IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 June 2022, accepted 11 July 2022, date of publication 21 July 2022, date of current version 8 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3192965

RESEARCH ARTICLE

A Two-Channel Chinese Enterprise Abbreviation Generation Method Based on an Enterprise Component and Single-Character Strategy

WEI AI[1], HONGEN SHAO[1], JIA XU[1], TAO MENG[1], AND KEQIN LI[2], (Fellow, IEEE)
[1]School of Computer and Information Engineering, Central South University of Forestry and Technology Changsha, Hunan 410082, China
[2]Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

Corresponding author: Tao Meng (mengtao@hun.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 69189338, in part by the Excellent Young Scholars of Hunan Province of China under Grant 20B625 and Grant 18B196, in part by the Changsha Natural Science Foundation under Grant kq2202294, and in part by the Program of Research on Local Community Structure Detection Algorithms in Complex Networks under Grant 2020YJ009.

ABSTRACT The automatic generation of Chinese enterprise abbreviations is a task that extracts enterprise abbreviations to represent the enterprise full name. Traditional methods do not divide abbreviations in detail, which leads to a poor generation effect of irregular Chinese enterprise abbreviations generation, and the best selection method of candidate abbreviations among traditional methods is still coarse-grained relationship modeling. To solve the problem of irregular abbreviation generation and abbreviation screening, this paper proposes a two-channel Chinese enterprise automatic abbreviation generation method. First, in the two-channel method, the enterprise component channel outputs regular candidate abbreviations, and the single-character channel outputs irregular candidate abbreviations to improve the processing effect of the method on irregular abbreviations. Then we design a Bayesian filtering model based on the position relationship of abbreviations in enterprise components to improve the final effect of the automatic generation of Chinese enterprise abbreviations. The results show that our effect is the best in the data performance of Chinese enterprises.

INDEX TERMS Bayesian, BERT-BiLSTM-CRF, CRF++, Chinese enterprise abbreviation.

I. INTRODUCTION

There are many enterprise names, and their numbers are constantly increasing with the development of society and economy. Because of people's language habits, enterprise abbreviations have gradually replaced the full names of the enterprises according to people's impressions [1]. With the wide application of abbreviation, standardizing and unifying enterprise abbreviations to their standard forms is very important in knowledge fusion, information retrieval, machine translation and other tasks [2]. Therefore, generating and maintaining an abbreviation library for Chinese enterprises is meaningful.

TABLE 1. Examples of enterprise abbreviations.

Index	Enterprise Name	Abbr	Abbr Type
1	北京矿冶科技集团有限公司	矿冶集团	Regular
2	北京沃盈投资有限公司	沃盈投资	Regular
3	中国石油化工集团有限公司	中石化	Irregular
4	国家电力投资集团有限公司	国家电投	Irregular

Chinese enterprise abbreviations have many different forms due to the characteristics of Chinese vocabulary. This is unlike english abbreviations, which are usually formed by the uppercase letters of each word [3]–[5]. We summarize the abbreviations of Chinese enterprises and divide them into regular abbreviation and irregular abbreviation. Table. 1 shows some examples of abbreviations. In examples 1 & 2, the abbreviation is a combination of some words in the enterprise name, which called regular abbreviation. In examples 3 & 4,

The associate editor coordinating the review of this manuscript and approving it for publication was Renato Ferrero.

the abbreviation combined with a single character and Chinese words in the enterprise name is irregular.

In the study of regular abbreviations, *Sun et al.* [6] generated candidate abbreviations based on enterprise component construction rules. *Chen et al.* [7] summarized the characteristics of full abbreviations and proposed a rule-based method of modeling long-distance constraints using the entity abbreviation generation method of first-order logic (FOL). These methods can accurately generate abbreviations based on the characteristics of full names; however, such methods rely too much on syntactic analysis, and the rules for mapping from full names to abbreviations is not exhaustive. In a study combining regular and irregular abbreviations, *Yang et al.* [8] converted the Chinese abbreviation generation process to a CRF tagging problem. Although this method can extract local information of the input sequence well, it cannot express complex rules in character long-distance feature extraction, so its performance in irregular abbreviation generation is still poor. *Zhang et al.* [9] introduced a filtering model that uses SVM to select the best abbreviations based on web data. In the same way, *Yang et al.* [8] used the length relationship between the full name and abbreviation to model when selecting the best abbreviation. However, these two screening methods are based on the coarse-grained relation of full abbreviations and cannot better screen the best abbreviation.

To solve these problems, we propose a two-channel Chinese abbreviation generation method, that includes enterprise component channel, single-character channel, and Bayesian filtering model. We construct regular abbreviation set, irregular abbreviation set, and enterprise component annotation set. We deal with regular abbreviations and irregular abbreviations separately to improve the effect of generating Chinese enterprise abbreviations. Specifically, we get regular candidate abbreviations in the enterprise component channel and irregular candidate abbreviations in the single-character channel. In addition, the Bayesian filtering model is designed to merge the candidate abbreviations generated by two channels and sort them according to the Bayesian probability of candidate abbreviations to filter out the best set of abbreviations.

Our contributions to this paper are summarized as follows:

1) We propose two channels for Chinese enterprise abbreviation generation based on an enterprise component channel and a single-character channel to improve the performance of irregular abbreviation generation.
2) We build a Bayesian filtering model based on the position of abbreviation characters in enterprise components to screen the best abbreviation.
3) We built an enterprise regular abbreviation set and an irregular abbreviation set [1]for the enterprise abbreviation generation task.
4) We verify the remarkable effect of the two-channel method proposed in this paper on Chinese enterprise abbreviation generation through a large number of experiments.

The rest of this article is organized as follows. In Section *II* we reviews related work. And in Section *III* introduces the method of Chinese abbreviation generation based on two-channel in detail. In Section *IV*, we will design comparative experiments and show the results. Finally, the conclusion of our contributions and a discussion of future work are presented in Section *V*.

II. RELATED WORKS

In terms of abbreviations, English abbreviations are usually formed by capitalizing the first letter of proper nouns. Due to the characteristics of the Chinese language, the formation of enterprise abbreviations is quite different in Chinese than in English. In recent years, many researchers have proposed solutions to the problem of Chinese abbreviation generation. These abbreviation generation methods are divided into three categories: rule-based methods [7], [10], statistics-based methods [11]–[13], and methods based on sequence annotation [8], [9], [14].

A. RULE-BASED

In rule-based abbreviation prediction methods, the specific abbreviation extraction rule set is established by analyzing the naming rule and the abbreviation characteristics of the enterprise full name. According to a series of factors, the corresponding weight of mapping rules is set to form an abbreviation generation model. *Qin et al.* [10] developed a mapping rule set and corresponding rule weights according to the abbreviations of Chinese device names, and exported the average weight of the mapping rule set to the abbreviation dictionary to determine the threshold of identifying abbreviated entities. *Chen et al.* [7] proposed a method to use *FOL* to model long-distance constraints and select the most appropriate mapping rule for each entity full name. The training dataset is expanded based on rules. Rule-based approaches make good use of people’s knowledge, but their rules are not exhaustive.

