Design of Intelligent Recognition English Translation Model Based on Improved GLR Algorithm

Yuezhou Wei and Lijun Huang(✉)

Guilin University of Electronic Technology, Beihai Campus, Beihai 536000, Guangxi, China
hlj1629@163.com

Abstract. We have developed an intelligent English translation model based on improved GLR algorithm. This algorithm is based on the improved version of GLR algorithm. The main difference between them is that in our algorithm, we use two types of data: (1) text data and (2) English translation data to calculate the similarity between text and English translation. A detailed description of the method used to calculate the similarity. The design process of the model includes the following steps: 1) The text is divided into three groups according to its difficulty level; 2) Each group contains a certain amount of text; The task of syntactic analysis is to automatically deduce the grammatical structure of a sentence according to a given grammar and method. The improvement of parsing performance will greatly promote the application of information retrieval and machine translation. This thesis mainly makes a comprehensive analysis of the related technologies of syntactic parsing, and on this basis, implements an English translation syntactic parsing system based on GLR algorithm. In order to deeply understand the application of syntactic analysis in practice; Firstly, this paper gives a detailed overview of the development and research background of syntactic analysis. It studies the basic concepts of syntactic analysis and some commonly used parsing algorithms. At the same time, it studies several popular parsing methods and compares various parsing algorithms; Secondly, based on the research of XUUPOS Corpus and rule base of Xinjiang Key Laboratory of Multilingual Information Technology, this paper extracts and improves the rules suitable for the research of English translation syntax analysis system based on GLR algorithm from three aspects: English translation word segmentation, part of speech tagging and rule base construction.

Keywords: Improved GLR algorithm · Intelligent identification · English Translation Model

1 Introduction

Natural language processing, also known as computer linguistics, is the process of processing and analyzing natural languages through the establishment of formal mathematical models, and the use of computer applications to achieve the processing and analysis

Y. Zhang and N. Shah (Eds.): BigIoT-EDU 2023, LNICST 580, pp. 340–349, 2024.
https://doi.org/10.1007/978-3-031-63130-6_38

process, thereby achieving the realization of computers to simulate some or even all of human language abilities. Natural language processing is one of the research branches of artificial intelligence [1]. After decades of exploration, the theory of computer processing natural language has become increasingly mature, and its application scope has become increasingly widespread. A technical system for various applications and research has initially formed.

Syntax analysis plays a very important role in natural language processing, and it is also recognized as a research challenge. It plays a connecting role in natural language processing, including machine translation systems. In particular, the quality of syntax analysis results directly affects subsequent natural language processing research. As a kind of natural language, English has the most essential things common to natural languages. Therefore, the study of English syntactic analysis must draw on and absorb mature linguistic theories [2]. At the same time, due to the characteristics of the English language itself, it is impossible to copy ready-made methods in English syntactic analysis. It is necessary to study and develop syntactic analysis techniques suitable for English based on the characteristics of English itself and under the guidance of advanced linguistic theories.

Due to the lack of word form conversion to express tense changes in Chinese, there are many challenges for Chinese English machine translation. First, it is almost impossible to judge the correct tense from only one verb. However, although the tense judgment in Chinese is not as simple as that in English, Chinese does contain the information of tense - not through the transformation of word form, but through time nouns and adverbs. There is a lot of useful information in Chinese context to help judge the tense of the whole sentence or a certain verb. For example, in the above example, the combination of "arrived", "arrived" and "a few days ago" can indicate that the whole sentence belongs to the past tense [3].

It is particularly important to maintain the temporal consistency between the source language and the target language in the translation process. However, due to the differences in temporal representation between Chinese and English, especially the characteristics of temporal representation in Chinese itself, the current machine translation system has brought many challenges. On the one hand, it is difficult to recognize Chinese tenses [4]. We cannot use the same method as English, which can judge tenses only through verb forms. We must fully consider the context information of the text; On the other hand, in Chinese English machine translation systems, it is also difficult to maintain the temporal consistency between the source language and the target language. Therefore, the study of Chinese tense recognition and maintaining the temporal consistency between Chinese and English in the process of translation is a very important aspect to improve the current Chinese English machine translation system.

