system, the lexical analysis algorithm not only needs to segment all kinds of sentence building units, but also needs to distinguish the root, affixes and their morphological features according to the rules of morphological changes. Through comparison in the program, the deformed words in the input string can be returned to the original words. This method can greatly reduce the number of dictionary entries and the storage space of the dictionary However, because the program is directly related to the affixes of specific natural languages, the algorithm is not easy to modify and maintain [1].

In this paper, a rule-based lexical analysis algorithm for German Chinese machine translation is proposed. By designing a lexical rule representation and corresponding rule processing mechanism, the dictionary only needs to include the definition of original words, and the morphemed words are reduced by using lexical rules. The algorithm realizes the independence of program and data, In order to provide deep grammatical information for the subsequent syntactic analysis mechanism, we add the corresponding deformed part of speech and morphological feature information to the lexical rules. Below, we will first summarize and summarize the rules of various morphological changes of German words, then give the computer internal representation of these rules, and finally give the specific lexical analysis algorithm [2].

M. A. Jan and F. Khan (Eds.): BigIoT-EDU 2021, LNICST 392, pp. 493–502, 2021.
https://doi.org/10.1007/978-3-030-87903-7_61

2 The Choice of Translation Methods

2.1 Literal Translation and Free Translation

In the 19th century, translators preferred free translation (meaning to meaning) rather than literal translation (word to word). Many translators and translation theorists supported the idea of free translation because if literal translation was too close to the original text, it would produce a strange translation, which would make the readers unable to understand the true meaning of the original text; while free translation could reproduce the content of the original text. Jerome once used a metaphor to describe the relationship between the source text and the Translation: the source text is like a prisoner, who is taken into the translation by the conqueror. Different text types should be classified and appropriate translation methods should be selected. Different translation methods can even be used in the same text to make the translation more "faithfulness, expressiveness and elegance" [3].

Free translation may be more fluent in sentences, but it may be different from the original in terms of content. For general literary works, these do not affect reading, but for scientific and technological literature, there is no tolerance for a single error, which is also the inevitable result of the accuracy, objectivity and preciseness of scientific and technological literature. Therefore, in general, literal translation (word to word) is the first choice for scientific and technological literature translation [4].

2.2 Semantic Translation and Communicative Translation

Peter Newmark, a British translation theorist, first introduced the concepts of semantic translation and communicative translation in his book "approaches to translation" published in 1981. Semantic translation means that the translator expresses the meaning of the original text as accurately as possible under the permission of the semantic and syntactic structure of the target text, so the target text is closer to the original text in form and style. Communicative translation, on the other hand, tries to make readers get the same effect as the original readers. It advocates adopting different translation methods according to different types of texts. On this basis, Newmark divides text types according to language functions. In his two books, approaches to translation and a textbook of translation [5], Newmark divides them into six types: expressive function, informative function, vocative function and aesthetic function (aesthetic function), phatic function and metalingual function [6].

Among them, information text emphasizes the information function of text, its content often involves a wide range of knowledge fields, and its form is generally very

standard [7]. The core of informative text is authenticity, where language comes second. Scientific and technological literature text is basically information type text. After comparing semantic translation with communicative translation, Newmark believes that communicative translation is more suitable for information texts (including technical texts, publicity texts and various standard texts).

3 The Rules if German Word Form Change

3.1 Lexical Analysis

The change of case is mainly the change of case such as noun, determiner, adjective and pronoun; the upgrade is the change of comparative and superlative of adjective and adverb; the change of morpheme refers to the variety of verb; the change of fusion form two prepositions merge into one preposition [8].

1) The change of rules is generally reflected by the change of regular affixes. It can be divided into the following categories: (1) the change of suffixes forms words with different lexical features through the regular change of suffixes, For example, Leib → Leiber (2) inflection + suffix change forms words with different lexical features through regular suffix change and inflection Bache (3) infix + suffix changes form words with different morphological characteristics through regular infix and suffix changes, such as aufmachen − aufgemacht (4) prefix + suffix changes form words with different morphological characteristics through regular prefix and suffix changes. For example, machen → gemacht (5) preposition fusion forms a new word through the fusion of two prepositions, such as an DEM bank [9].

