Improving Information Extraction using Knowledge Model

Yue Chen

School of Computer Science and Technology, Beijing Institute of Technology,
No.5, Zhongguancun Nandajie Haidian District, Beijing 100081, P. R. China
chenyue@bit.edu.cn

Abstract. This paper describes the work on automated Information Extraction that accepts arbitrary text and extracts information from the text. A new approach to implement Information Extraction system is proposed in this paper. Firstly, the article will be decomposed according to paragraph, sentence and phrase. Every sentence will be compared with the knowledge node, and then append the information extracted to the knowledge model. Finally, the answers are generated to the questions about the input text. With the experimental corpus the accuracy rate of knowledge matching is 63.5%, and accuracy rate of question answering is 65.0% with the system knowledge model.

Keywords: Natural Language Processing, Information Extraction, knowledge model, knowledge representation.

1 Introduction

This paper proposes an approach to establish an Information Extraction (IE) system by using knowledge model obtained from articles. A simple knowledge representation method, where knowledge is stored in a graph structure, is adopted for this system. IE system intends to test the ability of computer to extract the information from an article, and the main method of testing accuracy is to generate the answers automatically with the input text and questions. The objects of extraction in sentence include attributes, relationships and events. In the field of Natural Language Processing (NLP), the theory of knowledge model is conducive to a number of research directions, and many areas in NLP need to access the knowledge from article, which make the study of IE system has its important significance.

Questions can be formalized to query that related to the knowledge structure using the technology of knowledge model in QA system (McKeown, 1979). Machine translation (MT) system will benefit from knowledge model because it can generate more easily translated sentences for the input articles (Mitamura and Nyberg, 2001). IE System mainly use template method, while the knowledge model maintains the knowledge node which includes the characteristics of knowledge. Knowledge model can improve the accuracy and robustness of IE system (Shinyama *et al.*, 2002). In information Retrieval (IR), the document stored in database can be changed into knowledge structure with knowledge model technology so that improve the performance as well as precision of search result (Zukerman and Raskutti, 2002). In Summarization, the center of an article can be extract effectively by understanding the knowledge of article with knowledge model. It can extract a summary from the chapter, paragraph, or sentence level for different needs, and rely on natural language generation to get a more accurate digest (McKeown *et al.*, 2002). In paraphrasing, since that the process of paraphrasing can be considered as translation between the same languages, the part of natural language generation in knowledge node will improve the accuracy of paraphrasing through making the sentences and articles into knowledge representation (Iordanskaja *et al.*, 2002).

2 Related Work

The research of knowledge model technology originated in the seventies of last century, the introduction of this technology aims at trying to make a computer answer a number of questions according to an article. These questions designed manually, and there are clear answers to some questions in the original text, while some questions need knowledge or reasoning in order to get the results from the context. Hirschman (1999) studies on reading comprehension system using technology of natural language understanding. MITRE Corporation builds the Remedia corpus for reading comprehension system evaluation. Remedia contains 115 English articles, and divided into different grades according to the degree of difficulty. Each article contain 20 sentences and five questions (the type are who, where, when, what, why) on average, also defines the HumSent accuracy which means the proportion of the number of right answer for the questions. In Remedia corpus named entities, trunk of sentences and pronouns information have been tagged. Sentence is the unit in the corpus with a unique number, and each question has been marked the correct answer sentence.

Riloff and Thelen (2000) develop a rule-based reading comprehension system, and designed a large number of rules of thumb to determine the similarity between a candidate sentence and question. Ng (2000) use features to train a classification model with C4.5 learning algorithm. The feature includes “whether the sentence is the title”, “whether the sentence contains the names of persons, organization name, place name, date, time”, and “the number of words in sentence that matching with the question”.

Dagan and Glickman (2004) start the research from the relationship between the content of articles and semantic. Textual entailment is defined as a binary relation between a natural language text T and a hypothesis H. If H can be reason out from T, then T implies H. Recognizing Textual Entailment (RTE) begins in 2005 and has held five sessions so far (RTE1 - RTE5). The main method is to calculate the similarity between T and H, such as vocabulary similarity, syntactic similarity and so on (Ferrandez *et al.*, 2007).

