Research on English-Chinese Translation of Long and Difficult Sentences by Generalized Logistic Regression Parsing Algorithm Based on Neural Network

Liang Wu[1,*]

[1]Xi'an Peihua University, Xi'an, Shaanxi 710125, China
[*]*Corresponding author. Email: 396883523@qq.com*

ABSTRACT

One of the difficulties in E-C translation is long and difficult English sentence. An accurate and effective identification and parsing of long and difficult English sentences is directly related to the translation quality. This paper uses GLR syntax (Generalized Logistic Regression syntax) based on neural network to syntactically parse the long and difficult English sentences and find the backbone of the sentence, paving a way for the translation, so as to effectively improve the quality of E-C translation.

Keywords: *Neural network, GLR syntax, long and difficult sentences, E-C translation*

I. INTRODUCTION

Globally, English is used as official language by more than 50 countries; more than 82% countries use English to express scientific and technological information. Since the reform and opening up, with the continuous enhancement of China's comprehensive strength, Chinese is becoming an emerging popular language. However, the existing translators in China are far from meeting the needs of close communication and communication with the world, both in quality and quantity. In this case, the research based on E-C machine translation has great practical value. At the same time, as one of the most popular challenges in the world, making a research on machine translation is particularly urgent and important. If the research results can meet the high-quality requirements required by users, it will greatly boost the advancement of science and technology and the improvement of communication to a certain extent.

II. LONG AND DIFFICULT ENGLISH SENTENCES

A major difficulty in E-C translation is long and difficult English sentence having complex sentence structure, strict logical structure, and rich grammatical information and often consisting of multiple clauses in parallel or series. There are many components in the sentence, including attribute, adverbial modifier, complement, parenthesis, and so on. In a sentence, there may be up to several dozens or even hundreds of words.

The core of parsing is to analyze the sentence structure, context, lexical relationship, disambiguation, etc. The difficulty in E-C translation long and difficult sentences is how to determine the relationship between the subject-predicate structure, attribute, adverbial modifier and other modifiers such as attributes in series and parallel. The fundamental solution is to reasonably and efficiently split long and difficult sentences. The length of the fragment got after splitting is shortened and the structure is relatively simple, which provides favorable conditions for parsing long and difficult English sentences.

When dealing with English sentence with few vocabularies, single structure, and fixed sentence pattern, the traditional parsing method can play a good role. However, in the face of sentence with long length and complex structure, it is often in the cart, not available to make effective sentence analysis and obtain high-quality translation.

The difficulties in parsing long English sentence mainly include the following aspects:

- The concept of segmentation of long and difficult sentences is ambiguous, and there is no clear definition on long sentence. Generally speaking, long sentence has a large number of words and complex grammatical structure. However, it is very difficult to have a uniform standard as for how many words are included in the sentence and what kind of grammatical structure is complex structure. Many linguists both in China and foreign countries have

*Fund: This paper is funded by school-level scientific research project of Xi'an Peihua University "Research on E-C translation of long and difficult sentences by GLR Parsing algorithm based on neural network" (project No. PHKT19019).

proposed different methods for segmenting long and difficult sentences. However, there are different opinions on the segmentation result and its rationality. One of the difficulties in the study of long sentence segmentation is firstly reflected in the ambiguity and uncertainty.

- There is no effective criterion for judging the segmentation of long and difficult sentences. From different perspective of view, different people may have a more subjective view of a sentence. In other words, it is difficult to have standardized and objective standards for measuring and judging the sentence's segmentation.
- Most translations in Chinese version are point-to-point translations. The analysis and research on long English sentence serves the later production of Chinese translation. Therefore, for long English sentence processing, an important measurement index is to see whether the processed result is conducive to the production of a correct and high-quality Chinese translation. This should be regarded as one of the important indexes for judging the segmentation result.
- Feature selection needs to be unified. Long and difficult English sentence contains both deep and shallow linguistic phenomena. The linguistic connotation, especially its metaphor, is particularly prone to be mistranslated. The number and representative nature of features selected in the application will greatly interfere with the generation of correct translation, and it is hard to distinguish the features in the existing methods and on the basis of the current theories.

III. GLR PARSING ALGORITHM BASED ON NEURAL NETWORK

Based on the above difficulties, this paper introduces a GLR parsing algorithm based on neural network. This algorithm is built on a neural network with multi-layer propagation network. This is a one-way propagation network whose input and output vectors are often non-linear mapping. The neural network can be used to analyze the effect of E-C translation. When evaluating the translation quality, first of all, it can singularize the evaluation indexes of translation quality and get them applied to the neural network. Wherein the input vector and output vector are the result of normalization and the result of quantitative analysis of E-C long difficult sentences; secondly, it can get relevant experience effectively integrated with the database, rationally analyze the index weight, and continuously optimize long sentences in E-C translation till completing the translation; finally, in E-C translation of long and difficult sentences, by introducing debugged neural network model, the system can make rational analysis on each element of the long English sentence as collected to improve the translation accuracy.

