Analysis on Chinese Word Segmentation method of two international Chinese Teaching platform

Huanxin Dou

Bohai University, Jinzhou 121000, China

E-mail: douhuanxin@163.com

Abstract. Chinese Word Segmentation is a basic task of Natural Language Processing."Guidelines for International Chinese language teaching" and "Construction-Use Integrated Teaching Platform", both have word segmentation functions.The matching method should be used in the former, while the statistical method is used in the latter, so the results of word segmentation for the same paragraph are not the same. We believe that the future development direction should be deep learning methods.

Keywords: CWS; teaching platform; matching method; statistical methods

1 Introduction

Chinese Word Segmentation (CWS) is a basic task of Natural Language Processing (NLP), the result of which has a profound impact on international Chinese education.[1]At present, two platforms in the field of international Chinese education, "Guidelines for International Chinese language teaching (ICLT)" and "Construction-Use Integrated Teaching Platform (CUITP)", both have word segmentation functions.

ICLT is also known as "international Chinese teaching material writing guide" (http://www.cltguides.com/). The platform in harmony writing type, hot material, practical resources and evaluation tool four plates. Among them, "evaluation tool" can realize word segmentation, marking parts of speech and grade.

The main features of the CUITP(http://140.210.192.221:8888/teach/.)include "lesson preparation" and "analysis" tools. The "analysis" tool can implement word segmentation, marking parts of speech and rank.

The boundary of Chinese characters and words is blurred and the grammatical structure is complicated, which affects the direct processing performance of computer. For example, the above two platforms for the same paragraph segmentation, the results are not the same. This difference should be due to the use of different Chinese word segmentation methods. Looking at the whole development process of Chinese word segmentation methods, it can be roughly divided into three categories: matching, statistics and deep learning. [2] Below, we introduce various methods and explore the word segmentation methods of the above two platforms.

Y. Kuang et al. (eds.), *Proceedings of the 2024 5th International Conference on Education, Knowledge and Information Management (ICEKIM 2024)*, Atlantis Highlights in Computer Sciences 22,
https://doi.org/10.2991/978-94-6463-502-7_47

2 Matching Method and GUIDELINES FOR ICLT

The matching method mainly divides text and dictionary by various algorithms.Matching algorithm and dictionary construction are the core of this method, which directly affect the efficiency and performance of word segmentation.[3] The processing module of the algorithm is shown in the figure 1 below:

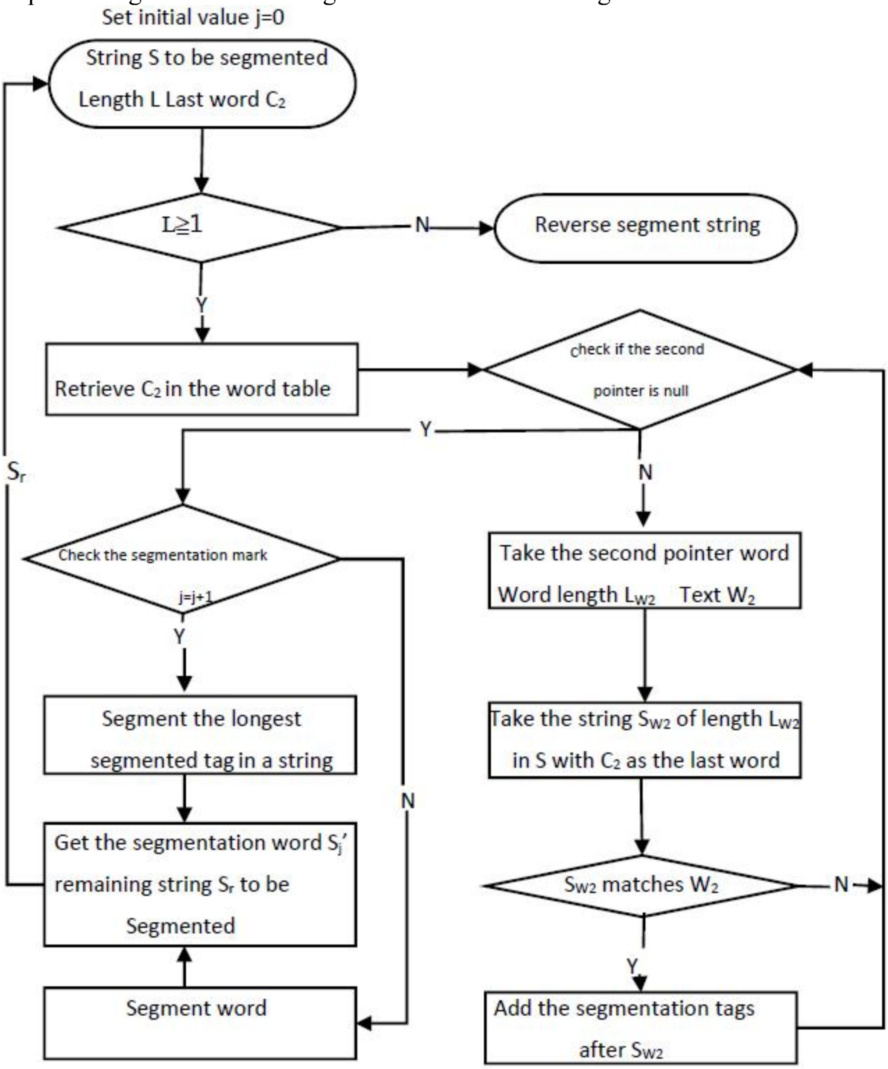
Set initial value j=0
String S to be segmented
Length L Last word C₂
L≧1
N
Reverse segment string
Y
Retrieve C₂ in the word table
check if the second
pointer is null
Y
N
Sr
Check the segmentation mark
j=j+1
Y
Take the second pointer word
Word length Lw2 Text W₂
Segment the longest
segmented tag in a string
Take the string Sw2 of length Lw2
in S with C₂ as the last word
N
Get the segmentation word Sj'
remaining string Sr to be
Segmented
Sw2 matches W₂
N
Y
Segment word
Add the segmentation tags
after Sw2