B. STATISTICS-BASED

In statistics-based methods, the full name and abbreviation of the enterprise are extracted from global resources, the probability of candidate abbreviation is calculated, and the candidate abbreviation with the highest probability is selected as the best abbreviation. *Kim et al.* [11] explored possible official candidates from various sources in biomedical texts and then selected the most acceptable candidate from the retrieved candidates through ranking. *Jian et al.* [12] investigated the problem of automatically extracting and ordering acronyms and related extensions on a large scale from web data and user interactions through web search engines as sources of information a large-scale search engine provides. *Liu et al.* [13] proposed a method to generate a candidate list by using the internet as the main information source, search engines, and learning “help words” to expand abbreviations

[1] https://github.com/HongEnShao/Two-channel-EC-SC.git

into full names and then select the best candidate based on the KNN ranking mechanism. Searching the full name and abbreviation pair in large-scale actual data can ensure the applicability of the result. However, processing large amounts of data will reduce the efficiency of abbreviation generation, and because there are not always full abbreviation pairs in limited resources, the accuracy is affected.

C. SEQUENCE ANNOTATION-BASED

In methods based on sequence annotation, the abbreviation generation task is transformed into a recognition task based on character sequence annotation. The sequence annotation model makes an abbreviation prediction by learning the constructed abbreviation annotation dataset. *Yang et al.* [8] used conditional random field (CRF) as the marking model. After the CRF is used to generate the abbreviated candidate list, a length model is constructed to filter the candidate abbreviation. Finally, the full name and abbreviation co-occurrence information of the network search engine further improves the effect. *Zhang et al.* [9] selected candidate pairs through sequence tagging (CRF) and then combined network data with search engine based on a support vector machine (SVM) to reorder candidates and select the best candidate pairs. *Zhang et al.* [14] introduced the minimum semantic unit based on the sequence annotation of word segmentation and used CRF for model training. An integer linear program with various constraints is then used to derive the best abbreviation from the generated candidate. Unlike previous character-marking approaches, this approach captures word-level information finer than characters but coarser than words in a sequence tag framework. Sequence annotation methods can learn the rules of abbreviation generation by themselves, but the existing sequence annotation methods of abbreviation generation tasks have a poor effect on irregular abbreviation generation.

After that, similar to the title abbreviation generation task [1], the Chinese abbreviation generation evolved into a task that expanded from abbreviations to full names [15], [16]. However, these methods are very scenario-specific and unsuitable for enterprise shorthand generation tasks.

Therefore, we propose a two-channel Chinese enterprise abbreviation generation method based on an enterprise component channel and a single-character channel. The two channels of our method are aimed at regular abbreviations and irregular abbreviatiosn respectively, and BERT-BiLSTM-CRF model is used to address the inaccurate generation of irregular abbreviations. Then we use the Bayesian filtering model to select the best abbreviation.

III. METHODOLOGY

In this chapter, we first explain the definition of the enterprise component in the two-channel method. Then, we outline the process of the Chinese enterprise abbreviation generation framework in detail and explain the reasons for such modeling.

TABLE 2. Sample presentation of enterprise components.

Company Name	Component
上海和舟机电设备有限公司	R(上海) U(和舟) I(机电设备) O(有限公司)
三鑫重工机械有限公司	U(三鑫) I(重工机械) O(有限公司)
中国电信股份有限公司	R(中国) I(电信) O(股份有限公司)
品牌中国集团有限公司	I(品牌) R(中国) O(集团有限公司)

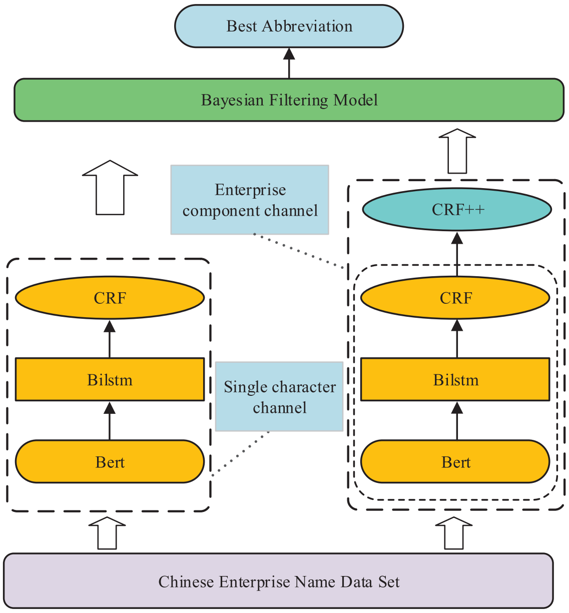


FIGURE 1. The illustration of our proposed two-channel framework.

A. ENTERPRISE COMPONENT DEFINITION

Enterprise names generally follow certain naming rules. The enterprise full name is usually divided into four categories: the region name is denoted R, the unique noun is denoted U, the industry type is denoted I, and the organization type is denoted O [17]. Table. 2 shows the composition of the enterprise name. For example, in "上海和舟机电设备有限公司", "上海" is the region name, denoted R, "和舟" is the unique noun, denoted U, "机电设备" is the industry type, denoted I, and "有限公司" is the organization type, denoted O.

B. METHOD FRAMEWORK DESCRIPTION

We describe a two-channel Chinese enterprise abbreviation generation method in this paper, termed EC-SC. That aims to improve the effect of Chinese enterprise abbreviation generation. Next, we describe the two-channel approach framework and model the method mathematically.

