Domain Unconstrained Language Understanding Based on How-net

*+Jhing-Fa Wang, *Hsien-Chang Wang and *Chin-Nan Lee
* Department of Computer Science and Information Engineering
+ Department of Electrical Engineering
National Cheng-Kung University, Taiwan, R.O.C.

Abstract

In this paper, we propose a method for domain unconstrained language understanding based on the How-net knowledge base. The goal is to construct a system that reads in an article and answers some related questions. For each sentence in the article, word segmentation is first applied. Then, the major components such as agent, theme, event, time, and place are extracted to construct a semantic-slot table and a semantic network. Answers of the questions are derived using two approaches, which are based on the relational and hierarchical relation among the major components.

Our method is applied to the understanding of the primary-school textbook, and it is able to answer questions in the exercise of the textbook.

1. Introduction

Currently, most available applications of natural language processing (NLP) are domain specific. In this paper, we propose a method for domain unconstrained language understanding. The knowledge base we used is the How-net knowledge base, which is constructed by researchers in Beijing. The description and content of How-net can be found in the following URL, http://www.how-net.com. How-net describes the relationship of objects using both concepts and attributes. Based on How-net knowledge base, we have some methods to analyze the sentences of an article. First, word segmentation is performed to find

the corresponding word sequence. Then, major components (agent, event, time, place, and theme) conveyed in the sentence are extracted. Finally, the semantic table and semantic network are constructed for the understanding of the article.

For the experiment, we try to answer the questions in the exercises of the primary-school textbook. Article of each lesson is used to construct the corresponding semantic table and semantic network. Answers of the questions are derived by measuring the likelihood of the major components.

This paper is organized as follows. In Section 2, we introduce basic concept of the How-net knowledge base. In Section 3, the method to analyze sentences is described. In Section 4, we show how to answer the questions of the primary-school textbook. The conclusion is given in the final Section.

2. The Structure of How-net

How-net is a Chinese knowledge base which describes the objects using concepts and attributes. The basic units in the How-net are physical and spiritual objects such as components, attributes, time, space, events, attribute values, and so on.

Common and individual characteristics of concepts are both recorded in How-net. Consider the words "*醫生(doctor)*" and "*患者(patient)*" as example, the common characteristic of them is "people". This common characteristic is recorded in the How-net. On the other hand, "doctor" has the individual characteristic being the agent of "cure"; and "patient" has the individual characteristic being the experiencer of "suffer".

The relations among different concepts and attributes are also described by How-net. The relations are shown below:

(1) Upper-Down relation.	Ex. Father-Son, Father-Daughter, ...
(2) Synonymous relation.	Ex. Good-Well, Big-Large, ...
(3) Antonymous relation.	Ex. Good-Bad, Large-Small, ...

(4) Attribute-Host relation. Ex. Age-Person, Color-Flower, ...

(5) Component-Entire relation. Ex. Leg-Body, Door-House, ...

(6) Material-Product relation. Ex. Rice-Wine, Sand-Glass, ...

(7) Event-Agent relation. Ex. Cure-Doctor, Build-Worker, ...

Figure 1 is an example showing relations among some objects (in rectangle shape) and actions (in round-rectangle shape).

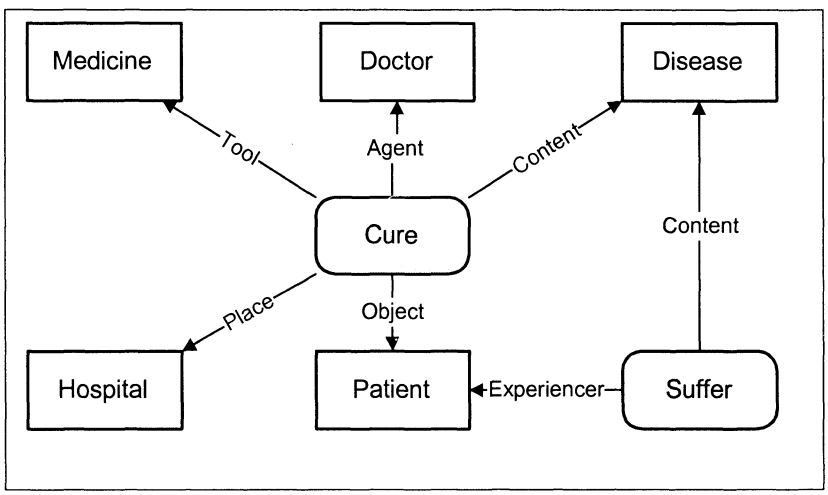

Figure 1. Example of relations among objects.

