2009 International Conference on Machine Learning and Computing
IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

Research on Sentence Relevance Based on Semantic Computation

Jinzhong Xu [1 +], Xiaozhong Fan [1], Jintao Mao [1]

[1] School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China

Abstract. In Automatic Question Answering System, one key of question parsing and answer extracting is relevance computation of sentences. This paper introduces an approach to compute sentence relevance. Using the semantic computation in HowNet, the relevance of sentences can be calculated. The relevance of sentences can be calculated through the computation of relevance between words of sentence and subject words. Experimental results show the effectiveness of the method.

Keywords: question parsing, sentence relevance, semantic computation, HowNet

1. Introduction

The research of Automatic Question Answering System has question match. In that treatment process of our system, the relevance between question and answer will be used. That research involves the sentence relevance, which is realized by computing the relevance of words semantic in sentences. The importance of the study is to have an appropriate calculation method of sentence relevance.

Semantic relevance and semantic similarity are two different concepts. But they are closely linked. Semantic similarity is a degree that two words in different contexts can be used to replace each other without altering the syntactic and semantic structure of text. Semantic relevance includes some concepts of semantic similarity, and the calculation method of similarity has reference value for the research of relevance. Semantic relevance is realized based on computing of semantic similarity with HowNet.

2. Common calculation method of relevance

At present calculation method sentence relevance mainly has two types: the calculation method based on similarity and the calculation method based on ontology.

2.1. Calculation method based on similarity

Calculation method based on similarity is generally to use of vector space model. It's very similar to the calculation method of word similarity based on statistical. Query sentence and sentences of sentence base are transformed into vectors of characteristic words space, and then the cosine angle between two vectors is used to describe the sentence relevance. This method is simple, but modelling and computing based on vector space are not an accurate reflection of semantic information of query sentence.

2.2. Calculation method based on ontology

Ontology has received more and more attention in computer science. A lot of research is tried to apply ontology to computing relevance. Sentence relevance computing based on ontology includes three parts: construct domain ontology, keywords weighting and relevance computing.

Ontology describes common concept of a field explicitly and formally, users and computers can accurately communicate based on semantic by the definition of shared domain-specific concept and words, and not just exchange the data of grammatical representation. Concept is the basic structural unit of ontology,

[+] Corresponding author. Tel.: + 86 13693330166; fax: +86 01068915944.
E-mail address: xujinzhong@263.net.

concept sets are organized by concept hierarchies. Concept has attribute, and the concepts are related to each other through the attributes. In addition, example and synonym are structured in ontology to describe the sentence contents adequately. Ontology theory has been widely used in knowledge engineering, natural language processing, digital library and other fields. Ontology describes the sentence words by a series of definition of attribute, relation and example, these are the basic resource to compute relevance.

3. Sentence relevance

3.1. HowNet

HowNet is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. Concept and sememe are two important parts of HowNet. Concept is a description of words semantic, a word can be described with several concepts, and concept is described by sememe. Description of concepts in HowNet is an attempt to present the inter-relation between concepts and that between their attributes.

HowNet has 1618 sememes, these sememes include 10 types: Event|事件, entity|实体, attribute|属性值, aValue|属性值, quantity|数量, qValue|数量值, SecondaryFeature|次要特征, syntax|语法, EventRole|动态角色 and EventFeatures|动态属性. The description of concepts in HowNet is necessarily complex. A concept is described by an example. It is found in the knowledge dictionary as:

NO.=023683
W_C=打
G_C=V [da3]
S_C=
E_C=~球，~网球，~篮球，~羽毛球，~牌，~扑克，~麻将，~秋千，~太极拳，球~得很棒
W_E=play
G_E=V
S_E=
E_E=
DEF={exercise|锻炼:domain={sport|体育}}

In the example, "No." is the entry number of the concept in the dictionary, "G_C" is the part of speech of this concept in Chinese, and "G_E" is that in English, "E_C" is the example of the concept, "W_E" is the concept in English, "DEF" is the definition.

Semantic computing has been introduced to computation of relevance with HowNet. First of all, calculative mechanism is set up between similarity and relevance of sememe. Second, according to the calculation results of sememe, similarity and relevance of words can realize. Last, sentence relevance can compute by similarity and relevance of words.

