BRIDGE KNOWLEDGE AND LANGUAGES: THE APPLICATION OF COMPUTATIONAL LINGUISTICS

ZHAO Jun[1] SUI Zhifang[2]
[1]The National Key Lab of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
[2]Institute of Computational Linguistics,
Peking University
jzhao@nlpr.ia.ac.cn, szf@pku.edu.cn

ABSTRACT

The paper discusses the methods of using natural language processing technologies for lexical semantic knowledge acquisition. The basic idea is to use shallow (while not complete) natural language processing technologies to acquire domain (while not general) lexical semantic knowledge. These acquiring approaches are based on domain corpus and are transportable among domains. As an application example, through an analysis to the need of 2008 Digital Olympics to multilingual language resources, we try to construct an Olympic-oriented lexical semantic knowledge base in the field of sports, where the above NLP technologies are used to build a human-machine interactive lexical semantic knowledge base construction platform.

Keywords: Multilingual Language Resources, Lexical Semantic Knowledge, Natural Language Processing, Domain Corpus, 2008 Olympics

1. INTRODUCTION

The inevitable trend of information processing is from surface-level processing to intelligent processing. The process needs powerful knowledge bases as resource support. Only with rich knowledge can computer systems to conduct intelligent analysis, reasoning and decision etc. One of the main objectives 2008 Digital Olympics is to construct a multilingual intelligent information service platform, which is depend on the powerful support of various types of multilingual language resources. The knowledge in the world is numerous and complicated, among which lexical semantic knowledge is one of the most fundamental and important types. The sources of lexical semantic knowledge are also varied, among which language as a kind of information carrier acts as a very important resource of lexical semantic knowledge. With the development of natural language processing technologies and the enhancement of the performance of computers, it becomes possible for us to automatically or semi-automatically acquire lexical semantic knowledge through analyzing natural languages texts using NLP technologies.

2. THE EXSISTING APPROACHES FOR LEXICAL SEMANTIC KNOWLEDGE BASE CONSTRUCTION

Currently, there are two main approaches for constructing lexical semantics knowledge bases, which are ①to manually construct lexical semantics knowledge bases; ② to automatically construct lexical semantics knowledge bases through automatic knowledge acquisition.

(1) To manually construct lexical semantics knowledge bases:

There are various types of manually constructed lexical semantic knowledge bases, such as WordNet, EuroWordNet, FrameNet, HowNet, Tong Yi Ci Ci Lin (Chinese Thesaurus), etc. These knowledge bases are concentrated reflection of human expert knowledge, and have been supplying very important

resource support for intelligent information processing. However, there are some limitations. ① The manually constructed knowledge bases involves too many artificial factors. However, there is a distance between the resources that can be well used by machine and the perfect knowledge system in the brains of linguists. So, we need to use data-driven, application-driven and computable approaches to build the knowledge bases oriented for practical applications. ② The manually constructed knowledge bases are often in the general domain. When these general resources are used to the applications in a specific domain, we often find that they are far from enough. Many domain-specific concepts can't be found in this kind of resources. Because of the characteristics of time-consuming and labor-consuming, it is not possible for us to manually build a specific knowledge base for each specific domain. Therefore, we need to study a series reusable and transportable approaches for acquiring lexical semantic knowledge in order to automatically or semi-automatically build domain-specific lexical semantic knowledge bases.

(2) To automatically construct lexical semantics knowledge bases through automatic knowledge acquisition

① Simple pattern matching approaches based on surface information:

Some researchers use simple surface-information-based pattern matching approach to automatically acquire lexical semantic knowledge from the corpus. However, because only surface information is used, some of the obtained knowledge is incomplete, inaccurate, even mistaken. Although we can resort to manual checking and confirmation, much knowledge has been lost in the process of knowledge acquiring, it is very difficult for human experts to find and compensate.

② The approaches based on complete NLP analysis: Some researches acquire lexical semantic knowledge based on the complete NLP analysis to the text. For example, for MindNet construction, the researchers got the syntactic trees and deeper logical forms for a text depending on a general syntactic parser, and generate the structures of the semantic relations through a package of manually written rules. However, because of the too deep analysis level, the approach replies too heavily on deep and complete NLP techniques, such as syntactic parser and semantic interpreter, it is hard for scale-up and transport from domain to domain.