Tomita has improved and extended the standard LR algorithm and proposed the GLR (Generalized LR) algorithm. This algorithm can identify context-free languages. It displays the analysis result by compressed shared sub-trees on the basis of standard LR analysis method. This way saves space structure, improves analysis speed, and resolves motion conflicts in the LR parsing table. In the absence of ambiguity, GLR is similar to the LR algorithm. When encountering reduction ambiguity, it can copy the relevant parsing stack and let each parsing stack individually form an action in the parsing table, thereby generating multiple results and performing independent parsing on subsequent symbols. When a parsing built on a parsing stack generates an error, the parsing stack will be discarded. With respect to an input string, the GLR algorithm gives more than one parsing results that are in line with the given syntax, and each result can get its corresponding syntax structure from the parsing stack.

This paper proposes a GLR parsing algorithm based on neural network to translate long and difficult English sentences. This method is based on the neural network structure model, and the neural network is established by simulating the reduced advancement of the parsing table. Once the network is formed, the E-C translation of long and difficult sentence complying with the syntax can be conducted by network output without the GLR algorithm. The parsing process of this method is stored in form of code, eliminating the storage of the graph structure stack.

IV. SPECIFIC IMPLEMENTATION

A. Tagging

The reflection of sentence structure is mainly embodied in parts of speech. Part-of-speech tagging is the act determining the grammatical function of each word in a sentence, determining the parts of speech, and tagging. In accurately grasping the structure of a sentence, part-of-speech tagging plays an important role in splitting long and difficult English sentence. This paper uses a part-of-speech tagging tool developed by Stanford University. This tool is based on statistical model, with accuracy of large-scale text tagging reaching more than 97%.

Given the input sentence:

"The above data demonstrates that of the antimicrobial agents tested.",

After making part-of-speech tagging:

"The picture above demonstrates the importance of social morality.

The_DE above_JJ picture_NN demonstrates_VBZ the_DE importance_NN of_IN social_JJ morality_NN."

The parts of speech are extracted to form a sentence pattern:

"DE JJ NN VBZ DE NN IN JJ NN"

The processing of long sentence starts from the parts of speech and pattern of the sentence. The sentence pattern is used as matching object for regular expression. If the regular expression in the rule matches some components in a sentence pattern, the operation of matching components will be implemented according to the parameter part of the rule. Since the words in a sentence correspond to the parts of speech in the sentence pattern, this paper not only operates the sentence pattern, but also performs corresponding operations on the source sentence.

B. Splitting

By a preliminary treatment of some factors that affect the rationality of splitting, long and difficult sentence is split into different levels of fragments based on the remaining splitting points. The hierarchical level of the fragment split from coordinating conjunction representing a logical relationship is higher than that of clause leading word. For clause that contain sub-clause, the corresponding sub-layers are divided according to the logical order of the sentence. Multi-fragment parallel operation is beneficial to improve the efficiency of E-C translation. However, it cannot be guaranteed that all the split fragments meet the requirement for being rational.

C. Merging

Splitting points are divided into three categories: clause leading word, punctuation, and coordinating conjunction. Clause leading words include Wh-initiated vocabularies, etc.; punctuations mainly include semi-colon, comma, and colon, excluding full stop, quotation mark, bracket, and exclamation point; coordinating conjunctions include conjunctions that indicate relationships such as juxtaposition, transition, and option. For fixed collocations, noun phrases, descriptive phrase and so on that shouldn't be split, they should be matched with the model base and converted into "word" with part-of-speech, thereby simplifying the sentence structure and reducing the probability of splitting error. For example, in "such | as_durian | coconut | jackfruit | and | tomato_NN", "as" can be used as conjunction should be a split point, and "such as" may be mistakenly split; "durian, coconut, jackfruit and tomato" also contain splitting-point conjunctions and commas and may be split because "such as" is a fixed collocation, the nouns connected by commas and "and" are noun phrases that should not be split, so they are combined into one word.

D. Identification analysis

1) Initialization: Push in the stack in state o, clear the termination mark till the parsing pointer points to the symbol of waiting for parsing.

2) Mapping: If there is no termination tag, use the mapping function to map the current input symbol to the parsing table terminator.

3) Refer to the ACTION table to predict what command will be executed in the next action: if it is moving in, pushing the current state and current symbol in stack, and the parsing pointer will move own; if it is contract, using constraint function to check whether it meets the contract provisions; if conditions, it is to construct the syntax tree composed of the nodes popped from the symbol stack and push it into the symbol stack, then pop up the intermediate states in the state stack and check the GOTO table, push the new state into the state stack, and adjust the center-word pointer pointing to corresponding center word; if not conditioned, it is to set up termination tag; if it is termination, it usually means that the pointer's "error report" on the terminal symbol of the parsing table, not meaning the parsing failure; in this case, it is to get the current input character remapped to the terminal symbol of the parsing table to continue the statistical parsing and set up termination symbol; if accepted, the identifiable phrase will be analyzed, the syntax tree at the top of the symbol stack will be popped up, and then returned; if an error occurs, it usually means that the pointer's "error report" on the terminal symbol of the parsing table, the parsing is failed and the parsing will return to the initial state; the next action is performed sequentially until the parsing ends and the result is generated.