3. Analysis of the Sentences

3.1 Word Segmentation

Unlike western language such as English, Chinese sentence is composed by characters. Since How-net is a word-based knowledge base, each sentence has to be segmented into word-sequence for further processing. Word segmentation can be done by several approaches. The simplest one is to use the greedy algorithm. This algorithm treats the input sentence as a large string, and then shrinks the string gradually to check whether the shrunk substring is a lexical term. The segmentation results may be different if the shrinking directions are different. For example, the sentence "*大學生活很快樂 (The life in college is very cheerful)*" may be segmented into "*大學生(college student) 活(live) 很(very) 快樂(cheerful)*" if the

string is shrunk from right to left. However, it may also be segmented into "*大學(college) 生活(life) 很(very) 快樂(cheerful)*", which is the correct segmentation, if shrunk from left to right.

In this paper, we employ the greedy algorithm for both shrinking directions. If the segmentation results are the same for both directions, then we are done. If not, the bigram scores of both word-sequences are calculated to determine the correct segmentation. This bigram information is trained using a large amount of news corpus. Since the bigram score calculation can be found in many lectures, it is abbreviated here.

Note that if the segmentation result contains dangling single-characters, these single-characters may be further combined with other words/characters to form a compound word using word-formation rules. For example, one of the rules is to deal with the Chinese naming principle. We try to combine the characters sequence "*李 金 男*" into one word "*李金男 (Lee Kin-Nan)*" since "*李 (Lee)*" is a Chinese surname. Another example is that the word sequence "*乾 (dry) 乾淨(clean) 淨(neat)*" is combined into the compound word "*乾乾淨淨 (clean)*" since it satisfy the adjective formation rule.

3.2 Extracting Major Components of the Sentence

Based on our previous study [2], we think that to understand a sentence is to know the answers of ***5W***, i.e., *who, what, when, where, and which*. In this paper, we define these answers as the major components of sentence. The major components represent the agent, theme, time, place, and event conveyed in the sentence and they contain essential information to understand the sentence.

For example, to analyze the sentence "*李金男打破教室的窗戶 (Lee Kin-Nan broke the window of the classroom)*", the sentence is first segmented into the word sequence: "*李金男 (Lee Kin-Nan) 打破(broke) 教室(classroom) 的(De) 窗戶(window)*". Then, the major components are determined by the part-of-speech tags found in How-net. In this example, we

have the words: Lee Kin-Nan (agent: n. name of people); break (event: v. break); classroom (place: n. place to conduct a lesson); window (theme: n. the hole in the wall to illuminate the house).

The major components in such a simple sentence can be easily determined. However, for some complex sentences, we may need extra information to properly catch the meaning of the sentences. So, the attributes of each component are attached to carry more information. Each component may have none or more than one attribute. The attribute of agent, for example, can be the height, weight, age, color, and so on.

After extracting the major components of each sentence, the semantic table that consists of the major components and their attributes can be constructed sentence by sentence. Figure 2 shows an example text chosen from the primary-school textbook. The corresponding semantic table is shown in Table 1. Note that the attribute of an event can be another event or sentence.

弟弟要上學了，他很高興。弟弟是一年級的新生，我上二年級了。
媽媽告訴我說：「今天是弟弟第一天上學，你要小心的帶他去。」
我和弟弟手牽手，一同去上學。
馬路上有很多車。有大汽車，也有小汽車。車子轉來轉去，弟弟很害怕。
我告訴弟弟說：「不要害怕，你看，前面是紅燈，不要走。等綠燈亮了，再過馬路。」
弟弟說：「好，我一定記住。」
(S_1) My brother is going to school, (S_2) He is very happy.
(S_3) He is in the first grade, (S_4) I am in the second grade.
(S_5) Mother says to me, (S_6) "today is the first day for your brother going to school, (S_7) you should take him to school carefully.
(S_8) I go to school with my brother hand in hand.
(S_9) There are many cars in the road: (S_{10}) big cars and small cars.
(S_{11}) The cars run far and near, (S_{12}) my brother is afraid of the running cars.
(S_{13}) I told my brother, (S_{14}) "Don't be afraid. (S_{15}) Look, the red light is on, (S_{16}) don't go. (S_{17}) Wait until the green light is on, (S_{18}) then cross the road."
(S_{19}) My brother says, (S_{20}) "OK, I will remember that."