Any kind of natural language has its own grammar rules, and English translation is no exception. The research theory of rule-based syntax analysis is relatively mature. The main methods are Earley algorithm, Tomita algorithm and Chart algorithm. The GLR algorithm mainly adopts two parts: the analysis table and the operation stack program. The analysis table is constructed through preprocessing, and the operation program is realized through the GLR algorithm through the graph stack structure [5]. The research on English translation syntax analysis is relatively few. We can only improve

the implementation of syntax analysis by studying Chinese and English syntax analysis methods, It will lay a foundation for natural language processing in future English translation. English translation belongs to Turkic language family. It is also an adhesive language [6]. This paper mainly discusses the syntax analysis of English translation based on GLR algorithm. As for the research of GLR algorithm, the current research is mainly based on DFA of finite automata and preprocessing based on LR analysis table. We adopt preprocessing based on LR (1) analysis table and LALR algorithm to generate the analysis table that cannot fully accept the expected analysis table of the generator; The analysis table generated by LR (O) algorithm is not very practical for natural language processing. In this paper, we mainly introduce the GLR algorithm to analyze English translation sentences when analyzing stack processing.

2 Related Work

2.1 Improved GLR Algorithm

Syntactic analysis plays an important role in the field of Chinese information processing, which is of great significance for the in-depth study of automatic Chinese English translation, natural language semantic processing, automatic summarization and other fields. At present, it has two main methods: rule-based and probability based parsing [7, 8]. English, as an international common language, has become increasingly prominent. With the support of intelligent recognition technology, automatic translation has become an important means of cross-language communication. However, there are many complicated situations in the process of English translation, such as grammatical structure differences, changeable word meanings and so on, which make the translation quality challenged. GLR (Graphical LR parsing) algorithm, as a general grammar parsing algorithm, has powerful parsing ability and flexibility, which provides a new way to solve the difficult problems in English translation. The purpose of this paper is to explore how to use GLR algorithm to optimize the process of intelligent recognition of English translation in order to improve the accuracy and fluency of translation (Fig. 1).

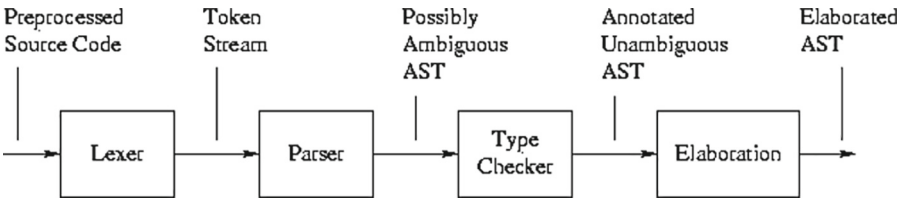

Fig. 1. GLR algorithm improvement process

In this process, this (P) is constantly multiplied into P (T).

$$\Delta w(i, y) = -\eta \frac{\partial e}{\partial w(i, y)} \tag{1}$$

$$D\left(x_i, x_j\right) = \sum_{l=1}^{m} d\left(x_{il}, x_{jl}\right) \tag{2}$$

In this way, in the idea of part of speech tagging, it can be assumed that W is a set of words and T is a set of part of speech tagging. For a given word sequence W = W1 W2 W3 …. n, from all possible tag strings generated by W, find the most suitable tag sequence for a specific word string T = t1 t2 t3 … tn. If Bayesian decision is used, that is, a posterior probability for each tag sequence T.

$$P(T/W) = \frac{P(T)P(W/T)}{P(W)} \tag{3}$$

Using the labeled English corpus, a method called relative frequency training can be used to obtain the part of speech probability and vocabulary probability parameters, namely, the formula is:

$$P_{i,j} = P\left(\frac{t^j}{t^i}\right) = \frac{c\left(t^i, t^j\right)}{c\left(t^i\right)} \tag{4}$$

$$p_{i,j,k} = P(t^j/t^i, t^k) = \frac{c(t^i, t^j, t^k)}{c(t^j, t^k)} \tag{5}$$

$$p_{j,k} = P(w^k/t^j) = \frac{c(w^k, t^j)}{c(t^j)} \tag{6}$$

Its program structure diagram is shown in Fig. 2 below.

The CLR algorithm uses the width first strategy to resolve the LR parsing table conflict. The GLR parser executes its words according to the predefined parsing table to "move" or "reduce". There is only one predicate verb in the English translation sentence, and the verb determines the necessary components in the whole sentence [10]. The verb is the most critical and decisive in a sentence.

2.2 Intelligent Identification of English Translation

Sequence to sequence model is a deep learning model, which has achieved good results in machine translation, text summary, English translation and translation tasks. Google Translation has also used the Seq2Seq model since 2016. Seq2Seq model can be seen in these two articles (Sutskever et al. 2014; Cho et al. 2014).