E. Error correction

Wrong splitting of long and difficult sentence mainly occurs at the boundary of reduced fragment. Such recognition error usually occur at the right boundary of clause, at conjunction, at punctuation, and at the place where there is "and" or "or". After categorizing and summarizing typical errors, seven main grammatical features are selected as the main basis for judgment, including whether it contains a predicate verb, whether it ends with punctuation, the length of the sentence, whether it contains "and" or "or", whether it starts with a conjunction, whether there is leading word and whether it contains common fixed collocations. In case of wrong splitting, the corresponding language segments will be merged to correct it, so as to improve the correctness and rationality of the parsing.

V. CONCLUSION

This paper proposes a GLR parsing algorithm based on neural network and attempts to apply it to E-C translation. In combination with the self-learning of neural network, the algorithm upgrades the parsing table of GLR algorithm on the basis of BP neural network structure model, simulates its reduction and advancement actions and makes syntax structure parsing by calculating the output. This method is mainly used for making simplified analysis on complex structure of long and difficult English sentence. Based on the neural network, GLR algorithm is used for identifying and parsing the phrases in the slices of the language fragment, determining the core words, laying a foundation for generating the translation; meanwhile, it uses model to automatically correct any wrong splitting case. It can greatly improve the parsing accuracy and rationality and lay a good foundation for the translation to generating a rational framework and meaning. English and Chinese sentence structures are very different. It is crucial to effectively regulate translations that conform to Chinese expression habits.

At the same time, it is also necessary to improve the regular matching rule base; especially the fragmentation method should be based on regular matching. This requires in-depth study of the regular expression related content, in order to be able to formulate a more normative rule base. Then, it is available to make researches on how to improve the coverage and how to avoid and deal with conflicts. Furthermore, it is needed to improve the error identification rule base. The long sentence fragmentation method proposed in this paper is based on error driving. By summarizing the errors in the fragmentation results, the methods for identifying and correcting errors are revised. These methods are fundamentally classified as rules. From the perspective of various linguistic features, formulating accurate expression can make the knowledge expression form more normative and perfect and further promote the application of the method. In the future, in-depth research will be conducted on the rules for generating translation of long and difficult sentences, with a view to further improving the formal theoretical system of parsing.

References

[1] Zuo Junjun. Research on English Long Sentence Partitioning for Machine Translation [D]. Liaoning: Journal of Shenyang Aeronautical University, 2013. (in Chinese)

[2] Zhu Jingguo, Turgun·Ibrahim, Zhang Lu. Research on Syntactic Analysis of the Uighur Based on GLR Algorithm [J]. Modern Computer, 2011 (4): 19-22. (in Chinese)

[3] Guo Yonghui, Wu Baomin, Wang Bingxi. A GLR-Based Shallow Syntactic Parser of English-Chinese Machine Translation [J]. Computer Engineering and Applications, 2004 (34): 124-129. (in Chinese)

[4] Zhi Fengning. Syntactic differences between English and Chinese [J]. Foreign Languages Research, 2012 (2): 305-306. (in Chinese)

[5] Wang Tao, Zhai Xuesong. Model of Recommended Courses Based on Neural Network [J]. Journal of Hefei University, 2016, (1): 30-34. (in Chinese)

[6] Wu Shuofeng, Yan Xuefeng. Integrated Neural Network Based on Just-in-Time Learning and Application on Dry Point Prediction [J]. Journal of East China University of Science and Technology (Natural Science Edition), 2016, (5): 696-701.

[7] Hu Li, Chen Bin, Lai Qiming. Improvement of BP neural network [J]. Computing Technology and Automation, 2015, (4): 86-89. (in Chinese)

[8] Tang Lili. The Evaluation System for Classroom Teaching Quality Based on the BP Neural Network [J]. Science and Technology of West China, 2014, (4): 103-105. (in Chinese)

[9] Doi Takao, Sumita Eiichiro. Input Sentence Splitting and Translating[C]. In: Proceeding of Workshop on Builing and Using Parallel Texts. USA: HLT-NAACL, 2003: 104-110

[10] Sheremetyeva Svetlana. Towards Designing Natural Language Interfaces[C]. In: Proceedings of the 4th International Conference. Mexico: Computational Linguistics and Intelligent Text Processing, 2003: 16-22

[11] E.F.T.K.Sang. Memory-based shallow parsing[J]. In: Journal of Machine Learning Research, 2002, 2:559-594

[12] Uwe Muegge. An Excellent Application for Crummy Machine Translation: Automatic Translation of a Large Database. In: Proceedings of the Annual Conference of the German Society of Technical Communicators. 2006

[13] Roh Yoon Hyung, Hong Munpyo, Choi Sung Kwon, et al. For the Proper Treatment of Long Sentences in a Sentence Pattern based English-Korean MT System[C]. In: Proceedings of MT Summit IX. New Orleans: MT Summit IX, 2003