Chinese Named Entity Recognition for Clothing Knowledge Graph Construction

Ming Zhu¹ and De-sheng Zhen¹,*

¹School of Computer Science and Technology, Donghua University, Shanghai 201620, China

*2397646126@qq.com

Abstract. Clothing knowledge graph is a kind of vertical domain knowledge base constructed for the description of clothing knowledge in the field of textile and apparel. In this paper, based on the limitations of the clothing knowledge graph in the effect of entity extraction, the deep learning model and the statistical model are combined. A Chinese named entity recognition method based on CNN-BiLSTM-CRF is proposed. Firstly, the convolutional neural network (CNN) is used to extract the text features, and the character-level vectors with morphological features of the words are trained. Then the bi-directional long short term memory networks (LSTM) is used to learn the context features, and the vector representation of the context of each word is output. Finally, the conditional random fields (CRF) model is used for self-learning. Get the best tag sequence for the sentence. The method can automatically recognize the text, and does not rely on the artificial feature to obtain the semantic category information. Finally, the experimental data and evaluation methods are introduced. The experimental results show that the Chinese named entity recognition method based on CNN-BiLSTM-CRF is superior to other models in all indicators, indicating the effectiveness of the method.

1. Introduction

In the context of the era of big data, with the emergence of massive data and the cross-application of multiple data sources, traditional data management modes and query methods are subject to certain constraints. In recent years, with the development of knowledge graph technology, the knowledge base with semantic processing capability and open interconnection ability has created great value in many fields such as intelligent search[1], knowledge quiz[2], recommendation system[3] and so on. Knowledge graph is a structured semantic knowledge base for describing various entities or concepts and their relationships in the real world. It constitutes a huge semantic network graph. Nodes represent entities or concepts, while edges are composed of attributes or relationships[4]. With the rapid development of e-commerce, online shopping has become an important channel for consumers to purchase clothing. Compared with the traditional market, there are more varieties and types of clothing sold online. Therefore, how to organize the clothing data and make it appear in front of the user in a structured and connected way is very important. Constructing knowledge graph in the field of clothing has an important research significance and application value for the realization of clothing knowledge quiz. Although the knowledge graph has developed rapidly, there is no mature and professional knowledge graph in the clothing field. Therefore, this paper constructs the conceptual layer of clothing knowledge graph, and takes knowledge acquisition as the starting point. A Chinese named entity
recognition method based on CNN-BiLSTM-CRF is proposed. The clothing entity and attribute can be automatically extracted from unstructured clothing text data. The experimental results show that the model achieves good results compared with other named entity recognition models without relying on any artificial features, and the F value can reach 91.2%.

2. Related Works
In the process of constructing the knowledge graph, it is necessary to perform physical extraction on the structured, semi-structured and unstructured text data of the clothing. For structured and semi-structured data, we only need to learn its fixed structure to extract entities and related relationships effectively, but extracting entities from unstructured text is a very difficult research task. Therefore, the rule-based approach gradually fades out the field of named entity recognition research. At present, there are many commonly used machine learning models to solve the problem of named entity recognition, such as maximum entropy model [5], hidden Markov model [6], conditional random field [7] and so on. Conditional random field is a discriminant model that does not require strict independence assumptions and overcomes the shortcomings of mark bias inherent in the maximum entropy model. Conditional random field is very suitable for sequence analysis. In recent years, it has been widely used in sequence labeling tasks, and it has outstanding performance in named entity recognition. The machine learning-based entity extraction method faces the problems of uneven data quality and low professional labeling when it is used in professional fields. Deep learning [8] is a new field in machine learning research. In recent years, deep learning has gradually been applied to the field of natural language processing. Based on the deep learning method, it can learn autonomously from the original data because it does not need to be artificially set. Therefore, it can reduce the interference of human beings on data, find deeper and more abstract features, and has become a research hotspot in recent years. Literature [9] uses a long short term memory neural network model to improve the performance of word segmentation, but this method can not obtain the semantic information behind the sentence. Literature [10] proposes to use CRF as the processing method of the output processing layer on the basis of bi-directional LSTM, which effectively improves the performance of the model.

