ATLANTIS PRESS
Advances in Engineering Research (AER), volume 131
3rd Annual International Conference on Electronics, Electrical Engineering and Information Science (EEEIS 2017)

Extracting Attribute Words for Domain Entity Knowledge Base Construction

Hong-Lin WU[1,a,*], Ruo-Yi ZHOU[2,b] and Ke WANG[1,3,c]

[1]College of Computer Science and Engineering, Northeastern University, Shenyang, China

[2]School of Information Engineering, Zhengzhou University, Zhengzhou, China

[3]Research Center for Artificial Intelligence, Shenyang Linge Technology Co., Ltd., Shenyang, China

[a]wuhl@mail.neu.edu.cn, [b]zhouruo.yi@qq.com, [c]flyingegg.ke@gmail.com

*Corresponding author

Keywords: Attribute, Entity, Extracting.

Abstract. One of the key problems in the construction of the entity attribute knowledge base for natural language understanding lies in domain attribute words acquisition. It is hardly to get these entity attribute word manually. This paper proposed a method of attribute words acquisition, which could acquire the entity attribute words from corpus automatically. The proposed method extracted a set of candidate attribute words based on the combination rules of part of speech; applied a series of queries in the search engine using the domain concept entity word or the candidate attribute word as the query term; calculated the mutual information values of all the domain concept entity words and the candidate attribute words; output the candidate attribute words whose mutual information value is greater than a specified threshold as the final attribute words. The experimental result showed that the proposed method performance well on the real corpus.

Introduction

One of the key problems in the construction of the entity attribute knowledge base for natural language understanding lies in domain attribute words acquisition. It is hardly to get these entity attribute word manually. This paper proposed a method of attribute words acquisition, which could acquire the entity attribute words from corpus automatically. The proposed method extracted a set of candidate attribute words based on the combination rules of part of speech; applied a series of queries in the search engine using the domain concept entity word or the candidate attribute word as the query term; calculated the mutual information values of all the domain concept entity words and the candidate attribute words; output the candidate attribute words whose mutual information value is greater than a specified threshold as the final attribute words.

Candidate Attribute Words Extracting

Recognizing the objective existence of those tangible objects is the first step of recognizing the real world. These objects may be called as entities. When recognize a large number of entities, they will begin to distinguish those entities. The distinction between entities is based on the structure and characteristics of the entity itself. We name these structure and characteristics as attributes. The structure of an entity is the components of the entity. An entity's attribute can be another entity or just a simple attribute. From the grammatical point of view, both the structural part and the character of the attribute words should belong to the grammatical unit of noun, such as: “PingMu(screen)/n” and “XianShi(display)/n + ZhiLiang(quality)/n”. But in the real corpus, after segmentation and part of speech tagging, there are many attribute words that are not the grammatical unit of noun. It could be a type of “v”, such as: “SanRe(heat dissipation)/v”. It could be a type of “v+n”, such as: “YunXing (running)/v + SuDu(speed)/n”. It could be a type of “n+v”, such as: “WaiGuan(appearance)/n + SheJi (design)/v”.

According to those attribute word combination types, we give priority to the recall rate by using such part of speech sequence template as much as possible to obtain the attribute words and put them

into the set of the candidate attribute words. And do simple filters for the set of the candidate attribute words. The procedure of the method is as follows.

The first step is matching the possible words or phrases from the annotated computer comment statements corpus as the set of the potential candidate attribute words according to the part of speech sequence templates.

Because the computer comment statements in the corpus are the direct review for the attributes, the second step is examining whether the last word of the potential attribute word which obtained by the part of speech sequence templates is an adverb or an adjective. The potential attribute words that associated with adjectives or adverbs with more than 4 times will be retained to the set of candidate attribute words.

According to the above extraction method, we use the part of speech template to find and extract the candidate attribute words from the corpus, and obtain 1125 candidate entries by using the initial filtering of word frequency and established knowledge resource. Although simple rule based filtering has eliminated most of the noise words, the existing candidate attribute words also contain many noise, such as "ShiHou(time)", "JianShao(reduction)" and other attributes which do not belong to the computer domain. We will use statistical approach to filter these noises.

Attribute Words Filtering Based on Mutual Information

We will use an entity attribute extraction strategy based on mutual information to acquire entity attribute words. Mutual information is the reduction in uncertainty of one random variable due to knowing about another, or in other words, the amount of information one random variable contains about another. The equation of mutual information is defined by Eq. 1.

$$I(x,y) = p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

There is a special kind of mutual information which called pointwise mutual information in practical applications. It is an information theoretically motivated measure for discovering interesting collocations which is defined by Eq. 2.

$$PMI(x,y) = \log\frac{p(x,y)}{p(x)p(y)} \tag{2}$$

In natural language processing, we often use mutual information to measure the degree of correlation between two language units. In general, the greater the relevance of the two language units, the greater the mutual information between them. Therefore, the value of the mutual information can be used to indicate the co-occurrence of two language units. In the acquisition of domain attribute words, the word is used as the unit of language. Attribute words and the feature words of the domain will have a higher co-occurrence rate. So in this paper, we will use the mutual information as a measure to achieve the filtering of candidate attributes to obtain the final attribute words.

We extracted a set of candidate attribute words based on the combination rules of part of speech. Many of the nouns, noun phrases and verb phrases in that set are not the attribute words of the computer domain. The set of candidate attribute words is mixed with a lot of noise. So we need to filter out the attribute words belong to the given domain.

