ICIC Express Letters
Volume **10**, Number **8**, August **2016**

 ISSN 1881-803X
pp. **1995–2001**

RESEARCH AND DEVELOPMENT OF AUTOMATIC SEMANTIC DICTIONARY CONSTRUCTION PLATFORM

Yao Liu, Baoqiang Qu and Yi Huang

Institute of Scientific and Technical Information of China
No. 15, Fuxing Road, Haidian District, Beijing 100038, P. R. China
liuy@istic.ac.cn

Received December 2015; accepted March 2016

Abstract. *With the development of computer and Internet technology, it offers promising opportunities for research into Chinese semantics. There is a significance and urgent need for semantic dictionary in natural language processing. In view of this, this paper proposes a framework to construct semantic dictionary, takes a deep research on hyponym and synonym extraction technologies, and builds a semantic dictionary construction platform. The experiments show that platform can effectively extract hyponymy and synonym relations, to provide basic data for the construction of domain semantic dictionary.*
Keywords: Semantic dictionary, Synonymy extraction, Hyponymy extraction

1. **Introduction.** Knowledge base is a basic and indispensable resource for a variety of natural language processing tasks. Most of the existing knowledge bases or ontologies are about generic domain knowledge, which hardly meet the needs of specific domain areas to conduct natural language processing tasks and provide knowledge services. The construction of semantic dictionary has become an essential job for natural language processing. The core of semantic dictionary lies in the hyponymy and synonym relations of its units. From the perspective of knowledge organization, hyponymy constitutes the basic structure of knowledge, and it is an essential step to transform unstructured information into structured information so as to provide some basic support for ontology, knowledge base construction, information retrieval, as well as obtaining additional information. Synonym has been widely used in the field of natural language processing, because it can provide extended query information for retrieval system, and also improve recall retrieval system. Currently, semantic dictionaries, such as WordNet [1], MindNet [2], FrameNet [3] in English and CCD [4], How-Net [5], HIT-CIR Tongyici Cilin [6] in Chinese, are manually developed and updated slowly with low coverage of words. Meanwhile, automatic construction of semantic dictionary in Chinese is still in the experimental stage in research papers, and there is no automated or semi-automated construction platform [7,8]. In addition, in the age of Internet, a variety of new terms not in the scope of traditional semantic dictionary continue to appear and it is difficult for those traditional semantic dictionaries to adapt the new terms. There is also a trend of domain semantic dictionary construction in every trade. In view of this, this paper takes a deep research on hyponymy, synonym relation extraction and tree structure generation technologies, and develops a platform to facilitate the semantic dictionary construction process.

The rest of this paper is organized as follows. Section 2 introduces idea and framework. Sections 3 describes the key technologies. Sections 4 describes the platform and its features. Experiment and analysis are shown in Section 5. Finally, a brief conclusion and future work are given in Section 6.

2. **Idea and Framework.** The data sources for hyponym extraction generally can be divided into three categories: structured, semi-structured resources, and unstructured text.

The methods of hyponym extraction can be divided into two categories: pattern-based method and statistical method, and pattern-based approach is in mainstream of current studies. Methods of synonym extraction can be divided into the following ones: dictionary definition based method, literally and similarity based method, large-scale corpus-based method, syntactic dependency structure based method and pattern matching method [9]. Thus, there are two aspects we should carefully examine: on the one hand, from the perspective of discovery, it is to get hyponym and synonym without external input; on the other hand, from the perspective of query, it is to obtain synonym and hyponym by fully utilizing existing resources. In view of the analysis above, with some research results our team have archived in recent years, this paper proposes a framework shown as Figure 1 [10-12].