However, I found that to understand and apply the Seq2Seq model, we need to understand a series of concepts that are built on each other. This is often a difficult thing for beginners. I think if we can visualize them, it will help us understand these complex concepts [11]. This is my motivation for writing this article. To understand this article,

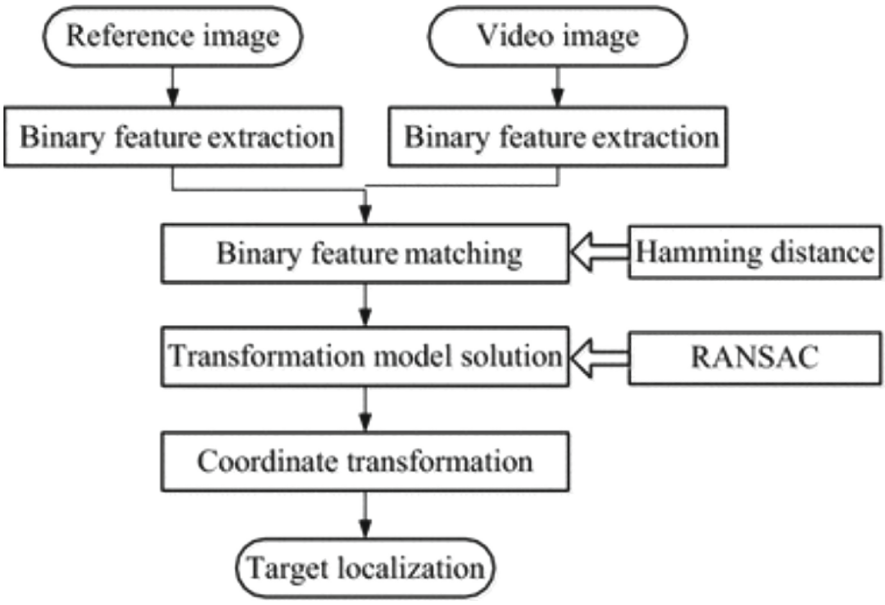

Fig. 2. Rule extraction flowchart

you need some basic knowledge of deep learning. I hope this article can help you in your learning process (and the two articles mentioned above).

The input of the Sequence to Sequence model is a sequence of objects (words, letters, features of English translation, etc.), and its output is another sequence of objects. A trained model works like this: in the French English machine translation task, a sequence is a string of words, in this case, the French Jesus etudiant. The input word string will be processed by the model in turn. Similarly, the output is also a string of words. In this example, it is English l'm a student [12].

The ultimate goal of English syntactic analysis is to be able to efficiently and accurately analyze large-scale real text sentences. Although rule based English syntax analysis can abstract English syntax rules, it is difficult to handle various complex phenomena in analyzing real texts, and it is difficult to achieve good analysis results. The main goal of this system is to analyze sentences to obtain a reasonable syntactic structure, and to be able to process open corpus and relatively stable sentences [13]. The text unit we deal with is a single sentence. The design purpose of this system is based on the following principles: rationality, practicality, and versatility, which can provide a foundation for further research on natural language in the future; And provide a good syntax analysis tree structure for future research; It can be easily combined with other systems to provide syntactic support for solving more related problems [14, 15].

Data collection and preprocessing: A large number of English corpora are collected and preprocessed, including word segmentation, part-of-speech annotation, syntactic analysis, etc., to construct high-quality training datasets.

Training and building models: Using GLR algorithms and deep learning technology to build English translation models. By training a neural network model, the model has the ability to understand and process the characteristics of the English language.

Model evaluation and optimization: The trained model is evaluated, and metrics such as accuracy, fluency, and recall are calculated by comparing the human translation results with the machine translation results to measure the performance of the model. Optimize and adjust the model based on the evaluation results.

Deployment and application: Deploy the optimized model to the intelligent recognition English translation system. In practice, the system can automatically complete the translation task of English text, and at the same time provide a human-computer interface for users to make modifications and feedback.

Continuous improvement and updating: Regularly collect user feedback and corpus update information, and retrain and optimize the model to improve translation quality and system performance. At the same time, we pay attention to the development trend of the English language, and constantly adjust and improve the translation strategy.

