PROCEEDINGS OF SCIENCE

Concept Discovery of Specific Field based on Conditional Random Field and Information Entropy

Jing Wan[1]
College of Information Science and Technology, Beijing University of Chemical Technology
Beijing, 100029, China
E-mail: wanj@mail.buct.edu.cn

Lidong Xing
College of Information Science and Technology, Beijing University of Chemical Technology
Beijing, 100029, China
E-mail: 2564849287@qq.com

Shuwu Zhang[2a]**; Wei Liang**[b]
Institute of Automation Chinese Academy of Sciences
Beijing, 100190, China
E-mail: shuwuzhang@ia.ac.cn; weiliang@ia.ac.cn

In order to detect the concept automatically in the specific field to obtain knowledge unit and relationship to create a knowledge map. This paper proposes a method based on conditional random field and information entropy, which divides the problem of concept discovery into two parts: the concept recognition and the new concept discovery. Firstly, the boundary of the text sequence is forecasted by the conditional random field. Compared with the concept in the dictionary, the candidate of the new concept is selected, the approximate position of the concept is found and the conceptual internal consistency is judged by mutual information. Determining the concept of boundary freedom is to carry out concept discovery by information entropy. Finally we can get the new professional concepts of the field of construction project. Experiments show that this method can effectively improve the accuracy, recall rate and efficiency of the concept identification and discovery in the field of specialization. It is also an effective alternative to the cascaded conditional random fields.

ISCC2017
16-17 December 2017
Guangzhou, China

[1]Speaker

This work is part of the research achievements of the Key Laboratory of Digital Rights Services, which is one of the National Science and Standardization Key Labs for Press and Publication Industry.

[2]Conrrespongding Author

PoS(ISCC 2017)037

1.Introduction

With the country's continuous investment in science and culture, China's scientific and cultural studies have achieved fruitful results, created many new theoretical methods. These theoretical methods contain a wealth of domain concepts different from the daily communication of people. The biggest method is professional. The domain concept is the smallest knowledge unit. The domain concept discovery is the process of extracting domain concepts by identifying the boundaries of domain text sequences by means of certain methods. It is used to construct the domain ontology and relationship. The domain concept discovery is the basic step of constructing the knowledge map, and it is also an important research content of natural language processing. Discoverying the concept of professional areas,it is a very meaningful topic[1] both for industry and academic arena. The term mentioned herein refers to the proprietary vocabulary and terminology, etc., in the field of specific areas or more in the form of proprietary or compound vocabulary, such as "the medium-term corporate bonds" and "the overall pre-stressed plate-column structure", etc.. The methods of domain concept discovery are mainly statistical-based and rule-based methods .

Peng combined domain knowledge features and lexical features into the model. The conditional random field training is used to identify words and new words are added to the dictionary to enhance the segmentation effect. Fu et al. annotated concepts manually, and then trained Hidden Markov Models (HMMs) with annotated corpus, then predicted the annotated results of test corpus text to discover concepts.

The rule-based approach requires manual construction of the rule. The concept of professional discovery rules need experts to establish the rules, which means greater costs while the domain concept updates fastly, requiring experts to follow up. In case of new fields of expertise, experts in the field need to re-establish the rules, which is obviously not realistic for the actual work. This paper presents a new statistical-based concept discovery approach. Firstly, the corpus is preprocessed and the existing corpus is used to annotate the corpus, and the marked corpus annotation set is used to train the conditional random field, and then the domain concept recognition model to mark the new field of corpus text sequence boundary, get the candidate string; then the candidate strings have some complete domain concept, some incomplete domain concept and some non-domain concept. The first to identify the existing concepts, the remaining strings are spliced by mutual information and filtered through left and right entropy to obtain new concepts.

In this paper, we first described the basic concepts of conditional random field and information entropy and the methods of using them in experiment. Then we introduced the experimental data labeling method and splicing method. Finally, we used conditional random field to combine information entropy method. Lastly, we carried out experimental results analysis.