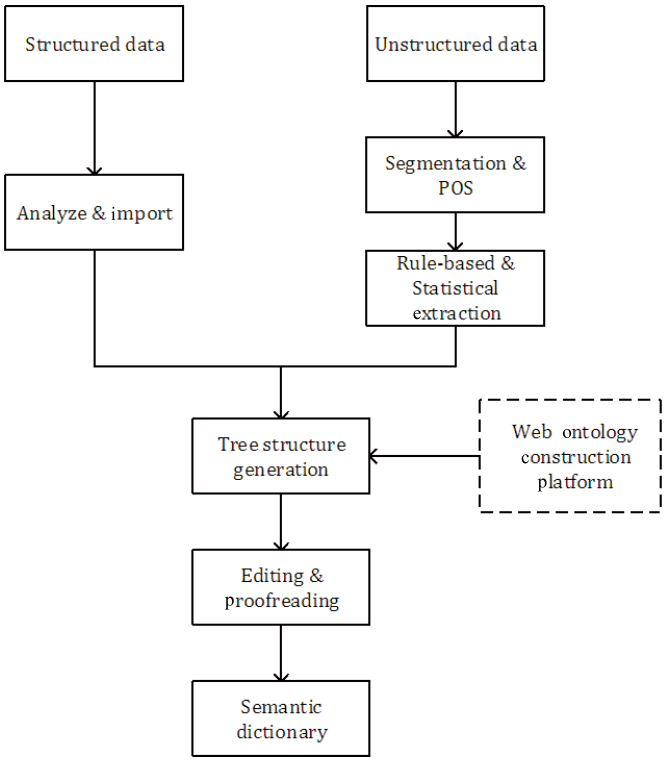

FIGURE 1. Semantic dictionary construction framework

Structured data, such as existing subject headings, can be imported as an important resource to construct semantic dictionary; for unstructured data, segmentation and POS tagging are being conducted, rule-based and statistical methods are used to extract hyponym and synonym relations, and semantic analysis between the relation pairs is processed to choose those with a good score as source data to generate tree structure; then, with some editing and proofreading, a semantic dictionary is well constructed.

3. Key Technologies.

3.1. Hyponym extraction.

Here are some rules used to extract hyponym relation.

(1) Using comma as mark to cut sentences into units, when there is 是一种 (ISA) pattern, the word before ISA is a subclass, shown as E1. However, when there is a 的 (DE) after ISA, the word after DE is a parent class, and the word before ISA is a subclass, shown as E2.

E1. 普通晶体管是一种电流控制元件。

(Common transistor is a current controlling element.)

parent class: current controlling element

subclass: common transistor

E2. 转换插头是一种对输入/输出插头类型（如立体声转卡侬）进行转换的接插件。
(Converter plug is a kind of input/output connector (such as stereo turn XLR) for conversion.)
parent class: connector
subclasses: converter plug
(2) Using comma as mark to cut sentences into units, there is 由/主要由/一般由/多数由/通常由 (BY) 组成/构成/合成/制成/组装/而成/组装而成/组合而成/粘合而成 (CMP) pattern, and the words before CMP are subclass, and the word before BY is a parent class, shown as E3, except for the case when there is a 的 (DE) shown as E4.
E3. 压电蜂鸣片由压电陶瓷片和金属振动板黏合而成。
(The piezoelectric buzzer is composed by piezoelectric ceramic sheet and metal vibrating plate.)
parent class: piezoelectric buzzer
subclass: piezoelectric ceramics, metal vibrating plate
E4. 它将线圈置于由永久磁铁、铁芯、高导磁的小铁片及振动膜组成的回路中。
(It placed the magnetic circuit in the magnet routine made up of permanent magnet, core, small iron piece and the diaphragm.)

3.2. **Synonym extraction.** Here are some rules used to extract synonym relation.
Using comma as mark to cut sentences into units, there is 也称/也称为/被称为/也被称为/俗称/简称/简称为/又称/又称为/可称为/可称/称为/又名 (AKA) pattern, and the word before and the word after AKA are synonym pair, shown as E5, except for the case shown as E6 and E7.
E5. 半导体三极管简称晶体管。
(Semiconductor triode is also called transistor.)
synonym pair: semiconductor triode, transistor
E6. 这种电容器即被称为反交联电容器。
(This capacitor is also known as anti-cross-linked capacitor.)
E7. 被称为“增益”。
(called “gain”.)

3.3. **Tree structure generation.** Based on our previous research results [13] relevant to tree structure generation, user only needs to follow the instruction to upload a text file in the format (parent class and subclass in a hierarchy structure separated by tab character) shown below, a tree structure can be generated automatically.

