Domain-unconstrained language understanding
Based on CKIP-AutoTag, How-net, and ART

WANG Jhing-fa, WANG Hsien-chang, LEE Kin-nan, and HUANG Chieh-yi

Department of Computer Science and Information Engineering, Tainan

Department of Electrical Engineering, Tainan

ABSTRACT

In this paper, we propose a method for domain unconstrained language understanding based on the linguistic toolsets CKIP-AutoTag and How-net, and the devised Acting Role Table (ART).

In our approach, the analysis of article is performed sentence by sentence. For each sentence in the article, word segmentation is first performed using CKIP-AutoTag. Next, the semantics of each sentence is represented in the devised Acting Role Table (ART). The ART consists of acting roles of the sentences, i.e., action, agent, instrument, theme, location, and time, together with their associated modifiers. In this step, we use verb-driven syntax analysis to determine the acting roles in the sentences; and use semantic analysis to constrain the acting roles based on the features defined in How-net. Finally, the semantic network is constructed to record the relationship of each sentence.

To test whether our approach for text understanding is feasible, an auto reading comprehension system is built trying to answer the exercises of the primary school textbook. The exercises contain many questions related to the article. Seven textbooks ranged from third-grade to ninth-grade are used as our testing articles. Most of the questions in the exercises can be answered by our system if the answers can be derived from the article.

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 approach we used integrated the CKIP-AutoTag [6] and the How-net knowledge base [1]. They are free linguistic toolkits and can be downloaded via the Internet.

The CKIP AutoTag, which is developed by the Chinese Knowledge Information Processing Group in Academia Sinica, Taipei, is able to segment the input Chinese sentence and output the POS (part of speech) of each word. The following example shows how a Chinese sentence is tagged by the AutoTag. The detail POS listing can be found in [5].

Input sentence: “王教授在大學教自然語言課程” (Professor WANG teach natural language in college.)

AutoTag output: “王(Nb) 教授(Na) 在(P) 大學(Nc) 教(TC) 自然(Na) 語言(Na) 課程(Na)”

On the other hand, the How-net knowledge base is developed by professor DONG Zhen-dong et al. in Mainland China; it is a common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese. How-net also contains the relationships of words such as synonymy (many words for the same concept), polysemy (many concepts for the same word), antonym (words with converse meaning), and so on. These information can be used to measure the similarity between two Chinese words.

Based on the CKIP-AutoTag and 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, acting roles (agent, event, time, place, and theme) conveyed in the sentence are extracted. Finally, the Acting Role Table (ART) 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 ART and semantic network. Answers of the questions are derived by measuring the likelihood of the acting roles.

This paper is organized as follows. In Section 2, the method to analyze sentences is described. The experiments are described in Section 3. Section 4 gives a conclusion remark of this paper.

2. ARTICLE ANALYSIS

For an article to be analyzed, we are interested in the agent, theme, event, location, and time, etc., i.e., the acting roles. In our approach, we use the acting role table to store the semantic information of an article. The flowchart of article analysis is shown in Figure 1.
2.1 Building the Acting Role Table (ART)

The acting roles represent the agent, theme, time, place, and event conveyed in the sentence which are essential information to understand the sentence.

For example, the sentence "李金男打破教室的窗户 (Lee Kin-Nan broke the window of the classroom)" is first analyzed to determine the agent, theme, time, place, and event. After word segmentation, we have the word sequence "李金男 (Lee Kin-Nan) 打破 (broke) 教室 (classroom) 的 (De) 窗户 (window)". 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 part of speech tags can be obtained by the CKIP AutoTag.

The acting roles 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, we attach the attributes to each acting role to carry more information. Each acting role 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 acting roles of each sentence, the acting role table (ART) that consists of the acting roles and their attributes, can be constructed sentence by sentence. Figure 2 shows an example text chosen from the primary-school textbook. The relative ART is shown in Table 1. Note that the attribute of an event can be another event or sentence.

