Advances in Engineering, volume 126

5th International Conference on Machinery, Materials and Computing Technology (ICMMCT 2017)

Research on Neologism Detection in Entity Attribute Knowledge Acquisition

Ke Wang[1,2], Honglin Wu[1,*]

[1] College of Computer Science and Engineering, Northeastern University, Shenyang, 110169, China

[2] Research Center for Artificial Intelligence, Shenyang Linge Technology Co., Ltd., Shenyang, 110004, China

[*] Corresponding Author: wuhl@mail.neu.edu.cn

Keywords: Neologism Detection; Entity Attribute; Knowledge Acquisition

Abstract. According to the requirements for the construction of the knowledge system of the entity attribute framework, the acquisition of attributes is extracted from large-scale real corpus. Real corpus must contain neologisms which cannot be identified by word segmentation program. This paper proposed a method of Chinese neologism detection which can discovery new words in real corpus, and can be used to revise the initial results of the word segmentation. The experimental result showed that the proposed method performance well on the real corpus of different fields, and may provide more accurate input for subsequent processing.

Introduction

According to the requirements for the construction of the knowledge system of the entity attribute framework, the acquisition of attributes is extracted from large-scale real corpus. Real corpus must contain neologisms which cannot be identified by word segmentation program. We need to re-analyze these strings. In the process of re-analyzing, the main task is the detection of the neologisms. The result of neologism detection can be used to revise the initial results of the word segmentation, and may provide more accurate input for subsequent processing.

A neologism is a new word or expression in a language, or a new meaning for an existing word or expression. There are a number of domain-specific vocabularies in the corpus for a particular domain. These words may not be included in the existing segmentation lexicon or the original training corpus of the word segmentation model. The segmentation system may segment these words in a wrong way. This will affect the identification of entities and attributes. For example, the Chinese word XianKa (video card) will be segmented into two Chinese characters Xian/v Ka/n. We must detect the neologism XianKa from the wrong segmentation Xian/v Ka/n. So that the recall of the entity attributes recognition process can be ensured.

Neologism detection is a basic study of natural language processing. At present the main methods of neologism detection are divided into two types: rule-based and statistical-based. Some methods were proposed to detect neologisms. Such as through analyzing real corpus from the Internet, build a large string set, and detect neologism by filtering rules. By using the word formation rules of the neologisms, established the regular word library and the special word formation rule base, and combine the relevant filter conditions to identify new words from the corpus. Consider the probability that a new word will consist of certain Chinese characters in the view of in-word probability in position. Carry out the detection by combine probabilistic statistical techniques and the rules. By calculating the covariance frequency among the entries, the candidate words are preferentially obtained with the frequency threshold, and then the final new words are determined by rule filtering and artificial decision.

Most of these methods need to combine the rules of the auxiliary, and have poor promotion of transplantation. In this paper, we combined the characteristics of the domain text, proposed a method of Chinese neologism detection based on the statistical language model which can discovery new words in real corpus. The proposed method performance well on the real corpus of different fields, and may provide more accurate input for subsequent processing.

The Process of Neologism Detection

Before put N-gram statistics on corpus, we firstly define the rules of neologism candidates. The rules specify which words may become neologism candidates. These rules can be simply described as:

a) Generate N-grams of the neologism candidates with Chinese characters as the processing element;

b) Never Generate N-grams crossing the sentence;

c) According to the regularity of the length of Chinese words, the length of neologism candidates is no more than five Chinese characters;

d) N-grams of the neologism candidates cannot be made by the combination of the terms in the non-extended word set and the context words. Non-extended word set includes: particle, conjunction, degree adverb and pronoun.

Under the constraint of the rules of neologism candidates, we find the bi-grams, the tri-grams and the four-grams which can became neologism candidates from the corpus. For a neologism of certain field, it will appear many times in the real text of the field. We can filter the neologism candidates by word frequency.

