A Military Named Entity Recognition Method based on pre-training language model and BiLSTM-CRF

Yiwei Lu, Ruopeng Yang, Xuping Jiang, Changsheng Yin, Xiaoyu Song

College of Info. Communications National Univ. of Defense Tech, Wuhan, China
yiweilu_89@163.com, yiweilu89@nudt.edu.cn

Abstract. Military named entity recognition is the basis of the military intelligence analysis and operational information service. In order to solve the problems of inaccurate word segmentation, diverse forms and the lack of corpus in military texts, the author proposes a method of military named entity recognition based on Pre-training language model. On this basis, and taking advantage of Bi-directional Long Short-Term Memory (BiLSTM) neural network in dealing with the wide range of contextual information, the BERT-BiLSTM-CRF named entity recognition model was constructed. The experimental results on the tagged military text corpus show that the extraction effect of this method is better than that of the traditional methods.

1. Introduction

In recent years, with the continuous development of natural language processing technology, the research on military named entities has gradually become a hot spot. In the process of identifying military named entities, there are three main difficulties:

1) There are a large number of abbreviations, combinations, nesting and other forms of military named entities. As individual language styles and habits are different, there is no fixed mode for the language expression of military texts. Even the military has only standardized some contents, so it is difficult to construct comprehensive and reasonable entity characteristics;

2) The existing word segmentation tool is mainly used in general, the military in the field of word segmentation accuracy is not high, especially professional military terminology in the field of general rare, even join the army language dictionary is also difficult to include all military entity, as a result, this method of participle dependence is stronger, recognition effect is difficult to break through the bottleneck at present.

3) The available corpus is small in scale. At present, there is no open military domain corpus, which makes it difficult to identify military named entities with data.

In the military field, a common method is to extract the entity of meaning from the military text based on rules and dictionaries. A conditional random field (CRF) [1] model can also be used to learn text features and identify named entities in the military field. A combination of models (CRF with rules, CRF with dictionaries and rules [2]) can also be used to identify military named entities. The traditional method has some shortcomings in recognition effect and extensibility. With the development of character vector, Feng [3] proposed a method based on BiLSTM, manual extraction can not only avoid the characteristics of high cost and high complexity of military named entity recognition, but also effectively reduce the dependence on word segmentation, greatly improve the effect of military named entity recognition and reduce the cost of system development, but its shortcoming is that this method has always been difficult to extract and use human features.
Therefore, in this paper, based on Bi-directional Long Short-Term Memory (BiLSTM) model, the pre-training network is used to train the character vector, which can better understand the sentence structure of military text, extract the features of military text, and finally improve the performance of military named entity recognition. Using the BERT-BiLSTM-CRF method, the named entity recognition efficiency is better than the traditional rule-based and deep learning method through the established military Corpus linguistics. Further analysis shows that the proposed method can identify more entities that do not appear in the training set.

2. Military Named Entity Recognition Model Based on BERT-BiLSTM-CRF

The BERT-BiLSTM-CRF model proposed in this paper uses the pre-training language model(Bert) to train the military text character vector, which preserves the semantic information of the military text more completely, and improves the two-way context feature extraction ability of the model, compared with the traditional named entity recognition model, it makes full use of the semantic information and improves the recognition rate of the model to military named entity, as shown in Fig.1.

![Model structure of military named entity recognition based on BERT-BiLSTM-CRF](image)

**2.1 Character vector representation layer based on Bert**

The first layer of the algorithm is the representation layer of the character vector, the representation of the character vector is the core of the military named entity recognition, the representation method of the character vector is more, the main generation models of the character vector and character vector are Word2Vec and GloVe[4] etc., these language models are usually static and cannot express the character features of the military text context effectively. Therefore, this paper uses the Bert pre-training language model to construct the character vector representation layer.

![The language model based on Bert](image)
2.2. BiLSTM Layer

BiLSTM layer is the core of the whole algorithm, and its main function is to extract semantic features of the text layer by layer. The basic idea is taking into account the information that each input character vector propagates backwards and forwards. During the training, the forward LSTM sequence trains an LSTM network that propagates information from front to back, and the backward LSTM sequence trains an LSTM network that propagates information from back to front. The hidden layer of the two LSTM networks is taken as the feature representation based on LSTM, and the feature representation of the two LSTM bases at the same time is combined as the input of the CRF at the current time. Compared with LSTM, BiLSTM must finish the bidirectional LSTM sequence training separately to take the result of the LSTM sequence as the input of the next layer, while LSTM takes the previous output of LSTM as the input of CRF during the training process, which greatly extends the training time. However, in terms of the model effect, BiLSTM takes into account the information propagated from the front to the back and the information propagated from the back to the front. The information BiLSTM considers is more comprehensive. It learns the features in front of the current text and behind the current text.