3. The BASIC PRINCIPLES FOR USING NLP TECHNOLOGIES TO ACQUIRE LEXICAL SEMANTIC KNOWLEDGE FROM TEXT

Based on the analysis in Section 2, we determine the basic principles for using NLP technologies to acquire lexical semantic knowledge from text as follow.

(1) Domain lexical semantic knowledge vs. General lexical semantic knowledge

General lexical semantic knowledge describes the generally applicable conceptual structures based mostly on philosophical and logical point of view, while domain lexical semantic knowledge describes a particular model of the world that is focused on applications. Domain lexical semantic knowledge bases have an enough high coverage to the domain concepts in the specific domains, and can be practically used in the applications in the intelligent text processing applications in the specific domains. SUMO is created by merging publicly available ontological content into a single comprehensive and cohesive structure, it provides definitions for general purpose terms and can be taken as a foundation for the specific domain ontologies. Our goal is to create domain ontologies that are aligned with the SUMO, the domain ontology inherits the broad conceptual distinctions of the SUMO and specifies the concepts and axiomatic content of a particular domain.

(2) Transportable vs. intransportable:

In practical application, domain lexical semantic knowledge bases are need for many domains. It is expensive for us to develop a knowledge acquisition system for each specific application and each specific domain. We wish to develop a series of approaches that can be transport to every specific domain through a simple tuning. For this goal, the lexical semantic knowledge acquisition approaches we study should have the following characteristics: corpus-based, data-driven and relying on machine learning algorithms.

(3) Surface Analysis vs. Deep Analysis:

The knowledge we get through surface analysis to natural language texts is often incomplete and inaccurate. On the other hand, although correct deep analysis results of natural language texts can give a guarantee for us to get accurate and complete knowledge, it is difficult for the deep analysis to natural language texts get a satisfied accuracy, and these deep analysis algorithms are difficult to be scale-up and transported from domain to domain. Therefore, we should leverage the above two approaches to develop a shallow syntactic and semantic analysis technology oriented for domain lexical semantic knowledge acquisition, which can not only make full use of the surface linguistic information, but also use some relatively mature NLP technologies to discover more lexical semantic knowledge hidden under the surface information.

(4) Automatic vs Semi-automatic:

We believe that it is impossible for computers to entirely replace human experts to construct lexical semantic knowledge bases in the current situation that the NLP technologies and machine learning technologies are far from mature. However, computers have the advantages of instantly find out the various types of linguistic regulations from the available large-scale corpus, while it is hard for human experts to complete the same thing is a short period. We should make good use of this advantage of computers to supply rich and reliable materials and references to human experts.

Therefore, the objective of our research are semi-automatic lexical semantic knowledge acquisition, that is to use adequate NLP and machine learning technologies to discover various types of linguistic regularities and submit them to human experts as rich and reliable materials and references. The human experts will semi-automatically construct the domain lexical semantic knowledge bases in a computer assistant fashion.

4. THE METHOD FOR ACQUIRING LEXICAL SEMANTIC KNOWLEDGE FROM TEXT USING NLP TECHNOLOGIES

4.1 The representation framework for lexical semantic knowledge

Lexical semantic knowledge denotes the set of the concepts represented by terms and the relations between these concepts. Therefore, the knowledge representation framework should reflect the above contents.

- **The determination of the representation unit for concepts:**

The principle for determining the representation units of lexical semantic knowledge is as follows. The units can represent independent concepts, and have as least semantic ambiguities as possible. According to the principle, and because we try to get the lexical semantic knowledge in a specific domain, therefore, we select the Chinese-English bilingual translation term pairs as the representation units for concepts. Because the standardized terms have a 1-to-1 mapping to the concepts and can cover the specific concepts in the domain, the term system can be regarded as the concept system's mapping onto the linguistic level. Terms are composed of words and phrased that have domain features.

- The determination of the relations between the concepts

For the convenience of automatic acquire concept relations, we determine the selection principles for concept relations as: with high coverage and adequate granularity. According to this principle, we select the following types of concept relations.

①Clustering relation: It denotes the logic relation between different concepts, including hyponymy, part-whole relationships (meronymy and holonymy), synonymy and antonymy, etc.

② Combinational relation: It can be also called as optional restriction relations. It denotes the semantic dependency relations among the different concepts, including Agent, Object, Dative, Tool, Location, etc.