Researchers in Chinese Academy of Sciences develop a QA system on people relationship (Wang *et al.*, 2007). Through logical reasoning mechanism, the system output description about people relationship using the knowledge from articles.

Knowledge model techniques is discussed in this paper, approaches on building knowledge structure mainly include knowledge-matching, knowledge markup, semantic reasoning of vocabulary. This article will introduce the main process of reading comprehension system first, and then describe the experiment with results analysis, and finally is the conclusion.

3 Knowledge Model in IE System

3.1 Knowledge Model

The IE system describes the contents of the article through knowledge model. So the sentence in the article is necessary to corresponds to a knowledge node in the knowledge Model. Knowledge node is the formal description of knowledge, it consists of knowledge feature, tag set and knowledge description. Knowledge structure is composed of a collection of knowledge nodes.

Theorem 1: Suppose F is the set of knowledge feature, M is the set of knowledge tag, the knowledge node Ks is a triple Ks=<F, M, G> where:

F is a subset of knowledge feature,

M is a subset of knowledge tag,

G is the description of the knowledge node.

Knowledge model is designed for storing information, unlike other methods that store information as a record in database, knowledge model is like the brain of human beings. The knowledge model is designed by the data structure of direct multi-graph. The Knowledge model is shown in Fig.1.

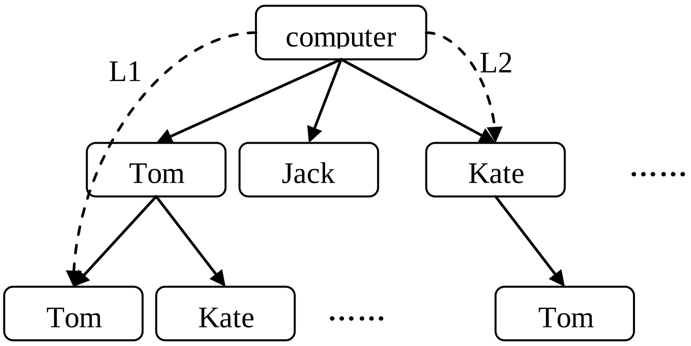

Figure 1: Knowledge model

As shown in Fig.1. The root node of the graph is the system itself, contains some properties used to describe the system. The child nodes of root are the knowledge nodes which are its views on other things. Also there are some properties contained in the child nodes. Every child node is the view of its parent. For example, the line "L1" which is connecting "computer → Tom → Tom" in the graph means that what is the view for the root that "Tom" view himself. And the line "L2" which is connecting "computer→ Kate" means that what is the view of "Kate" for "computer".

The system can only understand the sentences that match to knowledge node in knowledge structure, so the knowledge node need to include a description of knowledge feature. Knowledge feature can be characterized as many forms, the particle size of feature can be a single character, or a word. Take knowledge of "COLOR" for example. "COLOR" refers to the property of color of object; the corresponding knowledge features include "color", "red", "black", "blue" and so on. A sentence that response to the age property clearly will be included a few elements of the feature set of "COLOR". For example, "The color of the flower is red." contains the feature of "color", "red", while the phrase "What is the color of the flower?" contains the characteristics of "color".

In IE system we use knowledge node to indicate the properties of object, the relationship between objects, or event. The tag set in the knowledge nodes is used to reflect objects, attributes, relationships and other information. Take knowledge "AGE" as example. Knowledge feature in "AGE" is consisting of "<NAME>", "<AGE>". If a sentence represents the age property of object according to knowledge feature matching, the knowledge tag of "AGE" will be used to mark the sentence. So the sentence "1980年9月12日，姚明出生于上海市(Yao was burn in Shanghai in September 12, 1980)" will be marked as "AGE:<name>姚明 (Yao) </name><age>30</age>".

When users need the description information or the answers of questions about knowledge, the system will reply these queries with some sentences. So it is necessary to include the information of natural language generation. The part of NLG in knowledge node use manual rules and fixed templates, coupled with knowledge tag, to generate the answer as output. For example, knowledge description of "AGE" is "<NAME>[<TIME>]<NUMBER>岁(year)". Assuming that user query the question of "姚明多大岁数了? (How old is Yao?)", the output is "姚明今年 30 岁(Yao is 30 years old)."

