A Semantic Parsing Model Applied into Search Engine

LI Ying[1, a], WEI Xiang Feng[2,b] and CHI Yu Huan[2,c]

[1]Academy of Armored Force Engineering, China

[2]Institute of Acoustics, Chinese Academy of Sciences, China

[a]lypublic@hotmail.com, [b]wxf@mail.ioa.ac.cn , [c]cyh@mail.ioa.ac.cn

Keywords: Natural language processing. Semantic parsing model. Conceptual Structures. The theory of HNC. Search Engine.

Abstract. It is a tough work to select accurate web pages for search engine. This paper applied a semantic parsing model into search engine by comparing the conceptual structures between user's input and text. The words, sentences and sentence groups can be mapped into conceptual symbols, semantic categories, and contextual elements based on the theory of Hierarchical Network of Concepts. The approach will improve the intelligence and flexibility of the search engine in future.

Introduction

With the development of computer application science, many new types of computer are unceasingly developing. The functions of computer become stronger than before, and the speed of computer is higher and higher. Although the computer is widely utilized in many professional domains, it is still a tough task for computers to process natural language as easy as human brain.

Natural language processing involves three dimensions of language: grammar, semantics and pragmatics. Many mature linguistic grammar theories did not work well in computer science because of massive fuzzy semantics. Subsequently, the computer experts intensely propose the method of using statistical algorithms, such as massive stochastic processes and the information encoding theory. The statistical model achieved some excellent result in linguistic high frequency, but it is still not a good model for processing semantics of natural language.

The theory of Hierarchical Network of Concepts (HNC) [1,2] focuses on natural language understanding. Natural language understanding is considered as a mapping process from the actual language such as Chinese to the conceptual language space (see Fig. 1).

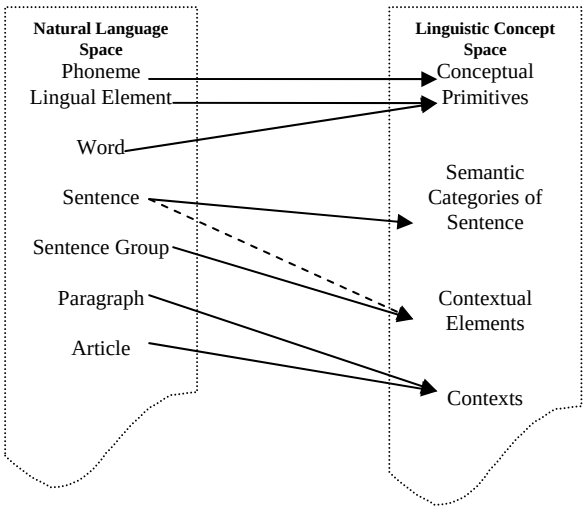

Figure 1. Natural language space mapped into linguistic concept space

Linguistic concept space is a redesigned symbolic system for classifying concepts. The contextual element which mapped from sentence group consists of three elements: domain, situation and background. Domain is the most essential. This paper will introduce how to parse sentences by computer program with the model of semantic categories (SC) of sentences and how to map sentence

Published by Atlantis Press, Paris, France.

groups into contextual elements (conceptual structures) in the conceptual language space. The semantic parsing model can be applied into search engine by comparing the conceptual structures and symbols between the user's input and the text of web pages.

The Semantic Parsing Model of a Sentence and a Text

Parsing a Sentence. A sentence can be mapped into a semantic category in the linguistic conceptual space. The semantic categories of sentences are defined as seven basic types: action, process, transfer, effect, relation, state and judgment. Each type has its sub-types. All types and their sub-types are the primitives of the semantic category of a sentence. Two primitives of semantic category of a sentence can be mixed into a compound type. All real sentences in language space can be mapped into the semantic categories of sentences.

A semantic category of a sentence (SC) is constructed of the main semantic chunks. There are four types of main semantic chunks: Actor(A), oBject(B), Content(C), eigen(E). The eigen semantic chunk is divided into two types. One is global eigen semantic chunk (Eg); the other is local eigen semantic chunk (El). Eg is the eigen semantic chunk at the top layer of a sentence like the head verb in a sentence. El is the eigen semantic chunk in clauses or nonfinite structures of a sentence. For example, the sentence '*He gave me a book*.' can be mapped into a sub-type of Transfer semantic category T0J (T0J=TA+T0+TB+TC). 'T0' means the concept of generic transfer (*gave*); 'TA' means the Actor of transfer (*He*); 'TB' means the oBject of transfer (*me*); 'TC' means the Content of transfer (*a book*).

