Research on concept-sememe tree and semantic relevance computation

GuiPing Zhang [1], Chao Yu[1], DongFeng Cai[1], Yan Song[1], JingGuang Sun[1]

1 Natural Language Processing Laboratory, Shenyang Institute of Aeronautical Engineering
P.O.box 118, No.52 North Huanghe Street, Shenyang, Liaoning, China, 110034.
Email: zgp@ge-soft.com, yc089067@sina.com, cdf@ge-soft.com,
mattsure@gmail.com , sunjingguang@gmail.com

Abstract. In this paper, we parse the hierarchy relation of concepts in HowNet, with the parsing we design a concept-sememe tree structure, which can make it easy to understand the relations between the sememes in concepts. The tree structure can easily describe the relationship between sememes in concept and make it more convenient to process by computer. The steps of building the tree are also presented in this paper. We then discuss the relevance computation based on HowNet. The preliminary experiment shows the relevance computation method can achieve satisfying results.

Keywords: HowNet; relevance; concept-sememe tree.

1 Introduction

The computation of similarity and relevance between words has wide application in the fields of machine translation, sense disambiguation, IR etc. In some cases, similarity and relevance may be confused. Similarity between words refers to the feature of clustering while relevance refers to the extent of association [1]. The two words with similar semantic relationship may have similar associated words such as the words “医生(doctor)” and “护士(nurse)”, both have the same associated words as “医院(hospital), 病人(patient), 打针(have an injection), 吃药(drug)” etc. However, associated words usually have no similarity such as the word pair “吃 (eat)” and “食物(food)” etc.

Researches on computation of similarity and relevance are widely spread in the world. Recently, there are two semantic similarity computation methods based on HowNet: one is described by Professor Liu Qun in [2] and the other is described by professor Dong ZhenDong in [3]. In this paper, we describe how to build concept-sememe tree on the basis of decomposing word concept definition (DEF). The concept-sememe tree illustrates DEFs of the words and makes them easy to process by computer.

2 Concept similarity computation and the implementation of the concept-sememe tree

Concept is a kind of description of word sense. Polysemous words have several different concepts. Because polysemous words contain several concepts, we must ascertain the DEFs of the polysemous words before we compute the semantic similarity of the word pairs. Thus the computation of the semantic similarity is, in fact, to compute the similarity of the concepts.

2.1 HowNet based concept similarity computation

Number of the same concept node determines the similarity of the concept of the words. The more concept nodes two concepts share, the more similar the two concepts are. Here the concept node is a pair of “semantic role=sememe”. Same concept nodes refer to the same description of sememes in character form and the same structure of these concepts. In practical computation, it is not easy to determine whether two concept nodes are with the same structure. In this paper we design the concept-sememe tree which translates concept DEF into tree structure, it can make us convenient to understand the hierarchy structure of concept DEF and also can make it easy to process by computer. For example, the concept DEF of the concept “医生（doctor）” and “护士（nurse）” are shown as:

“医生(doctor)”：DEF={human|人:HostOf={Occupation|职位},domain={medical|医},{doctor|医治:agent={~}}};“护士(nurse)”：DEF={human|人:HostOf={Occupation|职位},domain={medical|医},{TakeCare|照料:agent={~}}}

Through the analysis of the concept DEF we develop a program to decompose concept DEF and build concept-sememe tree as fig1 and fig2:

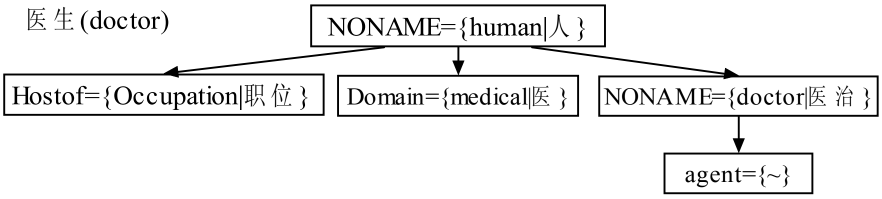

Fig. 1. The concept-sememe tree of concept “医生(doctor)”