3 Design of Intelligent Recognition English Translation Model Based on Improved GLR Algorithm

For any sentence, get the sequence by word segmentation and looking up the dictionary: S1_ W1 S2_ W2… Sn.. Wn (S represents part of speech, W represents word) uses subtree sharing and local ambiguity compression technology to analyze English translation. For example, an English translation sentence men yiraqtiki bir qizni ksrdum (I saw a girl in the distance) is expressed as: Pron by looking up the dictionary_ men N_ yiraqtiki M_ bir N_ qizni V_ ksrdum. The analysis process of GLR algorithm is: first initialize a graph stack structure and shared forest, then find the part of speech of each word of this sentence through the dictionary and store it in the buffer, so that we can analyze faster, and then analyze it according to GLR algorithm, In case of conflict (“reduction - moving”, “reduction - reducing”), take the exploratory method; in case of “moving - reducing” conflict, use “reduction” first This path is used for analysis. If the sentence can be analyzed completely, the analysis is considered correct. Continue to analyze the conflict from the place where the conflict was analyzed just now; If an error occurs, return to the conflict and continue to make other choices. Until the end of the sentence analysis. If the analysis is not successful, we will display the largest subtree of the sentence analysis. Otherwise, we will output the one with higher probability by calculating the probability of the analysis tree. However, the parallel LR analysis algorithm is split into two trees when conflicts occur, and the result may be more than one tree. We also use the same method to output the optimal rule tree. The process of English translation model is shown in Fig. 3 below.

The advantage of GLR algorithm is that its running time is shorter than that of the parallel LR parsing algorithm, and it can parse all syntax trees that conform to the syntax structure, but its disadvantage is that it does not give how to judge which parsing tree is better.

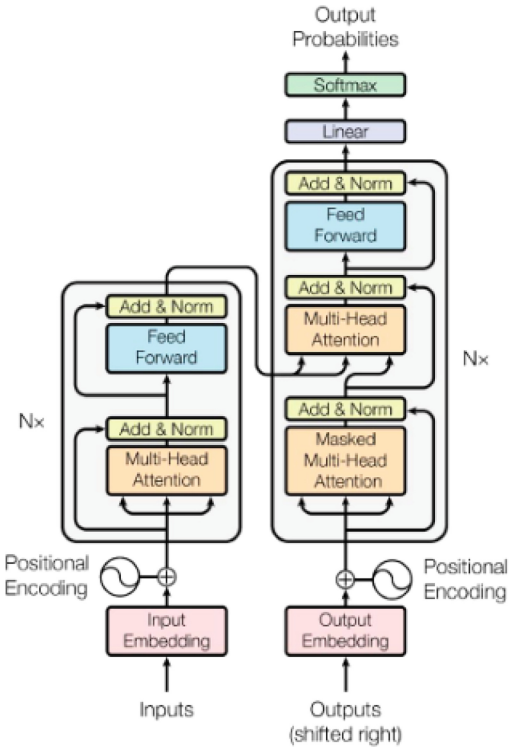

Fig. 3. English translation model process

The biggest difference between the NMT model based on attention mechanism and the above common "encoder decoder" structure is that the context vector is constructed in a different way, and its essence is a weighted average of the source input. And the NMT model based on attention mechanism uses another encoder structure: the common "encoder decoder" model usually uses forward RNN as the encoder, encoding the input source language sentences to form a fixed length context vector; The NMT model based on attention mechanism uses bidirectional RNN as the encoder to encode the input source language sentences forward and backward respectively, and finally, the two are spliced to form a variable length context vector. The biggest advantage of bidirectional RNN encoder is that the representation of each word in the source language sentence in the context vector already contains the semantic information on the left and right sides of the source language sentence, rather than the semantic information on the single side. Figure 4 shows the flow of phrase corpus information.

Construct a high-precision English corpus: When using the GLR algorithm to process English translation, it is necessary to construct a high-precision English corpus as training data. The corpus should cover a variety of English contexts and domains to improve the adaptability of the translation model to different linguistic features.