The concepts can be divided into 3 types: OBJECT, EVENT and PROPERTY. The concepts of the OBJECT type denote the agent, object, location or orientation etc. of the concepts of EVENT type, including human, concrete object, abstract object, time and space etc. The concepts of EVENT type reflect the action, spirit activity, state of the OBJECTs and the relationships between the OBJECTs. The concepts of the PROPERTY type are used to describe the features and characteristics of the concepts of the OBJECT type and the EVENT type.

For different types of concepts, the emphasis of the concept relationships to be studied should be different too. For the concepts of OBJECT type, we should pay emphasis on the hyponymy relationships, such as the hyponymy relationships between "**田径**/Track and field" and "**马拉松**/Marathon". For the concepts of EVENT type and PRPERTY type, we should study emphatically on the semantic restriction relationships between the EVENT/PROPERTY concept and the related OBJECT concepts, such as the AGENT of the EVENT concept "**发球**/service" can be "**主攻手**/spiker".

- The determination of the framework for lexical semantic knowledge representation

After the content for knowledge representation is determined, we need to further determine the formalized representation of the knowledge structure. We use the structure of semantic network to represent the lexical semantic knowledge, which is composed of a set of nodes and a set of arcs attached with tags, where a node denotes a domain concept, an arc denotes the relationship between the two concepts it connect, maybe a clustering relationship or a combinational relationship. The following is an example of semantic network.

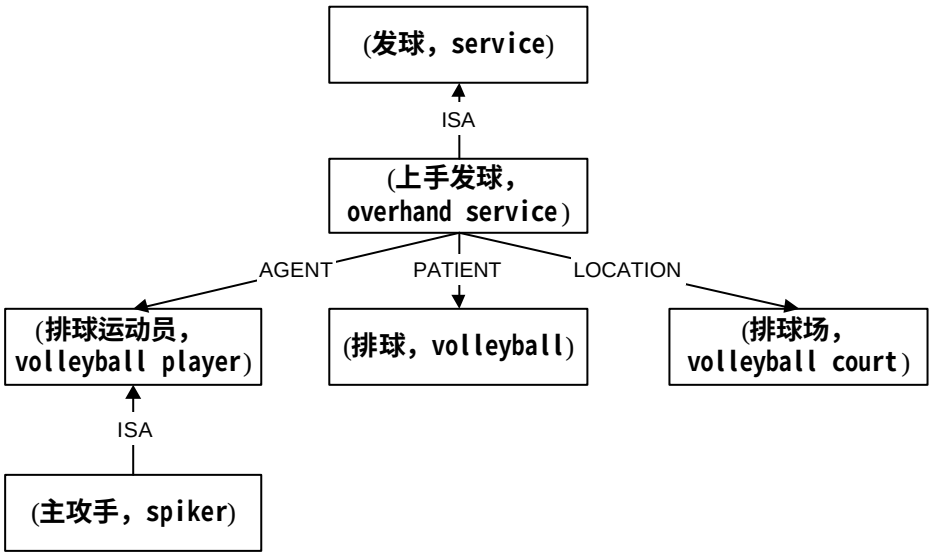

Figure 1: An example of semantic network

4.2 The automatic acquisition of domain lexical semantic knowledge

- The acquisition of knowledge representation units: Bilingual Term Pair Recognition

The critical problem of term recognition is to recognize the words and phrases with the domain specific features. The traditional term recognition methods include: pure linguistic method, pure statistical method and the mixture method combining grammatical structures with the statistical regulations. The results extracted are usually specific collocations, key words and base noun phrases. Because these units have no domain specific feature, they are not terms in strict sense.

Our method for term recognition is designed according to the characteristics of terms, i.e. closely combined, linguistically complete, and domain specific. First of all, according to the statistical association information between the words, we extract the linguistic fragments whose component words are closed combined as the candidates of terms. Then, according to the structural regularities of phrases and the structural regularities of terms, we eliminate the term candidates without linguistic completeness. Finally, we further eliminate the candidates without the domain specificity according to their frequencies and inverse document frequencies. The remaining term candidates are closely combined, linguistically complete, and domain specific, and are regarded as terms in the domain.

- **The Designing of the shallow syntactic and semantic analysis technology oriented for domain lexical semantic knowledge acquisition**

The precondition of knowledge acquisition from text is the necessary linguistic structure analysis to the text corpus, including chunking and lexical dependency relationship assignment used for extracting the linguistic representation pattern of conceptual semantic relationships. In the condition that the complete syntactic and semantic analysis technologies are far from mature, the above tasks are completed by partial and shallow syntactic and semantic analysis technologies, including baseNP recognition, shallow syntactic and semantic analysis, and surface pattern recognition, etc. The combination of these technologies not only meet the necessary need for lexical semantic knowledge acquisition, but also do not introduce too much noises because of the ambiguities and the analysis errors.