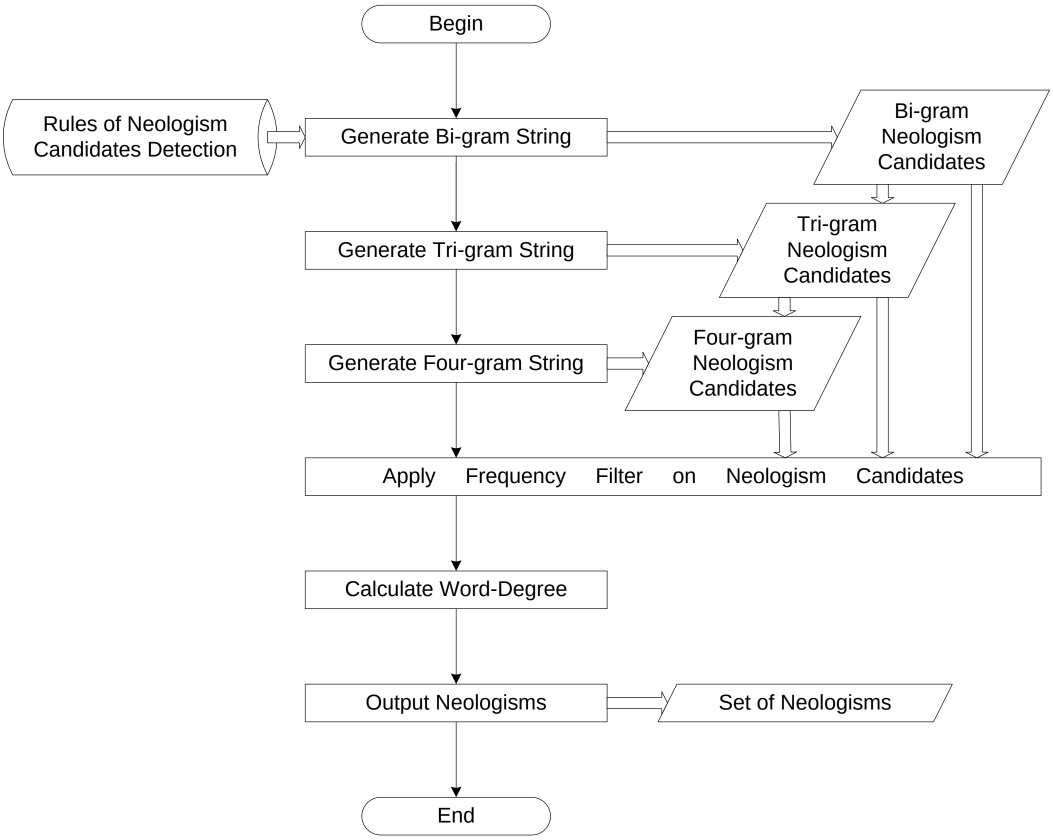
Begin
Rules of Neologism Candidates Detection
Generate Bi-gram String
Bi-gram Neologism Candidates
Generate Tri-gram String
Tri-gram Neologism Candidates
Generate Four-gram String
Four-gram Neologism Candidates
Apply Frequency Filter on Neologism Candidates
Calculate Word-Degree
Output Neologisms
Set of Neologisms
End

Fig.1. The process of neologism detection

The Word-Degree of a Neologism Candidate

We use word-degree to judge whether a neologism candidate is a real neologism. If the word-degree of a neologism candidate is high than a threshold, we select it and put it out as a real neologism. For a N-grams of the neologism candidate $w_1w_2\ldots w_n$, the word-degree will be defined as:

$$WordDegree(w_1w_2...w_n) = freq(w_1w_2...w_n)/\max\{freq(Substr(w_1w_2...w_n)\} \quad (1)$$

The numerator is the frequency of the word string $w_1w_2…w_n$ in the corpus. And the denominator is the maximum number of the frequency of all sub-strings of the string $w_1w_2…w_n$. Here the sub-strings of the string $w_1w_2…w_n$ do not contain itself.

We believe that the neologisms have strong field characteristics. In the text of a certain field, neologisms and their sub-string will be significantly different from common words. The sub-strings of a neologism will appear together with the neologism itself. So we define the denominator as the maximum number of the frequency of all sub-strings. In the actual processing, we can set a lower limit for the denominator via the statics of the real corpus.

Experimental Results

We collected a corpus with 13250 computer commentary sentences as the testing set. Under the constraint of the rules of neologism candidates, we found the bi-grams, the tri-grams and the four-grams in the testing set. Then we used the frequency as the threshold of the first step filter. The neologism candidates with a frequency greater than 5 were reserved. For the 1714 gotten neologism candidates, calculated the word-degree of them. Fig.2 showed the distribution of precision on neologism detection experiment. From the statistical analysis we could find that for the top 200 neologisms three is a sharp reduction at the range of 60 to 100. So the threshold is set accordingly. Finally we got the set of neologisms.