Figure 2. The article chosen from the first grade textbook.

3.3 Constructing the Semantic Network

In addition to the semantic table, semantic network [3,4] also records the relationship of

each sentence. The nodes of the semantic network are the major components and their attributes of the sentences. The edges are the relations of major components and attributes. The semantic network is constructed gradually while reading each sentence of the text. When the major components and attributes are extracted, if the component already exists in the semantic network, the attributes and relations are updated using current ones. Otherwise, new nodes and edges are added into the semantic network. Figure 3 ~ Figure 6 illustrate the construction of semantic network when input the sentences S1~S4 in Figure 2. Note that new nodes and edges of the semantic network are painted in white color.

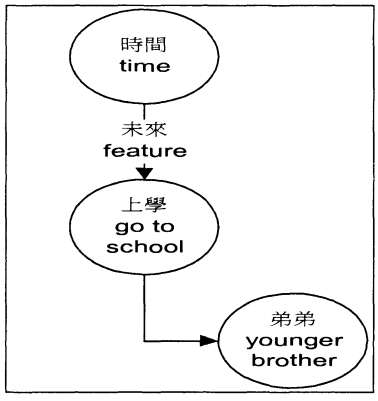

Figure 3. Semantic network construction, step 1:
Input the sentence "*弟弟要上學了(My brother is going to school)*"

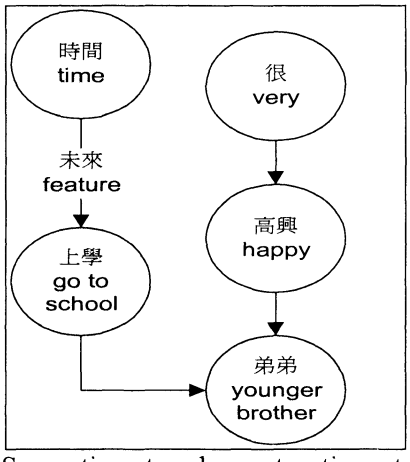

Figure 4. Semantic network construction, step 2:
Input the sentence "*他很高興(He is very happy)*"

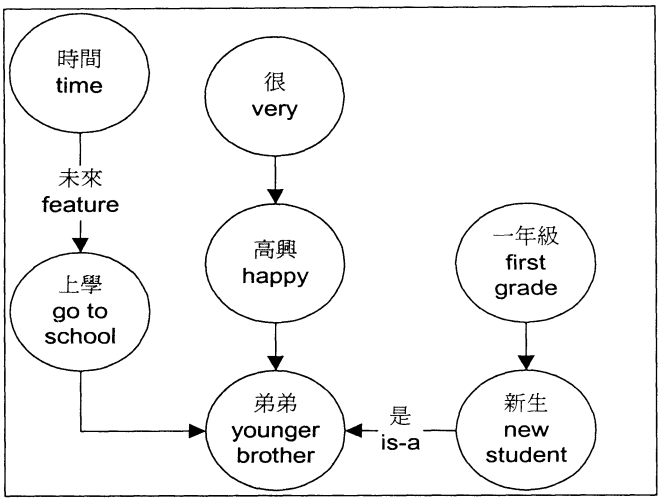

Figure 5. Semantic network construction, step 3:
Input the sentence "*弟弟是一年級的新生*"
(My brother is a new student in first grade)