The LSTM element in this algorithm is composed of four parts: one is the input gate, the weight matrix corresponding to the input gate is, and the bias; The other is the forgetting door. The weight matrix corresponding to the forgetting door is as follows; Three is the output door, the output door corresponding to the weight matrix, and bias. The above gate takes the current input, the state generated in the previous step, and the current state of the current unit as the input, and by calculation produces some intermediate state, which is used to determine which of the input information should be adopted, which of the stored historical information should be forgotten, and which of the output states generated by the unit. The detailed expression is as follows:

\[
\begin{align*}
    i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
    f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
    g_t &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + W_{cc}c_{t-1} + b_c) \\
    c_t &= i_t g_t + f_t c_{t-1} \\
    o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\
    h_t &= o_t \tanh(c_t)
\end{align*}
\]

Therefore, the state of the current cell will be weighted by the state of the previous cell and the current cell information. Since the single-layer LSTM processes the sequence in a chronological order, the semantic information provided below cannot be effectively utilized. Therefore, the model adopts the two-layer LSTM, namely BiLSTM, as shown in Fig.1. Among them, the horizontal connections of neurons in the two-layer network are in opposite directions: LSTM in the first layer is the forward channel, and LSTM in the second layer is the backward channel. With the above design, the model can use both the semantic information above and below to improve the recognition accuracy of entity relationship. Thus, the output produced by the \(i\) word is shown in the following formula:

\[
h_t = [h_t \oplus h_t^\dagger]
\]

In the above formula, ‘\(\oplus\)’ is used to represent the combination of forward channel output and reverse channel output.

2.3. CRF Layer

The CRF layer mainly combines the feature sequence output from the BiLSTM feature extraction layer with the final given labelling sequence training output sequence and the conditional probability model of the input sequence. Condition Random Field (CRF) [5] model is a conditional probability model used to label and segment ordered data. This model combines the advantages of hidden markov model and maximum entropy model, avoids some disadvantages of these models, and can effectively solve the problem of sequence annotation and transform entity recognition task into a sequence annotation task. In the judgment, given the sequence of output vectors of the bidirectional LSTM hidden layer, the probability vector of each word belonging to each entity label is obtained.

For a given military text to sentence as the basic unit, the input sequence of text segmentation, a sentence containing \(n\) characters, using to represent character sequences corresponding to the input,
assuming that $P$ is the size of BiLSTM layer score matrix, the $k$ is the number of different labels, corresponding to score of the $i$ word of the $j$ label. For a label prediction, define its score as:

$$S(x, y) = \sum_{i=1}^{N} P_{y_i} + \sum_{i=1}^{N} A_{y_{i-1}y_{i}}$$  \hspace{1cm} (8)

In the above formula, $A$ is the transfer fraction matrix, representing the fraction transferred from label $i$ to label $j$, and $y_1$ and $y_N$ are the tags added for the position at the beginning and end of the sentence. Therefore, $A$ is a square matrix of size $k + 2$. For sequence $Y$, this paper uses softmax to generate all:

$$p(y \mid X) = \frac{e^{s(X, y)}}{\sum_{y \in Y_X} e^{s(X, y)}}$$ \hspace{1cm} (9)

In the above formula, $Y_X$ is all possible tag sequences for input sentence $X$. It can be clearly seen from the above formula that when BiLSTM neural network produces effective output tag sequences for final decoding, the output sequence with the maximum score is predicted by the following formula:

$$y^* = \arg\max_{y \in Y_X} s(X, y)$$ \hspace{1cm} (10)

3. Experiment and analysis

3.1. Construction of experimental corpus

The existing military text corpus is the source of the experimental data set of military entity recognition. Firstly, we need to pre-process the existing military text corpus and label the processed text data to generate the data set for model training and testing. In view of the characteristics of more entity terms and less ambiguity in military texts, a simple and efficient bio annotation method is used to label 14 types of military entities in sequence. In the process of sequence annotation, manual annotation and automatic annotation are combined. First, the original text data is pre-processed by manual means, and then the Open Source Toolkit Yedda [6] is used for automatic annotation.

