Advanced Materials Research Vols 143-144 (2011) pp 668-671 *Online: 2010-10-28*

doi:10.4028/www.scientific.net/AMR.143-144.668

A model for Chinese sentence similarity computing

Yang Xian-feng[1,a] ,Yu Zhou[1,b] ,ZHANG Pei-ying[2,c], Gao Guo-hong[1,d]

[1] the school of information technology, Henan Institute of Science and Technology, Henan, China

[2] College of Computer & Communication Engineering, China University of Petroleum, Shandong, China

[a]yxf110@hist.edu.cn, [b]yuzhou@hist.edu.cn, [c]smartfrom1024@126.com, [d]gaoguohong@hist.edu.cn

Key words: Natural Language Processing; Sentence Similarity; Word Form Similarity; Word Order Similarity; Semantic Similarity

Abstract. Sentence similarity computation is very important in the field of case-based machine translation. Through the in-depth analysis of sentence and the sentence similarity computing method based on the similarity computation of the word form feature, the word order feature and the semantic feature, we propose a sentence similarity computing model based on the multi-featured weight. By fusing the three features, giving different feature different weight to adapt the contribution of each feature to the sentence similarity computation, make sentence similarity computation more accurate. Experiment result shows that this approach has better accuracy in sentence similarity computation than the others.

Preface

Sentence similarity computing is of great importance in the various branches of natural language processing. For example, in case-based machine translation, search the most similar examples as in-put sentences from samples gallery using sentence similarity computing; in information retrieval, find the similar sentences as user requirement using sentence similarity computing; in question an answer, what sentence similarity reflects is the accouplement degree between question and answer; in multiple text automatic summarization, sentence similarity can reflect the fitting degree of partial theme information.

With the rapid development of these fields, many approaches about sentence similarity computing appear. They are mainly classified as, according to the analysis degree of sentences, the one based on statistics and the one based on comprehension. The one based on statistics mainly uses the key words appeared in the sentence and the times of N—Gram to compute sentence similarity. The more representative includes the method based on vector model [1], sentence similarity model and search algorithm of most similar sentence [2]; the one based on comprehension mainly uses semantic knowledge to compute sentence similarity. The more representative includes Chinese sentence similarity computing based on semantic dependency[6], Sentence similarity computing based on multi-hierarchies combination[7], An improved model for sentence similarity computing[8].

Sentence similarity computing model based on multi-featured weight proposed in the paper, mainly describes the features of sentences from word form, word order and word meaning, which are particularly focused and complementary when expressing sentence information. Experiment data shows, the method has a higher accuracy in sentence similarity computation.

Sentence similarity computation of multi-featured weight

Similarity in word form

The number of same words included in two sentences is used to reflect the similarity in word form in the two sentences. Here stop words must be cancelled when computing. Set S1, S2 as two sentences, the similarity in word form of S1 and S2 is：

Sim1(S1,S2)=2*(SameWord(S1,S2)/(Len(S1)+Len(S2)))

In it, Same Word(S1,S2) is the number of same words in S1 and S2;Len(S)is the number of words in sentence S.

Similarity in word order

what similarity in word order reflects is the word similarity in position relation in the two sentences. Because of the various expressing forms of chinese sentence, different word order denotes different meaning. Sentence is denoted as three vectors：

V1={d11,d12,…,d1n1}

V2={d21,d22,…,d2n2}

V3={d31,d32,…,d3n3}

In it, each dimension d1i in vector V1 represents tf×idf value of a word; each dimension d2i in vector V2 represents whether a 2-gram appears in the sentence（0 denotes doesn't appear，1 denotes appear）; each dimension d3i in vector V3 represents whether a 3-gram appears in the sentence.

The similarity in word order of two sentences is：

Sim2(S1,S2)=λ1*Cos(V11,V21)+λ2*Cos(V12,V22)

+λ3*Cos(V13,V23)

In it：λ1+λ2+λ3=1. λi represents proportionality factor occupied by each weight.

Semantic similarity

Semantic similarity represents the similarity in word meaning in two sentences. Here the similarity computation is based on HowNet. The similarity WSSim between word W and sentence S is defined as the maximal value of similarity of all words between word W and sentence S, the specific computation formula is as follows：

WSSim(W,S)=max{Sim(W,Wi)|Wi∈S}

In it：Sim(W,Wi) is the similarity between word W and Wi.