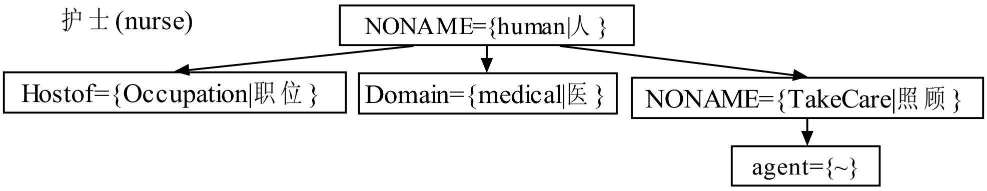

Fig. 2. The concept-sememe tree of concept “护士(nurse)”

We can see there are five concept nodes in the concept-sememe tree of “doctor” and “nurse” respectively. Seen from the character form, there are four same pairs of the concept node, they are: “NONAME={human|人}”, “Hostof={Occupation|职位}”, “Domain={medical|医}” and “agent={~}”. But we will find from the concept-sememe tree of concept “医生(doctor)” that the father node of concept node “agent={~}” is “NONAME={doctor|医治}”，and so to the concept-sememe tree “护士(nurse)”, the father node of “agent={~}” is “NONAME={doctor|照顾}”，that is ,their corresponding father nodes are different, so we consider the concept node “agent={~}” of concept “医生(doctor)” is different from that of concept “护士(nurse)” in structure. That is, in these two concepts with five concept nodes, the same number of concept nodes is three, which is the most important reference to similarity computation between concepts. The computation is described in detail in [3].

2.2 The process of building concept-sememe tree

1：We categorize the concept node into two types, one type is the “dynamic role = {value}” which has been fully described. The partly described ones are classified to the other type. For example, in the concept DEF of “医生(doctor)” “HostOf={Occupation|职位}” is the first type and “human|人”, “doctor|医治” is the second type.

2：The method of searching the father node to one concept node of the first type: to any concept node (*Node(i)*), find the nearest colon *j* which is ahead of *Node(i)*, if the number of the symbol “{” equals to the number of “}” between the area of *Node(i)* and colon *j*, so the concept node which is ahead of colon *j* is the father node of *Node(i)*. Otherwise find the next position of the colon which is ahead of colon *j*, continues the same judgment.

3：The method of searching the father node from concept nodes of the second type: to any concept node (*Node(i)*), find the nearest colon *j* which is ahead of *Node(i)*, if the number of the symbol “{” is one more than the number of “}” between the area of *Node(i)* and colon *j*, so the concept node which is ahead of colon *j* is the father node of *Node(i)*. Otherwise find the next position of the colon which is ahead of colon *j*, continues the same judgment.

4：When father nodes of all concept nodes are found, we can easily describe the relationship of the sememes through concept-sememe tree.

Now we make an example to illustrate the use of concept-sememe tree: The DEFs of the words “洗衣(wash clothes)” and “洗衣机(washer)” are “{wash|洗涤:patient={clothing|衣物}}” and “{tool|用具:{wash|洗涤:instrument={~},patient={clothing|衣物}}}” respectively. According to procedure mentioned above we can build concept-sememe tree as fig 3, 4 shows:

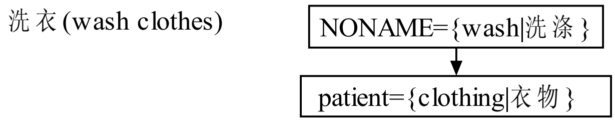

Fig. 3. Concept-sememe tree of “洗衣（wash clothes)”

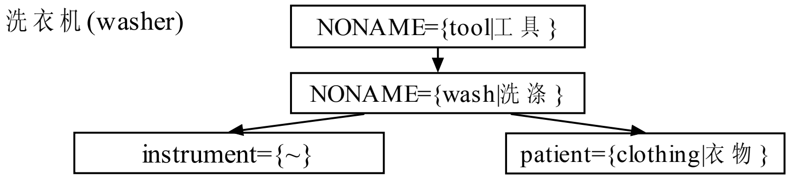

Fig. 4. Concept-sememe tree of “洗衣机（washer)”

It is easily seen from fig 3,4 that the concepts “洗衣(wash clothes)” and “洗衣机(washer)” have two same concept nodes which are concept node “wash|洗涤” and concept node “patient={clothing|衣物}” in the form of character. But the ancestral nodes of the two concept-sememe trees are “wash|洗涤” and “tool|工具” respectively. Then there won’t be any same concept nodes between them. Thus the value of similarity between concepts “wash clothes (洗衣)” and “washer (洗衣机)” is small in HowNet.