The bio sequence annotation method is to label each element in the dataset as "B-X", "I-X" or "O". Where "B-X" indicates that the fragment of this element is of type X and this element begins this text fragment, "I-X" indicates that the fragment of this element belongs to type X and this element is in the middle or end of this fragment, "O" indicates that it does not belong to any type. The specific marking method is shown in Tab.1:

| Entity type  | Begin       | Inside and End | Entity type  | Begin       | Inside and End |
|--------------|-------------|----------------|--------------|-------------|----------------|
| Force        | B-FORCE    | I-FORCE        | Rank         | B-RANK     | I-RANK         |
| Organization | B-ORG      | I-ORG          | Position     | B-POS      | I-POS          |
| Facilities   | B-FACI     | I-FACI         | Direction    | B-DIREC    | I-DIREC        |
| Location     | B-LOC      | I-LOC          | Active       | B-ACT      | I-ACT          |
| Weapons      | B-WEAP     | I-WEAP         | Pre-time     | B-PRET     | I-PRET         |
| Environment  | B-ENVI     | I-ENVI         | Out-time     | B-OUTT     | I-OUTT         |
| Services     | B-SERVI    | I-SERVI        | Number       | B-NUM      | I-NUM          |

Specific annotation shall be made according to the two parts of determining entity type and entity boundary:

1) Determine the type of entity; For example, "the eastern theatre of the blue army" should be labelled as an organization in the sentence "offensive combat force of the eastern theatre of the blue army".

2) Determine the boundary of the entity; "Defense department of the blue army" should be labelled as one institutional entity, not as two entities of the "blue army" and "Defense department"; "Early June" should be labelled as a general time entity, not as "early June" and "early June".

The structure of the experimental data set is shown in Tab.2. Restricted by artificial conditions, the overall size of the data set is not large, with a total of 3,516 marked sentences and 10,918 entities.
Tab.2 Military NER Dataset

| Item                  | Training set | Develop set | Test set |
|-----------------------|--------------|-------------|----------|
| Num of sentences      | 2,462        | 527         | 527      |
| Num of characters     | 124,681      | 27,254      | 26,360   |
| The marked Number     | 7,643        | 1,638       | 1,637    |
| Proportion            | 6.13%        | 6.01%       | 6.21%    |

The experiment uses the accuracy P, the recall rate R and the F value to evaluate the results, where the F value can reflect the overall test results. The calculation formulas of the three evaluation indicators are as follows:

\[
P = \frac{\text{Number of entities correctly identified}}{\text{Number of entities identified}} \times 100\% \quad (11)
\]

\[
R = \frac{\text{Number of entities correctly identified}}{\text{Total number of sample entities}} \times 100\% \quad (12)
\]

\[
F = \frac{2 \times P \times R}{P + R} \times 100\% \quad (13)
\]

3.2. Parameter settings

The model was evaluated by using the official evaluation index in evaluation task 8 of semeval-2010, which was based on the macro average F1 scores of 14 entities. When training the model, the learning rate is set to 0.01, and the mini batch size is set to 10. By default, the Bert pre training language model uses a transformer with 12 attention mechanisms, and the vector length of the pre training words is 768 dimensions; the length of a single military text is mostly 160-300 words, so the sequence length of each read is 256, and the size of each batch is 64. The dropout rate of the BiLSTM layer and CRF layer is set to 0.5 and 0.6 respectively, and other parameters of the model are initialized randomly.

3.3. Experimental results and analysis

(1) Comparison of different model effects

In the experiment, three methods are used for comparison and verification, including method 1 based on CRF++ model[7], method 2 based on word vector matrix initialized by conventional dictionary, feature extraction using BiLSTM neural network[8], using CRF as classifier, method 3 based on word vector matrix generated by pre training speech model based on Bert, and other methods are the same as method 2. The experiment is carried out in the same corpus and entity classification, and three methods are compared. The test results of the three methods are shown in Tab.3.