2) Irregular changes the irregular morphological changes of German words are not regular. Generally, they can not be summarized by the rules of morphological changes. They can only be listed one by one. For example, essen-a 3. Wenig ± minder. In addition, sometimes a German prototype needs to be deformed many times to produce a specific word. Generally, it needs to be changed several times. Typical examples are (1) adjective upgrading and then changing case, such as EB liebst -Liebsten (2) verbs change irregularly, ten people call the ending change. For example, sprechen → sprach → sprach (3) verb first participle + case change. For example, machen → machender (4) verb second participle + case change. The translation process of German Chinese translation system is shown in Fig. 1 below:

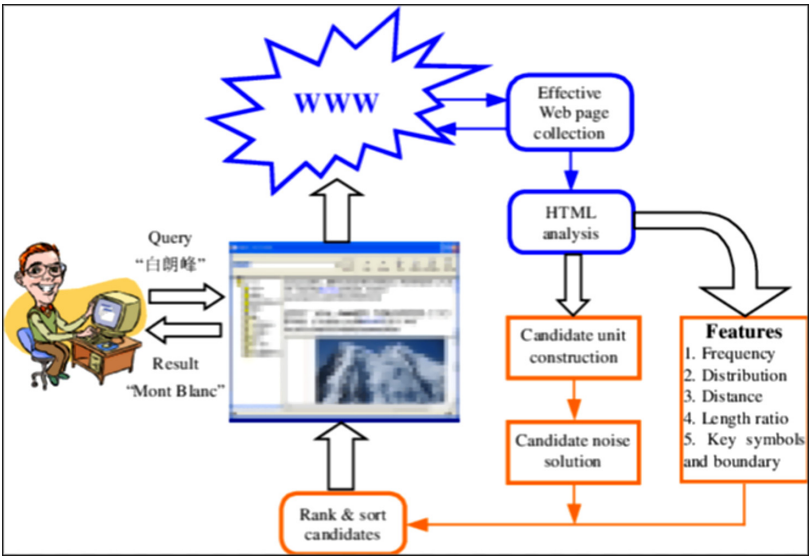

Fig. 1. Translation process of German Chinese translation system

3.2 Apriori Algorithm

Apriori algorithm accounts for a large proportion in association rules algorithm. Apriori algorithm is usually divided into two steps: the first step is to select the frequent itemsets that meet the user's requirements, and select the frequent itemsets that are larger than the user's agreed minimum support threshold in all databases. The second step is to summarize the expected association rules in the process of using frequent itemsets. The confidence level must be greater than or equal to the total critical value of the minimum confidence level agreed by the user, which is the most fundamental condition to find out the association rules [10].

$$\sup port(X \Rightarrow Y) = P(X \cup Y) \tag{1}$$

$$confidence(X \Rightarrow Y) = P(Y|X) \tag{2}$$

4 The Construction of a Small Parallel Corpus Between German and Chinese

Because of the difficulty in collecting and aligning bilingual corpus and the special characters different from English in German, the construction of German Chinese parallel corpus is much harder than that of English-Chinese parallel corpus. Therefore, it is necessary to solve the technical problem that the retrieval software can not recognize both German and Chinese bilingual. After trying to overcome the technical problems, the author has built a small German Chinese one-way parallel corpus with a scale of

600000 words. Parallel corpus consists of two parts: parallel corpus stored in computer in electronic text form and positioning retrieval software for managing and retrieving these corpora, Corpus construction is also carried out from these two aspects [11].

4.1 Selection and Collection of German and Chinese Bilingual Corpus

First, we should select the corpus according to the research purpose and research needs. Scherer pointed out that when creating a corpus, four principles should be considered, namely representativeness (Germany]repr □ sentimental □ T), persistence ([Germany] best □ ndigkeit), scale ([Germany]gr9 □ E) and content ([de] inhalt), among which representativeness is the highest goal of creating a corpus. Therefore, the constructors of corpus should take representativeness as the basis for corpus selection, and usually choose random sampling or stratified sampling. Of course, some scholars believe that all corpora can not be completely objective and not subject to the subjective influence of researchers. Therefore, the most common method of corpus is to choose corpus appropriately, which is also the development trend. Under the balance of the two, the author believes that as long as the selection of corpus can achieve the predetermined research purpose and meet the predetermined research needs, this is the proper process of corpus selection [12].

In the process of collecting data, the builders of corpus can use different ways, such as downloading German and Chinese reference materials through the Internet. In order to collect specific text and improve the quality of the corpus, scanners can also be used. Because most of the corpus needs to be saved as TXT format at last, the corpus builder should master different format conversion methods [13].

4.2 The Arrangement of German and Chinese Bilingual Corpus

Before using paraconc to search the German and Chinese bilingual corpus, we must deal with the corpus to meet the requirements of the software. The author briefly describes the process of corpus arrangement according to the experience of self-construction of German Chinese parallel corpus.