Fig. 1. Reverse maximum match word segmentation module

The execution steps of the reverse maximum matching word segmentation module are as follows:

(0) Set the initial value j for segmentation counts, making j = 0.

(1) Determine the length L of the string S to be segmented. If L is less than 1, the module segmentation ends; If L is greater than or equal to 1, skip to (2).

(2) Retrieve the last word C_2 of S in the word index table and skip to (3).

(3) Check whether the second pointer is null, if it is skip to (7); if it is not, skip to (4).

(4) To take the word in the pointer, get the word length Lw_2 and text W_2, skip to (5).

(5) If Lw_2 is greater than the length of L, skip to (7); if no, go to (6).

(6) Take out the character string Sw2 in S with C2 as the last word and Lw2 as the length, and match Sw2 with W2. If it is matched, the segmentation mark is added to, skip to (3); If it is unmatched, go to (3).

(7) Check if there is a segmentation mark in the string and make j = j+1. If so, the sentence is segmented according to the longest segmentation mark; if there is no segmentation mark, the word is segmented. Get the remaining string Sr to be segmented and the segmented word Sj ', skip to (1) with Sr as the new S.

The GUIDELINES FOR ICLT should adopt this algorithm, and the dictionary uses the existing "International Chinese Teaching general Curriculum Outline", "Syllables, Chinese characters, lexical classification for Chinese International Education" and "vocabulary hierarchies".Use this platform to analyze the paragraph "bu xing que yi xia zi luo dao ta tou shang lai(means'misfortune suddenly fell on her head')", the analysis results are as the Table 1:

Table 1. The GUIDELINES FOR ICLT's word statistics results

Word statistics details				
Serial number	Word	Frequency	"International Chinese Teaching general Curriculum Outline"	"Syllables, Chinese characters, lexical classification for Chinese International Education" "Vocabulary hierarchies"
1	buxing	1	beyond the syllabus	Level2
2	que	1	Level4	Level2
3	yixiazi	1	beyond the syllabus	Level1
4	luo	1	Level4	Level2
5	dao	1	Level2	Level1
6	ta	1	Level1	Level1
7	tou	1	Level5	Level1
8	shang	1	Level1	Level1
9	lai	1	Level1	Level1

The matching algorithm is convenient and efficient to use, but the performance of this method is highly dependent on the dictionary, and it cannot deal with the words that do not appear in the dictionary and many possible segmentation situations.[4]

3 Statistical Methods and CUITP

The basic idea of statistical method is to determine whether to divide into word boundaries according to the probability of word combination.This method can solve the ambiguity of word segmentation well to some extent, and can identify the unknown words. In current CWS research, common statistical methods include N-gram Model, Hidden Markov Model (HMM), Conditional Random Fields model (CRF)[5] , etc.

CRF is an undirected graph model, which considers the global distribution of data in normalization and uses context to achieve global optimal word segmentation. [6]It has become the most widely used model in statistical methods. Suppose X is the marked observation sequence, Y=P (y_1, y_2, y_3 ... Y_{n-1}, y_n|X) is the joint distribution function corresponding to the labeled sequence. Then CRF (X, Y) is an undirected graph model with X as the condition. When CRF is simplified and the same features are added at different positions, the local feature function is transformed into the global feature function, and CRF is expressed as the inner product form before the weight vector and the relative feature vector. Assuming that the number of transition features and state features are k_1 and k_2 respectively, let $k=k_1+k_2$, then:[7]

$$\text{make}\, h_k(x,y)=\sum_{i-1}^{n} f_k(y_{i-1},y_i,x,i), \tag{1}$$

w_k is the weight of $h_k(x,y)$, namely

$$w_k=\begin{cases}\lambda_k, k=1,2,\cdots k_1\\ s_k, k=1,2,\cdots k_2\end{cases}. \tag{2}$$

