Research on Chinese Event Extraction Method Based on HMM and Multi-stage Method

Liying Zhan¹*, Xuping Jiang¹ and Qian Liu¹
¹College of Information and Communication, National University of Defense Technology, Wuhan, China

*Corresponding author e-mail: 1241761652@qq.com

Abstract. The existing Chinese event extraction methods are mostly one-stage methods. Even if there are two-stage methods, the correlation between the stages is not high, the extracted event trigger words and event elements have low matching degree, and there is a class imbalance problem in the training process. This paper proposes a Chinese event extraction method based on HMM and multi-stage method. The first stage of the method recognises the event trigger word positive examples s in the text; the second stage classifies the trigger word positive examples recognised in the first stage, determines the event type, and forms the sequence of event trigger words; the third stage matches the event elements, and forms the sequence of event elements, according to the first two stages of event trigger word extraction results. This method effectively alleviates the class imbalance problem in training, improves the matching degree between event trigger words and event elements extraction, and obtains better extraction performance. The accuracy, recall rate and F value of Chinese event extraction are all More than 78%.

1. Introduction

Chinese event extraction, mainly researches on extracting event information of interest from Chinese texts, and present them in a structured form for use in other information extraction services [1]. This paper proposes a Chinese event extraction method based on HMM and multi-stage method for Chinese event extraction, which improves the accuracy and recall rate of event extraction.

This paper chooses Hidden Markov Model, which is a machine learning model. It has been widely used in event extraction because of its simplicity, easy establishment and strong adaptability [2]. The model is relatively fast in training and recognition. The Viterbi algorithm is more efficient in solving sequence annotation, and the model is more suitable for some applications that require real-time and need to process large amounts of text [3].

In this paper, the method of event extraction is improved, and an event extraction method based on multi-stage method is proposed. It alleviates the class imbalance problem in the training process, improves the matching degree between event trigger words and event elements extraction, and effectively improves the accuracy of the event extraction.

2. Chinese event extraction method

2.1. Chinese event extraction process design

Event extraction is the process of extracting events in the text, mainly includes three parts: pre-processing, model training, and information extraction. Among them, the pre-processing includes
Chinese clauses, word segmentation and corpus annotation, and corpus annotation forms three types of data sets. Model training and information extraction process are shown in Fig. 1.

2.2. Machine Learning Model and Algorithm

In the model training part, the HMM machine learning model is encoded by the event extraction training set, and the trigger word extraction parameter $\lambda_1 = (N, M, A, B, \pi)$ and the event element extraction parameter $\lambda_2 = (N, M, A, B, \pi)$ are obtained. Firstly, the event trigger word recognition training data set and event element recognition formed by the pre-processing, is digitally converted, and then the machine learning model is trained.

The machine learning training process based on the HMM model can be described as using the existing training data set to separately train the trigger word recognition parameter $\lambda_2$ and the event element recognition parameter $\lambda_2$ in the HMM model. The parameter $\lambda$ can be expressed as: $\lambda = (N, M, A, B, \pi)$ (See the reference [4] for details of parameter values). In this paper, the maximum likelihood estimation method (MLE) is used to estimate the value of the HMM model parameter $\lambda$. (See the reference [5] for details).

2.3. Information extraction model and algorithm

2.3.1. Multi-stage method. In the first stage of the information extraction part, the event extraction test set is input, and the event trigger word positive examples is recognised in base of the event trigger word extraction parameter $\lambda_1$.

In the second stage, in base of the event trigger word extraction parameter $\lambda_1$, the trigger word positive examples recognised in the first stage are classified, the event type is determined, and the event extraction test set is decoded by the Viterbi algorithm to obtain the "optimal" sequence of event trigger word.
In the third stage, the event trigger word extraction sequence is used to determine the trigger word label of each sentence line by line, according to the judgment conclusion, the type event element extraction parameter \( \lambda_2 \) is called, and the information extraction test set is decoded by the Viterbi algorithm to obtain the “optimal” event element extraction sequence.

The information extraction part is composed of two parts: event trigger word recognition module and event element recognition module. The event trigger word extraction module: in base of the event trigger word extraction parameter \( \lambda_1 \), the event recognition test data set consisting of the word sequence, the event trigger word labelling sequence, and the event element labelling sequence is decoded by the Viterbi algorithm to generate the event trigger word recognition result including the word sequence, the event trigger word labelling sequence, and event trigger word recognition sequence, and calculate \( P, R, F \) values.

Event element extraction module: in base of the event element extraction parameter \( \lambda_2 \), the event trigger word recognition result including the word sequence, the event trigger word labelling sequence, the event element labelling sequence, and the event trigger word recognition sequence is decoded by the Viterbi algorithm to generate event element recognition result including the word sequence, event trigger word labelling sequence, event element labelling sequence, event trigger word recognition sequence, event element recognition sequence, and calculate \( P, R, F \) values.