3.2. Words semantic similarity

The sememe classification tree is given in HowNet, upper and lower semantic relation exists between parent node and child node, so we can compute the semantic similarity by using sememe classification tree.

$$Sim(p_1, p_2) = \frac{2 \times Spd(p_1, p_2)}{Depth(p_1) + Depth(p_2)}$$

In the formula, p_1 and p_2 are two sememe, $Spd(p_1, p_2)$ is the coincidence degree of p_1 and p_2, $Depth(p)$ is the depth of sememe in sememe tree.

In HowNet, concept has 4 parts: 1) first basic sememe description, the first sememe in DEF; 2) other basic sememe description, other sememes except first sememe in DEF; 3) relation sememe description, parts of concept is described by "relation sememe=basic sememe" or "relation sememe=(specific word)" or "(relation sememe=specific word)" in DEF; 4) symbol sememe description, parts of concept is described by "relation symbol basic sememe" or "relation symbol (specific word)" in DEF. The two concept similarity of the 4 parts are $Sim_1(C_1, C_2)$, $Sim_2(C_1, C_2)$, $Sim_3(C_1, C_2)$ and $Sim_4(C_1, C_2)$. So the whole similarity of concept is as follow:

$$Sim(C_1,C_2)=\beta_1 Sim_1(C_1,C_2)+\sum_{i=4}^{4}\beta_1\beta_i Sim_i(C_1,C_2)$$

In the formula, $\beta_i(1\leq i\leq 4)$, and $\beta_1+\beta_2+\beta_3+\beta_4=1$, $\beta_1\geq\beta_2\geq\beta_3\geq\beta_4>0$.

In the two words W_1 and W_2, if W_1 has n concepts: c_{11}, c_{12}… $c_{1\mathrm{n}}$, W_2 has m concepts: c_{21}, c_{22}… $c_{2\mathrm{n}}$, the similarity of W_1 and W_2 is the maximum similarity of concepts, it is as follow:

$$Sim(W_1,W_2)=\max_{i=1\cdots n,\, j=1\cdots m} Sim(s_{1i},s_{2j})$$

3.3. Words semantic relevance

Semantic relevance is a vague concept, there is no specific objective criteria can be measured. In syntactic analysis, the relevance of two words is higher; their distance is shorter in syntactic tree. Relevance of words involves morphology, syntax, semantic, even pragmatic and others. In these, relevance of the words the greatest impact is semantic relevance. The definition of relevance is a real number between 0 and 1.

Definition 1: In syntactic analysis, semantic relevance is the degree of modify relation, subject-predicate relation and co-referential relation of two words in a phrase structure.

Definition 2: In HowNet, for W_1 and W_2 are any two words, W_1 has n meanings: S_{11}, S_{12},…, S_{1n} , W_2 has m meanings: S_{21}, $S2_2$,…, S_{2m}. If there exists $S_{1i}= S_{2j}$, $1\leq i\leq n$, $1\leq j\leq m$, then relevance of W_1 and W_2 is 1.

If similarity of two words is high, their relevance is high, but relevance of two words is high, their similarity is not high. Semantic is described by sememe in HowNet. The sememe has 6 classes in HowNet, each class is a tree structure, and these classes are related to each other by explanation sememe. Hyponymy of sememe tree constitute the relevance of sememe, relation between sememe and explanation sememe constitute the relevance of sememe.

In the system consist of sememe, each sememe may also have a certain relation with the sememe which is in different tree. This adds lateral ties to tree hierarchy structure of sememe, so the system of sememe becomes a network structure. According to inheritance, hypogynous sememe inherits explanation sememe of upper sememe, and explanation sememe also has certain hierarchical structure, so it has lateral relevance of sememe. The relevance between two sememes is $S\mathrm{Re}l$, the formula as follow:

$$S\mathrm{Re}l(S_1,S_2)=\max\left(1-\frac{d(p_i,p_j)}{D}\right)$$

$$(1\leq i,j\leq 2, i\neq j)$$

In the formula, p_i and p_j are the first basic sememe of concept S_{i} and S_{j} respectively; D is the degree of lateral relation, it is the difference value of the layer explanation sememe influencing a sememe and its layer, and exceeding the layer, the influence can be ignored, the value of D is 10 suitably; $d(p_{\mathrm{i}}, p_{\mathrm{j}})$ is the difference value of the layer which p_{i} appears in the explanation sememe of p_{j}.