3. Definition And Construction of the Concept Layer
The construction of the concept layer is the construction of the entire clothing knowledge graph framework. The construction of the knowledge graph concept layer is equivalent to the establishment of Ontology [11]. The most basic ontology includes concepts, concept levels, attributes, attribute value types, relationships, relationship definition fields, concept sets and relationship values, concept sets. This paper uses the protege5.2.0 tool and uses the five-tuple form to describe the clothing knowledge ontology model, defined as $\text{CMO} = \{C, P, H^c, I, R\}$, where $C$ represents all conceptual collections representing clothing, and $P$ represents all the set of attributes, $H^c$ represents the hierarchical structure of the ontology concept, that is, the classification structure of the concept in C, $I$ represents the instance set, and $R$ represents the relationship rule. First, give a hierarchical definition of the concept of the clothing field. Followed by the concept of attribute definitions and constraints. For example, in the clothing ontology, the domain of "isPartof" is the tops and bottoms, the value domain is the full body; the domain of "isFitfor" is the clothing, and the value domain is the human. These attributes can be used to relate concepts in the ontology. Such classes and classes can be associated by attributes. Then there is the definition of rules or axioms. According to the rules, some scattered and implicit knowledge can be inferred, and the original information can be supplemented and improved. Finally, between each stage and stage of the ontology life cycle, a certain reference frame is used to make technical judgments on the ontology and software environment and documents, to acquire clothing knowledge, and finally to generate a clothing knowledge base.

4. Chinese Named Entity Recognition Based on CNN-BiLSTM-CRF
In the process of constructing the clothing ontology knowledge base, a large number of clothing entities and their attributes need to be acquired. For structured and semi-structured data, lack of clothing expertise leads to incomplete clothing knowledge. This part of the knowledge can be obtained from unstructured text data such as documents. Therefore, using deep learning methods, features in unstructured text can be extracted automatically without manual participation. Therefore, this paper proposes a Chinese named entity recognition method based on CNN-BiLSTM-CRF, which can automatically extract clothing named entities from unstructured data. Based on BiLSTM-CRF Chinese named entity recognition method, the combination of deep learning model and statistical learning model can effectively extract text context information. At the same time, in order to make use of these spatial semantic information, a convolutional neural network is introduced to extract spatial semantic features. The model framework is shown in Figure 1. From the figure, the model is divided into four parts: the word vector representation part, the CNN part, the BiLSTM part and the CRF part.

![CNN-BiLSTM-CRF structural model](image)

### 4.1. Convolutional Neural Network Model Layer
This layer uses the convolutional layer to extract features from the input matrix vector. Using CNN to combine the word vector of the input word with the CNN window sliding to connect the current word with the preceding and following words, calculate the influence of the words before and after the current word, and the generated words are used to represent the feature of the word. After the convolution is completed, the word is extracted. The context information between the character and the character constitutes a new word and sentence representation feature, and the word vector is expanded in detail, and then input into the lower layer neural network.

The convolution operation uses a filter $w = R^{h \times k}$ to join the words with $h$ as a window to form a new feature. For example, $x_{i:j+h-1}$ words are combined into a new feature $c_i$.

$$c_i = f(w \cdot x_{i:j+h-1} + b)$$ (1)

Where $b \in R$ represents the bias term, $f(\cdot)$ represents the nonlinear excitation function, and filter $w$ combines each possible window word to form a new feature map $c = [c_0, c_1, ..., c_{n-h+1}]$. Then, we apply the maximum pooling operation on the feature map, and take the maximum value $c = \max\{c\}$ as the feature corresponding to this particular filter. This step can obtain the most important features and can adapt to the variable. The length of the sentence.

### 4.2. Bi-directional Long Short Term Memory Networks Model Layer
The one-way LSTM neural network model can only obtain the above information of the sentence, and the context information of the sentence after the sentence cannot be obtained. In order to compensate for the fact that the one-way LSTM cannot obtain the context information behind the sentence, this
layer uses the bi-directional long short term neural network (BiLSTM) model. The BiLSTM receives the character vector as an input and obtains a hidden layer representation of the character by encoding. The hidden layer representation consists of the output of the forward LSTM and the output of the backward LSTM. The forward sequence of the forward LSTM received character vector sequence is input, \( c = (c_0, c_1, c_2, ..., c_n) \), and the forward representation sequence is obtained by forward coding \((\vec{H}_t, \vec{H}_{t-1}, ..., \vec{H}_0)\). \( \vec{H}_t \) denotes the forward representation of the \( t \)-th character, and the formula is as follows:

\[
\vec{H}_t = \text{LSTM}([c_0, c_1, c_2, ..., c_n])
\]

Where \( \text{LSTM}(\cdot) \) represents the calculation of the LSTM model. The backward LSTM receives the reverse sequence of the sequence of character vectors as input, \( c = (c_n, c_{n-1}, ..., c_1, c_0) \), and obtains the backward representation sequence \( (\vec{H}_t, \vec{H}_{t-1}, ..., \vec{H}_0) \). \( \vec{H}_t \) denotes the backward representation of the \( t \)-th character, and the formula is as follows:

\[
\vec{H}_t = \text{LSTM}([c_n, c_{n-1}, ..., c_1, c_0])
\]

Where \( \text{LSTM}(\cdot) \) represents the calculation of the LSTM model. The forward and backward directions are combined to obtain the output \( H_t \) of the BiLSTM layer, where \( H_t = [\vec{H}_t \cdot \vec{H}_t] \).

