IMPLEMENTATION PRINCIPLES OF KNOWLEDGE ACQUISITION FOR INTELLIGENT SYSTEM

Qian Longwei, Li Wenzu
Department of Intelligent Information Technology,
Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus
E-mail: qianlw1226@gmail.com, wzzggml@gmail.com

The article is dedicated to ontology-based implementation principles in component development of natural language interface, which is able to extract specific construction represented in internal language of intelligent system from unstructured data (i.e., natural language free texts). The main feature of implementation principles is the unified semantic model to integrate different aspects of linguistic knowledge (e.g., syntactic level, semantic level) and various problem solving models oriented on solving relation extraction problems.

Introduction

The natural language interface is a key component of any intelligent system. It plays the role of information exchange between users and intelligent systems, and realizes the conversion function between external languages and internal languages of intelligent systems. The natural language is one of the most frequently used external language. The natural language text is also the main representation form of unstructured data. For knowledge-based intelligent system, the knowledge base is one of the most key components of intelligent system, it's a finite information structure represented in internal language form. Unstructured data is not organized in any predefined way. It contains a wealth of knowledge of various subject domains. The automatic knowledge acquisition from unstructured data (i.e., natural language text) can effectively improve the efficiency of intelligent system development. The information extraction is the key technology to the implementation of automatic knowledge acquisition.

I. Direction of Information Extraction

From the perspective of knowledge acquisition, in this article information extraction is a technology that automatically extracts structured information represented in internal language of intelligent system from unstructured data. For processing of unstructured data, information extraction can be considered as a simplified natural language understanding technology. Information extraction from natural language texts can be divided into the following two directions:

- traditional information extraction focuses on closed domain, i.e., small homogeneous corpora;
- open information extraction, a new domain-independent knowledge discovery paradigm.

In the early stages, there are just limited natural language text about a certain of specific domains. The traditional information extraction methods just extract predefined relation sets including entities, relation and so on. With the advent of the Internet and Wikipedia, there are massive, multi-source and heterogeneous corpora such as various websites. The target of open information extraction is to extract various relation sets without requiring a pre-specified vocabulary to specify relation type from massive and heterogeneous text corpora, where target relations are unknown in advance. In order that information extraction methods are able to solve this problem, recently many efforts have been invested in exploring methods for open information extraction, which aims to discover new relation types from open-domain corpora.

II. Overview of existing approaches

Generally speaking, the procedure of information extraction is divided into entity extraction, relation extraction, attribute extraction and entity linking and so on. In this article, we just consider the open relation extraction. The open relation extraction approaches directly extract relation sets that contains entities and relations from open-domain text corpora.

Several open information extraction systems had developed to solve the problem. For example, TextRunner [1] and other open information extraction system achieve promising performance in open relation extraction on English sentences. They realize relation extraction which verb phrases or non-verb phrases can be considered as, and the binary relation, multiple relation extraction. For Chinese sentences that are very different from English, according to the characteristics of the Chinese language, there are several systems developed to extract relation sets from Chinese sentences, such as CORE [2]. However, The development of these systems lack a unified basic to achieve reasonable compatibility between relation extraction components and intelligent system.

III. Proposed approach

As described above, the main difficulties of information extraction come from two aspects:

- the structure analysis of external language texts;

– the analysis of fragment of knowledge base represented in internal language.

In order to solve these two aspects problems, we proposed use OSTIS Technology [4] to develop relation extraction component, which is able to extract sc-construction from natural language texts. The ostis-system is a knowledge-based intelligent system being developed based on OSTIS technology. The sc-construction represented in SC-code, which is an internal language used for knowledge representation, is a fragment of the knowledge bases of ostis-systems. SC-code is the form of semantic network language with the basic set-theoretic interpretation. There are several ways of external representation such as SCg, SCn, SCs, which will be shown blow.

By analyzing the structure of the knowledge base, we can clarify our extraction targets. In order to convert constructions of the external language into the sc-constructions, it is necessary to describe the syntactic and semantic knowledge for text analysis, as well as the construction of extraction rules. Since each external language can have its own set of unique features, within the framework of OSTIS Technology, concept **Subject Domain** is proposed to describe knowledge of a certain of specific domain. In our proposed implementation principles, we attempt to construct ontologies of specific natural language to extract open relations from open text corpora.

Based on the principles of OSTIS Technology, An ontology is interpreted as a specification of the system of concepts of the corresponding subject domain. In each subject domain various distinguished ontologies are described to reflect a certain set of the concept features of the subject domain, for example, terminological ontology, logical ontology, set-theoretic ontology, etc. The general structure of **Subject domain of Chinese language texts** is presented in SCn-language.

Section. Subject domain of Chinese language texts
⇒ *section decomposition**:
- ***Section. Subject domain of lexical analysis***
- ***Section. Subject domain of syntactic analysis***
- ***Section. Subject domain of semantic analysis***

The *Section. Subject domain of lexical analysis*, the *Section. Subject domain of syntactic analysis* and the *Section. Subject domain of semantic analysis* describe specification of a system of concepts, logical rules (e.g., extraction rules) and even other knowledge from the lexical aspect, syntactic aspect and semantic aspect of the Chinese language, respectively.

Aspect that construction of ontologies, the problem solvers of ostis-system also need to be considered, it can utilizes the knowledge base of linguistics to extract sc-construction from natural language free texts based on a series of text processing technologies, i.e., lexical analysis, syntactic analysis, semantic analysis, and extraction rules. The problem solver is a hierarchical system of agents of knowledge processing which oriented on solving specific problem. The general structure for developed problem solver for translating external texts into the fragment of knowledge base is represented below in SCn-language.

Problem solver translating external texts into the fragment of knowledge base
⇐ *decomposition of an abstract sc-agent**:
{
- *Abstract sc-agent of lexical analysis*
- *Abstract sc-agent of syntactic analysis*
- *Abstract sc-agent of semantic analysis*
- *Abstract sc-agent of generating sc-construction*
- *Abstract sc-agent of verifying the knowledge base*

}

The structure of developed problem solver corresponds to conversion steps from natural language sentence to sc-construction represented in SC-code.

IV. Conclusion

The article proposes a unified semantic model to solving open relation extraction problem. According to the constructed knowledge base and problem solvers, proposed implementation principles can integrate various types of knowledge and problem solving models. The constructed knowledge base contains various linguistic ontologies that used to analyse natural language texts, not just natural language texts of specific domain.

V. List of references

1. Banko, M. Open Information Extraction from the Web / M. Banko, M. J. Cafarella, S. Soderland, et al. // Proceedings of the 20th International Joint Conference on Artificial Intelligence. –Hyderabad, 2007, –P. 2670–2676.
2. Tseng, Yuen-Hsien Chinese Open Relation Extraction for Knowledge Acquisition / Yuen-Hsien Tseng, Lung-Hao Lee, , Shu-Yen Lin, et al. // Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers. –Gothenburg, 2014, –P. 12–16.
3. Golenkov V.V., Gulyakina N.A.: Project of open semantic technology of component designing of intelligent systems. Part 1 Principles of creation. / V. V. Golenkov, N. A. Gulyakina // Ontology of designing. – 2014. № 1. – c.42–64.