The knowledge structure of "AGE" is shown in Fig. 2.

KNOWLEDGE NODE
Name:AGE
Feature: 年龄(year) 岁(age) NUMBER TIME
Tag: <name></name> <age></age>
Description: <name>[<time>]<num>岁(age)

Figure 2: Style of knowledge node.

3.2 Similarity between Sentence and Knowledge Node

The IE system analyzes the input sentence to find the collection of related knowledge nodes. After the process of marking the sentence with knowledge tag, the formatted statement will be getting which can be handled by system. In this process, we need to settle the following problem:

Finding features from the input sentence. After the process of word segmentation the sentence can be considered as a linear sequence or words. Some of the words are the feature of the sentence due to their ability to distinguish the semantic information, while there are still some words due to non-semantic information affect the system uptime.

Multiple results are returned when calculate the knowledge nodes that correspond to the sentence. The sentence commonly corresponds to a knowledge node when dealing with fewer areas; while the sentence corresponds to some knowledge nodes when increase the areas. It is unclearly which is the right knowledge node unless make further judgments.

Mark the sentence with knowledge tags. In the process of sentence analysis, the input sentence should be marked according to the knowledge tags in knowledge node if the sentence is related with certain, or uncertain knowledge node.

Therefore, in order to solve these problems, it is necessary to enhance the use of sentence features while dealing with the similarity judgment with knowledge nodes, the methods are as follows:

Only to extract the keywords in the input sentence, that is, just find out the words exists in F, the set of knowledge feature, and remove the words which is not feature as stop words.

Add the knowledge node into a temporary collection if the similarity rate is greater than the threshold when calculate the similarity between the feature vector of knowledge node and sentence with key words extraction. The sentence will be marked according to the set of knowledge tag in every node in the temporary collection. It will be the right knowledge node that matching the sentence if marked the sentence successfully. If all the knowledge nodes in temporary collection failed to mark on the sentence, that is, the sentence cannot be identified in this system.

The similarity algorithm is shown as follows.

```
Input: sentence
Output: Knowledge representation of sentence
S → w1w2…wn
for i=1 to length(Ks)
  do
  sim = Similarity (S, Ksi)
  if ( sim > threshold ) then
    if ( tag(S) ==ture ) then
      return Ksi.Name + tag
    end if
  end if
end for
```

```
return NULL
```

3.3 System Knowledge Model

Articles are expressed in formal semantics in IE system. An articles can be decomposed into four levels, including articles, paragraphs, sentences, and phrases. The system starts from the phrase level and generates semantic knowledge of phrase, and then generates the knowledge of the sentence, paragraph, and article using recursive method. Whenever obtain certain knowledge, it will be added to the system knowledge model which is build as a direct multi-graph. Initially, there is only one graph node, store the information that the system itself, which can be considered as a central node. When there are new knowledge come, it will update the map data, using multiple graph structure can handle the multiple relationships between objects. The algorithm of adding knowledge to system model is shown as follows.

```
Input: Collection of knowledge representation.
Output: The updated model of systematic knowledge
for i=0 to size(knowledge collection)
  do
    get objects Oi in knowledge Ki
    if( Oi ∉ model ) then
       add Oi to model
    end if
    if ( type( Ki ) =  attribute ) then
       add attribute to Oi
    end if
    if ( type(Ki ) =  relationship ) then
       add relationship to Oi
    end if
    end for
    return model
```

4 Experiment

4.1 Experiment Setting

PFR corpus is used for our experiment, which is made of People’s daily with segmentation and part of speech (POS) tagging in the first half of 1998. Each word in the corpus has a POS tag. Besides the 26 basic parts of speech marks, the corpus has increased some proper nouns (names, places, organizations, and other proper nouns) for the perspective of applications. Now the number of mark in this corpus is more than 40 with linguistic marks. The experiment corpus consists of 400 sentences which are extracted manually from the PFR corpus, while the process of anaphora resolution for pronouns has been achieved taking advantage of corpus tags. Every sentence has been given its knowledge representation. For each sentence, we designed a question that used to ask the relevant knowledge of the contents. The type of knowledge includes object attributes, object relations and description of events. The distribution of knowledge is shown in the table 1.