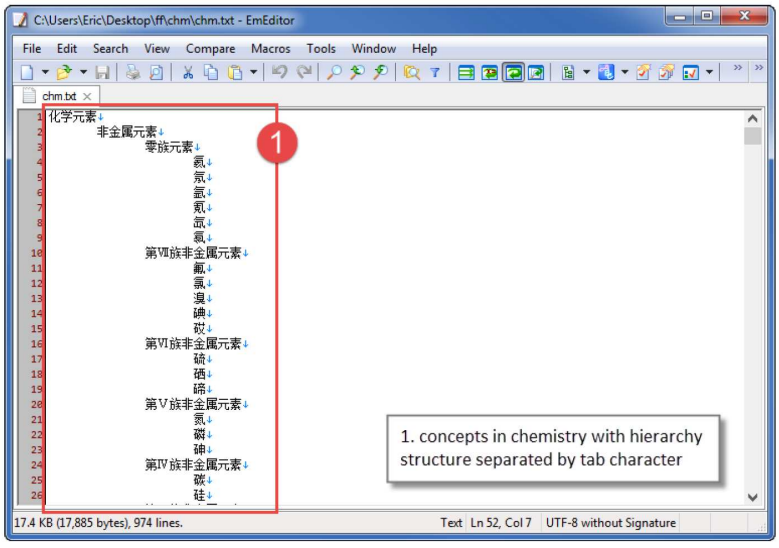

FIGURE 2. A text file with content in hierarchy structure separated by tab character

4. **Platform and Its Features.** Based on B/S architecture, the platform can be accessed directly via browser. The architecture of the platform is shown as Figure 3: the view layer displays the information of the platform related to such services as browsing a dictionary, editing and proofreading terms in a dictionary; the application layer controls the platform's functionality by conducting detailed processing; and the data layer includes the data persistence mechanisms of the platform.

Its main features include:

(1) Upload an existing dictionary to the platform shown as Figure 4;

(2) Upload a domain document to the platform to conduct hyponym and synonym extraction shown as Figure 5;

(3) Manage extraction rules shown as Figure 6;

(4) Generate a tree structure shown as Figure 7;

(5) Edit and proofread terms in a semantic dictionary shown as Figure 8.