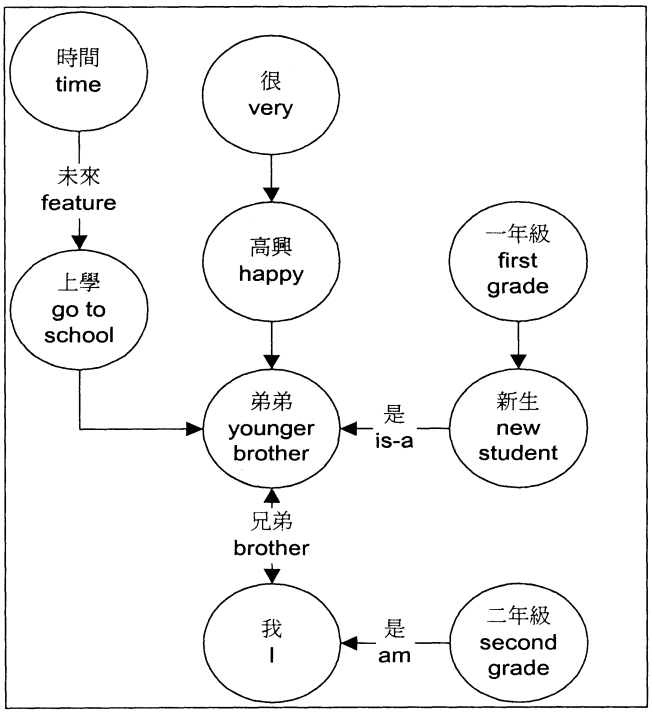

Figure 6. Semantic network construction, step 4:
Input the sentence "我上二年級了 (I am in second grade)"

4. Answering the questions

The questions we deal with are those in the exercises of primary-school textbook. Each question in the exercise provides several candidate answers for the reader to choose the right

one. Some examples of the questions are shown below.

Q1: 教室是（上課 / 上車）的地方。
The classroom is the place to (study / ride).

Q2: （紅燈 / 綠燈）亮了才可以過馬路。
If the (red / green) light is on, you can cross the road.

Q3: 勇敢就是（害怕 / 不害怕）。
Brave is to be (afraid / not afraid).

Q4: 弟弟是一年級的（新生 / 教室）。
My brother is a new (student / classroom) in the first grade.

Q5: 學校裡有許許多多的（學生 / 汽車）。
There are many (students / cars) in the school.

Q6: 我上（一年級 / 二年級）了。
I am in the (first / second) grade.

To answer the questions, we have developed two methods. The first one tries to find answer directly from the input text. The other one calculates the relation score among major components extracted from the questions and the article to find the proper answer.

4.1 Answering the Question Directly from the Text

Some questions of the primary-school textbook are easy to answer. The answer can be found by scanning the article to match the sentence that has the same components in question. For example, Q6 is a question of this kind. To answer question Q6, we first extract the components of the question, i.e., "我(I)". This component is used to active the corresponding subgraph in the semantic network. Then, the answers "一年級(first grade), 二年級(second grade)" are determined by choosing the one which matches the components best. In this case, the answer will be "二年級(second grade)".

4.2 Answering the Question According to the Similarity Measure

If the method described in Section 4.1 does not work, we use another approach to find the right answer. This approach checks the similarities of the major components in the question. The similarities are measured by the relation derived from the How-net. Two kinds of similarities, i.e., relational and hierarchical similarity, are used to calculate the overall

similarity measure.

4.2.1 Measuring Relational Similarity

In the How-net knowledge base, each word has its corresponding definition(s). These definitions are recorded in How-net as the field "DEF". For example, the words "*學校(school)*", "*學生(student))*", and "*汽車(car)* have the DEFs as shown below. Note that a word may have several meanings, thus it has many DEFs.

```
W_C[49479]=學校 // word index
G_C = N   // part of speech tag
DEF[0]=InstitutePlace|場所 // definition of the word's concept
      Feature of noun: *engage|從事#affairs|事務
      Feature of verb: engage|從事{agent,content}
DEF[1]=@teach|教
      Feature of verb: teach|教{agent,content,target}/{agent,patient,ResultEvent}
DEF[2]=@study|學
      Feature of verb: study|學{agent,content,source}
DEF[3]=education|教育
      Feature of noun : education|教育
```

```
W_C[49451]=學生
G_C = N
DEF[0]=human|人
      Feature of noun: N.1.1.1.1.1.1!name|姓名!wisdom|智慧!ability|能力!occupation|
      職位*act|行動
DEF[1]=*study|學
      Feature of verb : study|學{agent,content,source}
DEF[2]=education|教育
      Feature of noun : education|教育
```

```
W_C[34290]=汽車
G_C = N
DEF[0]=LandVehicle|車
      Feature of noun: N.1.1.1.2.2.7.3.1#land|陸地*VehicleGo|駛
      Feature of verb: VehicleGo|駛{agent,direction,LocationIni,LocationFin}
```