First, the preprocessing of corpus includes the unification of format and the removal of various impurities. Especially, the files downloaded from the Internet have different text formats such as font, paragraph arrangement and document format. Sometimes there are redundant spaces, broken lines, random codes, unnecessary or unrecognized graphics and symbols, which have no significance for the research. Therefore, it can be used in Microsoft Word and powergrep or “text finisher”, Sometimes it is necessary to assist manual proofreading again [14].

Next, the segmentation of bilingual corpus, parallel alignment between Chinese and German corpus are followed. Corpus alignment is the key to deal with bilingual or multilingual parallel corpus by using paraconc software. It refers to the source language corpus and its target language corpus being stored in different texts respectively, and the corpus in two texts is aligned according to the relationship between paragraphs, paragraphs or sentences, sentences or words and words. At present, parallel alignment at the lexical level is almost difficult to achieve, and sentence level alignment needs to be combined with software application and manual intervention. The author has realized

sentence level alignment in the corpus built by myself. Before parallel alignment of the corpus, the German and Chinese materials are divided into sentence units. The steps are as follows: create a new office word document, copy and paste the text file, select all the text, and then "in the Chinese corpus". Replace all with ". ^P^P", to avoid four newline characters between some sentences, you need to set the retrieval item to "^p^p^p^p", and replace it with "^p^p". Similarly, the German text is also processed accordingly, only the Chinese period is replaced by the period in the German text. However, considering that there are some abbreviations in the original German, such as bzw., etc., it is still necessary to manually check after segmentation. Because Chinese is a word-based writing unit, there is no obvious distinguishing mark between words, and paraconc can not recognize and calculate Chinese as English. Therefore, it is necessary to divide Chinese text before aligning the corpus. The author suggests that the Chinese word segmentation system ictcola of the Institute of Sciences be used to segment Chinese. Although paraconc software has align format drop-down menu, there are four alignment options: not aligned, new line delimiter, delimiter, starter/stop tags, but the author finds that manual adjustment is still needed after automatic alignment through paraconc, which is very difficult. Therefore, it is recommended to use Excel table to manually align text (sentence alignment), and then load it into paraconc software. The advantage of this operation is that the corpus can be automatically aligned after loading, and the actual operation is more convenient and practical than relying on software. The specific operation is to copy the corresponding German and Chinese bilingual corpus to an excel file as required, adjust it to align the sentence and sentence, and then copy and paste them into different word files after the alignment is completed. In this process, the builders of corpus need to pay special attention to the fact that there may be less translation and additional translation in the target language corpus [15].

5 Regular Expression of Lexical Rules

In order to realize the independence of lexical analysis algorithm and specific data, and at the same time, it is not necessary to provide dictionary definition for each morpheme, we propose a rule-based lexical analysis algorithm, in which rules are used to represent various morphological changes and their corresponding morphological features.

On the other hand, although the complex morphological changes of German words bring complexity to German lexical analysis, these rich morphological changes can also provide very useful deep grammatical information for German syntactic analysis, We should not only consider how to recognize each word string in the input sentence, but also consider how to transfer the feature attributes of the word string obtained in the process of recognition to the later syntactic analysis module, It can provide useful grammatical information for syntactic analysis [16–19]. Considering that the morphological change of the original word only corresponds to one part of speech, the corresponding deformed part of speech information is also given in the lexical rules. When looking up the dictionary, only the dictionary entry definition of the original word with the corresponding part of speech features is taken, so as to facilitate the processing of word classification [20].

6 Lexical Analysis Algorithm

Based on the above lexical rule representation, In order to save the result of lexical analysis, we design a rule-based lexical analysis algorithm. For each word, we use a lexical analysis information table to save the result of lexical analysis. Each element is represented by the structure of (prototype word string, part of speech/feature mark, morphology feature table) [21]. The prototype word string is analyzed by lexical analysis mechanism and lexical rule library The string representation of the original word after deformation processing: the part of speech/feature mark is used to represent the part of speech mark corresponding to the corresponding deformation of the analyzed word or the mark of the component of the special sentence (such as punctuation, number, etc.); the form feature table is used to store all kinds of form feature information corresponding to the deformation of the word, Then the original word string is set to empty [22].

6.1 Algorithm Flow

The flow chart of lexical analysis algorithm is shown in Fig. 1. This algorithm first divides the input of source text into sentences and their components, and then carries out different processing according to different types of sentence components. The specific algorithm is described as follows.

