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

ICIC International ©2016 ISSN 1881-803X
pp. 1901–1906

RESEARCH ON SEMANTIZATION OF UNSTRUCTURED TEXT IN MEDICAL FIELD

Yao Liu[1], Zhifang Sui[2] and Yi Huang[1]

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

[2]Institute of Computational Linguistics
Peking University
No. 5, Yiheyuan Road, Haidian District, Beijing 100871, P. R. China

Received January 2016; accepted April 2016

Abstract. *In recent years, with the rapid development of computer technology and the popularity of the Internet, huge information began to emerge in the form of unstructured text in various fields. Faced with such a huge and rapidly growing amount of data, how to quickly and effectively obtain specific knowledge of the real needs out from the mass is a challenge faced by researchers. Researches in medical field are comparatively few than other fields. Chinese medical text has its own unique characteristics, and we cannot copy the foreign Natural Language Processing methods and technologies. This paper presents a method of processing medical unstructured text. The application results show that the proposed methods can effectively process semantization of medical unstructured text.*

Keywords: Unstructured text, NLP, Semantic role labeling, Discourse annotation

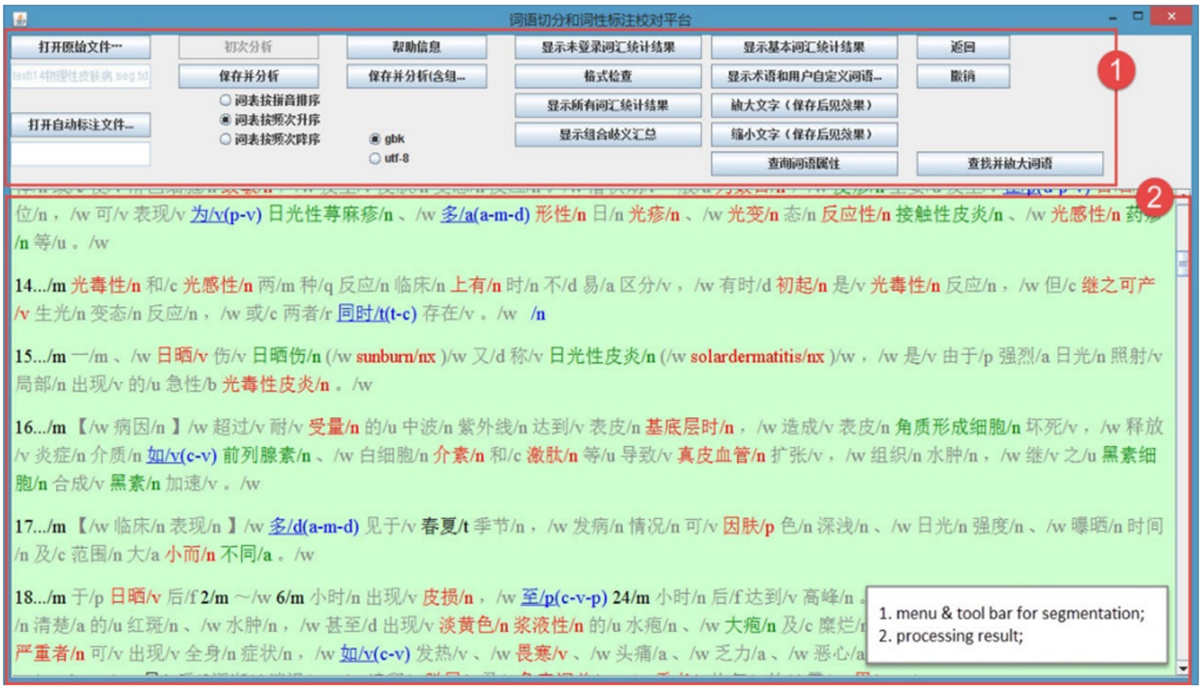
1. **Introduction.** In recent years, with the rapid development of computer technology and the popularity of the Internet, lots of information began to emerge in the form of unstructured text in various fields. As one of the top research fields, medical research delivered a huge amount of data online. More than 600,000 papers are published in medical field every year. There are more than 20 million papers in MEDLINE [1]. With serious challenges brought by information explosion, we are in urgent need of some technologies and tools to help us quickly find valuable information out of the mass. Nowadays, semantization is becoming area of research focus, being widely used around the globe. In China, Ou [2] investigated the semantic representation of Chinese Thesaurus. Liu et al. [3] proposed a method of using semantic annotation and ontology to conduct semantization of library resources. Bai and Qiao [4] studied the semantization of bibliography data based on ontology and linked data. However, there are relatively few studies in Chinese medical literature than those in other fields, for the text is often with longer sentences and more complex structures. Also, we cannot copy methods and technologies in medical field used in foreign languages [5]. Therefore, the need to work out a systematic method of processing Chinese medical unstructured text is demanding. Based on the characteristics of unstructured Chinese medical text, this paper proposes a method of processing unstructured Chinese medical text by conducting segmentation, POS tagging, semantic annotation, semantic role labeling, and discourse annotation to achieve semantization of unstructured text in medical field.