The basic idea of this kind of partial and shallow analysis is described as the following example.

Input Sentence:

这里介绍的是排球技术中最基础的一种拦网技术----单人拦网。
Here, we introduce one of the most fundamental blocking technologies in volleyball technologies ----single blocking.

Domain term recognition:
这里介绍的是[排球技术]中最基础的一种[拦网]技术----[单人拦网]。
Here, we introduce one of the most fundamental [blocking] technologies in [volleyball technologies] ----[single blocking].

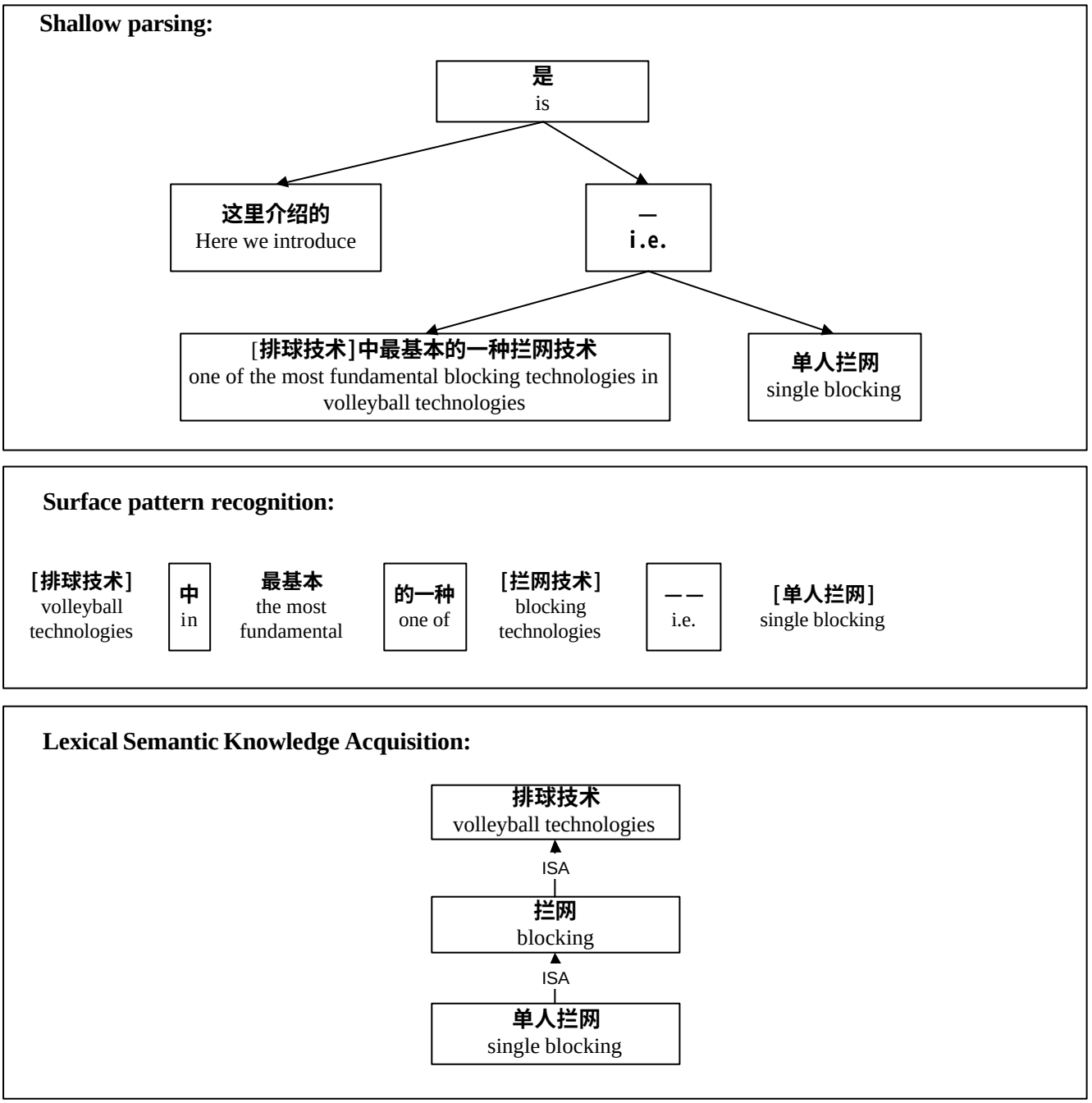

Figure 2: The basic idea of partial and shallow analysis

(2) **The iterative learning technology for acquiring concept relations:**

In the process of automatic acquisition of lexical semantic knowledge from corpus, the resources we can used include: ① the existent lexical semantic resources, such as manually constructed ontologies (HowNet and Tong Yi Ci Ci Lin etc.), the paraphrase texts in the lexicons, and encyclopedia, etc. ② large-scale real domain text corpus. The former can intuitionally supply large quantities of semantic relations between concepts, but cannot cover the specific concepts and concept relations in the specific domains. The latter may cover the domain concepts and concept relations in a relatively entire way, but the concepts and concept relations are often hidden behind the complicated linguistic phenomena that are hard to be found out.

Based on the above analysis, we use the following strategies to automatically extract the concept relations from the corpus. First of all, we use the concept relations from the manually constructed ontologies as seeds to discover the linguistic representation patters from the knowledge rich texts like the paraphrase texts of lexicons and encyclopedia, such as “x 即 y”;“所谓 x 就是 y”;“x 是一种 y” etc. Apply these patterns to the large-scale domain corpus to iteratively learn more linguistic representation patterns of concept relations in a bootstrapping way. The updated linguistic representation pattern set is used to more extended domain corpus in order to find more domain concept relations.

4.3 The organization of knowledge

The organization of knowledge involves the manipulation of merging, inheritance, inference and duplication-check of the scattered knowledge learned from the corpus. Currently, we organize the knowledge in the following two aspects.

(1) To assign terms to concepts and merge concepts using the existent domain semantic resources or conceptual hierarchies. For example,

If we have automatically obtained the following knowledge as