As shown in Fig. 1, the two-channel approach consists of two channels and a Bayesian filtering model. We enter the enterprise name into two channels. In the enterprise component channel, we construct an enterprise component prediction model to identify the component label of the whole enterprise, then CRF++ obtain regular candidate

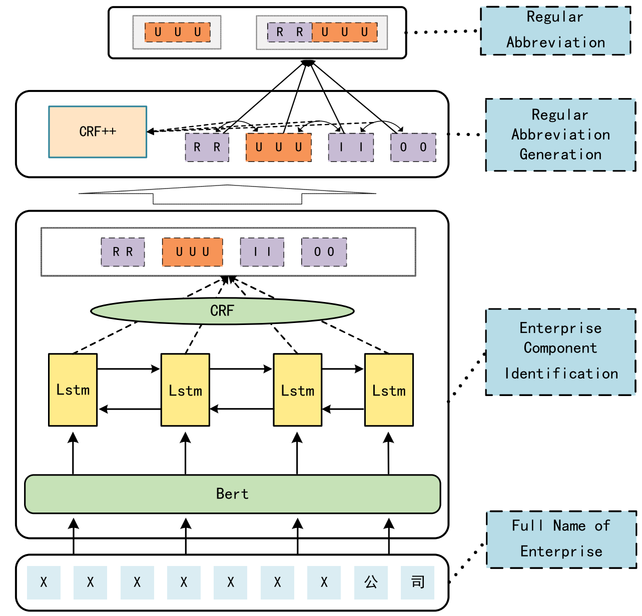


FIGURE 2. Enterprise component channel framework interpretation diagram.

abbreviations of the enterprise through the enterprise component. In the single-character channel, we use BERT-BiLSTM-CRF to construct an irregular abbreviation generation model to obtain irregular candidate abbreviations. With two channels, the effectiveness of irregular abbreviation generation is improved. In the Bayesian filtering model, its input is candidate abbreviations set for the two channels, which exploits the position of abbreviation characters in enterprise names to obtain the best abbreviations.

C. ENTERPRISE COMPONENT CHANNEL

The enterprise component channel consists of two parts: enterprise component identification and regular candidate abbreviation generation. As shown in Fig. 2, in enterprise component identification, we construct an enterprise component prediction model that is composed of a feature vector generation model (BERT) [18], bidirectional long short-term memory (BiLSTM) [19], and a conditional random field (CRF) [20]. Through the enterprise component prediction model we can obtain the enterprise component label. In regular candidate abbreviation generation, we use CRF++ to construct a regular abbreviation prediction model to obtain the regular candidate abbreviation. In this channel, the enterprise name is first tagged by the enterprise component prediction model, and then, the regular candidate abbreviation is generated.

1) ENTERPRISE COMPONENT IDENTIFICATION

To predict the regular abbreviation more accurately, we construct an enterprise component prediction model to obtain the composition pattern of the enterprise name. The BERT pre-training model has a bidirectional function and self-attention mechanism that can extract the feature vector better. Then, BiLSTM extracts the features required for component sequence annotation from the output vector to obtain the enterprise component probability vector. For enterprise component prediction, consistency of labeling within ingredients is needed. Therefore, a CRF is used to calculate the global optimality of the entire sequence.

2) REGULAR CANDIDATE ABBREVIATION GENERATION

After marking the components of the enterprise's full name, we use CRF++ to select whether the characters are retained in the abbreviation based on the components to make the final abbreviation prediction. *Chang et al* [21]. initially used HMM model for abbreviation prediction, but did not consider the case of "Word to Null". Compared with HMM model, CRF [22]–[24] has been widely used in sequence labeling tasks such as medical named entity recognition after it was proposed, which can extracts more sequence information by introducing feature functions. In addition, [25]–[27] *et al.* put CRF at the end of the model to decode and label the optimal global sequence, which also achieved good results. A large number of experiments have proved that the CRF model has a good performance in sequence labeling tasks. Therefore we choose CRF++ to obtain the regular enterprise abbreviations and use enterprise component annotations to enhance the learning of rules by the model.

We input the enterprise component labels into CRF++, defined as $x = \{x_1, x_2 \ldots, x_m\}$. The label value set $Y = \{K, D\}$, where K means "Keep" and D means "Delete". The full name of the prediction label is $y = \{y_1, y_2 \ldots, y_m\}$, and x_i in the input sequence corresponds to y_i in the prediction label set. CRF++ can perform label prediction on new data by learning character retention rules in the label. Finally, according to the predicted label sequence, the character combination corresponding to the full name of the enterprise is selected as the enterprise regular candidate abbreviation.

D. SINGLE-CHARACTER CHANNEL

In the single-character channel, we label each character in the full name of the enterprise with K and D labels and construct an irregular abbreviation prediction model based on the BERT-BiLSTM-CRF model. The input sequence of the model is the enterprise name. It obtains the feature vector of each character combination by BERT. BiLSTM performs selective extraction from the feature vector output by BERT to get the corresponding label information. Then, the CRF layer learns the relationships between adjacent characters and decodes them to improve the performance of the character-level generation of abbreviations. Finally, the enterprise irregular candidate abbreviation is extracted from the enterprise name according to the output sequence.

E. FILTERING MODEL

In essence, an enterprise abbreviation is generated by extracting a character from the enterprise full name [28]. Therefore, the abbreviations and full names of enterprises are strongly

TABLE 3. Sample of filtering model character prior probabilities.

Constitute pattern	Number	Prior probability of position information
RRUUIIIIOOOOOO	1652	R_1: 0.29, ..., I_2: 0.80, U_2: 0.99, ..., O_6: 0.08
UUUUOOOOOO	122	U_1: 0.96, ..., U_4: 0.93, O_1: 0.33, ..., O_6: 0.16
UUIIOOOOOO	349	U_1: 0.96, ..., I_2: 0.87, O_1: 0.27, ..., O_6: 0.25
RRUUOOOOOO	106	R_1: 0.64, ..., U_1: 0.99, O_1: 0.59, ..., O_6: 0.08

related in many aspects. In the process of Chinese enterprise abbreviation generation, the position of the selected character in the full name has a certain degree of influence on the generation of the correct abbreviation. Due to the regularity of enterprise full names, the position of characters in each component of the full name also determines whether they remain abbreviated.

In this paper, each character in the full name of the enterprise is position-coded, and the constituent information of the full name is integrated into the position code. A filtering model based on Bayesian probability is designed. We use the position information of reserved characters in candidate abbreviations to filter the model to obtain the best abbreviation set. The filtering model is divided into the following three steps:

1) First, the characters in the full name of the enterprise are encoded in position based on their constituent components. For example, when a character appears in the first position in the constituent R of the enterprise, it is encoded as R_1.
2) Second, the number of each constituent mode is counted, and the prior probability of the position code retained in the abbreviation is calculated. The details can be shown in Table. 3.
3) Third, the character position-coding calculates the Bayesian probability of the candidate abbreviation in the candidate abbreviation and the prior probability of the formation mode of the position-coding in the full name of the enterprise.