In HowNet, the concept is described by sememe, so the concept relevance must be computed by sememe relevance. If a concept C_1 is expressed by n sememes, concept C_2 is expressed by m smemes, and then concept relevance is approximate to the max value of the relevance between a sememe in C_1 and a sememe in C_2. It is as follow:

$$C\mathrm{Re}l(C_1,C_2)=\max \quad S\mathrm{Re}l(S_i,S_j)$$

In that, S_i is the ith smeme in C_1, S_j is the jth sememe in C_2, and $1\leq i\leq n$, $1\leq j\leq m$.

In HowNet, a word can has several concepts, so the words relevance can be computed by concept relevance. If word W_1 has x concepts, word W_2 has y concepts, and then words relevance is approximate to the max value of the relevance between C_1 and C_2. It is as follow:

$$W\mathrm{Re}l(W_1,W_2)=\max \quad C\mathrm{Re}l(C_i,C_j)$$

In that, C_i is the ith concept in W_1, C_j is the jth concept in W_2, and $1\leq i\leq x$, $1\leq j\leq y$.

According to these researches above, the semantic relevance of two words consist of semantic similarity and words relevance. So define words semantic relevance is $WS\mathrm{Re}l(W_1,W_2)$:

$$WS\mathrm{Re}l(W_1,W_2)=\eta_1\times Sim(W_1,W_2)+\eta_2\times W\mathrm{Re}l(W_1,W_2)$$

In that, $Sim(W_1,W_2)$ is the similarity between word W_1 and word W_2, $W\mathrm{Re}l(W_1,W_2)$ is the words relevance, η_1 and η_2 are the weighting of similarity and relevance.

3.4. Keywords weighting computation

A sentence is described by a set which consists of words, but some of words are the keywords which have high weight in sentences. Computing keywords weighting of sentence is start before computing relevance. This implies that the distance between keywords is short, keywords weighting of corresponding is high, and vice versa.

Keywords weighting computation adopt TFIDF, the formula as follows:

$$w_{ik}=\frac{tf_{ik}\times idf_k}{\sqrt{\sum_{k=1}^{n}[tf_{ik}\times idf_k]^2}}=\frac{tf_{ik}\log(N/n_k+0.01)}{\sqrt{\sum_{k=1}^{n}[tf_{ik}\times\log(N/n_k+0.01)]^2}}$$

In that, w_{ik} is a weighting of words in sentence, tf_{ik} is occurrence frequency of words of sentence in corresponding document, id_{fk} is quantization of distribution which keywords are in all documents of the document set. $idf_k=\log(N/n_k+0.01)$, N is the total number of document, n_k is the number of document which appears keywords, n is the number of sentence. The denominator of formula is normalization processing for every component.

Based on this assumption, occurrence frequency of words of sentence in corresponding document must be high, but it is low in other documents of all documents. So, weighting is the product of tf_{ik} and id_{fk}. From this, quantization of weighting is based on occurrence frequency of words and documents.

3.5. Sentence relevance computation

Sentence relevance is the relevance between query sentence and sentences of sentence base. Query sentence has more keywords of sentence base and weighting of keywords in query sentence is high, the relevance is high. So sentence relevance is measured by words and keywords weighting in sentence.

So definition as follow:

1 a set of words in sentence A:

$WordSetA=\{Word_1,Word_2......Word_n\}$

2 a set of words in sentence B in sentence base:

$WordSetB=\{Word_1,Word_2.....Word_r\}$

3 similarity between $Word_i$ and $Word_j$:

$Sim(Word_i,Word_j)\ \ 1\le i\le n,1\le j\le r$

4 a set of weighting relevance:

$Weight=\{Weight_1,Weight_2......Weight_r\}$

5 relevance between A and B:

$St\mathrm{Re}l(A,B)$

Taking a $Word_{\mathrm{i}}$ from $WordSetA$, if $1\le j\le r$ and $j\ne p_{\mathrm{i}}$, then $Sim(Word_i,Word_{p_i})>Sim(Word_i,Word_j)$. If $Sim(Word_i,Word_j)$ is more than threshold tmp, then weighting relevance $Weight_{p_i}=1$. Every word in $WordSetA$ is process as above, the weighting relevance is $WeightA$.