（**主攻手** ace-spiker，**AGENT**，**发球** service）
（**二传手** setter， **AGENT**，**发球** service）

The system has got the knowledge as

(**主攻手** ace-spiker **ISA** **排球运动员** volleyball player)
(**二传手** setter **ISA** **排球运动员** volleyball player)

Then we can get the following more abstract knowledge through semantic merging operation.

（**排球运动员** volleyball player，**AGENT**， **发球** service）

(2) Use the logic relationships of the semantic relations to conduct rational inference of the semantic relations in order to infer the unknown semantic relations from existent semantic relations. For example,

If we have automatically obtained the following knowledge as

(**跳发球** jump-serving ISA **发球** service)

The system has got the knowledge as

（**排球运动员** volleyball player， **AGENT**，**发球** service）

Then we can get new knowledge through inference as

（**排球运动员** volleyball player，**AGENT**，**跳发球** jump-serving）

5. THE REQUEST OF 2008 BEIJING OLYMPIC GAMES FOR MULTILINGUAL LANGUAGE RESOURCES AND A COMPUTER-ASSISTANT PLATFORM FOR DOMAIN LEXICAL SEMANTIC KNOWLEDGE BASE CONSTRUCTION IN THE FIELD OF SPORTS

5.1 The Objective of 2008 Digital Olympic Games and its Request for Multilingual Language Resources

The Programme of Action for Beijing Olympic Games points out: “Till 2008, we will basically realize different, affordable and non-language-barrier personalized information service, which can be supplied to anyone at any time and place in a secure, convenient, prompt and effective manner.”. For the above goal, “We should make good use of the modern information technologies, especially natural language processing technology in the field of artificial intelligence, to basically solve the language barrier problem of Olympic Games, to convenience the communication among the people from different countries and to promote the understanding and friendship, which is also one of the most important goals of 2008 Olympic Games: Human and Cultural Olympic Games.

For the above goals, a multilingual information

service platform is being constructed, including various applications of information issuance, information retrieval, question and answering, machine translation etc. The platform will supply various types of multilingual and customized information service for the athletes, pressmen and audiences from different countries so that Olympic-related information can be obtained conveniently by anyone, at any time and place, and by various types of means.