(1) The input sentences of the source text are segmented by components (such as words, punctuation, numbers, etc.)

(2) Take a component in a sentence and deal with it differently according to its type: if it is a special symbol or expression, turn it.

(3) If it's a word, turn to (4) (3) if it is a special symbol, the punctuation mark and the string representation of the special symbol are directly added to the lexical analysis information table; if it is a number, date and other special representation forms, such as 321.3112194, which are not easy to be defined directly in the dictionary, the number mark and its string representation are added to the lexical analysis information table [23].

4) Search the dictionary with candidate morpheme string and transfer.

The flow of lexical analysis algorithm is shown in Fig. 2 below.

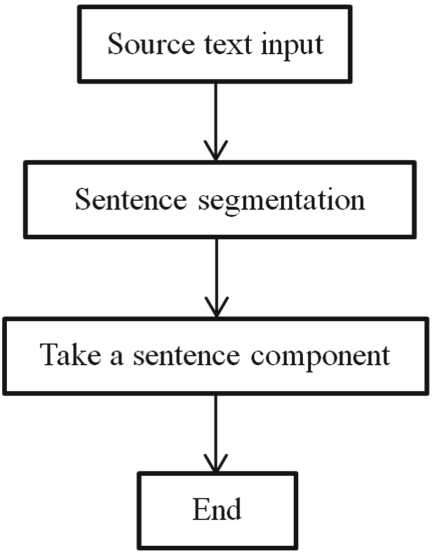

Fig. 2. Lexical analysis algorithm flow.

6.2 Lexical Analysis Information

When querying the dictionary according to these lexical analysis information, we will extract the corresponding part of speech information in the lexical analysis information table of each word, and only extract the entry definition corresponding to the part of speech information in the dictionary meaning and the part of speech information in the lexical analysis information table [24], In this way, the morphological features are only associated with the dictionary definition of the corresponding part of speech. Therefore, this lexical rule representation and the corresponding lexical analysis algorithm can comprehensively and accurately analyze the word class/features and morphological features of each sentence component, In the lexical analysis stage [25], it provides useful information for the following translation processing mechanism to deal with the multi category of words. From the above algorithm, we can see that all kinds of morphological deformation and feature information related to specific languages are expressed by the rules in the lexical rule base. The lexical analysis algorithm is only related to the expression form of lexical rules, but has nothing to do with the specific content of lexical rules, As long as we use the same rule form to represent the morpheme rules in different languages, this algorithm is applicable (see Fig. 3).

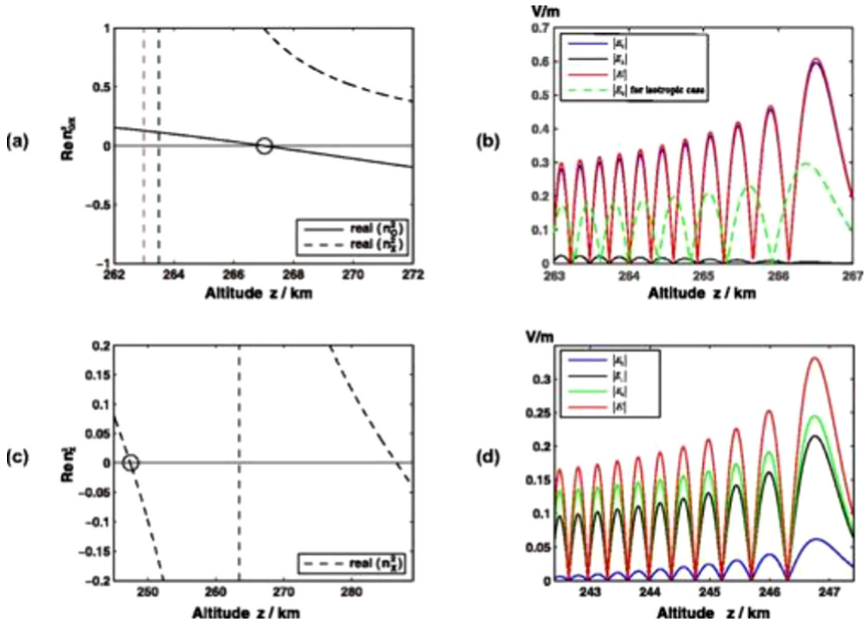

Fig. 3. Simulation for analysis information

7 Concluding Remarks

We present a rule-based lexical analysis algorithm for German Chinese machine translation. By designing a lexical rule representation and the corresponding rule processing mechanism, the algorithm solves the problems existing in the traditional descriptive lexical analysis algorithm and procedural lexical analysis algorithm, and also helps to deal with the dual category of words in machine translation processing, so as to achieve the high efficiency of the lexical analysis algorithm.