2.2 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 acting roles and their attributes of the sentences. The edges are the relations of acting roles and attributes. The semantic network is constructed gradually while reading each sentence of the text. When the acting roles and attributes are extracted, if the acting role 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)

### 2.3 Lexical Semantic Similarity Measurement

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. We show two similarity score measurement below, the detailed calculation of the similarity score can be found in [7].

![Image of a diagram](image-url)

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

#### 2.3.1 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 "學生(student)" has the DEFs as shown below. Note that a word may have several meanings, thus it has many DEFs.

```
W_C[49451]=學生
G C = N
DEF[0]= human|Feature of noun: N.1.1.1.1.1!name|姓名|wisdom|智慧|ability|能力|occupation|職位
*act
DEF[1]= *study|學|Feature of verb: study|學
{agent, content, source}
```

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:

```
Relational Similarity ("學校(school)" & "學生(student)") >
Relational Similarity ("學生(student)" & "汽車(car)")
```

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

#### 2.3.2 Hierarchical similarity

How-net also specifies the hierarchical relation of entities. Hierarchical relations of noun and verb are all recorded in How-net. Figure 7 shows a partial hierarchical relation of noun. In this paper, the hierarchical similarity of two
objects is defined as the shortest-path distance between these objects.

3. EXPERIMENTS

To test whether our approach for text understanding is feasible, an auto reading comprehension system is built trying to answer the exercises of the primary school textbook. The exercises contain many questions related to the article. In our experiments, eight textbooks ranged from third-grade to ninth-grade are used as our testing articles. The topics of the article are all different, i.e., domain unconstrained. For each article, CKIP AutoTag is first applied to segment the sentences. Then, the ART of the article is built by the method described above.

Method to find the answer is described below. The interrogative of the question is first analyzed to know the intention. Then, process similar to that of article is performed to build the partial-acting-role-table (PART) of the question. The PART is then compared, based on the features found in How-net and the semantic network, with the full ART to choose the best one as the answer. The comparison uses the semantic similarity between the acting roles. The similarity measure is performed by the approaches described in Section 2.3. Table 2 shows the experimental results of the auto reading comprehension system.

4. CONCLUSION

As a conclusion, this paper proposes a good direction for language understanding based on semantics. The experimental result encourages us that the proposed ART is suitable for domain-unconstrained language understanding. This paper also provides a good demonstration for the possible utilization of the How-net knowledge base. Since How-net is now available only in plaintext format, we have developed many toolsets such as MFC (Microsoft Fundamental Class) classes and APIs (Application Program Interface) for further researches based on How-net to promote the using and sharing of the free linguistic resources. Readers interested in this work can refer to the master thesis of the authors to find more detailed information in [8].

<table>
<thead>
<tr>
<th>Primary School Textbook</th>
<th>Article tested</th>
<th>Number of questions</th>
<th>Correctly answered questions</th>
<th>Correct rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book 3</td>
<td>14</td>
<td>60</td>
<td>30</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>Book 4</td>
<td>24</td>
<td>101</td>
<td>78</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td>Book 5</td>
<td>18</td>
<td>74</td>
<td>28</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.45</td>
</tr>
<tr>
<td>Book 6</td>
<td>18</td>
<td>104</td>
<td>52</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.64</td>
</tr>
<tr>
<td>Book 7</td>
<td>19</td>
<td>75</td>
<td>42</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.64</td>
</tr>
<tr>
<td>Book 8</td>
<td>21</td>
<td>83</td>
<td>43</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.62</td>
</tr>
<tr>
<td>Book 9</td>
<td>20</td>
<td>86</td>
<td>53</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.72</td>
</tr>
<tr>
<td>Total</td>
<td>134</td>
<td>583</td>
<td>326</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>19</td>
<td>83</td>
<td>47</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.65</td>
</tr>
</tbody>
</table>

Table 2. The correct rate of the auto reading comprehension system.

5. REFERENCES