Tab.3 Experimental results

| Method                | Evaluation indicators |
|-----------------------|-----------------------|
|                       | P   | R   | F1   |
| CRF++                 | 76.1| 75.2| 75.6 |
| BiLSTM-CRF            | 87.2| 88.3| 87.7 |
| BERT-BiLSTM-CRF       | 92.2| 91.3| 91.7 |

It can be seen from table 3 that method 2 has better recognition effect than method 1, indicating that using BiLSTM as feature extractor can more effectively classify the results of military named entity recognition; method 3 has better recognition effect than method 2, indicating that using pre training language model based on Bert can more effectively guarantee the feature extraction of military named entity in the process of entity recognition Take. This is because the manual selection of military text data features has certain limitations. On the one hand, it is difficult to exhaust all features. On the other hand, there may be overlaps or contradictions between features. The recognition model based on BERT-BiLSTM-CRF solves the above problems by automatically learning text features layer by layer.

(2) Recognition results of different types of military named entities

Based on the BERT-BiLSTM-CRF model, 14 Military named entities are identified, shown in Tab.4.
Tab.4 Recognition rate of different named entities

| Entity type | P   | R   | F1  | Entity type | P   | R   | F1  |
|-------------|-----|-----|-----|-------------|-----|-----|-----|
| Force       | 93.2| 92.5| 92.8| Rank        | 77.6| 78.4| 78.0|
| Organization| 91.8| 90.9| 91.3| Position    | 75.1| 76.2| 75.6|
| Facilities  | 86.5| 87.2| 86.8| Direction   | 79.1| 79.2| 79.1|
| Location    | 94.6| 95.1| 94.8| Active      | 85.9| 86.3| 86.1|
| Weapons     | 89.1| 88.9| 89.0| Pre-time    | 88.2| 87.9| 88.0|
| Environment | 78.1| 80.3| 79.2| Out-time    | 87.3| 86.9| 87.1|
| Services    | 76.1| 77.2| 76.6| Number      | 92.1| 91.7| 91.9|

The model proposed in this paper has a high recognition rate for four military named entities: Force, Organization, Location and Number, but the recognition rate of Environment, Services, and Position is low. This is mainly because these entities are described in too many ways, and it is difficult to find a general rule expression. For example, the method of finding a general rule is not ideal for environment, including temperature, humidity, weather and other types. These problems can be further subdivided by each entity category solve.

(3) Comparison experiment of different training epoch

With the increase of training rounds, the F1 values of the three methods change as shown in Fig. 3. during the 1st to 10th epoch training, the F1 values have changed significantly, and after the 20th epoch, the F1 values basically tend to be stable, which shows that the convergence of the three methods is relatively consistent, and it also shows that setting the epoch to 20~30 during training can achieve better training results.

![Fig.3 Distribution of F1 under different epoch](image)

4. Conclusion

Aiming at the problem that the traditional methods cannot fully represent the semantic features of military texts, this paper proposes a military named entity recognition method based on BERT-BiLSTM-CRF. The experimental results of this method on the constructed military NER dataset show that the method is effective and reaches a comprehensive evaluation value of 91.7%. In the future, the military NER corpus will be expanded to integrate more external resources for training, so as to further improve the efficiency of military named entity recognition.

REFERENCES

[1] Li L S, Zhou R P, Huang D G. Two-phase biomedical named entity recognition using CRFs[J]. Computational Biology 8. Chemistry, 2009, 33(4) ;334-338.

[2] Zheng X Q, Chen H Y, Xu T Y. Deep learning for Chinese word segmentation and POS tagging[C] // Proc of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 647-657.
[3] Soutner D. Continuous distributed representations of words as input of LSTM network language model [C] / /Proc of International Conference on Text, Speech, and Dialogue, 2014: 150-157.

[4] Pennington, SochilR, Manning C D. Gloves: Global vectors for word presentationz [C]// proceedings of 2014 Conference on Empirical Methods of Natural Sciences Language processing (EMNLP). Doha, Qatar: AssociationComparative Linguistics, 2014:1532-1543.

[5] PETERS M E, AMMAR W, BHAGAVATULA C, et al. Semisupervised sequence tagging with bidirectional language models[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2017: 1756-1765.

[6] GRAVES A, JAITLY N, MOHAMED A R. Hybrid speech recognition with deep bidirectional LSTM[C]// Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Piscataway: IEEE, 2013: 273-278.

[7] H. Y. Shan, H. S. Zhang, et al. A Method of Military Named Entity Recognition Based on CRFs under Small Granularity Strategy. Journal of Armored Force Engineering Institute[J]. Beijing. 2017.2

[8] Mcfallum A, Li Wei. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C] / / Proc of NAACL-HLT 2003,2003: 188-191.