Acknowledgements. Metaphor in the Language of Science and Technology——A Comparison between Chinese and German.

References

1. Gao, T.: Design and application of web based collaborative translation system. Electron. Des. Eng. **28**(19), 85–88+92 (2020)
2. Hanji, L., Haiqing, C.: Philosophical reflection on the dilemma of machine translation technology. J. Dalian Univ. Technol. (Soc. Sci. Ed.) **41**(06), 122–128 (2020)
3. Bin Bin, G., Huang, Z.S.: Design and application of web based collaborative translation system in Colleges and universities. Comput. Program. Skills Maintenance **2020**(08), 23–25 (2020)
4. Qiang, H., Ruili, H.: Translation studies from the perspective of modern science and technology: an introduction to the future of translation technology: towards a world without Babel Tower. Orient. Translation **04**, 92–95 (2020)

5. Jing, Z.: Poetry translation from the perspective of Benjamin's translation Transcendence Theory: a case study of Pound's translation of Chinese poetry into English. J. Chizhou Univ. **34**(01), 91–94 (2020)
6. Tingting, L.: Comparison and translation of animal metaphors between German and Chinese. J. Jilin Radio TV Univ. **02**, 90–92 (2020)
7. Zhiyong, Z., Fenglan, G.: On the basic unit of German Chinese Translation: Yuyuan. Fudan Foreign Lang. Lit. **01**, 125–130 (2017)
8. Shangke, G.: Comparison of German and Chinese idioms in cross cultural translation mode [J]. Anhui Lit. (Second Half) **08**, 76–78 (2017)
9. Lei, L.: Research on German translation of science and technology. New Campus (Reading) **05**, 184 (2017)
10. Ge, N.: Research on German Chinese translation of patent documents supported by corpus. Beijing Foreign Studies University (2017)
11. Lan, G.: A study on the course of Dehan. Shanghai Normal University (2017)
12. Nannan, G.: Construction and application of small German Chinese parallel corpus. German Humanities Res. **4**(02), 44–52 (2016)
13. Zhou, T.: Translation of cultural image differences between German and Chinese under the guidance of functional translation theory. Chin. J. **2016**(11), 4–5+61 (2016)
14. Wang, Y.: German Chinese patent translation practice report - Taking word and sentence translation as an example. Xi'an Foreign Studies University (2016)
15. Li Xiang, Li Dongliang. German Chinese science and technology translation and auxiliary translation in the new situation. China Sci. Technol. Translation **29**(02), 30–31+29 (2016)
16. Li, D., Li, X., Wu, M.: Research on the course of German Chinese general practical translation. Teach. Chin. Univ. **11**, 66–69 (2015)
17. Liyan, H.: Application of functional translation theory in German Chinese translation. Test Weekly **93**, 11–12 (2014)
18. Yin, W.: Corpus assisted contrastive study on the translational accents of Jude's nameless Chinese versions. Dalian Maritime University (2014)
19. Bo, L.: The application of structural adaptation in German Chinese translation. J. Jilin Radio TV Univ. **2013**(05), 80–82+111 (2013)
20. Jing, J.: Grasp the cultural factors of professional text categories and improve the translation skills of German and Chinese science and technology – Taking the instructions of German and Chinese household appliances as an example. Innov. Appl. Sci. Technol. **08**, 281–282 (2013)
21. Gao, F., Huang, X.: Translation of synonyms in German scientific articles: a case study of standards of German automobile industry association. J. Tonghua Normal Univ. **34**(01), 83–87 (2013)
22. Xiaoqing, W.: On the curriculum design of German Chinese translation. Xueyuan (Educ. Sci. Res.) **19**, 69–70 (2012)
23. Juan, S., Chao, Y.: Discussion on the interpretation of German neologisms: a case study of German Chinese neologisms DICTIONARY. J. Neijiang Normal Univ. **27**(05), 60–62 (2012)
24. Xingzhou, H.: Reflections on the current German translation textbooks in colleges and universities in China – comments on Professor Wang Jingping's new German Chinese translation course. Orient. Translation **02**, 22–27 (2012)
25. Hui, C.: An analysis of German English appellation of some linguistic terms. Chin. Terminology Sci. Technol. **14**(01), 7–13 (2012)