3 Semantic relevance computation

Semantic relevancy refers to how close of the relationship between two words is. In this paper, we decide to implement concept relevance computation by using HowNet. The sememes in concept of the words and related concepts field[4] in HowNet provide an approach for relevance computation.

Relevant concepts are the concepts which are associated with the concept of a given word. Related concepts field is a set of relevant concepts, which is figured by words.

3.1 Compute the relevance Rel_1 of the sememes of DEFs between two words

The expression of the sememes of the word concept provides clues for us to build the association relationship of the words. We separate the DEF into the set of sememes. The overlap of the sememes indicates the extent of semantic relevancy, here we describe it as Rel_1. See the following formula (1):

$$Rel_1(A,B) = \frac{|Sememe_A \cap Sememe_B|}{|Sememe_A \cup Sememe_B|} \qquad \textbf{(1)}$$

Here Sememe_A and Sememe_B refer to the sememe set of concept A and sememe set of concept B respectively. The numerator refers to the number of the same sememes in concept A and concept B while the denominator refers to the number of the sememes in the union of sememes in concept A and sememes in concept B. For example the concept of the word “报纸(paper)” and “新闻(news)” is “{publications|书刊:{publish|出版: ContentProduct={news|新闻}, LocationFin={~}}}” and ” {news|新

闻}” respectively. We can get the number of sememes of the two concepts is 4 and 1, the same sememe between them is only “{news|新闻}”. Then the Rel_1(“报纸(paper)”, “新闻(news)”) is 0.25.

3.2 Compute the complete containing degree *Rel₂* of the sememes of DEF

Two related words may have no same sememe in their DEFs. The sense of two words is related because they contact with intermediate entities in some context. For example, “食物(food)”, “鱼(fish)” and “海货(seafood)”, their DEFs are “{food|食品}”, ” {fish|鱼}” and “{food|食品:material={fish|鱼}}” respectively. It is easy to see that the DEF of “海货(seafood)” describes the relationship between “食物(food)” and “鱼(fish)”. Because “鱼(fish)” can be regarded as material of “食物(food)”, then the relevancy of “食物(food)” and “鱼(fish)” can be built through the DEF of the word “海货(seafood)”.

Take another example, the words “吃(eat)” and “面包(bread)”, their DEFs are “{eat|吃}” and “{food|食品}” respectively. It is hard to find the association from their DEFs since they are too simple, but they should be relevant from common sense. Then we should ascertain their relevancy by the description of the sememes of related words in their related concepts field. There are 2554 words in the relate concepts field of “{eat|吃}” of the word “吃(eat)” while there are 116 of them contain the DEF “{food|食品}”. And there are 857 words in the relate concepts field of “{food|食品}” of the word “面包(bread)” while there are 13 of them contain the DEF “{eat|吃}”. Thus we count the number of words which completely contain the concept of the word to illustrate the extent of containing DEF as Rel_2, see the formula (2):

$$Rel_2(A,B) = MAX(\sqrt{\frac{Num(A \subset b_j)}{N_B}}, \sqrt{\frac{Num(B \subset a_i)}{N_A}}) \qquad \textbf{(2)}$$

In the formula (2), A and B refer to different concepts, a_i refers to the concept of $word_i$ in the related concepts field of concept A while b_j refers to the concept of $word_j$ in the related concepts field of concept B. N_A refers to the number of the words in the related concepts field of word A and N_B refers to the number of the words in the related concepts field of word B. Supposing concept A is “{eat|吃}” and concept B is “{food|食品}”, we apply formula (2) as following:

$$Rel_2(A,B) = MAX(\sqrt{\frac{13}{857}}, \sqrt{\frac{116}{2554}}) = MAX(0.123163, 0.213117) = 0.213117$$