4.3. Conditional Random Fields Model Layer
The CRF layer assigns a marker to each character, and calculates the score of the entire sequence. The sequence score consists of the character marker score and the marker transition score, and the label sequence with the highest score is selected as the final prediction result. The hidden layer output of BiLSTM, \( H = (\vec{H}_1, \vec{H}_2, ..., \vec{H}_n) \), is the input to the CRF, where \( N \) is the length of the input sequence and \( H_i \) is the \( i \)-th word of the input vector. Then \( Y = (y_0, y_1, y_2, ..., y_n) \) is the output tag sequence corresponding to \( Z \). For a given input sequence \( Z \) with a value of \( z \), the conditional probability of taking \( y \) on the label sequence \( Y \) is \( p(y|z) \). The specific formula is as follows:

\[
p(y|H) = \frac{1}{S(H)} \exp(\sum_{t,k} \lambda_k t_k(y_{n-1}, y_n, H, n) + \sum_{n,t} u_t s_t(y_n, H, n))
\]

\[
S(H) = \sum_y \exp(\sum_{n,k} \lambda_k t_k(y_{n-1}, y_n, H, n) + \sum_{n,t} u_t s_t(y_n, H, n))
\]

Where \( n=1,2,...,N, t_k(\cdot) \) and \( s_t(\cdot) \) are eigenfunctions, \( \lambda_k \) and \( u_t \) are the weights corresponding to \( t_k(\cdot) \) and \( s_t(\cdot) \), respectively, and \( S(z) \) is the normalization factor. In the training phase, the maximum likelihood estimation is used to calculate the optimal labeling sequence. The likelihood logarithm of the training set is \( \sum_n \log p(y|z) \), and the \( y \) is finally obtained by training to obtain the highest conditional probability. The input sequence is annotated to get the best tag sequence for the sentence and the formula is as follows:

\[
y = \arg\max_{y \in Y} p(y|H)
\]

5. Experiment

5.1. Experimental Data
The experimental data comes from 12,000 articles of clothing products obtained by a clothing store in Taobao through crawling technology. Among them, 8400 pieces of comment data are randomly selected as the training set, and 3600 pieces of comment data are randomly selected as the verification set. The sequence is marked for the task, and each character in all the comment sentences has a corresponding label correspondence. This experiment uses the "BIEO" labeling system to identify whether each word in the sentence is a component of the named entity and distinguish the entity boundary. Where "B" represents the first word of the entity, "I" represents the other part of the entity, "E" represents the last word of the entity, and "O" represents the word that is not an entity. The main identification task of this experiment is to identify the named entities in the clothing ontology knowledge base in the clothing review, such as style, layout, color and other named entities.

5.2. Evaluation Methods
In this experiment, the precision (P), recall rate (R) and F value (F) were used to measure the model. Among them, the accuracy rate refers to the ratio of the number of correctly identified entities to the total number of identified entities. The recall rate refers to the ratio of the number of correctly identified entities to the total number of entities, and the F value is the harmonic mean of the accuracy and recall rate. The specific formulas of each indicator are as follows:

\[
P = \frac{n}{M} \times 100\% \quad (7)
\]

\[
R = \frac{n}{N} \times 100\% \quad (8)
\]

\[
F = \frac{2PR}{P+R} \times 100\% \quad (9)
\]

Where M represents the number of entities identified, N represents the total number of entities in the test set, and the number of entities correctly identified in n generations.

5.3. Model Comparison Analysis

In order to have a more comprehensive evaluation of the performance of the model in this chapter, this experiment designed a number of models for comparative analysis. The experimental results are shown in Table 1.

| number | model name     | precision | recall | F value |
|--------|----------------|-----------|--------|---------|
| 1      | CRF            | 75.68     | 79.83  | 77.70   |
| 2      | BiLSTM         | 83.36     | 86.68  | 84.99   |
| 3      | BiLSTM-CRF     | 88.34     | 91.26  | 89.78   |
| 4      | CNN-BiLSTM     | 85.74     | 89.52  | 87.59   |
| 5      | CNN-BiLSTM-CRF | 88.95     | 93.56  | 91.20   |

It can be seen from the experimental results in Table 1 that although the single models of CRF and BiLSTM can complete the entity recognition tasks to a certain extent, on the whole, these models are inferior to other combined models, indicating that the single model has been increasingly, it is not possible to meet the needs of actual named entity recognition tasks. By comparing the model 4 and the model 5, it can be seen that the effect of introducing CRF on the basis of BiLSTM is slightly better than that of introducing the convolutional neural network, which explains in a certain sense that the introduction of CRF has a recognition effect on the named entity. Promotion has a significant effect. Finally, Model 5 is 1.42% higher than the F value of Model 3, which indicates that feature extraction using convolutional neural networks can better represent text and improve model performance. In summary, CNN-BiLSTM-CRF can achieve better experimental results than the general deep learning model.