2.3.2. Viterbi algorithm. The Viterbi algorithm is a dynamic programming algorithm that uses the idea of dynamic programming to solve the optimal Viterbi path [6]. Let the Viterbi algorithm variable be \( \delta_t(i) \) and the memory variable be \( \varphi_t(i) \). The specific process of the Viterbi algorithm:

**Step1. initialization:**

\[
\delta_t(i) = \pi_i b_i(q_1), \quad i = 1,2,\cdots N \\
\varphi_t(i) = 0, \quad i = 1,2,\cdots N
\]

**Step2. recursive calculation:**

\[
\delta_t(i) = \max_{1 \leq j \leq N} \{ \delta_{t-1}(j) a_{ij} \} b_i(q_t), \quad i = 1,2,\cdots N \\
\varphi_t(i) = \arg \max_{1 \leq j \leq N} \{ \delta_{t-1}(j) a_{ij} \}, \quad i = 1,2,\cdots N
\]

**Step3. end:**

\[
P = \max_{1 \leq i \leq N} \delta_T(i) \\
i_T = \arg \max_{1 \leq i \leq N} \{ \delta_T(i) \}
\]

**Step4. path backtracking:**

\[
i_t = \varphi_{t+1}[i_{t+1}], \quad t = T - 1, T - 2,\cdots, 1
\]

**Step5. optimal path:**

\[
I = (i_1, i_2,\cdots, i_T)
\]

3. Experiment and analysis

3.1. Corpus and evaluation methods

In the research process, the manual annotation method is used to construct Chinese event extraction data set, including three types of data sets: event trigger word recognition training data set, event element recognition training data set, and event recognition test data set.

The experiment used \( P \) (accuracy rate), \( R \) (recall rate) and \( F \) value three evaluation indicators to evaluate the event extraction results.

\[
P = \frac{\text{Correctly recognized number of event trigger word(event element) belonging to Type}_i}{\text{recognized number of event trigger word(event element) belonging to Type}_i}
\]

\[
R = \frac{\text{Correctly recognized number of event trigger word(event element) belonging to Type}_i}{\text{The number of event trigger word(event element) that actually belong to Type}_i}
\]

\[
F = \frac{2 \times P \times R}{P + R}
\]
Among them, Type\textsubscript{i} is the event trigger word (event element) type to be evaluated, and i is a certain type event trigger word (event element) in the event.

3.2. Experimental results and analysis

In this experiment, HMM and multi-stage method are used to extract event trigger words and event elements, and to obtain the event trigger word and event element extraction accuracy, recall rate and F value, as shown in Table 1.

| Table 1. Chinese event trigger word and event element recognition experiment result |
|-----------------|---------|--------|---------|
|                 | P(%)    | R(%)   | F(%)    |
| Event trigger word | 87.03   | 78.65  | 82.63   |
| Event element     | 83.10   | 82.64  | 82.87   |

The experimental results show that the Chinese event extraction accuracy, recall rate and F value based on HMM and multi-stage method are more than 78%, and a better extraction effect is obtained. The multi-stage method has is greatly improved compared with the one-stage method, and more suitable for event extraction.

4. Conclusion

This paper proposes a Chinese event extraction method based on HMM and multi-stage method for Chinese event extraction task, and obtains better performance on self-built text corpus. The main conclusions are as follows:

In the Chinese event extraction task, the shallow machine learning method requires a large amount of corpus, which is highly dependent on the corpus and is prone to data sparseness. The self-built training corpus is prone to class imbalance problems, and the existing one-stage method has the problem that the event trigger word and the event element have low matching degree. Therefore, this paper proposes a Chinese event extraction method based on multi-stage method.

The method adopted in this paper is aimed at the characteristics of Chinese event description. Under the condition of limited text corpus, it can alleviate the class imbalance problem in the training process and improve the matching degree between event trigger words and event elements. Experiments show that the multi-stage method has better effect on Chinese event extraction, and it is feasible and effective.

References

[1] Xuyang Xu, Yongfeng Han, Wenzheng Song. Review and prospect of event extraction technology[J]. Journal of Information Engineering University, 2011, 12(01):113-118.

[2] Huixing Jiang, Xiaojie Wang, Jilei Tian. Second-order HMM for event extraction from short message [J]. Lecture Notes in Computer Science, 2010, 6177: 149-156

[3] Jiangde Yu, Xinfeng Xiao, Xiaozhong Fan. Chinese Text Event Information Extraction Based on Hidden Markov Model [J]. Microelectronics & Computer, 2007, 24(10):92-94+98.

[4] Zhe Shuang. Research and application of hidden Markov model in web page information extraction [D]. East China Normal University, 2016.

[5] Bjorne J. Biomedical event extraction with machine learning[J]. TUCS Dissertations, 2014 (178):1-121.

[6] Li K, Chen G, Cheng J. Research on Hidden Markov Model-based Text Categorization Process [J]. International Journal of Digital Content Technology & Its Applications, 2011, 5(6):244-251.