In the filtering model, the input sequence is the coding sequence $L = \{l_1, l_2 \ldots, l_n\}$, and l is the component position label of the abbreviation character. The output $P(L)$ is the Bayesian probability value of the candidate abbreviation. $A_{n-1} = (l_1, l_2, \ldots, l_{n-1})$, indicating that $l_1, l_2 \ldots, l_{n-1}$, the characters corresponding to the l_{n-1} sequence tag, exist at the same time. Event B_n is reserved for the character corresponding to the label l_n. The Bayesian probability of event B_n occurring when event A_{n-1} occurs is defined as follows:

$$P(B_n|A_{n-1}) = \frac{P(B_n)P(A_{n-1}|B_n)}{\sum_{m=1}^{n} P(B_m)P(A_{n-1}|B_m)}. \tag{1}$$

The Bayesian probabilities of each character in the abbreviation are accumulated to obtain the final candidate abbreviation Bayesian probability. The formula for $P(L)$ is defined as follows:

$$P(L) = \sum_{2}^{n} P(B_n|A_{n-1}) + P(B_1). \tag{2}$$

As shown in Fig. 3, we obtain regular candidate abbreviations and irregular candidate abbreviations for Chinese enterprises through two channels. In the enterprise component channel, the composition pattern of the enterprise name and the position probability of each character in this pattern are obtained. R_1 is the first character of the local component in the combination mode, $P(R_1)$ is the probability of the character appearing in the abbreviation, and the probability is the prior probability. In the filtering part, we find the corresponding position information and prior position probability according to the candidate abbreviation and calculate the bayesian probability of each character in the candidate abbreviation by using Equation 1. The final probability of candidate abbreviations is then obtained by Equation 4. Finally, the candidate abbreviations are sorted according to Bayesian probability to obtain the best abbreviation set.

Algorithm 1 Two-channel(ES-CS) Algorithm

```
Input: Chinese enterprise name $C_{name}$;
Output: Enterprise abbreviation $En$;
Procedure:
Regular abbreviation prediction:
  $word\_vector \leftarrow Bert(C_{name})$;
  $bi\_vector \leftarrow Bilstm(word\_vector)$;
  $component\_label \leftarrow CRF(bi\_vector)$;
  $regular\_abb \leftarrow CRF++(component\_label)$;
Irregular abbreviation prediction:
  $irregular\_abb \leftarrow Bert\_Bilstm\_CRF(C_{name})$;
Fusion candidate abbreviation:
  $all\_abb \leftarrow regular\_abb \mid irregular\_abb$;
Screening best abbreviation:
for $j$; $j \leq length(all\_abb)$ do
    for $l_n \in all\_abb_j$ do
        let: $A_{n-1} \leftarrow l_1, l_2, \ldots, l_{n-1}$;
        let: $B_n \leftarrow l_n$;
        $P(B_n|A_{n-1}) \leftarrow \frac{P(B_n)P(B_n|A_{n-1})}{\sum_{m=1}^{n} P(B_m)P(A_{n-1}|B_m)}$;
    end for
    $P(L_j) \leftarrow \sum_{2}^{n} P(B_n|A_{n-1}) + P(B_1)$;
end for
$storted(all\_abb)$;
$E_n \leftarrow Top(all\_abb)$;
Return $E_n$
```

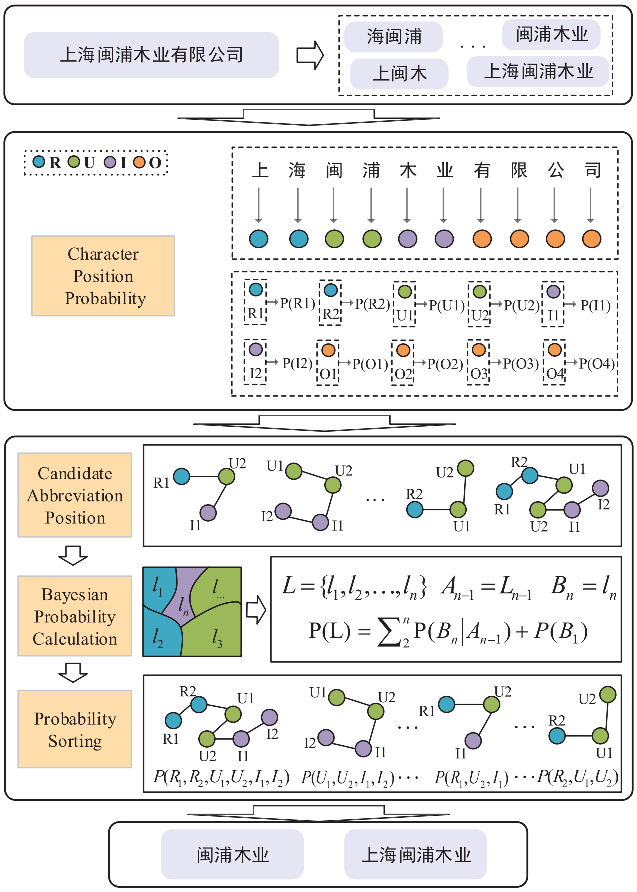

上海闽浦木业有限公司
海闽浦
...
闽浦木业
上闽木
上海闽浦木业
R U I O
Character Position Probability
上 海 闽 浦 木 业 有 限 公 司
R1 P(R1) R2 P(R2) U1 P(U1) U2 P(U2) I1 P(I1)
I2 P(I2) O1 P(O1) O2 P(O2) O3 P(O3) O4 P(O4)
Candidate Abbreviation Position
R1 U2 I1 U1 U2 I2 I1 ... R2 U2 U1 R2 U1 R1 U2 I1 I2
Bayesian Probability Calculation
l_1 l_n $l_{...}$ l_2 l_3
$L=\{l_1,l_2,\ldots,l_n\}$ $A_{n-1}=L_{n-1}$ $B_n=l_n$
$P(L)=\sum_2^n P(B_n|A_{n-1})+P(B_1)$
Probability Sorting
R2 U1 R1 U2 I1 I2 U1 U2 I2 I1 ... R1 U2 I1 ... U2 R2 U1
$P(R_1,R_2,U_1,U_2,I_1,I_2)$ $P(U_1,U_2,I_1,I_2)\cdots$ $P(R_1,U_2,I_1)\cdots P(R_2,U_1,U_2)$
闽浦木业
上海闽浦木业

FIGURE 3. The illustration of the filtering model.

The abbreviation generation algorithm with the two-channel method is shown in Algorithm 1. First, the regular candidate abbreviation is predicted according to the enterprise name. Specifically, we use BERT to learn and obtain feature vectors of enterprise names, and then extract the features required for entity recognition from these feature vectors by BiLSTM. The CRF layer is decoded to obtain the sequence component label. CRF++ takes the enterprise component label as input and outputs the best regular candidate abbreviation set. Second, BERT-BiLSTM-CRF is used to predict irregular abbreviation candidate sets based on enterprise names. Third, the candidate abbreviations of both channels are merged. Fourth, the Bayesian probability is obtained according to the position information of each character of the abbreviation, and the best abbreviation set is obtained by sorting.

IV. EXPERIMENTS

This section describes the dataset construction method, framework evaluation metrics. The evaluation results are compared with the existing methods of automatic abbreviation generation.