2.Related Model

2.1 Conditional Random Fields

Conditional random field (CRF) is derived from the hidden Markov model. As a kind of discriminant probability model, it is a kind of random fields. In the natural language processing, the most commonly used linear chain condition random field (Linear Chain Conditional Random Field). It can be trained and predicted by using complex, non-independent, and overlapping features, and can take advantage of contextual information as a feature to add other

external features. The model can obtain rich feature information and solve the problem of labeling bias of the maximum entropy model.

CRF model, as shown in Fig. 1, the inter-vertex connection represents the dependency between random variables, satisfying Markovian. In the conditional random field, the distribution of the random variable Y is the conditional probability and the given observed value is the random variable X.

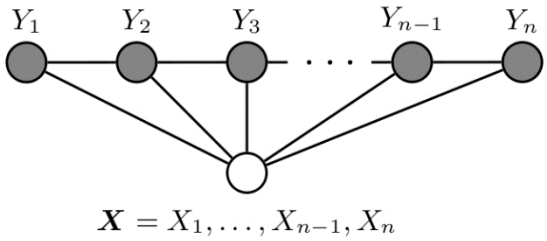

Figure 1: Graphic of CRF

Let the observation sequence $X=\{X_1, X_2, X_3, \cdots, X_n\}$, where the input data can be characters or words in the text. The corresponding state sequence is $Y=\{Y_1, Y_2, Y_3, \cdots, Y_n\}$. When the observation sequence X is given, the conditional probability calculation method of the state sequence Y *shall be*:

$$p(y|x)=\frac{1}{Z(x)}\prod_{t=1}^{T}\exp\left\{\sum_{k=1}^{K}\theta_k f_k(y_t,y_{t-1},x_t)\right\} \tag{2.1}$$

Which θ_k is the corresponding weight function of the corresponding function f_k, the function of the text y_{t-1}, y_t, respectively, is an output state and the current output state of the text, x_t is the current input state and the normalization factor for the Z (X). The calculation is shown as follows:

$$Z(x)=\sum_{y}\prod_{t=1}^{T}\exp\left\{\sum_{k=1}^{K}\theta_k f_k(y_t,y_{t-1},x_t)\right\} \tag{2.2}$$

f_k, where y_{t-1} is the output at time t, y_t is the context at time t, as shown in the corpus example: func1 = if (output = S and feature = "U02: MOS tube") return 1 else return 0。

In the process of applying the conditional random field, the selection of the feature will directly affect the validity of the characteristic function, which is directly related to the performance of the model. The choice of features is not a fixed form. It needs to be based on text language, target areas, textual features and other aspects of comprehensive consideration. Normally, the characteristics of the input state sequence are superimposed.

2.2 Mutual Information

Mutual information is often used to measure the degree of interdependence of two signals, and also measure the degree of intimacy of the binary. The internal consistency of the word is a measure of the degree to which two Chinese characters are closely integrated and used to measure the possibility of two Chinese characters that constitute words. For the integration of

the binary, the greater the likelihood that the binary will be part of the new word.. The mutual information is defined as:

$$MI(x,y)=log_2\frac{p(xy)}{p(x)p(y)} \tag{2.3}$$

Where p (xy) is the probability, in which, x and y occur simultaneously in the corpus; p (x) is the probability, in which x is present separately, and p (y) is the probability that y occurs alone.

2.3 Entropy

Studies have shown that a meaningful concept in a corpus usually has a higher frequency and will appear in a different document, and that word also has a higher degree of flexibility. The more types of characters that are adjacent to the string, the more flexible the string and the higher the degree of freedom of the border will be. Therefore, this paper introduces the left and right entropy as the quantification method of the degree of freedom of the new word boundary. The entropy is defined as

$$H_l = -\sum_{wl\in sl} p(wl|w)log_2p(wl|w) \tag{2.4}$$

$$H_r = -\sum_{wr\in sr} p(wr|w)log_2p(wr|w) \tag{2.5}$$

Where sl is the set of left adjacent words of the candidate word w. The right information entropy.

From the definition of the left and right entropy, it can be seen that if the left and right entropy of the candidate string is large, the number of word strings adjacent to the candidate word is more and the adjacent frequency distribution is more uniform, then the probability of words is greater.