The similarity of different words is measured by comparing their DEFs. Words with similar DEFs will result in higher similarity measure. For example, the relational similarity of the words "*學校(school)*", "*學生(student))*", and "*汽車(car)*" are calculated as:

Relational Similarity (No.49479 '學校' & No.49451 '學生') = 12
Relational Similarity (No.49479 '學校' & No.34290 '汽車') = 2
Relational Similarity (No.49451 '學生' & No.34290 '汽車') = 3

From above result, we can conclude that the word *"學校(school)"* is more similar to *"學生(student)"* than *"汽車(car)"*.

4.2.2 Measuring Hierarchical Similarity

How-net also specifies the hierarchical relation of entities like a tree structure as shown in Figure 7. In this paper, the hierarchical similarity of two objects is defined as the shortest-path distance between these objects. For example, the similarity measure of "human" and "animal" is two, as shown is the upper-right part of Figure 7.

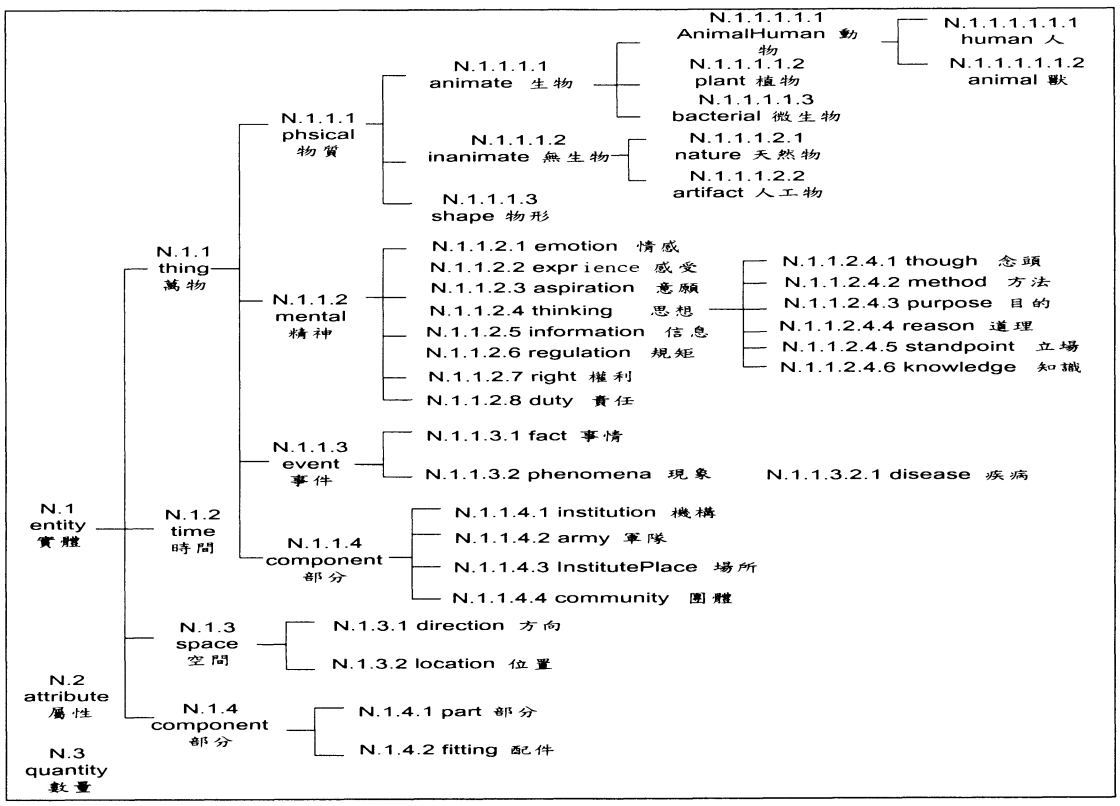

Figure 7. The hierarchical relation of objects.