A. DATA PREPARATION

In terms of data acquisition, this article crawls the full names and abbreviations of companies from Baidu entries and Wikipedia's corporate introductions, and crawls news from some financial news websites (*Financial News*, *Sina News*, *Toutiao*), and uses NER to identify the entities in the news and extract full-abbreviation pairs in the knowledge base. To improve the accuracy of the dataset, manpower will be added to check that each obtained full-abbreviation pair is correct.

TABLE 4. Data sets used for training the models.

Data set name		Train	Dev	Test
Abbreviated-annotation	Regular-abbreviation	5760	500	500
	Irregular-abbreviation	2240	500	500
Component-labeling		8000	1000	1000

In our experimental data, a full name of an enterprise corresponds to one or more abbreviations. We classify the obtained full abbreviation pairs according to abbreviation types. In irregular abbreviations, our dataset still includes regular abbreviations to verify that the single character channel can maintain high accuracy in regular abbreviations and improve the generation effect of irregular abbreviations. We also annotate the component of each character in all enterprise names and construct the enterprise component annotation dataset.

As shown in Table. 4, through data collection, the experiment prepares 8000 full abbreviation pairs. The enterprise abbreviation prediction model dataset is divided into two parts, a regular abbreviation set and an irregular abbreviation set, of which 75% are regular abbreviations and 25% are irregular abbreviations. In addition, all enterprise names are annotated with RegionName (R), UniqueNoun (U), IndustryType (I), and OrgType (O) sections to construct enterprise component dataset. When selecting the enterprise name, the enterprise component set contains as many enterprise name composition patterns as possible.

The statistics of the datasets are detailed in Table. 4, our dataset is divided into two parts, Abbreviated annotation and Component labeling. Abbreviated annotation is constructed by marking whether the characters are reserved in the abbreviation, divided into regular and irregular abbreviation sets. Component labeling is constructed by labeling the enterprise components with the characters in the full name of the enterprise. Then Dev represent the validation set.

B. EXPERIMENTAL SETTING

Standard measurement methods are used for model evaluation, such as accuracy, recall and F1-score. Each experiment is run 10 times with four datasets of different sizes, and the average result is taken as the final result of the model.

C. ALGORITHM DESIGN FOR COMPARISON

This section introduces the comparative ideas of generating experiments. We compare the experimental results of each

TABLE 5. Introduction of comparison method.

Methods	Condition
Rules+MLN [7]	based on rules
CRF [8]	character sequence annotation
ILP [14]	Minimum Semantic Unit and Global Constraints
CRF+Rules [6]	character sequence annotation and rules
CNN-BLSTM-CRF [29]	character sequence annotation

part of the method to highlight the role of each subprocess in this method. In terms of the experimental data, we use the data we constructed to calculate the comparison effect achieved by restoring the comparison experimental algorithm as much as possible. We choose the methods shown in Table. 5 for experimental comparison.

1) Rules+MLN

This method defines global character deletion rules and local character deletion rules and considers that words provide important information for the generation of abbreviations. Because of the flexibility of the Markov logical network (MLN) in capturing local and global language features, this method uses the MLN to select local or global deletion rules for words.

2) CRF

In this method, the abbreviation generation problem is transformed into a character sequence annotation task, and each character in the full name is sequentially annotated to construct an abbreviation generation model. It is necessary to build a certain number of datasets and to tell the model the features to be learned to train the model. After the candidate abbreviation is obtained, the model is modeled by the length relation between the full name and the abbreviation, and the best abbreviation is selected by combining the co-occurrence times of the full name and the abbreviation in the web search engine. This method reduces some workload and only needs to annotate some data.

3) ILP

The sequence labeling method performs poorly when the "character duplication" phenomenon exists. To solve the character duplication problem in Chinese abbreviation prediction, this method also uses a substring tagging strategy to generate local substring tagging candidates. This method uses an integer linear programming (ILP) formulation with various constraints to globally decode the final abbreviation from the generated candidates.

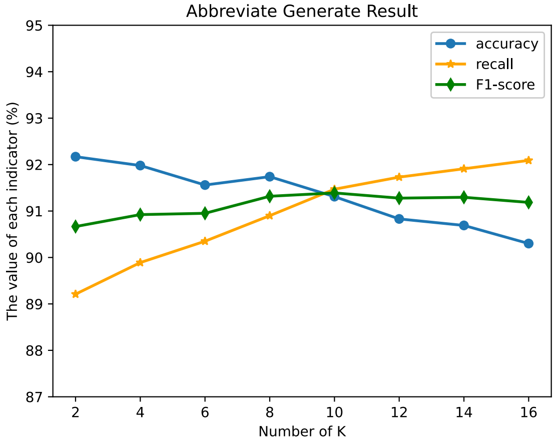


FIGURE 4. Comparison of parameter *K* in terms of accuracy, recall, and F1-Score.

4) CRF+Rules

This method combines the characteristics of rules and character sequence annotation to generate abbreviations, formulates certain rule sets, achieves a self-learning effect of rules through an algorithm model, and uses the CRF model to generate abbreviations to make up for the abbreviations not covered by rules. This method combines the high accuracy of rules with the high coverage of models to achieve better results.

5) CNN-BLSTM-CRF

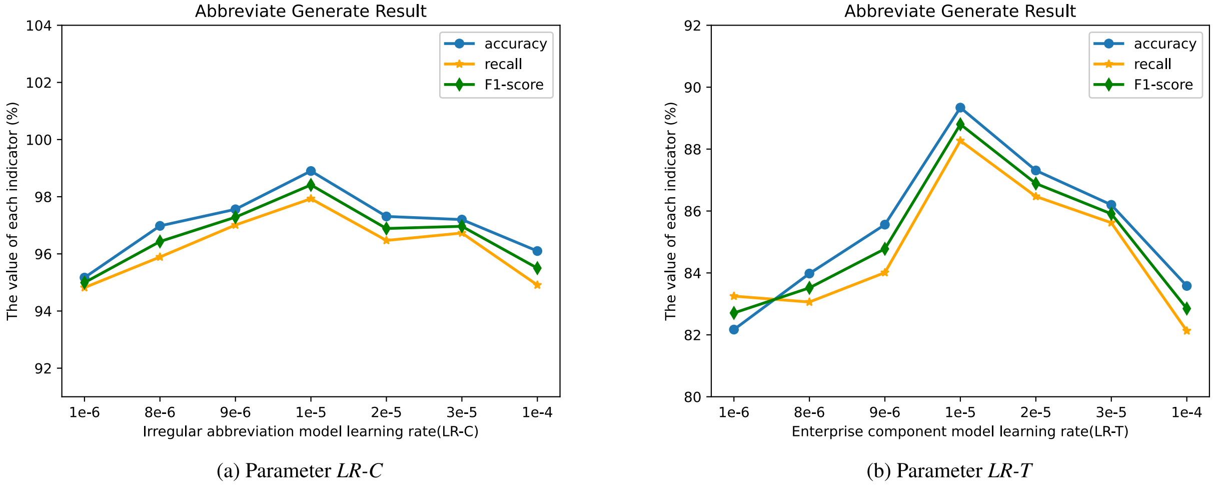
The neural network-based abbreviation generation model avoids a lot of manual annotation and feature engineering and can predict abbreviation better without much dependence on the feature system. CNN-BLSTM-CRF neural network model extracts the local word and Chinese character information from the full name through the convolutional neural network that uses The BLSTM model to retain the context information selectively. Finally, CRF obtains the optimal sequence by learning the information of adjacent tags.