The similarity of word meaning between sentence S1 and S2 is defined as：

$$\mathrm{Sim}_3(\mathrm{S}_1,\mathrm{S}_2)=\frac{\sum_{w_i\in S_1} WSSim(w_i,S_2)+\sum_{w_j\in S_2} WSSim(w_j,S_1)}{|S_1|+|S_2|}$$

In it：|S| is the number of words in sentence S.

Sentence similarity

Sentence similarity reflects the similarity between two sentences. It is usually expressed as a numerical value ranging from 0 to 1, 0 denotes not similar, 1 denotes totally similar, the larger numerical value is, the more similar two sentences are. Set S1 and S2 as two sentences, thus the similarity of two sentences is：

Sim(S1,S2)=λ1*Sim1(S1,S2)+λ2*Sim2(S1,S2)

+λ3*Sim3(S1,S2)

In it：λ1、λ2、λ3 are constants, and meetλ1+λ2+λ3=1. in the paper,λ1=0.2, λ2=0.1, λ3=0.7.

Experiment result and analysis

100 Chinese sentences which are artificially segmented are taken as testsuite, whose average length is 13.6. They are classified as 20 categories according to similarity, each of which includes 4-6 similar sentences between each other. TF-IDF method, semantic dependency method and multi featured weight method are adopted to measure data, experiment result is as Table 1:

Table 1 experiment result

Computation method	Sentences measured (number)	Sentences with correct results(number)	Accuracy rate(%)
TD-IDF	100	43	43%
semantic dependency	100	82	82%
multi featured weight	100	86	86%

The analysis of experiment result: experiment result above shows that the method adopted in the paper is better than semantic dependency method in accuracy rate, which mainly dues to the consideration of the three features as word form, word order and word meaning. By analyzing the 14 sentences which are false in similarity computation, the reasons causing errors lie in the longer length of the sentences, unlisted words contained, errors appearing when the key words extracted are semantically computed, thus accuracy rate is decreased accordingly.

Summary

A Chinese sentence similarity computing method based on the multi-featured weight is adopted in the paper, which combines word form, word order and word meaning in sentences, and effectively describes the expressive meaning of sentences. According to the contribution of the three features to sentence similarity, the three features are given different proportionality factor, thus make sentence similarity computation reach its optimum. Although a higher accuracy rate in experiment result is obtained, the method in the paper is affected by computation result of word similarity. If the accuracy rate of similarity computation between words is further improved, a better result will be achieved.

References

[1] Zhang Qi, Huang Xuan-jing, Wu Li-de:*A new method for calculating similarity between sentences and application on automatic* text summarization(Journal of Chinese information processing, 2004,19(2)), p. 93-99

[2] Lv Xue-qiang, Ren Fei-liang, Huang Zhi-dan et al: *sentence similarity model and search algorithm of most similar sentence [J]* (Journal of Northeastern University(Natural Science),2003,24(6)), p. 531-534

[3] Liu Qun, Li Su-jian. [A]. Taipei: The 6th Chinese Lexical Semantics Workshop,2002

[4] Jiang Min, Xiao Shi-bin, Wang Hong-wei et al. *An improved word similarity computing method based on HowNet[J]*(Journal of Chinese information processing, 2008,22(5)), p. 84-89

[5] Dong Zhen-dong. HowNet[OL].http://www.keenage.com

[6] Li Bin, Liu Ting, Qin Bing et al:*Chinese sentence similarity computing based on semantic dependency relationship analysis[J]*(Application Research of Computers. 2003,20(12)), P. 15-1.

[7] Nan Xuan-guo, Cui Rong-yi. *Sentence similarity computing based on multi-hierarchies combination [J]*. (Journal of Yanbian University(Natural Science),2007,33(3)), p. 191-194

[8] Yang Si-chun. *An improved model for sentence similarity computing[J]*(Journal of University of Electronic Science and Technology of China,2006,35(6)), p. 956-959

Smart Materials and Intelligent Systems, SMIS2010
10.4028/www.scientific.net/AMR.143-144

A Model for Chinese Sentence Similarity Computing
10.4028/www.scientific.net/AMR.143-144.668