One of the most important difficulties of *The Programme of Action for Beijing Olympic Games* is "Language Barrier". In order to solve this problem, we should construct integrated multilingual language resources, such as Multilingual Lexicons, Multilingual Corpora, and Multilingual Ontologies, etc. The collecting, construction, organization, storage, merge of the resources are all the fundamental tasks for research and programming. These resources are constructed for the purpose of Digital Olympics, they have the characteristics of Domain-dependent and Application-dependent. Therefore, it is a very important and urgent task to construct a series of Olympic-oriented multilingual language resource bases to support the multilingual information service platform. Among them, multilingual lexical knowledge base in the domain of sports is a very important fundamental resource.

5.2 A Computer-Assistant Platform for Domain Lexical Semantic Knowledge Base Construction in the Field of Sports

"Digital Olympics" is an important objective of 2008 Beijing Olympic Games. For this goal, a multilingual intelligent information service platform is needed, including different applications of information release, information retrieval, question and answering, machine translation, etc. All these applications will be focused in the fields of sports, tourism, etc. To construct domain lexical knowledge bases are critical for all these above applications. Here we discuss the computer-assistant methods for lexical semantic knowledge base construction in the field of sport. These methods can be used for constructing the domain ontologies in the domain of sports and tourism etc, which will be important fundamental resources for the information service related to the 2008 Olympic Games.

The following is the architecture figure of a computer-assistant platform for domain lexical semantic knowledge base construction in the field of sports. In the platform, it includes an human-machine iterative learning mechanism and a human editing interface for domain ontology construction, in addition to the NLP-based knowledge acquisition technologies that have been introduced in Section 4.2. The following two sections will introduce the two additional functions in detail.

(1) the human-machine iterative learning mechanism: In the process of human-machine coordination, on the one hand, human experts will proof-check the results from machine learning; On the other hand, the machine will gradually absorb the human knowledge that has been used by experts in proof-checking so that the automatic semantic relation acquisition mechanism can be gradually extended and improved with the human-machine coordination process. First of all, we design a set of proof-check patterns. Then we use a error-driven learning mechanism to learn the results from human experts' proof-checking and generate a series of proof-checking rules which will be added into the automatic extraction mechanism later. In the next learning cycle, when we have automatically got the semantic relations, before human experts' proof-check, we will use these proof-checking rules to automatically proof-check the automatically obtained semantic relations and their linguistic representations. The automatic extraction mechanism will be gradually optimized through the iterative learning approach.

(2) The human editing interface for domain ontology construction

A graphic and friendly human editing interface will

be developed for computer-assistant ontology construction. Using the interface, the domain experts can conveniently editing the concept semantic relations obtained automatically by machine to construct the domain ontology. In the editing process, the machine can supply a series of functions to the human experts, such as the retrieval of relevant contents, the display of examples and consistency checking, etc, and can timely update and maintain the domain ontology.