D. EXPERIMENTAL PARAMETER SELECTION

This section introduces the related parameter selection of the experiment. The core parameter candidates of this method are the referred to as coverage parameter *K*, enterprise component prediction model learning rate *LR-C*, and irregular abbreviation prediction model learning rate *LR-T*. We describe the selection of each core parameter separately; see the following subsections for details.

1) PARAMETER *K*

The candidate coverage parameter *K* is an important parameter affecting the recall rate of the model. Liberalizing the number of predicted abbreviations for the two channels in the framework model can increase the probability of correct abbreviations being predicted by the model. As shown in Fig. 6, as the *K* value increases, the probability of candidate abbreviations being the correct abbreviations increases.



(a) Parameter *LR-C*

(b) Parameter *LR-T*

FIGURE 5. Comparison of parameter *LR-C*, *LR-T* in terms of accuracy, recall, and F1-Score.

At the same time, the ever-increasing number of interference abbreviations makes the filtering model face more significant difficulties, which leads to a decline in the final extraction accuracy. Through the experimental results, the value of parameter K is 10, which is the number of candidate abbreviations needed to balance the accuracy and recall rate of the model.

2) PARAMETERS *LR-C* AND *LR-T*

By adjusting the enterprise component prediction model learning rate parameter *LR-C* and the irregular abbreviation prediction model learning rate parameter *LR-T*, the prediction effect of the enterprise composition and irregular abbreviation can be improved. The learning rate affects the convergence time of the model and whether the model fits the training data in the training model. Therefore, we find a suitable value to improve the effect of the two-channel abbreviation generation method. We use regular and irregular abbreviation sets to adjust parameters based on the enterprise component prediction model, and use an irregular abbreviation set to adjust parameters based on the character sequence annotation model. We set the learning rate parameter in the range $[0.1 - 10] * E^{-5}$.

As shown in Fig. 5a, after training, the enterprise component prediction model achieves a good effect overall. The curve shows an upward trend from $1E^{-6}$ to $1E^{-5}$, and reaches the optimal effect at $1E^{-5}$. Then, with the increase in the learning rate, the prediction effect of enterprise components decreases. Therefore, we take $1E^{-5}$ as the optimal parameter of the enterprise component prediction model.

As shown in Fig. 5b, the effects of the irregular abbreviation prediction model are significantly different at different learning rates. When the value of the learning rate is $1E^{-6}$, the F1-score value of the model is only 82.71%. The model is too fit to verify the data, resulting in an unsatisfactory effect on the test data. At $1E^{-5}$, the optimal effect is achieved for all indexes. With the increase in the learning rate, the prediction effect of the irregular abbreviation decreases. Therefore, in the final experiment of this paper, we choose $1E^{-5}$ as the learning rate parameter for the irregular prediction model.

E. EFFECTIVENESS ANALYSIS

In this section, we first conduct a comparative experiment on each link of enterprise abbreviation generation based on two channels, reflect the experimental effect through the accuracy, recall, and F1-score values, use Brier fraction and logarithmic loss to evaluate the filtration model. Then, the experimental results are compared with those of several abbreviation generation methods. Finally, the experimental conclusion is drawn.

1) ENTERPRISE ABBREVIATION FORECAST RESULTS

In the enterprise component channel (EC) we only test the effect on the regular abbreviation set, while in the single-character channel (SC) we test the effect on each dataset and the combination of irregular and regular abbreviations. Finally, the Bayesian filtering model is connected to each channel to observe the role of each link in the two-channel method.

As shown in Table. 6, the experimental results show that the F1-score value of the EC channel is 1.76% higher than that of the SC channel on the regular abbreviation set. Thus, the feasibility of the CRF++ regular abbreviation generation method is proved, and the effect of abbreviation prediction improves after the component division of the enterprise name. In addition, the F1-score values of the two channels in the regular abbreviation set increase after the filtering model is used. In the irregular abbreviation set, the F1-score value of the SC model with the filtering model is improved by 1.30%. This shows that the filtering model can select the best abbreviation from the candidate abbreviations. As the number of fused candidate abbreviation increases, the effect of the two-channel method decreases in the regular and irregular abbreviations, but improves by 1.54% in the total abbreviations compared with the F1-score value of the SC channel.

TABLE 6. The test accuracy, recall, and F1-score of all self-comparison on three datasets.

Method	*Data sets*	*accuracy*	*recall*	*F1-score*
EC	*Regular-abb*	**93.52%**	91.37%	92.43%
EC+Filter	*Regular-abb*	**93.81%**	91.40%	92.59%
SC	*Regular-abb*	91.76%	91.28%	91.52%
	Irregular-abb	88.69%	87.93%	88.31%
	All-abb	89.45%	88.80%	89.12%
SC+Filter	*Regular-abb*	**92.13%**	91.81%	91.97%
	Irregular-abb	90.71%	89.16%	89.93%
	All-abb	91.29%	89.57%	90.42%
EC-SC	*Regular-abb*	**93.57%**	92.61%	**93.09%**
	Irregular-abb	90.13%	88.87%	89.50%
	All-abb	**92.22%**	91.71%	91.96%

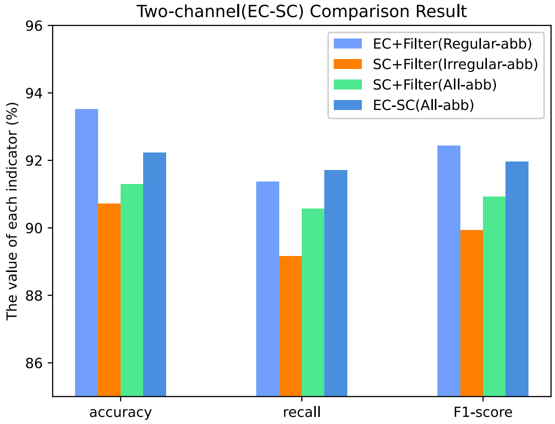


FIGURE 6. Partial results presentation of EC-SC at accuracy, recall, and F1-score.

In general, the EC channel can deal with regular abbreviations well, the SC channel has good performance in irregular abbreviations, and the filtering model can select the best abbreviations. Fig. 6 shows some comparative experimental results.