Table 1: Distribution of knowledge.

Type of knowledge	Amount	Percentage
Attribute	240	60%
Relationships	80	20%
Events	80	20%

4.2 Experimental Results and Analysis

In IE system using knowledge model, the first step of sentence analysis is knowledge identification, which means that matching the input sentences with knowledge nodes. We take separate experiment using two kinds of feature representation of single character and word. The recognition precision rate indicates the proportion of sentences which can be identified clearly in the system in all input sentences.

$$\text{precision}=\frac{\text{the number of identified sentences}}{\text{the total number of sentences}} \tag{1}$$

The recognition precision of two kinds of knowledge representation in respective application is shown in Table 2 and Fig.3.

Table 2: Precision of sentence recognition.

Knowledge	Single character		Word	
	Sentence	Percentage	Sentence	percentage
Attribute	129	53.8%	163	67.9%
Relationships	41	51.3%	43	53.8%
Events	40	50.0%	48	60.0%
Totals	210	52.5%	254	63.5%

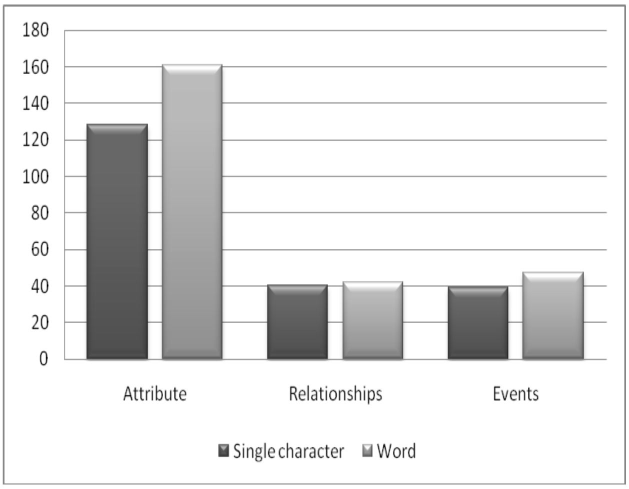

Figure 3: Precision of sentence recognition.

Experimental results show that the precision using words as particle size of characteristics is better than using single character, especially in dealing with the sentences belongs to presentation of object attributes. From the kind of knowledge type, the sentences of object attribute have highest precision, while the recognition precision of the sentences regarding object relationships is low.

Some characters may appear in a number of features of knowledge nodes when using single character as the particle size. For example, character “好(good)” appears in knowledge “AGE”, “HEIGHT”, “SOUND”, “COLOR” and other knowledge, which makes the sentence related to many knowledge nodes, so the system need more works to deal with the knowledge tags, and may lead to a recognition failure when the tag does not match with the meaning of the sentence. Table 3 shows the relationship between the characteristics of feature in knowledge node and the number of knowledge nodes which are relevant to the sentence.

Table 3: Sentence and knowledge nodes correlation table.

Nodes count	Number of sentence	
	Single character	word
0	0	0
1	31	195
2	33	88
3	80	61
4	78	39
>4	168	17

As shown in Table 3, every input sentence has at least one knowledge node that associated with it. And the sentence can be quickly positioned to the relevant knowledge node when using word as the feature.

After the analysis of sentence, system will add the representation of knowledge to the system knowledge model. We use a number of questions to test the performance of IE system. Different from most of IE system, the output of our system is precise answer for each question, while other systems return the original sentence or a sentence number from the article. Because of the sentences that without correct identification, we only use the sentences that identified correct in the system for the experiment.

Table 4: Precision of question answering.