3. Conceptual Discovery based on CRF and Information Entropy

The domain concept discovery can be viewed as a process of predicting the boundary of textual text sequences in the field of expertise. The conceptual discovery is integrated into the process of word segmentation and the new concept is found in the existing vocabulary. The concept discovery algorithm based on conditional random field and information entropy includes four parts: corpus preprocessing, corpus annotation training, new corpus recognition and candidate word splicing.

After the conditional random field model is identified, the candidate concept words are obtained and incorrect new concept words are mostly due to incomplete string . This paper proposes an algorithm to edit these incorrect concepts. The basic idea is to summarzie the left and right candidate words of the concept of incorrect. The corresponding mutual information is calculated, the new word is obtained by splitting the mutual information, and then the left and right information entropy of the new vocabulary is calculated, and the smaller value of the left and right entropy is taken as the information entropy, and so recursive to get the maximum entropy of the information under the constraints of the new words when the new concept is found, for example, through the conditional random field to identify the "construction process", which is an incomplete word; through the mutual information to splice the right word after the "construction process simulation analysis" and then calculate the "construction process

simulation analysis" left and right information entropy. The minimum entropy is its information entropy, and the information entropy is found in the process of splicing. It is found that the information entropy of the "construction process simulation analysis" is the largest, that is, the "construction process simulation analysis" is the new concept. The specific algorithm is shown as follows:

Input： Candidate Concepts S={ S_1， S_2， S_3， …， S_n}

Output： Concept words with maximum information entropy S_i^j=argmax(H_i)

1. **foreach** i {1,2,3…n} do
2. // K_i is the word frequency of Si
3. **if** K_i=1
4. //pos(Si) is the part of speech distribution splicing method
5. **return** S_i^j = pos(Si);
6. **else** K_i>1
7. //wl,wr is adjacency of S_i^j , mi(w) is mutual information method
8. **foreach** j {1,2,3…n} do
9. **if** mi(wl)>mi(wr)
10. S_i^j ←wl S_i^{j-1} ;
11. **else**
12. S_i^j ← S_i^{j-1} wr;
13. //set(s) is the method of save S_i^j
14. set(S_i^j);
15. **end for**
16. **return** S_i^j=¿ argh(set(S_i^j));
17. **end for**

4. Experimental Design

In order to establish a professional field of knowledge map and facilitate the professional areas of editing publications, we need to extract the field of expertise in the field, extract the concept of related terms. The book of professional journals is an accurate textual description, which contains a wealth of relevant professional concepts. The experimental data set is derived from 405 specific fields of expertise, a total of 26.85 million words, corpus size 418M, the use of domain dictionary 70962 domain concept, automatically marked 405 specific areas of expertise books, automatically identified a total of 66,704 old concepts in the domain dictionary Concept, automatically found a total of 20,271 complete and correct new concept, which condition random field found 7683, adding mutual information and left and right entropy found 12588. In this paper, CRF ++ 0.58 as a CRF implementation tool, and the experimental set to do ten-fold cross validation.

MOS pipe nz S	DC nx M	, w N

Figure 2: Character-based Domain Concepts Labeling

In this paper, the experimental data is sorted in the format shown in Fig. 2, in which the first column of each annotation is the word to be recognized, the second column is the penultimate column and the middle part is the part of speech feature. The design algorithm realizes the automatic annotation of field corpus based on vocabulary. The basic idea is to firstly use the dictionary tree to store the domain vocabulary, and then the word for word segmentation, word segmentation query dictionary tree, according to the query results to label.

Other,feature templates are used to generate feature functions, where each row is a template. Each template is specified by% x [Row, Column] for each unit in the input data. Row represents the row offset of the current cell, and Column represents the column position. Each line% x [#, #] generates a state function in the CRFs: f (s, o), where s is the output of the time t and o is the context at time t.

5.Analysis of Results

As we can see, when using words as the basic unit of concept recognition, accuracy and recall are both higher than when words are used as basic units. When using the CRF model with words as its basic unit, its recall rate is 3.56% or 5.7%, higher than the best recognition level for new word recognition or concept recognition, respectively by using words as the basic unit. The accuracy of word-based model increased by 20.06% after mutual information and entropy were added, while the recognition rate of word-based model increased by 46.54% after adding mutual information and entropy. The effect is very significant.