Experimental results show that using enterprise components to generate abbreviation can better capture the rule of regular abbreviation generation. The location information of th enterprise component is used in Bayesian probability to filter the best abbreviation from the candidate abbreviations.

2) FILTERING MODEL EVALUATION EXPERIMENT

We use the Brier fraction and logarithmic loss index for the filtering model to measure the difference between the probability of candidate abbreviation predicted by the model and the labeled abbreviation. The two indexes are negatively correlated with the model effect. The expression for the Brier fraction is as follows:

$$B_{score} = \frac{1}{N}\sum_{L}^{n}(P(L) + O_L). \tag{3}$$

where N represents the number of samples, $P(L)$ represents the probability value predicted by the filtering model, and O_L represents the real label information. If the candidate abbreviation is in the set of original labeled abbreviations, which indicates that the abbreviation event occurs, $O_L = 1$; otherwise, $O_L = 0$. The difference between the predicted and the original labeled abbreviation is accumulated to obtain the final Brier fraction model measurement index.

When calculating logarithmic loss, we set that there is a classification label for the prediction candidate abbreviation. The logarithmic loss formula is as follows:

$$L_{loss} = \frac{1}{N}\sum_{L}^{n} -(L_{true} * log(P(L)) \\ +(1 - L_{ture}) * log(1 - L_{true})). \tag{4}$$

where N is the number of samples, L is the sequence label of candidate abbreviation, L_{true} is the true label of candidate abbreviation, and $P(L)$ represents the probability value predicted by the filtering model. If the candidate abbreviation appears in the original labeled dataset, $L_{true} = 1$; otherwise, $L_{true} = 0$.

We evaluate the filtering model by comparing the results of the two-channel method with and without the filtering model on the Brier fraction and logarithmic loss index. As shown in Table. 7, in each dataset, the values of the two indexes decreased significantly after adding the filtering model. Due to the negative correlation between the two indexes and the filtering model, it can be observed that the filtering model plays a good role in the two-channel method.

3) COMPARATIVE EXPERIMENT ANALYSIS

Due to the problem of rule set construction, we reproduce only the CRF and ILP methods on our own dataset. In the other two methods, we directly compare the experimental results in the paper.

As shown in Fig. 7, the MLN method has good performance in terms of recall by identifying as many correct abbreviations as possible by the rules. Compared with the CRF method, the ILP method with a minimum semantic unit and global decoding can better extract full name feature information. In the CRF-Rule method, the rule and sequence annotation methods complement each other to improve the accuracy of abbreviation generation. However, the Rule-based approach consumes many resources in practice. The performance of the CNN-BLSTM-CRF model in enterprise abbreviation generation is significantly improved compared with CRF and other methods. The model rarely relies on feature engineering, which proves the neural network model's feasibility in the abbreviation generation task. The two-channel approach is similar to the sequential rule

TABLE 7. B_{score} **and** L_{loss} **of our EC-SC method with and without filtering model.**

Indicators	*EC-SC (No Filter)*			*EC-SC*		
	Regular-abb	*Irregular-abb*	*All-abb*	*Regular-abb*	*Irregular-abb*	*All-abb*
B_{score}	0.29	0.68	0.52	0.17	0.34	0.31
L_{loss}	0.36	0.72	0.61	0.23	0.57	0.45

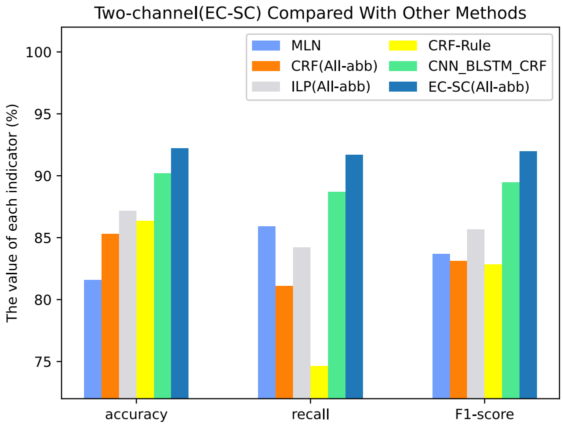


FIGURE 7. **The test accuracy, recall, and F1-score of all methods.**

approach, but we use CRF++ for regular abbreviation prediction based on the enterprise component. Moreover, our neural network-based two-channel model performs better than the CNN-BLSTM-CRF model in the abbreviation generation task. From the F1-score value, the overall effect of the two-channel method is better than that of the other methods.

From Fig. 7, we can see that compared with other methods, ES-CS achieves the best results on all indexes. It also shows that the two-channel method has the characteristics of the end-to-end neural network to avoid some cumbersome process of feature engineering and can also make full use of the enterprise name's information to improve the effect of the abbreviation generation task.

V. CONCLUSION

This paper design a two-channel enterprise abbreviation generation method based on an enterprise component channel and a single-character channel. In the subprocess based on the enterprise component channel, first, we use BERT-BiLSTM-CRF to label the full name of the enterprise to obtain the composition of the enterprise and then use CRF++ results to make predictions for regular candidate abbreviations based on labeling. In the single-character channel, we directly use BERT-BiLSTM-CRF to generate irregular candidate abbreviations. Then, we input the candidate abbreviations of the two-channel output into the Bayesian filtering model to obtain the best abbreviation set.

The experimental results show that our two-channel method improves the effect of irregular abbreviation generation, and the Bayesian filtering model based on the position of abbreviations in enterprise components is more fine-grained. In our future work, we will search for more comprehensive abbreviations through web and manual accumulation and try to increase the amount of training data during its construction to observe the prediction effect of the method.

ACKNOWLEDGMENT

The authors would like to thank the experimental equipments provided by National Super Computing Center of Changsha, Hunan, China.

REFERENCES

[1] J. Zhang, Y. Sun, S. Huang, C.-T. Nguyen, X. Wang, X. Dai, J. Chen, and Y. Yu, "AGRA: An analysis-generation-ranking framework for automatic abbreviation from paper titles," in *Proc. IJCAI*, 2017, pp. 4221–4227.

[2] Y. Jiao, H. Wang, H. Wang, and L. Zhang, "Abbreviation prediction using conditional random field and web data," *J. Chin. Inf. Process.*, vol. 26, no. 2, pp. 62–69, 2012.

[3] S. Pakhomov, "Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 160–167.

[4] H. Yu, W. Kim, V. Hatzivassiloglou, and J. Wilbur, "A large scale, corpus-based approach for automatically disambiguating biomedical abbreviations," *ACM Trans. Inf. Syst.*, vol. 24, no. 3, pp. 380–404, Jul. 2006.