3.3 Compute the containing degree of the words in the related concepts field of two words (*Rel₃*)

By using the method mentioned above, we still find it is hard to establish relationship between some words such as the words “鱼(fish)” and “水(water)”. Their DEFs are “{fish|鱼}”and “{water|水域}” respectively. We can see that there is no same sememe and mutual comprisal relationship between sememes but the word “水(water)” still appears in the related concepts field of the word “fish|鱼”. This is because the first sememe of DEF “鱼(fish)” is described as “fish|鱼” and its sememe frame is “{animal|兽：MaterialOf={edible|食物},{alive|活着：experiencer={~},location={waters|水域}},{eat|吃：patient={~}},{swim|游：agent={~}}}”. Professor Dong ZhenDong extracts the DEF segment “location={water|水域}” to make a fuzzy search and embodies all the words whose DEF contains the whole segments into the related concepts field of word “鱼(fish)”. This makes an approach to build the relationship between words “鱼(fish)” and “水(water)”. We use the number of the same words in related concepts field of two words to indicate the extent of the semantic relation of the words. See formula (3):

$$Rel_3(A,B) = \frac{|W_A \cap W_B|}{|W_A \cup W_B|} \qquad \textbf{(3)}$$

Here W_A and W_B refers to word sets of concept A and concept B respectively, numerator refers to the intersection of two words’ related concepts field while denominator refers to the union of them.

4 Experiment

After the computation of Rel_1（A, B）、Rel_2（A, B）、Rel_3（A, B）, we can use formula (4) to compute the concept relevance between words A and B.

$$Rel(A,B) = \beta_1 Rel_1(A,B) + \beta_2 Rel_2(A,B) + \beta_3 Rel_3(A,B) \tag{4}$$

Here β_i is an adjustable weight coefficient and the sum is 1. In our experiment, these parameters are: β_1＝0.3，β_2＝0.2，β_3＝0.5. Here we list some experimental results in table 1:

Table 1. relevance computation results

Word 1	Word 2	similarity (HowNet)	relevance
鱼(Fish)	水(Water)	0.016667	0.3686129
吃(Eat)	食物(Food)	0.000624	0.2339492
吃(Eat)	水(Water)	0.000624	0.0454115
吃(Eat)	报纸(Newspaper)	0.000624	0.002011
新闻(News)	报纸(Newspaper)	0.116667	0.76
新闻(News)	记者(correspondent)	0.118605	0.733
新闻(News)	传播(Disseminate)	0.000624	0.1710146
警察(Policeman)	法官(Judge)	0.825000	0.7278904
警察(Policeman)	警衔(Police rank)	0.101247	0.2913509
警察(Policeman)	治安(public order)	0.001247	0.0540706
医生(Doctor)	护士(Nurse)	0.620000	0.6752305
医生(Doctor)	手术(Operation)	0.000624	0.5999999
护士(Nurse)	手术(Operation)	0.000624	0.4780876

From the table we can see that most experimental results are satisfying. The words with strong similarity also get high relevance value such as the words “警察(policeman)” and “法官(judge)”. The words with strong relevancy usually do not show strong similarity such as words “新闻(news)” and “记者(correspondent)” etc.

5 Conclusion

The experimental result is acceptable and conforms to human’s intuition. In the future work, we will make some further researches on semantic information of concept of word and we will classify the concepts to make the relevance value more reasonable.

References

1. Dagan I., Lee L. and Pereira F. (1999), Similarity-based models of word cooccurrence.
2. LiuQun, LiSuJian, word similarity computation based on HowNet http:/ /www.keenage.com, 2002.
3. Dong ZhenDong, HowNet and Computation of Meaning[M] Singapore：World Scientific press, p197-206 2006.
4. Dong Qiang, Dong ZhenDong, related concepts field’s building based on HowNet[J] Language Computing and Text Processing based on context p364-369 2003