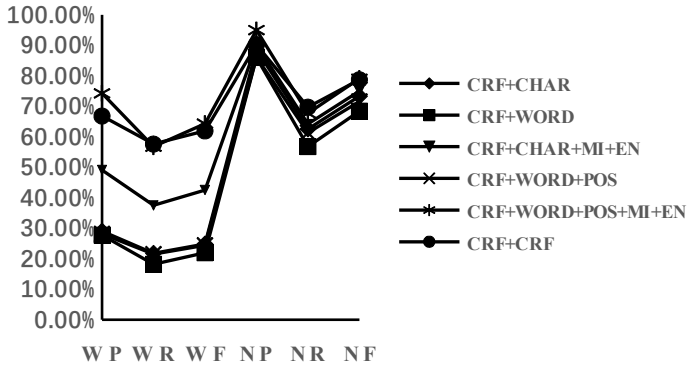

Figure 3: The Fusion Experiments

Experiments show that the conditional random field recognition and posterior processing can effectively improve the accuracy and recall rate of concept discovery by using mutual information and left and right entropy. The role of conditional random location recognition is to find out the position of the concept. Based on this method, the information entropy method can be used to extract the complete and accurate domain concept, and the effect of recognition is improved. At the same time, the method of mutual information and the effectiveness of the left and right entropy is improved. Therefore, the word-based recognition method is compared with the word-based recognition method. The conceptual position of the method is basically the same as that of the conditional random field model. The accuracy rate of the recognition method is lower than that of the latter after information entropy. The recall rate is based on the word. But also the method based on word is low, and the word-based model with the addition of part of speech is the best result of adding mutual information and the left and right entropy after recognition.

In addition, the analysis process and the experimental results show that the inclusion of mutual information and left and right entropy processing based on conditional random field recognition has the effect of replacing the stacking random condition field. The stratified condition field is trained by using the low-level condition random field recognition result to annotate the corpus and then use the high-level conditional random field to get the model. Based on the conditional random field and information entropy method, the recognition effect can be effectively improved and the difficulty of training can be ameliorated based on the method of stacking random field, it is a simple and effective alternative.

6. Conclusion

In this paper, the concept discovery and recognition of professional domain are studied from the aspects of word size, feature selection, mutual information, left and right entropy postprocessing. The concept recognition method based on conditional random field and information entropy is proposed in combination with the advantages of each experiment. This method can improve the accuracy and recall rate of concept recognition and discovery without increasing the manual labeling, and can improve the efficiency of recognition. The quality of the label set, the performance of the machine will greatly affect the identification of the training model, and thus affect the effect of concept recognition.

References

[1] Zheng YB, Liu ZY, Sun MS, Ru LY, Zhang Y. *Incorporating user behaviors in new word detection.* In: Proc. of the IJCAI 2009.San Francisco: Morgan Kaufmann Publishers, 2009. 2101-2106.

[2] Li H, Huang C N, Gao J, et al. *The use of SVM for Chinese new word identification*[M]//Natural Language Processing-IJCNLP 2004. Springer Berlin Heidelberg, 2005:723-732.

[3] Feng H, Chen K, Deng X, et al. *Accessor variety criteria for Chinese word extraction*[J].Computational Linguistics, 2004, 30(1):75-93.

[4] Peng F C, Feng F F, McCallum A. *Chinese segmentation and new the word detection using conditional random fields*[C]/ /Proc of the 20th lnternational Conference on Computational Linguistics. 2004:221-227.

[5] Fu G H, Luke K K. *Chinese unknown word identification as known word tagging*[C]//Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on. IEEE, 2004, 4: 2612 — 2617.

[6] Ye Y, Wu Q, Li Y, et al. *Unknown Chinese word extraction based on variety of overlapping strings*[J]. Information Processing&Management, 2012.

[7] Sun X, Wang H, Li W. *Fast online training with frequency — adaptive learning rates for Chinese word segmentation and new word detection*[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 .Association for Computational Linguistics, 2012: 253 — 262.