[5] Y. HaCohen-Kerner, A. Kass, and A. Peretz, "Combined one sense disambiguation of abbreviations," in *Proc. ACL HLT*, 2008, pp. 61–64.

[6] S. Liping, G. Yi, T. Wenwu, and X. Yongbin, "Enterprise abbreviation prediction based on constitution pattern and conditional random field," *J. Comput. Appl.*, vol. 36, no. 2, pp. 449–454, 2016.

[7] H. Chen, Q. Zhang, J. Qian, and X.-J. Huang, "Chinese named entity abbreviation generation using first-order logic," in *Proc. 6th Int. Joint Conf. Natural Lang. Process.*, 2013, pp. 320–328.

[8] D. Yang, Y.-C. Pan, and S. Furui, "Automatic Chinese abbreviation generation using conditional random field," in *Proc. Hum. Lang. Technol., Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2009, pp. 273–276.

[9] L. Zhang, S. Li, H. Wang, N. Sun, and X. Meng, "Constructing Chinese abbreviation dictionary: A stacked approach," in *Proc. COLING*, 2012, pp. 3055–3070.

[10] J. Qin, L. Ma, and Y. Wang, "A generation method for Chinese equipment name abbreviations based on rules," in *Proc. 2nd Joint Int. Inf. Technol., Mech. Electron. Eng. Conf. (JIMEC)*. Amsterdam, The Netherlands: Atlantis Press, 2017, pp. 1–5.

[11] J.-B. Kim, H.-S. Oh, S.-S. Nam, and S.-H. Myaeng, "Using candidate exploration and ranking for abbreviation resolution in clinical document," in *Proc. IEEE Int. Conf. Healthcare Informat.*, Sep. 2013, pp. 317–326.

[12] A. Jain, S. Cucerzan, and S. Azzam, "Acronym-expansion recognition and ranking on the web," in *Proc. IEEE Int. Conf. Inf. Reuse Integr.*, Aug. 2007, pp. 209–214.

[13] H. Liu, Y. Chen, and L. Liu, "Automatic expansion of Chinese abbreviations by web mining," in *Proc. Int. Conf. Artif. Intell. Comput. Intell.* Berlin, Germany: Springer, 2009, pp. 408–416.

[14] L. Zhang, L. Li, H. Wang, and X. Sun, "Predicting Chinese abbreviations with minimum semantic unit and global constraints," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1405–1414.

[15] K. Gorman, C. Kirov, B. Roark, and R. Sproat, "Structured abbreviation expansion in context," in *Proc. Empirical Methods Natural Lang. Process. (EMNLP)*, 2021, pp. 995–1005.

[16] Q. Jin, J. Liu, and X. Lu, "Deep contextualized biomedical abbreviation expansion," in *Proc. BioNLP Workshop Shared Task*, 2019, pp. 88–96.

[17] L.-W. Zhong and T. Zheng, "Study on approaches to retrieval of Chinese organization name based on its abbreviated name," *J. Chin. Inf. Process.*, vol. 21, no. 1, pp. 38–42, 2007.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Assoc. Comput. Linguistics (NAACL)*, vol. 1, 2018, pp. 4171–4186.

[19] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 207–212.

[20] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1529–1537.

[21] J.-S. Chang and Y.-T. Lai, "A preliminary study on probabilistic models for Chinese abbreviations," in *Proc. 3rd SIGHAN Workshop Chin. Lang. Process.*, 2004, pp. 9–16.

[22] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[23] R. Tang, Z. Yu, Y. Ma, Y. Wu, Y.-P. P. Chen, L. Wong, and J. Li, "Genetic source completeness of HIV-1 circulating recombinant forms (CRFs) predicted by multi-label learning," *Bioinformatics*, vol. 37, no. 6, pp. 750–758, May 2021.

[24] H. Shah, T. Xiao, and D. Barber, "Locally-contextual nonlinear CRFs for sequence labeling," 2021, *arXiv:2103.16210*.

[25] X. Li, H. Shu, Y. Zhai, and Z. Lin, "A method for resume information extraction using BERT-BiLSTM-CRF," in *Proc. IEEE 21st Int. Conf. Commun. Technol. (ICCT)*, Oct. 2021, pp. 1437–1442.

[26] Y. Qin, G.-W. Shen, W.-B. Zhao, Y.-P. Chen, M. Yu, and X. Jin, "A network security entity recognition method based on feature template and CNN-BiLSTM-CRF," *Frontiers Inf. Technol. Electron. Eng.*, vol. 20, no. 6, pp. 872–884, Jun. 2019.

[27] Y. Jia and X. Xu, "Chinese named entity recognition based on CNN-BiLSTM-CRF," in *Proc. IEEE 9th Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Nov. 2018, pp. 1–4.

[28] L. Zhang, H. Wang, and X. Sun, "Coarse-grained candidate generation and fine-grained re-ranking for Chinese abbreviation prediction," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1881–1890.

[29] J. Zheng, X. Xiao, B. Wang, Y. Zhu, and L. Yang, "A new method for abbreviation prediction via CNN-BLSTM-CRF," *J. Phys., Conf. Ser.*, vol. 1267, no. 1, Jul. 2019, Art. no. 012001.

WEI AI received the Ph.D. degree from the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. Her research interests include date mining, big data, cloud computing, and parallel computing.

HONGEN SHAO is currently pursuing the master's degree with the School of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha, China. His research interests include natural language processing and named entity recognition.

JIA XU is currently pursuing the master's degree with the School of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha, China. Her research interests include natural language processing and big data analysis.

TAO MENG received the Ph.D. degree from the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. His research interests include date mining and network analysis.

KEQIN LI (Fellow, IEEE) is currently a SUNY Distinguished Professor in computer science with the State University of New York. He is also a National Distinguished Professor with Hunan University, China. He has authored or coauthored over 850 journal articles, book chapters, and refereed conference papers, and has received several best paper awards. He holds over 70 patents announced or authorized by the Chinese National Intellectual Property Administration. He is among the world's top five most influential scientists in parallel and distributed computing based on a composite indicator of Scopus citation database. His current research interests include cloud computing, fog computing, mobile edge computing, energy efficient computing and communication, embedded systems and cyber-physical systems, heterogeneous computing systems, big data computing, high-performance computing, CPU-GPU hybrid and cooperative computing, computer architectures and systems, computer networking, machine learning, intelligent, and soft computing. He is a Member of Academia Europaea (Academician of the Academy of Europe). He has chaired many international conferences. He is currently an Associate Editor of the *ACM Computing Surveys* and the *CCF Transactions on High Performance Computing*. He has served on the editorial boards of the IEEE Transactions on Parallel and Distributed Systems, the IEEE Transactions on Computers, the IEEE Transactions on Cloud Computing, the IEEE Transactions on Services Computing, and the IEEE Transactions on Sustainable Computing.

• • •