Design and Implementation of Word Segmentation System for Ideological and Political Education Based on Unsupervised Learning

Peng Shao(✉) and Junjie Guo

Jiangxi Vocational College of Tourism and Commerce, Nanchang 330100, Jiangxi, China
tyu78945@126.com

Abstract. Chinese word segmentation is an important technology in Chinese natural language processing. The quality of Chinese word segmentation results will affect the effect of text processing. The relationship between the two is very close. In the field of ideological and political education, these words have the characteristics of fast birth, wide coverage and large vocabulary.

Keywords: Chinese word segmentation · ideological and political education · language model · unsupervised learning · natural language processing

1 Introduction

The components of China's spiritual civilization construction include ideological and political education, and ideological and political education is also an important factor in solving social contradictions and problems [1]. With the development and progress of the times, the level of ideological and political education has gradually improved, and there are more and more documents related to ideological and political education [2]. How to better obtain information on a large number of ideological and political education documents is the main task of the research. It is a very efficient solution to summarize and generalize a large number of documents through computer-related knowledge and relevant tools in the field of natural language processing.

2 System Structure Design

According to the analysis of the project background and system requirements, the ideological and political education word segmentation system can be divided into four levels: presentation layer, business layer, support layer and data layer according to the structural characteristics [5].

The political education word segmentation system is divided into three functional modules: corpus training module, Chinese word segmentation module, and feature extraction module [6]. The main module can also be divided into several sub-modules according to different functions [7]. The overall function diagram of the system is shown in Fig. 1.

Z. Zhan et al. (Eds.): EIMSS 2022, AHCS 7, pp. 792–799, 2023.
https://doi.org/10.2991/978-94-6463-024-4_83

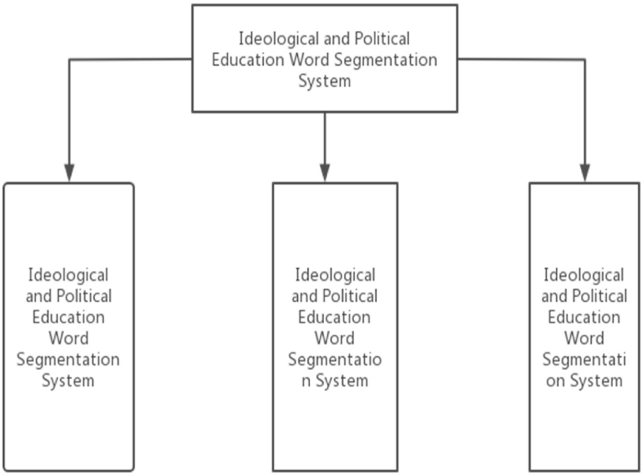
Ideological and Political Education Word Segmentation System
Ideological and Political Education Word Segmentation System
Ideological and Political Education Word Segmentatio n System
Ideological and Political Education Word Segmentati on System

Fig. 1. Overall function diagram of the system

3 Word Level N-Gram Language Model

Compared with the traditional word-level N-gram language model, the word-level N-gram language model has different training set requirements. The word -level N-gram training set is trained based on a large-scale annotated corpus [8]. The training set of the word-level N-gram language model has almost no requirements for labeling, but requires a large-scale corpus for model training, and the corpus can only be used after a standardized process [9]. The standardization process means that the text in the corpus needs to be separated by newlines and words with spaces, so as to ensure that the text in the corpus is in word units [10]. The word-level language model uses the word-level language model to predict the nth word according to the first (n−1) words. In a sequence M composed of n words, the word-level N-gram language model The calculation formula is:

$$p(wn|wn1, wn2, \ldots wn - 1) \tag{1}$$

where: w1, w2,…,wn−1 is the (n−1) word in the sentence, wn is the nth word, according to the conditional probability formula:

$$P(B|A) = p(A, B)P(A) \tag{2}$$

The segmentation path probability of sequence M can be obtained as:

$$P(w1w2 \ldots wn) = P(w1) * P(w2|w1) * \ldots * P(wn|w1w2 \ldots wn - 1) \tag{3}$$

Because the length of the sequence M is not fixed, when faced with a short Chinese text sequence, the corresponding result can be obtained quickly with the help of the above formula [11]. However, when the length of the sequence M is relatively long, the algorithm will have a large amount of computation. Therefore, the Markov hypothesis needs to be introduced. In layman's terms, the probability of word occurrence is only related to the first m words[J]1989. When m = 1, it is a 1-g language model. The probability of the split path of the sequence M can be corrected as:

$$P(w1w2 \ldots wn) = P(w1) * P(w2|w1) * \ldots * P(wn|wn - 1) \qquad (4)$$

where $P(wn|wn - 1)$ can be obtained by maximum likelihood estimation, which is equal to

$$\mathrm{Count}(wn, wn - 1)\mathrm{Count}(wn - 1) \qquad (5)$$

And so on, when m = 2, it is a 2-g language model, and the N-gram model includes 1-g, 2-g, 3-g, 4-g, 5-g and other language models. The above represents the longest length of the field sequence in the N-gram. The high-order N-gram language model contains the low-order N-gram language model. For example, a 4-g language model actually contains 1-g, 2-g, 3-g, 4-g language models [10]. When n is relatively small, the size of the corpus required to train the language model is also small, the time is also shorter, and