6. Discussion

In this paper, the task of identifying the entity in the clothing domain knowledge graph, the character vector is obtained by convolutional neural network to supplement the word vector, and then the CNN-BiLSTM-CRF neural network model method is constructed. The experimental results show that the model achieves the best results. The main discussions are as follows:

First of all, in the knowledge graph of the clothing field, the knowledge of the artificial feature domain has a great influence on the result of the entity extraction, which leads to the increase of the cost of the system and the decline of the generalization ability. Therefore, the constructed CNN-BiLSTM-CRF deep neural network model does not use any In the case of artificial features, it is better to obtain machine learning methods than using a large number of rich features and domain knowledge. Secondly, this paper uses CNN network to obtain character vectors representing word morphological features to supplement the deficiencies of word vectors. The model can identify the special characters more effectively, which improves the performance of the model. Finally, in order to obtain a more accurate recognition result, the output of the CNN-BiLSTM network is decoded by CRF to obtain an optimal marker sequence. The integration of CRF improves the recognition performance of clothing
named entities with multiple modifiers and blurred boundaries. Therefore, in the process of entity extraction of clothing knowledge map, the CNN-BiLSTM-CRF fusion model proposed in this paper is an effective way to effectively improve the recognition performance.

7. Conclusion
This paper defines and constructs the conceptual layer of the knowledge graph of the clothing field, and studies an entity automatic extraction method based on the knowledge graph of the clothing field, aiming to build a relatively high quality knowledge base for the clothing field. Firstly, the concept layer is constructed in a top-down manner, and the whole clothing knowledge graph framework is constructed. In the process of construction, named entity recognition and extraction are performed from unstructured clothing data. And focus on how to automatically extract garment entities from unstructured text data through deep learning models. Experiments show that the model effect is better than other models, and the F value reaches 91.20%.

Building clothing knowledge graph is a complex and challenging task. There are many more areas where the clothing knowledge graph needs improvement. For example, the knowledge of the clothing field is not rich enough, and it is also necessary to seek a multi-source heterogeneous data source to expand the knowledge base. Knowledge fusion, including entity alignment and entity linking, automatically updates the established clothing knowledge graph. In general, this paper studies the clothing knowledge graph based on unstructured entity extraction, which fills the lack of knowledge graph in the domestic clothing field. For the intelligent question and answer in the field of clothing, accurate search has laid the foundation and is of great significance.

References
[1] David Amerland’s. Google Semantic Search[M]. 2013.
[2] Rao Z Y, Zhang Y X. Research on Intelligent Customer Service System Based on Knowledge Map[J]. Electric Power Information & Communication Technology, 2017.
[3] Zhang J, Pourreza M, Ramachandran R, et al. Facilitating Data-Centric Recommendation in Knowledge Graph[C]// 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC). 2018.
[4] Zeng-Lin X U , Yong-Pan S , Li-Rong H E , et al. Review on Knowledge Graph Techniques[J]. Journal of University of Electronic Science & Technology of China, 2016.
[5] Guo J, Zhao J, Zeng S. Structural reliability analysis based on analytical maximum entropy method using polynomial chaos expansion[J]. Structural & Multidisciplinary Optimization, 2018(1):1-17.
[6] Leos-Barajas V, Photopoulou T, Langrock R, et al. Analysis of animal accelerometer data using hidden Markov models [J]. Methods in Ecology and Evolution, 2017, 8 (2): 161-173.
[7] Teichmann M T T, Cipolla R. Convolutional CRFs for Semantic Segmentation[J]. 2018.
[8] Lecun Y, Bengio Y, Hinton G. Deep learning.[J]. Nature, 2015, 521(7553):436.
[9] Chen Xinchi, Qiu Xipeng, Zhu Chenxi, et al. Long short-term memory neural networks for Chinese word segmentation[J]. Conference on Empirical Methods in Natural Language Processing, 2015, 295(1):1197-1206.
[10] Dong C, Zhang J, Zong C, et al. Character Based LSTM- CRF with Radical- Level features for chinese named entity recognition[C]. International Conference on Computer Processing of Oriental Languages. Springer International Publishing. 2016:239—250.
[11] Rodriguezgarcia M Á, Hoehndorf R. Inferring ontology graph structures using OWL reasoning[J]. Bmc Bioinformatics, 2018, 19(1):7.