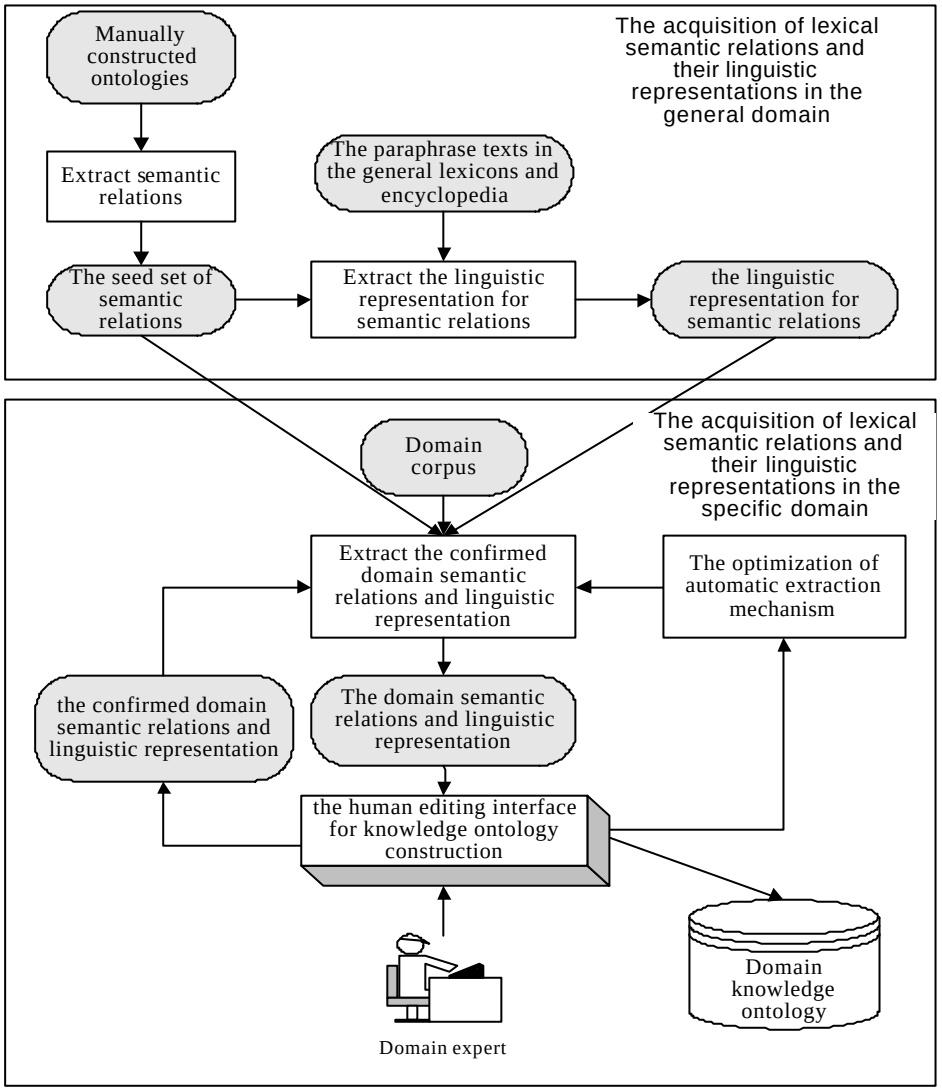

Figure 3: the architecture of a computer-assistant platform for domain lexical semantic knowledge base construction in the field of sports

6. CONCLUSION

The paper discusses the application of NLP technologies in the acquisition of domain lexical semantic knowledge and introduces a human-machine coordination platform for domain ontology construction. In the platform, the human experts coordination with machine, the human knowledge and the knowledge learned by the automatic learning mechanism are well combined to conveniently construct domain ontologies for different domains and different applications. The lexical semantic knowledge acquisition technologies we studied are based on large-scale corpus and are

transportable among different domains. Using the technologies, it is possible for us to construct domain ontologies for different domains in a unified framework. These domain ontologies can be aligned with top ontology like SUMO to supply critical resource support for various kinds of intelligent information processing applications and the next generation network – SemanticWeb.

7. ACKNOWLEDGEMENTS

The research work in this paper is supported by the Scientific Research Foundation for Returned Overseas Chinese Scholars, State Education Ministry, and HIMALAYA Foundation.

REFERENCES

1. Miller. G., Introduction to WordNet: an on-line lexical database. In: *International Journal of Lexicography*, Vol. 3. No.4. 1990
2. Piek Vossen, *EUROWORDNET: A multilingual database with lexical semantic network*, Kluwer Academic Publishers, 1998
3. Fillmore, Charles J. and Collin F. Baker, FrameNet: Frame semantics meets the corpus. In: *Poster presentation, 74th Annual Meeting of the Linguistic Society of America*, 2000
4. DONG Zhendong, DONG Qiqang, *HowNet*, http://www.keenage.com/
5. MEI Jiajv, TONG YI CI CI Lin, The Dictionary Press of Shanghai , 1983
6. Richardson S. D., Dolan W. B., and L. Vanderwende, "MindNet: Acquiring and Structuring Semantic Information from Text", In: *COLING-ACL'98: 36th Annual meeting of the Association for Computaitonal Lingustics and 17th International conference on computational linguistics*, 1998
7. WANG Hui, A Syntagmatic Study of Noun Sememes in Contemporary Chinese, Doctoral Dissertation, Peking Univ., Beijing, 2002
8. Eduard Hovy, Building Semantic/Ontological Knowledge by Text Mining, In: *SEMANET: Building and using semantic network*, 2002
9. Michael Thelen and Ellen Riloff, A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts, In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002
10. ZHAO Jun, The Research on Chinese BaseNP Recognition and Structural Analysis Doctoral Dissertation, Tsinghua Univ. Beijing, 1998
11. ZHOU Qiang, The Research on Automatic Phrase Chunking and Tagging for Chinese Corpora, Doctoral Dissertation, Peking Univ., 1996
12. Niles, I., Pease, A. Toward a Standard Upper Ontology, in the Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001).