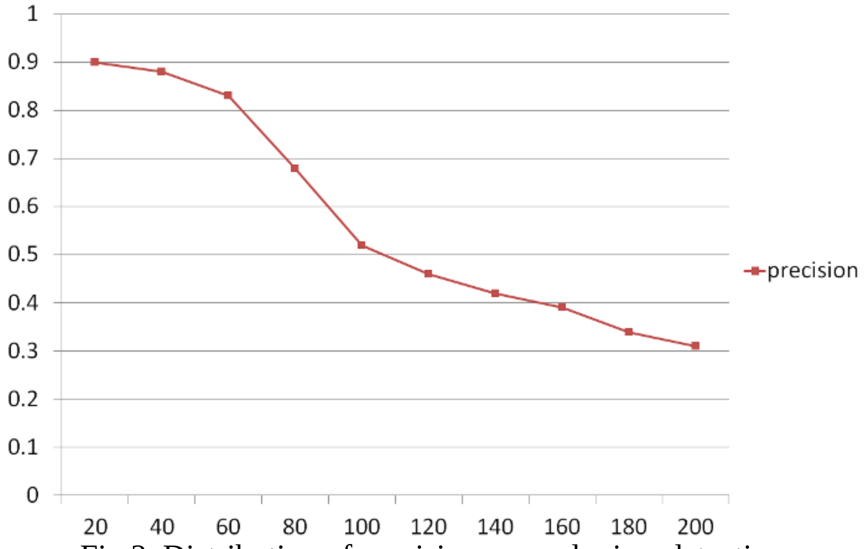

Fig.2. Distribution of precision on neologism detection

Here are the top 10 neologisms detected by the neologism detection system(include: Chinese Pinyin, English and word-degree): MoShou(World of Warcraft, 0.989), ShenZhou (a Chinese computer brand, 0.779), XianKa (video card, 0.625), GuiQi (Devil May Cry, 0.444), MoSha (scrub, 0.424), XunLei (a Chinese software name, 0.419), GuangQu (CD-ROM driver, 0.387), LiangLi (beautiful, 0.386), KaoQi (paint, 0.379), SheXiangTou (camera, 0.340).

It can be seen that the neologism detection method proposed in this paper is feasible. Since the neologism detection is applied for the subsequent process of the attribute acquisition and the extraction of the attribute value, we will re-marking the neologisms in correct form for the corpus to be analyzed.

Conclusion

According to the requirements for the construction of the knowledge system of the entity attribute framework, the acquisition of attributes is extracted from large-scale real corpus. Real corpus must contain neologism which cannot be identified by word segmentation program. This paper proposed a method of Chinese neologism detection which can discovery new words in real corpus, and can be used to revise the initial results of the word segmentation. The experimental result showed that the proposed method performance well on the real corpus of different fields, and may provide more accurate input for subsequent processing.

Acknowledgement

In this paper, the research was sponsored by the National Natural Science Foundation of China (Project No. 61370155).

References

[1] A.D Wu, Z.X. Jiang. Statistically-enhanced new word identification in a rule-based Chinese system[A]. Proceedings of the Second Chinese Language Processing Workshop[C]. 2000:46-51.

[2] A.M. Turing. On computable numbers, with an application to the Entscheidungs problem[A]. Proceedings of the London Mathematical Society[C]. 1937 (42) 230-265.

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

[4] C.E Shannon. Communication theory of secrecy systems[J]. Bell System Technical Journal. 1949(28) 656-715.

[5] S. Abraham, K. Ferenc Kiefer. A theory of structural semantics[M]. Mouton & Co., Hague. 1967.

[6] C. Fellbaum, M. Palmer, H.T. Dang. L. Delfs, S. Wolf. Manual and automatic semantic annotation with WordNet[A]. Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources[C]. 2001: 405-414.

[7] S. Shannon, E. Claude. Prediction and entropy of printed English[J]. Bell System Technical Journal. 1951(30)50-64.

[8] C. Manning, H. Schutze. Foundation of statistical Natural Language Processing[M]. Cambridge, MA:MIT Press. 1999.

[9] E.H. Yang, G.Q. Zhang, Y.K. Zhang. The research of word sense disambiguation method based on co-occurrence frequency of Hownet[J]. Journal of Computer Research & Development. 2001:138(7).

[10] K. Church, P. Hanks. Word association norms, mutual information, and lexicography[A]. Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics[C]. 1987:76-83

[11] M. Blaxquez, M. Fernandez, J. M. Garcia, A. Gomez. Building ontologies at the knowledge level using the ontology design environment[A]. Banff Knowledge Acquisition for Knowledge-Based Systems Workshop[C]. 1998.

[12] Y. Yin. GDC: A robust tag recommendation algorithm[j]. Journal of Computational Information Systems. 2015:11(22):8061-8069.

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

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