Introduce deep learning technology: Combined with deep learning technology, a large number of corpora are learned by training neural network models to extract the

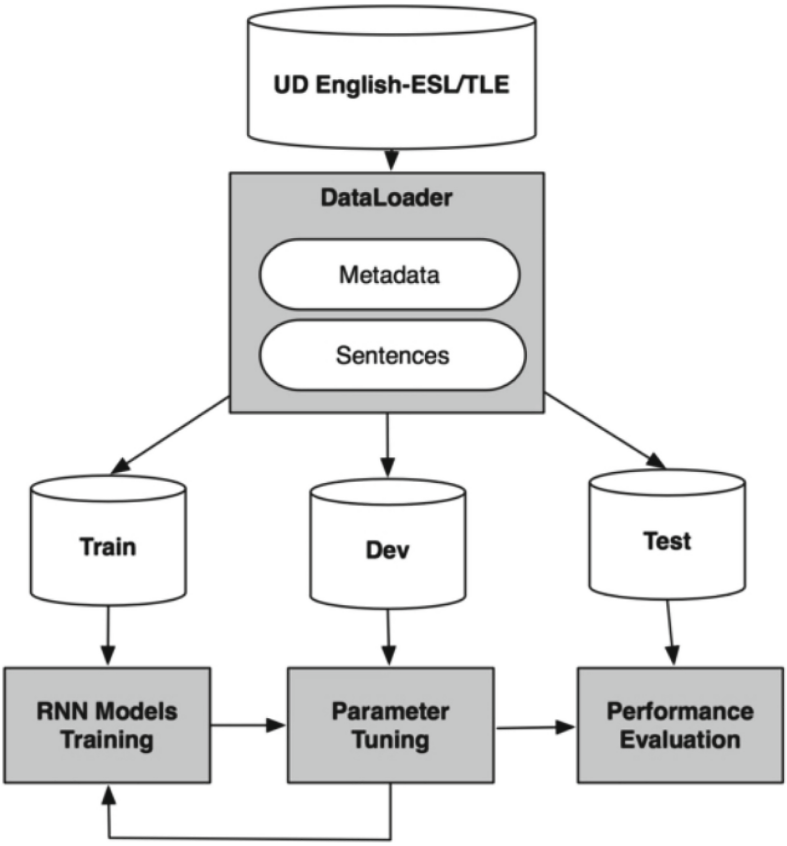

Fig. 4. Phrase Corpus System Architecture

internal rules and features of the English language. The combination of deep learning and GLR algorithms can help improve ambiguity processing and semantic understanding in the translation process.

Personalized translation strategy: Develop personalized translation strategies for different translation needs and application scenarios. For example, for specific fields such as law and medicine, a rigorous and accurate translation style is required, and for everyday communication scenarios, a translation style that pays more attention to fluency and legibility can be adopted.

Real-time optimization and updating: With the evolution of the English language and the emergence of new expressions, the intelligent recognition English translation system should have the ability to optimize and update in real time. Through continuous training and model tuning, we can adapt to changes in language development and improve the accuracy and timeliness of translation.

Strengthen human-computer interaction: In the process of intelligent recognition of English translation, user feedback and interaction should be emphasized. Users can make corrections and additions to the translation results to form a high-quality corpus.

At the same time, according to user feedback and needs, the translation algorithm and model are continuously optimized to improve the user experience.

4 Design and Implementation of a GLR Parsing System

GLR algorithm is a bottom-up recursive descent analysis algorithm, which represents the process of grammar analysis as a graph, and realizes grammar analysis through the reduction process of graph. The algorithm can deal with ambiguous syntax and has good error handling ability. In the process of English translation, GLR algorithm can better deal with complex grammatical structure and word meaning ambiguity, thus improving the accuracy of translation. The overall design structure of the system is shown in Fig. 5 below.