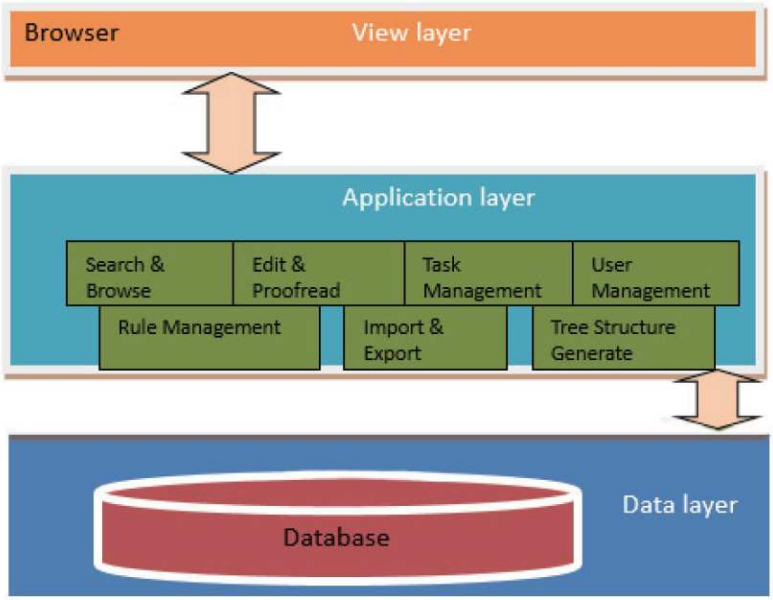

FIGURE 3. Architecture of semantic dictionary construction platform

FIGURE 4. Upload an existing dictionary to the platform

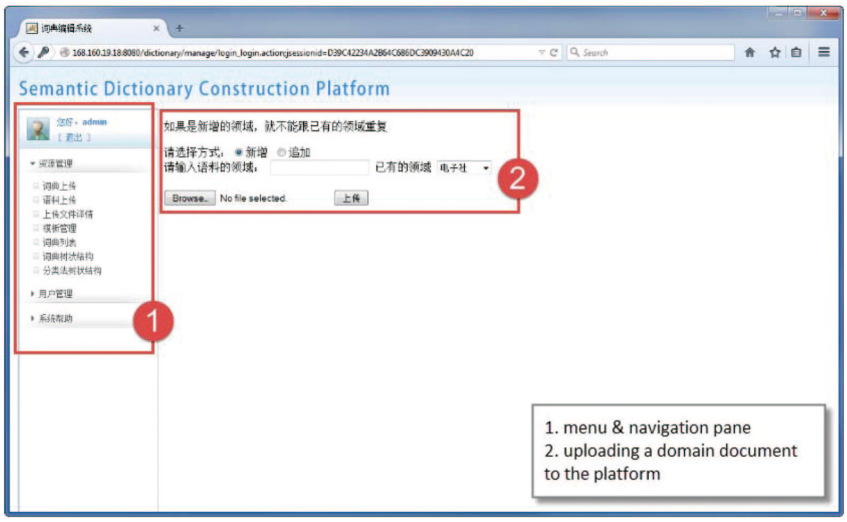

FIGURE 5. Upload a domain document to the platform

FIGURE 6. Extraction rules management

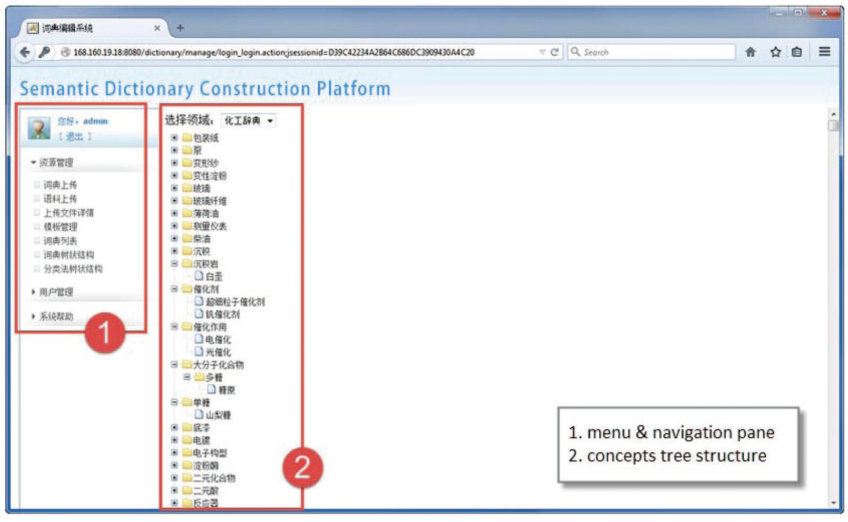

FIGURE 7. Tree structure generation

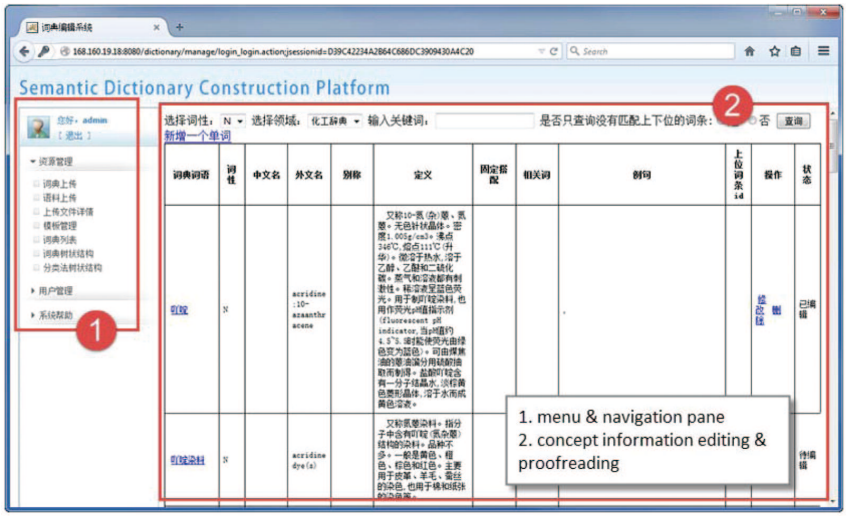

FIGURE 8. Edit and proofread terms in a semantic dictionary

5. **Experiment and Analysis.** We constructed two corpora in electronic field and chemical field to extract hyponym and synonym relations. Precision rate of those entries that the platform is capable of processing is chosen as the evaluation standard to verify the effectiveness of our approach.

(1) Data from PHEI (Publishing House of Electronics Industry, China)

Processing 37.3% of the electronic corpus (1438/3882 entries), we extracted 783 hyponym relations and 638 synonym relations, and the precision rate of this experiment is shown as Table 1.