So the relevance between A and B is the sentenc relevance $St\mathrm{Re}l(A,B)$:

$$St\,\mathrm{Re}\,l(A,B)=\sum_{k=1}^{r}w_k\times WeightA_k\times WeightB_k$$

The sentence relevance after normalization processing:

$$St\,\mathrm{Re}\,l(A,B)=\frac{\sum_{k=1}^{r} w_k \times WeightA_k \times WeightB_k}{\lambda+\sum_{k=1}^{r} w_k \times WeightA_k \times WeightB_k}$$

$\lambda>0$, λ is constant which value is determined according to the specific conditions.

4. Experiment

4.1. Sentence relevance computation

Word segmentation processing is the basic of this experiment system, the sentence segment independent words. Words relevance computation is the key of sentence relevance computation, it is the theoretical basis of query words matching, and its result has an important effect on relevance statistic. Sentence relevance statistic is the basic of matching, it is a relevance statistic of query sentence, and system output the sentence of maximal relevance and its relevance. The experiment is realized as Figure 1.

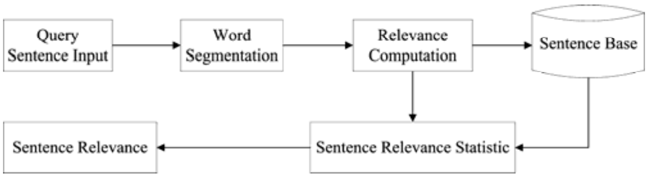


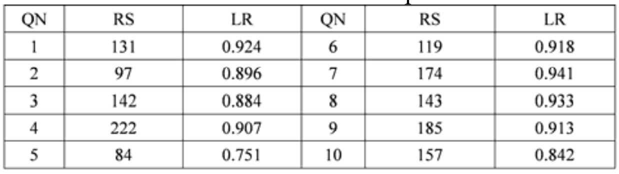
Fig. 1: Calculation process of Sentence relevance

4.2. Experimental analysis

In the experiment, we choose computer field and collect 10 Common computer failures from Web, magazine and books. These sentences are the query sentences. For examining the result of sentence relevance computation, we choose sentences as sentence base from a book which is Common Computer Problem and 1000 Cases of Failures.

In the experiment system, we have a statistic. The result is in table 1, the QN is the query No., RS is the number of relevant sentences of query sentence, and LR is the largest relevance value.

Table. 1: Result of experiment



QN	RS	LR	QN	RS	LR
1	131	0.924	6	119	0.918
2	97	0.896	7	174	0.941
3	142	0.884	8	143	0.933
4	222	0.907	9	185	0.913
5	84	0.751	10	157	0.842

Words semantic relevance computation is based on HowNet, and the sentence relevance is computed using words relevance. From experiment, the result is satisfactory, and the method is an effective method in Automatic Question Answering System. The method also can be applied to other fields, such as text categorization, text clustering, information retrieval and others.

5. References

[1] Qun Liu, and Sujian Li. Word similarity computing based on Hownet. *Computational Linguistics and Chinese Language Processing.* 2002, pp. 59-76.

[2] Zhendong Dong, and Qiang Dong. Hownet [EB/OL]. http://www.keenage.com,2000203210 2003201211.

[3] Tian Xia. Study on Chinese Words Semantic Similarity Computation. *Computer Engineering.* 2007, 33(6):191-194.

[4] K.W. Gan, and P.W. Wong, Annotation information structures in Chinese texts using Hownet. Charniak E.Second Chinese Language Processing Workshop. *HongKong: Hong Kong University of Science and Technology.* 2000, pp. 85-92.

[5] Sujian Li. Research of Relevance between Sentences Based on Semantic Computation. *Computer Engineering and Applications.* 2002, 38(7):75-76.