Knowledge	Single character		Word	
	Sentence	Percentage	Sentence	percentage
Attribute	77	59.7%	113	69.3%
Relationships	20	48.8%	23	53.5%
Events	22	55.0%	29	60.4%
Totals	119	56.7%	165	65.0%

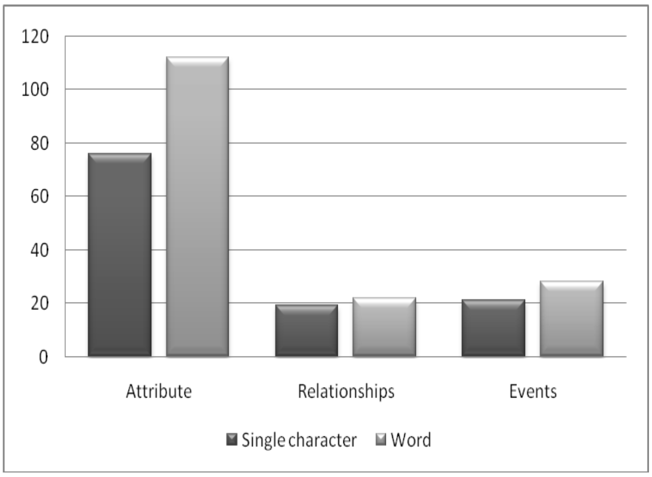

Figure 3: Precision of question answering.

As shown in Table 4, the precision is 65.0% when using words as the particle size of feature, which is higher than the method that using single character. Because of the analysis error of the sentence, system does not reply the correct answer.

To sum up, the method that using words as the particle size of feature get higher precision than using single character in knowledge matching and question answering in IE system. The method with characteristic of words can be quickly positioned to the knowledge node. Segmentation error is the main reason for the precision. And the step of segmentation requires large-scale lexicon and the corresponding program support to deal with the process which needs segmentation, such as the collection of knowledge features, article understanding and question analysis.

5 Conclusions and Future Work

This paper studies the IE system based on knowledge model. Through the analysis of sentences it can obtains the corresponding form of knowledge representation. And the system knowledge model uses a graph structure to store the knowledge that generated by sentence analysis. PFR corpus is used in our experiment, from which we extracted some sentences to form the text corpus, the type of knowledge includes object attributes, relationships, and descriptions of events. Experimental results show that the knowledge model can represent the article in semantic level.

References

Dagan I and U. Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. *Proc. of the PASCAL.*

Ferrandez O, D. Micol, R. Munoz and M. Palomar. A perspective-based approach for solving textual entailment recognition. *Association for Computational Linguistics*, pp. 66-71.

Hirschman L., M. Light, E. Breck, and D. John. 1999. Deep Read: A Reading Comprehension System. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* pp. 325-332.

Iordanskaja L., R. Kittredge and A. Polguère. 1991. Lexical selection and paraphrase in a meaning-text generation model. *Natural Language Generation in Artificial Intelligence and Computational Linguistics*. pp. 293-312.

McKeown K. R. 1979. Paraphrasing using given and new information in a question-answer system. *Association for Computational Linguistics*, pp. 67-72

McKeown K. R., R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman and S. Sigelman. 2002. Tracking and summarizing news on a daily basis with Columbia's newsblaster. *Morgan Kaufmann Publishers Inc.*, pp. 280-285.

Mitamura T, E. 2001. Automatic rewriting for controlled language translation. *In: Proc. of the NLPRS*. pp. 1-12.

Ng H. T., L. H. Teo, J. L. P. Kwan. 2000. A Machine Learning Approach to Answering Questions for Reading Comprehension Tests. *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* . pp. 124-132.

Riloff E. and M. Thelen. 2000. A Rule-based Question Answering System for Reading Comprehension Tests. *Proceedings of ANL P/NAACL 2000 Workshop on Reading Comprehension Test s as Evaluation for computer-Based Language Understanding Systems*. pp. 13-19.

Shinyama Y, S. Sekine and K. Sudo. 2002. Automatic paraphrase acquisition from news articles. *Morgan Kaufmann Publishers Inc.*, pp. 40-46.

Wang S. X., Q. Liu and S. Bai. 2003. An expert system about human relationship question answering. *Journal of Guangxi Normal University(Natural Science)*, 21 (1) , 31-36.

Zukerman I and B. Raskutti. Lexical query paraphrasing for document retrieval. 2002. *Association for Computational Linguistics*, pp. 1-7.