TABLE 1. Precision rate of electronic corpus relation extraction

	hyponym relations	synonym relations
PHEI	93.1%	88.23%

(2) Data from CIP (Chemical Industry Press, China)

Processing 35.27% of the chemical corpus (4290/12162 entries), we extracted 1523 hyponym relations and 2767 synonym relations, and the precision rate of this experiment is shown as Table 2.

TABLE 2. Precision rate of chemical corpus relation extraction

	hyponym relations	synonym relations
CIP	87.63%	96.45%

As can be seen from above, the precision rate for hyponym relation extraction is 93.10% for PHEI and 87.63% for CIP. Our approach performed better in PHEI than in CIP, which might be attributed to the following factors: 1) electronic corpus are relatively smaller than chemical corpus; 2) most of hyponym relations in PHEI are covered by our method. However, in synonym relation extraction, our method performed well in CIP, simply because most chemicals have alternative names. Overall, the platform is capable of processing more than 1/3 entries in both corpus, and the precision rate is above 87%. From the view of semantic dictionary construction, these data are an essential resource to build a domain semantic dictionary.

6. **Conclusion and Future Work.** This paper proposes a framework to construct semantic dictionary, takes a deep research on hyponymy, synonym relation extraction technologies, and develops a platform to facilitate the semantic dictionary construction process. The experiments show that platform can effectively extract hyponymy and synonym relations with precision rate above 87% in both electronic and chemical corpus, to provide basic data for the construction of domain semantic dictionary. In the future, we plan to introduce the results of domain ontology evolution to further improve the effective extraction of hyponym and synonym relations with the help of web ontology construction platform.

Acknowledgement. This work is partially supported by National Social Science Fund No. 12BTQ006, National Key Project of Scientific and Technical Supporting Programs No. 2013BAH21B02; the authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] C. Felbaum, *WordNet: An Electronic Lexical Database for English*, 1998.

[2] S. D. Richardson, W. B. Dolan and L. Vanderwende, MindNet: Acquiring and structuring semantic information from text, *Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, vol.2, pp.1098-1102, 1998.

[3] C. F. Baker, C. J. Fillmore and J. B. Lowe, The berkeley framenet project, *Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, vol.1, pp.86-90, 1998.

[4] J. S. Yu, Y. Liu and S. W. Yu, Chinese concept dictionary specifications, *Journal of Chinese Language and Computing*, vol.13, no.2, pp.177-194, 2003.

[5] Z. D. Dong and Q. Dong, Construction of a knowledge system and its impact on Chinese research, *Contemporary Linguistics*, 2001.

[6] J. J. Mei, Y. M. Zhu, Y. Q. Gao and H. X. Yin, *Tongyici Cilin*, Shanghai Lexicon Publishing Company, Shanghai, 1983.

[7] C. X. Zhang, C. G. Cao, L. Liu, Z. D. Niu and J. H. Lin, Extracting hyponymy relations from domain-specific free texts, *International Conference on Machine Learning and Cybernetics*, vol.6, pp.3360-3365, 2007.

[8] R. J. Fu, J. Guo, B. Qin, W. X. Che, H. F. Wang and T. Liu, Learning semantic hierarchies via word embeddings, *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics: Long Papers*, vol.1, 2014.

[9] Z. F. Sui, Y. Liu and Y. W. Hu, Extracting hyponymy relation between Chinese terms based on term types' commonality and sequential patterns, *ICIC Express Letters*, vol.3, no.4(B), pp.1233-1238, 2009.

[10] Z. F. Sui, Y. Liu, J. Zhao and H. Zhang, The development of an NLP-based chinese ontology construction platform, *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, vol.3, pp.190-194, 2010.

[11] J. Li, R. J. Wang and Y. Liu, Research on semantic metadata online auxiliary construction platform and key technologies, *Chinese Lexical Semantics*, Springer Berlin Heidelberg, 2013.

[12] Y. Liu, H. Q. Shi, D. J. Zheng and Y. Huang, Study on semantic annotation for professional literature, *ICIC Express Letters, Part B: Applications*, vol.5, no.5, pp.1383-1389, 2014.

[13] Y. Liu, Z. F. Sui, Q. L. Zhao, Y. W. Hu and R. J. Wang, On automatic construction of medical ontology concept's description architecture, *International Journal of Innovative Computing, Information and Control*, vol.8, no.5(B), pp.3601-3616, 2012.