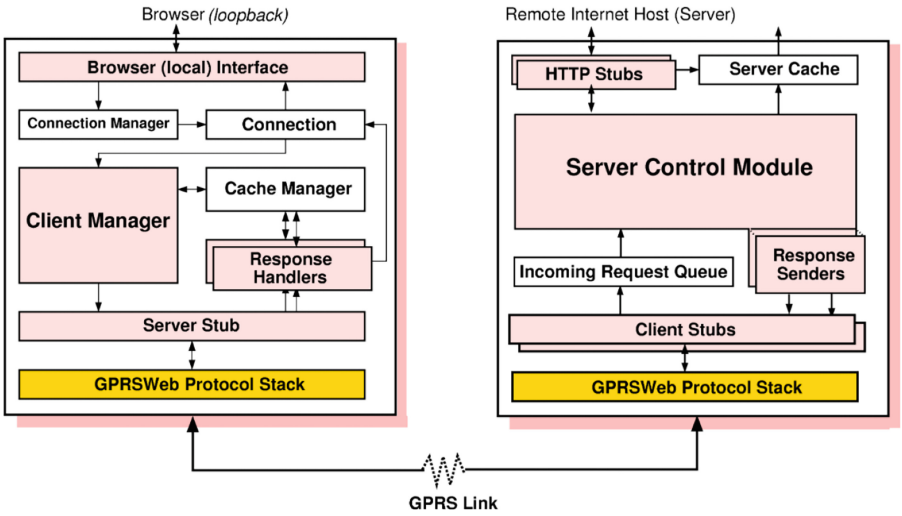

Fig. 5. Overall design structure of the system

By constructing high-precision corpus, introducing deep learning technology, formulating personalized translation strategies, optimizing and updating in real time, and strengthening human-computer interaction, the accuracy and fluency of translation can be effectively improved. In the process of implementation, we should pay attention to the steps of data preprocessing, model training and evaluation, deployment and application, and continuous improvement and update. With the continuous development of technology, intelligent recognition of English translation will be more accurate and efficient, providing better support for cross-language communication. The test of this system was conducted on the basis of the Xinjiang Multilingual Information Technology Key Laboratory's part of speech tagging corpus (XJUUPos Corpus) and rule base. 2000 manually annotated English sentences extracted from the rule base were analyzed, and the test results were satisfactory, with a result of 100%, and 17 English sentences with ambiguities were correctly analyzed.

5 Conclusion

Designing intelligent recognition model based on improved GLR algorithm is a method to classify English translation into different categories. This method uses features extracted from English translation to classify them into different classes. In this study, we applied the proposed method and used two types of features: color features and texture features. We use these two types of features for classification because they are more effective than other methods that use only one feature type. Using both types of functionality allows us to achieve better results than using only one type of functionality.

References

1. Sui, Y.: Computer Intelligent Proofreading Method for English Translation Based on Foreign Language Translation Model (2021)
2. Liu, X., Yang, Z.: Research on legibility of English text based on improved decision tree and intelligent interactive system. J. Intell. Fuzzy Syst.: Appl. Eng. Technol. **4**(Pt. 2), 39 (2020)
3. Zhang, F.: English machine translation model based on modern intelligent recognition technology. Mod. Electron. Tech. (2018)
4. Thomas, G.A.S., Robinson, Y.H., Julie, E.G., et al.: Intelligent prediction approach for diabetic retinopathy using deep learning based convolutional neural networks algorithm by means of retina photographs. Comput. Mater. Continua (2), 17 (2021)
5. Xu, X., Wang, C., et al.: A new deep learning model based on improved AlexNet for radiation source target recognition (2018)
6. Chen, T., Wang, L., Li, Y., et al.: Research and Implementation of Breast Cancer Intelligent Recognition Algorithm Based on Deep Convolutional Neural Network (2020)
7. Liang, Y., Fan, Y., Peng, Y., et al.: Smart Grid Project Benefit Evaluation Based on a Hybrid Intelligent Model (2022)
8. Yu, H., Ji, Y., Li, Q.: Student sentiment classification model based on GRU neural network and TF-IDF algorithm. J. Intell. Fuzzy Syst.: Appl. Eng. Technol. **2**, 40 (2021)
9. Zheng, X., Chunyao, M.A.: An intelligent target detection method of UAV swarms based on improved KM algorithm (2021)
10. Zhao, J., Gao, H., Liu, Y., et al.: Speech recognition algorithm based on neural network and hidden Markov model. **25**(4), 11 (2018)
11. Gu, M., Lyu, J., Li, Z., et al.: Research on fast recognition of vulnerable traffic participants in intelligent connected vehicles on edge computing. J. Circuits Syst. Comput. **32**(03) (2023)
12. Hmida, H., Hafsi, S., Bouani, F.: An embedded Hildreth-based model predictive control of an elbow joint orthosis robot. Trans. Inst. Meas. Control. **45**(5), 911–920 (2023)
13. Guo, Y., Wang, R., Zhao, D., et al.: Numerical Simulation of Vapor Dropwise Condensation Process and Droplet Growth Mode (2023)
14. Sun, X., Hu, W., Xue, X., et al.: Multi-objective optimization model for planning metro-based underground logistics system network: Nanjing case study. J. Ind. Manag. Optim. **19**(1), 170–196 (2023)
15. Agac, G., Baki, B., Ar, I.M., et al.: A supply chain network design for blood and its products using genetic algorithm: a case study of Turkey. J. Ind. Manag. Optim. **19**(7), 5407–5446 (2023)