The rest of this paper is organized as follows. Section 2 introduces idea and framework. Section 3 describes the key methods and technologies. Section 4 describes the applications of the proposed methods and technologies. Finally, a brief conclusion and future work are given in Section 5.

2. Framework. From a linguistic point of view, natural language understanding is a hierarchical process that includes the following five levels: speech analysis, lexical analysis, syntactic analysis, semantic analysis and pragmatic analysis. Among them, semantic information is the core part of the process, and it is one of the main objectives in the field of natural language processing research as well. Based on the characteristics of unstructured Chinese medical text, this paper proposes a systematic method of processing unstructured Chinese medical text by segmentation, POS tagging, semantic annotation, semantic role labeling, and discourse annotation to achieve semantization of unstructured text in medical field.

3. Key Methods and Technologies.

3.1. Word segmentation & POS tagging. Unlike English, there is no space to separate Chinese words in a sentence. Segmentation is the very first step to process any Chinese text. Since there are many long terms in medical field, we constructed a professional field dictionary containing about 40 thousand terms with the help of Online Semantic Dictionary Construction Platform (OSDCP) to help cut sentences into words. Then, word segmentation and POS tagging are conducted with WSPTT shown as Figure 1.



位/n ，/w 可/v 表现/v 为/v(p-v) 日光性荨麻疹/n 、/w 多/a(a-m-d) 形性/n 日/n 光疹/n 、/w 光变/n 态/n 反应性/n 接触性皮炎/n 、/w 光感性/n 药疹/n 等/u 。/w

14.../m 光毒性/n 和/c 光感性/n 两/m 种/q 反应/n 临床/n 上有/n 时/n 不/d 易/a 区分/v ，/w 有时/d 初起/n 是/v 光毒性/n 反应/n ，/w 但/c 继之可产/v 生光/n 变态/n 反应/n ，/w 或/c 两者/r 同时/t(t-c) 存在/v 。/w /n

15.../m 一/m 、/w 日晒/v 伤/v 日晒伤/n (/w sunburn/nx)/w 又/d 称/v 日光性皮炎/n (/w solardermatitis/nx)/w ，/w 是/v 由于/p 强烈/a 日光/n 照射/v 局部/n 出现/v 的/u 急性/b 光毒性皮炎/n 。/w

16.../m 【/w 病因/n 】/w 超过/v 耐/v 受量/n 的/u 中波/n 紫外线/n 达到/v 表皮/n 基底层时/n ，/w 造成/v 表皮/n 角质形成细胞/n 坏死/v ，/w 释放/v 炎症/n 介质/n 如/v(c-v) 前列腺素/n 、/w 白细胞/n 介素/n 和/c 激肽/n 等/u 导致/v 真皮血管/n 扩张/v ，/w 组织/n 水肿/n ，/w 继/v 之/u 黑素细胞/n 合成/v 黑素/n 加速/v 。/w