In order to get the right SC of a sentence, the approach of 'Hypothesis-Test' is adopted. When parsing a sentence, the parser firstly hypothesized the EK according to some special concepts. The second, the SC of the sentence is hypothesized according to the EK. The third, the parser tests the hypothesized SC of the sentence according to the knowledge base in the computer. If all semantic chunks in the sentence are corresponded with the transcendental concepts in the knowledge base, the SC is confirmed. Otherwise, the SC is rejected. The knowledge base [3] stores the conceptual categories, the conceptual symbols, the SC codes, the format codes, the used frequency, and the restriction in syntax of the words used frequently.

Parsing a Text. After parsing the semantic conceptual structure of a sentence, it is possible to parse the semantic conceptual structure of a text. Because the text is made up of sentences, parsing the text must be constructed on the results of parsing sentences.

In the linguistic concept space, contexts can be mapped from articles in the natural language space. But what are the elements of the context? The essential main elements of context include: domain, situation, and background. Domain is the most basic information of the context in sentence group. It describes any activities about human being, including the actives of instinct, disaster, and state. The situation is the dynamic description of events. It indicates the semantic relationship between the participants of the event. The background describes the subjective and objective conditions of the event, such as time and location.

To extract the context elements of the text, it is necessary to cut the text into sentences and then analyze the semantic structures of the sentences. There are five steps in the process of analyzing a sentence. The first, the sentence must be segmented into words, Chinese characters, and other symbol units. The second, the words are mapped into the corresponding conceptual symbols according to the knowledge base. The third, the semantic chunks are perceived from the conceptual symbols according to the formal rules, including the EK (Eigen chunK) and GBK (General oBject chunK). The fourth, the code and expression of SC will be tested and verified according to the results of the perception in the third step. The fifth is analyzing the inner structures of the semantic chunks.

The domain information is implicated in the conceptual concepts of some words. It is convenient to extract the domain information from the semantic chunks which is made up of words. If more than one semantic chunk contain the domain information, which domain should be extracted? There is a principle to tackle this problem [4]. Because the station and significance of semantic chunks in a sentence are different, the domain information in different semantic chunks possesses different

Published by Atlantis Press, Paris, France.

priority. The priority is revealed as the following: Eg>El>C>B or A. The different domains also possess their own priorities. So if there are more than one domain to be selected in different sentences, the highest priority domain must be selected. After the domain is extracted, the SC expression with domain can also be confirmed. According to the framework of the SC expression with domain, the framework units of the situation can be constructed. The framework unit of the situation must be described as the following: EK name[EK conceptual symbol] | GBKm name[GBK conceptual symbol], the value of m is from 1 to 3. If the same EK or GBK appears in more than one sentence, the framework unit of the situation must be combined according to the sequence of appearance. In general the background is extracted from the supplemental semantic chunks, such as time supplemental semantic chunks, space supplemental semantic chunks and so on.

Application in Search Engine

In the early period of web search engine, the technique is classifying and indexing all kinds of websites, so users can visit some websites or web pages following the index. This kind of approach is inefficient and dependent on manual work. Therefore, the technique of key word is developed. But how to select the exact web page that the user wanted from numerous web pages and how to rank them according the user's demand. The PageRank algorithm [5] is a revolutionary approach to rank trillions of web pages in the internet. However, there are still many faults of the current web search engines: the accurate answer is concealed in large numbers of useless result pages, the user can't input a question naturally in sentence mode to search. The more intelligent web search engine should allow the user find out the exact answer through his question.

As we have mentioned above, a sentence such as a question can be parsed into semantic chunks and their conceptual symbols. A text can be cut into paragraphs or sentence groups, which can be mapped into the contextual elements in the conceptual language space. The contextual elements of a sentence also can be extracted through the method discussed in the foregoing section. So it is possible to match the semantic structure of the user's question and the semantic structure of the text when searching the exact text starting with a sentence or question. On condition that their conceptual structures are the same, the wanted answer text is discovered.