4.2.3 Answering the Question According to the Similarity Measure

With the similarity measures obtained by the above methods (relational and hierarchical), we are able to answer questions that can not be answered using the method in Section 4. For example, the similarity measure of major components in Q4 and Q5 can be calculated as below.

Q4: 弟弟是一年級的 (新生 / 教室)。
My brother is a (new student / classroom) in the first grade.
Major components: "弟弟(brother)", "新生(new student)", "教室(classroom)"

Similarity measure (No.10017'弟弟' & No.48295'新生') = 10
Similarity measure (No.10017'弟弟' & No.22191'教室') = 0

The answer is : "新生(new student)"

Q5: 學校裡有許許多多的（學生 / 汽車）。
There are many (students / cars) in the school.
Major components: "學校(school)","學生(student)","汽車(car)"

Relational Similarity (No.49479 '學校' & No.49451 '學生') = 12
Relational Similarity (No.49479 '學校' & No.34290 '汽車') = 2

The answer is : "學生(student)"

5. Discussion and Conclusion

As the technique of speech processing improved, the success of speech applications, such as human-machine interactive system and spoken dialogue system, now depends on the success of natural language processing. Currently, few significant results are presented for Chinese NLP due to the lack of good knowledge base. In this paper, we employ the How-net to be our knowledge base for the domain unconstrained language understanding. Methods to answer the questions of the input article are proposed. Word segmentation, major component extracting, and semantic network construction of the input article enable us to derive the answer of the questions in the primary-school textbook.

Although the How-net plays an important role in our system, we find that some adaptations should be made to achieve better performance for the Chinese language understanding in Taiwan. The major shortcoming is that How-net is built in the Mainland China, thus some wording habits are different from that in Taiwan. For instance, the word "software" means "軟體" in Taiwan, however, it is written as "軟件" in Mainland China. This kind of word-disagreement should be solved before the How-net can be applied deeply in the natural language processing of Chinese in Taiwan. We are currently enhancing the How-net by using the word dictionary built by the Academic Sinica of Taiwan [5].

Our future research is to apply the techniques described in this paper to the task of web understanding.

6. References

[1] Zhen-Dong Dong, The How-net web site, http://www.how-net.com/

[2] Jhing-Fa Wang, Hsien-Chang Wang, Chin-Nan Lee and Mao-Sheng Hung, "On the Construction of the Knowledge Base for the Domain Unconstrained Spoken Dialogue System", Oriental COCOSDA 1999, pp.133-136.

[3] J. Allen, Natural Language Understanding, Second Edition, The Benjamin/Cummings Publishing Company, Inc. 1995.

[4] J. Rumbaugh, et. al. Object-Oriented Modeling and Design, Prentice Hall, 1991.

[5] 中文詞知識庫小組, "中文詞類分析 (三版)", 中央研究院資訊科學研究所, 1993.

Sentence	classification	Agent	attribute	Event		attribute	Theme	attribute	Time		attrib.	Place			attrib.
				Trans.	Intrans.				period	point		at	to	from	
S1	Declarative	Brother			Go to school										
S2	Declarative	He	Very happy												
S3	Declarative	Brother	New student, In first grade												
S4	Declarative	I	In 2nd grade												
S5	Subordination	Mother		Tell		S6, S7	Me								
S6	Declarative	Brother			Go to school	First time			(period)Today			(at) School			
S7	Imperative	You		Take		Carefully	Him								
S8	Declarative	I, Brother	Hand in hand		Go to school										
S9	Declarative						Cars	Many				(at) Road			
S10	Declarative						Cars	Big and small							
S11	Declarative				Run	Near and Far	Cars								
S12	Declarative	Brother			Is	Afraid									
S13	Subordination	I		Tell		S14, S15, S16, S17,	Brother								
S14	Imperative			Be	Afraid	Not									
S15	Declarative	You		Look			Light	Red, on							
S16	Imperative				Go	Not									
S17	Imperative				Wait		Light	Green, On							
S18	Adverbial clause				Cross							Road			
S19	Declarative	Brother		Say		S20									
S20	Affirmative	I		Remember			S13								

Table 1. The semantic-slot table for the text in Figure 2