We introduced the definition of mutual information and its application in natural language processing above. Now we decided to use the mutual information value as a measure to quantify the relevance between the domain concept and the candidate attribute words. The relevance is the degree of co-occurrence of attribute words and domain concepts. The definition of co-occurrence is as follows: if two words occurred in the same sentence at the same time, the two words are co-occurrence.

In order to calculate the co-occurrence of words better, we need statistics in a large scale of corpus. Our existing computer comment corpus could only be said to be good focus on the computer entity attribute word. The size and language characteristics of that corpus are not sufficient to calculate the co-occurrence of candidate attribute words and domain concepts. The calculation of mutual information should be carried out in larger and domain-specific corpus. Currently the Internet has become an explosive of resource library. The text of the webpages could be used as a huge corpus. The web pages which as searching result of search engine could provide the domain-specific text. We proposed a mutual information calculation method based on search engine and the Internet resource. The procedure of the method is as follows:

Input: The set of candidate attribute words.

Step 1: Apply a query in the search engine using the domain concept entity word as the query term. Record the total number of pages (Pages(w_{domain})) returned by the search engine.

Step 2: For each candidate attribute word in the set of candidate attribute words, apply a query in the search engine using the candidate attribute word as the query term. Record the total number of pages (Pages(w_{attri})) returned by the search engine.

Step 3: Connect the candidate attribute word to the domain concept entity word by a blank. Apply a query in the search engine using the new combined word as the query term. Record the total number of pages (Pages(w_{attri}, w_{domain})) returned by the search engine.

Step 4: Calculate the pointwise mutual information values of all candidate attribute words according to Eq. 3.

$$MI(w_{attr_i}) = \lambda_{attr_i} \log \frac{Pages(w_{attr_i}, w_{domain})}{Pages(w_{attr_i})Pages(w_{domain})} \tag{3}$$

Output the candidate attribute words whose mutual information value is greater than a specified threshold as the final attribute words.

This search engine-based mutual information calculation method converts the frequency required in the traditional point mutual information calculation formula into the number of search results. At the same time, we add a weight coefficient for the mutual information of each candidate attribute word in order to take into account the frequency of the candidate attribute word in the existing corpus. The coefficient is defined by Eq. 4.

$$\lambda_{attr_i} = \log(freq(attr_i)) \tag{4}$$

The coefficient is the logarithmic value of the frequency of the selected attribute word in the experimental corpus.

Experiment

In order to verify the feasibility of the search engine-based mutual information attribute words filtering method proposed above, we designed the following experiment. We extracted a set of 1125 candidate attribute words based on the combination rules of part of speech using as the testing set of the experiment, and manually determined a total of 268 right attribute words in the set as a standard answer. In order to have a relatively accurate search for the number of pages, we chose the “Sogou Search” (www.sogou.com) which provides a complete search page results as the experimental search engine.

The evaluation criteria for the experiment are precision, recall and F-1 measure. When the threshold was set as 52.5, the system identified 168 attribute words achieving the highest F1 measure of 60.11%. (The precision was 57.73%, and the recall was 62.69%.)

Conclusion

One of the key problems in the construction of the entity attribute knowledge base for natural language understanding lies in domain attribute words acquisition. It is hardly to get these entity attribute word manually. This paper proposed a method of attribute words acquisition, which could acquire the entity attribute words from corpus automatically. The proposed method extracted a set of candidate attribute words based on the combination rules of part of speech; applied a series of queries in the search engine using the domain concept entity word or the candidate attribute word as the query term; calculated the mutual information values of all the domain concept entity words and the candidate attribute words; output the candidate attribute words whose mutual information value is greater than a specified threshold as the final attribute words. The experimental result showed that the proposed method performance well on the real corpus.

Acknowledgement

This research was financially supported by the National Natural Science Foundation of China (61370155).

References

[1] K. Church, P. Hanks, Word association norms, mutual information, and lexicography, Proceedings of the 27th Annul Meeting of the Association for Computational Linguistics, (1987) 76-83.

[2] C. Manning, H. Schutze, Foundation of Statistical Natural Language Processing, Cambridge, MA: MIT Press, 1999.

[3] F. Zhang, Z.M. Ma, J.W. Cheng, A survey on fuzzy ontology for the semantic web, Knowledge Engineering Review, 3 (2016) 1-44.

[4] Y. Yin, GDC: A robust tag recommendation algorithm, Journal of Computational Information Systems, 22(2015)8061-8069.

[5] F. Smadja, Retrieving collocations from text: Xtract, Computational Linguistics, 19 (1993) 143-177.

[6] H.L. Wu, R.Y. Zhou, K. Wang, Knowledge representation of entity attribute frame for natural language understanding, Proceedings of the 2nd International Conference on Advances in Management Engineering and Information Technology, 2017, in press.

[7] K. Wang, H.L. Wu, Research on neologism detection in entity attribute knowledge acquisition, Proceedings of the 5th International Conference on Computer Science, Electronics Technology and Automation, 2017, in press.

[8] H.L. Wu, R.Y. Zhou, K. Wang, Template based attribute value words acquisition in entity attribute knowledge base construction, Proceedings of the 2017 International Conference on Computing Intelligence and Information System, 2017, in press.

[9] N. Chomsky, G.A. Miller, Introduction to the formal analysis of natural languages, Handbook of Mathematical Psychology, 2 (1962) 269-321.

[10] S. Abraham, K. Ferenc Kiefer, A Theory of Structural Semantics, Mouton & Co., Hague, 1967.

[11] F. Zhang, Z.M. Ma, J.W. Cheng, Enhanced entity-relationship modeling with description logic, Knowledge-Based Systems, 93 (2015) 12-32.