For example, the user wants to search with the sentence '*Please search diplomatic activities of any country*'. After extracting the context elements of the sentence, the domain, situation and background of the sentence are showed as the following:

{DOM//114： TBC }
{SIT//XT19[(9219,v93808)]|A[]|TBC[gv6500] TBCB[pj2] TBCC[ga114]}
{BAC// }

The domain(DOM) of the condition sentence is diplomacy(*114*). It is extracted from the content of the semantic chunk TBCC(*ga114*). The situation(SIT) is made up of EK(XT19), GBK1(A,omited), and GBK2(TBC,gv6500). The oBject of TBC is named TBCB. Its conceptual symbol '*pj2*' means '*country*'. The Content of TBC is named TBCC. Its conceptual symbol '*ga114*' means '*diplomatic activities*'. There is no background in the condition sentence.

If the text contains the following sentences, it will be an answer for the search condition because their DOMs are all *114*. But if the search condition is changed to 'Please search diplomatic activities of China', the last sentence will not be matched the condition, because the conceptual symbol of China is '*fpj2*01*'. It means that domain, situation and background are the semantic structures for search condition and the answer.

SENTENCE1: Chi Haotian met the army commander of South African.
{DOM//114： R0104 }
*{SIT// R0104 [va14;vc711]|RB1[(pea41/fpa018 +fpj2*01)]|*
*RB2 [(pea41/fpa018+ fpj2*635)] }*
{BAC// }
SENTENCE2: Hu Jintao met the chief of general staff of Russian.
{DOM//114： R0104 }

Published by Atlantis Press, Paris, France.

*{SIT// R0104 [va14;vc711]|RB1[(fpa1 +fpj2*01)]|RB2 [(fpa41e21 + fpj2*03)] }*
{BAC// }

SENTENCE3: Bush met the chief of general staff of Russian.
{DOM//114： R0104 }
*{SIT// R0104 [va14;vc711]|RB1[(fpa119+ fpj2*304)]|RB2 [(fpa41e21+fpj2*03)] }*
{BAC// }

Compared with the current technique of keyword search, the method based on semantic model and extracting has many advantages. The text is abstracted into three aspects which reflect the basic semantic information of the text. The semantic information is described as inter-correlative conceptual symbols. So it is easy extracted by computer program. The search answer is accurate with the conceptual relationship, not only the form of key word. The text search based on semantic model is more intelligent and flexible.

Summary

The approach of text search based on the semantic parsing model is a new pass to improve the accuracy of web search engine. A conceptual symbol system has been established to express all kinds of concepts so that the words can be mapped into conceptual symbols. Sentences can also be mapped into semantic structures. Text and sentence group can be mapped into the context elements which include domain, situation and background as the semantic framework. Therefore, the search result can be found by matching the conceptual structures between user's input and text. In this way, the algorithm of text search would be more intelligent and flexible.

The technology of analyzing the text based semantic framework has been applied into a practical system to filtrate the harmful information in the internet. The next work about improving the approach is to perfect the conceptual symbols, expand the knowledge data base and improve the matching algorithm in the conceptual structures.

References

[1] Z.Y. Huang. The Theory of Hierarchical Network of Concepts. Tsinghua University Press, Beijing, China, 1998.

[2] Z.Y. Huang. The Basic Theorem and Mathematic and Physical Expression in Lingual Concept Space. Ocean Press, Beijing, China, 2004.

[3] C.J. Miao. The Introduction of HNC Theory. Tsinghua University Press, Beijing, China, 2005.

[4] X.F. Wei, H.F. Zang, and Q. Zhang. A Clustering Approach of Conceptual Sentence Groups. Proceedings of The Seventh International Conference on Advanced Language Processing and Web Information Technology, 7(2008)38-42.

[5] Sergey Brin and Larry Page. The anatomy of a large-scale hypertextual WEB search engine. In: Proceedings of the Seventh International World Wide WEB Conference. Also on http://www-db. stanford.edu/~backrub/google.html. 1998.

Published by Atlantis Press, Paris, France.