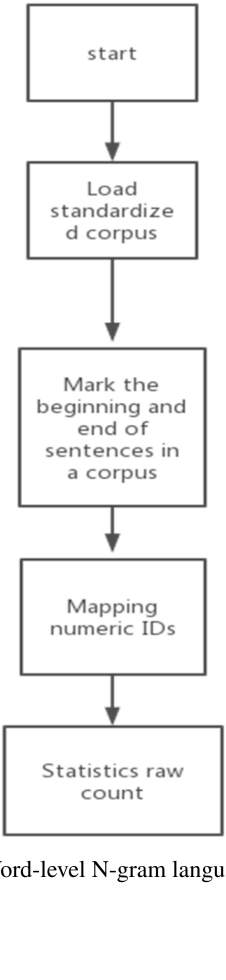
start
Load standardize d corpus
Mark the beginning and end of sentences in a corpus
Mapping numeric IDs
Statistics raw count

Fig. 2. Word-level N-gram language model

the constraint information is less, which will make the model's discriminative ability limited, especially the domain literature may not be applicable [13]. The word-level N-gram language model is shown in Fig. 2.

4 Calculate the Word Segmentation Path

For the text to be segmented with sequence length M, the path probability is

$$P = P(s1)P(s2)P(s3)\ldots P(sl) \tag{6}$$

s1, s2…sl are the words in the text to be segmented. In layman's terms, it is the result of segmentation according to the word combination in the language model, and the final word segmentation result is the optimal word segmentation path [14]. The path probability is the maximum of the path probabilities of all paths in the text sequence. The problem of Chinese word segmentation is transformed into a process of finding the maximum value of the path probability. Because the number of segmentation paths in the Chinese sequence increases exponentially with the increase in the length of the Chinese sequence, this is the reason why the algorithm faces a large amount of computation. Therefore, the Markov assumption is introduced in the step of calculating the word segmentation path, that is, the probability of the mth word of the Chinese sequence appearing is only related to the first (m−1) words. After research, it is found that the optimal word segmentation path is calculated using the Viterbi algorithm. As one of the widely used dynamic programming algorithms, the Viterbi algorithm is suitable for finding the -Viterbi path-hidden state sequence that generates the observed event sequence, and is used to find the optimal word segmentation path of the sequence [4]. Because there are multiple possibilities for the same Chinese text sequence The word segmentation result of, assumes that xij represents the jth possible value of state ci, and expands the state sequence. Due to the consideration of the model, the 4-g language model is finally used to solve the traditional word segmentation problem by marking the problem. The words in different positions in a single word are marked and calculated based on the words [3]. If b is a single-word word or the first character of a multi-word, c is the second character of a multi-word, d is the third character of a multi-word, and e is the rest of the multi-word, for a word ck in a sentence, as follows:

$$p(b) = P(ck) \tag{7}$$

$$p(c) = P(ck|ck - 1) \tag{8}$$

$$p(d) = P(ck|ck - 2ck - 1) \tag{9}$$

$$p(e) = P(ck|ck - 3ck - 2ck - 1) \tag{10}$$

The probability in the formula can be known with the help of the trained word-level language model: the only non-zero transition probabilities are P(b|b), P(c|b), P(b|c), P(d|c), P(b|d), P(e|d), P(b|e), P(e|e), with the help of the setting of the transition probability, the long words and short words in the final word segmentation result can be determined The frequency of occurrence, according to the transition probability, the optimal word segmentation path is obtained. In layman's terms, that is: the optimal word segmentation result.

5 Chinese Word Segmentation Optimization Method Research

With the help of the test of the original algorithm, we can know that the words with unsatisfactory word segmentation results generally have the characteristics of low frequency and high aggregation. In order to deal with the characteristics of the word segmentation results, this method uses the language model and the Viterbi algorithm to perform preliminary word segmentation on the text to be segmented and obtain the preliminary word

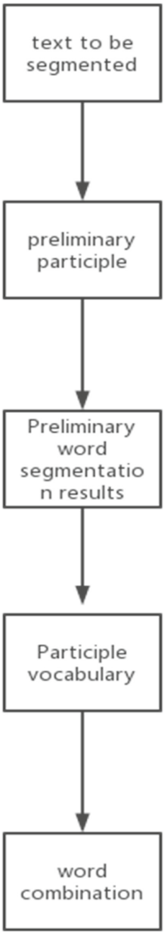

Fig. 3. Chinese word segmentation optimization algorithm flow chart

segmentation results, and then memorize the preliminary word segmentation results and the corpus in some corpora to calculate the corresponding word frequency deviation (Term Frequency Deviation, TFD.), and calculate the ranked term frequency deviation index (anked TermFrequency Deviation, rTFD) according to the word frequency deviation, and then the preliminary word segmentation results can be combined with word combination, which is similar to the word segmentation problem. In order to calculate the optimal word combination merging path, the Viterbi algorithm is used to solve the optimal word combination merging path, and the final word segmentation result is obtained. The algorithm flowchart is shown in Fig. 3.