Let vector $\overline{w}=(w_1,w_2,\dots,w_k)$ represents the weight vector, vector

$$\overline{F}(x,y)=(h_1(x,y),\dots,h_k(x,y))^T \tag{3}$$

represents the global eigenvector, then the conditional probability formula of CRF is:

$$P(y\mid x,\overline{w})=\frac{\exp(\overline{w},\overline{F}(x,y))}{Z(x)}, \tag{4}$$

where:

$$Z(x)=\sum\exp(\overline{w},\overline{F}(x,y)). \tag{5}$$

When using CRF for Chinese word segmentation, the purpose is to obtain the maximum value of P= (y|x), which can be obtained through the Veterbi algorithm.[8]

CUITP should adopt this algorithm. Use this platform to analyze the paragraph "bu xing que yi xia zi luo dao ta tou shang lai", the analysis results are as the Table 2:

Table 2. The CUITP's word statistics results

List of standard words	List of non-standard words
Buxing Level5 adj. que Level4 adv. yixiazi Level5 ta Level1 pron. Lai Level1 v.	luodao toushang

By comparing The segmentation functions of CUITP and The GUIDELINES FOR ICLT, we can see that the segmentation results of the two platforms are different for the same paragraph "bu xing que yi xia zi luo dao ta tou shang lai". The GUIDELINES FOR ICLT divide this paragraph into nine words, while CUITP divides it into seven words, with the main differences being "luo dao" and "tou shang". The GUIDELINES FOR ICLT divide it into 4 words, while CUITP divides it into 2 words. This divergence is due to different algorithms. [9]The GUIDELINES FOR ICLT should adopt the matching method, which is highly dependent on dictionaries. Since "luo dao" and "tou shang" are not the same word in the dictionary, they are divided into four words respectively. However, CUITP should adopt a statistical method, which can better solve the ambiguity in word segmentation. For example, although "luo dao" and "tou shang" do not appear in the dictionary, they are still divided into words and listed in the List of non-standard words.

4 Conclusion

We give a brief introduction to CWS methods. The matching method should be used in The GUIDELINES FOR ICLT, while the statistical method is used in CUITP, so the results of word segmentation for the same paragraph are not the same.

At present, the construction of the dictionary of the international Chinese Education platform is mostly a combination of manual and statistical methods, and even depend mainly on the manual construction of experts, which requires a lot of manpower and time. Maybe the deep learning algorithm is the future direction of international Chinese education.[10]

Deep learning is essentially an emerging machine learning algorithm, which enables computers to simulate human learning and word segmentation processes through various types of neural network models. Current deep learning approaches are mostly built on variations of basic neural network models such as convolutions and loops. Convolutional Neural Network (CNN) is a feedforward neural network based on convolutional computation, including convolutional, pooling, full connection and other structures, which is widely used in the field of image recognition and processing. Bidirectional Long Short-Term Memory (Bi-LSTM) is developed from Recurrent Neural networks (RNN), It can obtain the actual word segmentation of text context well and has good learning ability for the information with long interval or delay, which is widely used in the deep model research of CWS tasks. Bidirectional Gated Recurrent Unit (Bi-GRU) not only has simpler structure, but also has faster segmentation efficiency without loss of segmentation accuracy. BERT pre-training model. Based on bidirectional Transformer, BERT realizes the full utilization of text information on both sides of the word and can dynamically generate word vectors and word vectors, providing higher efficiency for the next application.

Although language teaching requires standardization of pronunciation, vocabulary, grammar and Chinese characters, language processing technology in international Chinese education should also keep pace with The Times.

References

1. Xiong,W.X.(2014) Does Chinese really need inter-word space? -- Doubts about Chinese word segmentation and joint writing. Language Science, 13:655-669.
2. Zhong,X.Y., Li,Y.(2023)Summary of the research progress of Chinese word segmentation technology.Software Guide,22(02):225-230.
3. Zhang, D.Y., Hu,S., Xu, A.P.(2019) Joint learning method based on BLSTM for Chinese word segmentation. Application Research of Computers,36(10): 2920−2924.
4. Tu, W.B., Yuan, Z.M., Yu,Ki.(2020) Convolutional neural networks without pooling layer for Chinese word segmentation. Computer Engineering and Applications, 56(2): 120−126.
5. Li,L.X., Zhang,Y.T.(2022) Research on Chinese word segmentation based on Conditional random field. Information Technology and Informatization, 08:116-118+122.
6. Charles,S.,Andrew,M.(2010) An Introduction to Conditional Random Fields. Foundations and Trends in Machine Learning.18-26.
7. John,D.L.,Andrew,M.,Fernando,C.N.Pereira.(2001)Conditional Random Fields: Probabilistie models for segmenting and labeling sequence data.Proceedings of ICML-2001. Morgan Kaufmann Publishers Inc.,San Francisco, CA,USA.282-289.
8. Yuan,L.C.(2023) A joint method for Chinese word segmentation and part-of-speech tagging based on BiLSTM-CRF. Journal of Central South University(Science and Technology), 54(8): 3145−3153.
9. Mikolov,T., Chen, K., Corrado,G. et al.(2013) Efficient estimation of word representations in vector space. Proceedings of The 1st International Conference on Learning Representations. Arizona, USA. 1388−1429.
10. Hu,X.H., Zhu, Z.X.(2020) Research on Chinese word segmentation based on deep learning. Computer & Digital Engineering,48(3): 627−632.