6 The Role of the Word Segmentation System in Ideological and Political Education

The key to the effective integration of education and artificial intelligence technology is the effective integration of teachers and artificial intelligence technology. Facing the rapid development of artificial intelligence technology in the information age, ideological and political teachers in colleges and universities must quickly adapt to the requirements of this situation. Be good at learning and mastering relevant knowledge in the field of artificial intelligence and information technology, and constantly improve the ability to use new technologies to reflect and transform technological advantages into Their professional advantages and teaching ability make advanced information technology such as artificial intelligence an important means to improve their ideological and political education level. This situation puts forward higher requirements for the comprehensive quality of ideological and political teachers in colleges and universities. However, teachers of ideological and political courses have insufficient proficiency in artificial intelligence technology, which prevents the advantages of artificial intelligence technology from being fully and effectively displayed in the educational process. At present, there are relatively few artificial intelligence and information technology platforms for ideological and political education, which also limits the degree of integration of the two. The platform for ideological and political teachers to learn new knowledge in the field of artificial intelligence is not perfect. In addition, individual ideological and political teachers have insufficient personal active learning. Therefore, how to improve the ability of ideological and political teachers to use artificial intelligence technology and promote the organic integration of teachers and artificial intelligence technology to become a Suck the problem to be solved.

It is precisely because of the current development of the Internet that great changes have taken place in the ideological and political learning environment of students. In order to adapt to this change, schools and education departments can start from computer technology to find more suitable ideological and political education methods for students, and give full play to the power of computers advantage. The word segmentation system for ideological and political education proposed in this study can allow computers to intelligently analyze ideological and political factors in articles, help students learn, and give full play to the positive guiding ability of computers.

7 Conclusion

And test results of the ideological and political education word segmentation system are described in detail. At the beginning, the key functions of the system are analyzed, and the layout of the system interface is explained by taking the home page as an example. According to the different key functions, the following three modules are analyzed: corpus management module, Chinese word segmentation module, and feature extraction module. The basic principles and implementation effects are introduced according to the interface. The following three aspects of the system are tested: system security, system function and system performance, and the final statistical results are obtained and the test results are displayed. It also adds a lot of workload to word segmentation and work in this field. For these difficulties, this paper designs a word segmentation system for ideological and political education, which builds a corpus based on domain literature and trains a word-level language model based on statistical ideas. The preliminary Chinese word segmentation results are obtained with the help of the Viterbi algorithm, and finally the preliminary Chinese word segmentation results are optimized with the help of the Chinese word segmentation optimization algorithm based on word frequency deviation. The system then provides users with the following functions for word results: extracting keywords, word frequency statistics, drawing word cloud graphs, etc., and completing Chinese word segmentation and text analysis of field documents.

References

1. Huang C, Zhao H (2007) A decade-long review of Chinese word segmentation. Chinese J Inf 03:8–19
2. Wang J (2004) Research on some key technologies in Chinese information processing. Fudan University, Shanghai
3. Cheng Y, Shi Y (2018) Chinese word segmentation method for professional fields. Comput Eng Appl 54(17):30–34+109
4. Li D, Wang P, Zhang G (2020) Research on multi-model Chinese word segmentation method based on word cluster. Comput Appl Res 37(02):355–359+374
5. Liang N (1984) Automatic word segmentation and an automatic word segmentation system in written Chinese—CDWS. J Beijing Inst Aeronaut 04:97–104
6. Wang X, Wang K, Li Z et al (1989) The problem of least word segmentation and its solution. Sci Bull 13:1030–1032
7. Huang C (1996) Research and development of natural language processing. Japan Stud :10
8. Chang J, Shen W (2016) Research on Chinese word segmentation algorithm based on string matching. Ind Control Comput 29(02):115–116+119
9. Li H, Chen PH (2014) Improved backtracking-forward algorithm for maximum matching Chinese word segmentation. Appl Mech Mater 536–537:403–406
10. Wang T (2012) Research on Chinese word segmentation algorithm based on dictionary and its application in Nutch system. Jilin University, Changchun
11. Zhou Q (2015) Research and implementation of Chinese word segmentation based on the combination of statistics and dictionary. Harbin Institute of Technology, Harbin
12. Network hidden Markov model approach for automatic story segmentation. J Ambient Intell Hum Comput 8(6):925–936

13. Zhang X, Low YM, Koh CG (2020) Maximum entropy distribution with fractional moments for reliability analysis. Struct Saf 83
14. Zou J, Wen H, Wang T (2019) Research on Chinese word segmentation algorithm based on statistics. Comput Knowl Technol 15(04):149–150+153