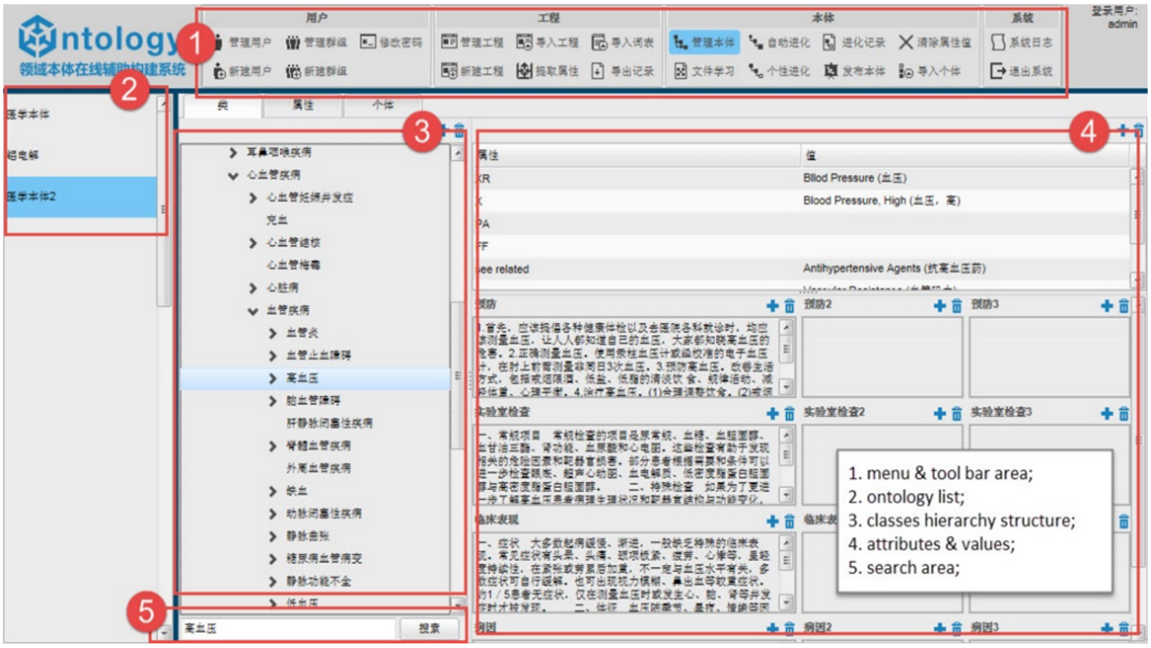
17.../m 【/w 临床/n 表现/n 】/w 多/d(a-m-d) 见于/v 春夏/t 季节/n ，/w 发病/n 情况/n 可/v 因肤/p 色/n 深浅/n 、/w 日光/n 强度/n 、/w 曝晒/n 时间/n 及/c 范围/n 大/a 小而/n 不同/a 。/w

18.../m 于/p 日晒/v 后/f 2/m ～/w 6/m 小时/n 出现/v 皮损/n ，/w 至/p(c-v-p) 24/m 小时/n 后/f 达到/v 高峰/n 。
/n 清楚/a 的/u 红斑/n 、/w 水肿/n ，/w 甚至/d 出现/v 淡黄色/n 浆液性/n 的/u 水疱/n 、/w 大疱/n 及/c 糜烂/
严重者/n 可/v 出现/v 全身/n 症状/n ，/w 如/v(c-v) 发热/v 、/w 畏寒/v 、/w 头痛/a 、/w 乏力/a 、/w 恶心/a

1. menu & tool bar for segmentation;
2. processing result;

Figure 1. GUI of WSPTT and its output results

3.2. Semantic annotation. In order to make computer understand the unstructured medical text at the content level, we need to process the text fragments at a semantic level. Semantic annotation cannot be effectively conducted without domain ontologies, because in a domain ontology there are many concepts as well as relations between them, which can be made good use of in the annotating process. By using Online Ontology Construction Platform (OOCP), we construct a medical ontology. It is assumed that there are text fragments related to a concept in those paragraphs. Semantic score is being used to measure how closely a sentence or paragraph is related to a concept in the domain ontology. Then, based on the occurrence of a concept and its properties in a sentence or paragraph, semantic annotation is performed [6].



1. menu & tool bar area;
2. ontology list;
3. classes hierarchy structure;
4. attributes & values;
5. search area;

FIGURE 2. GUI of OOCP and its results

3.3. **Semantic role labeling.** Semantic role labeling is another important factor in the semantization process. Most existing Chinese semantic role labeling systems are generally dependent on complete syntactic analysis. Since the performance of Chinese parser is still relatively low, especially in medical field, the result of automated semantic role labeling is unsatisfactory [7]. In view of this, shallow syntactic parsing is introduced instead of complete syntactic analysis as the input for Chinese semantic role labeling. In order to fully reflect the characteristics of the Chinese language with fine-grained description of the words in particular, we refer to the "*Modern Chinese Lexicon Structure* [8]", to extract feature characters from verbs and nouns as well as relevant features of word formation.

Character set in Chinese is relatively closed collection with long-term stability. Word formation in Chinese is much like syntactic structure. With the information, it is much easier to find relations between words and new word discovering. For example, in dealing with a new word, the POS and semantic role of word in a sentence can be inferred from the preceding word and back word. Based on the analysis above, the features having a positive effect in syntactic analysis and semantic role labeling include: center morpheme, morpheme type, word formation, and sequence information. In the parsing stage, in order to facilitate a fairer semantic role labeling, there is no introduction of external resources, and we use rules to generate "pseudo center morpheme" as a feature to provide the basis for parsing [9]. Chunk definition [10] is adopted, and a total of 12 types of chunk are chosen.

The following is a detailed description of the method:

(1) Input: correct word segmentation and POS tagging result;

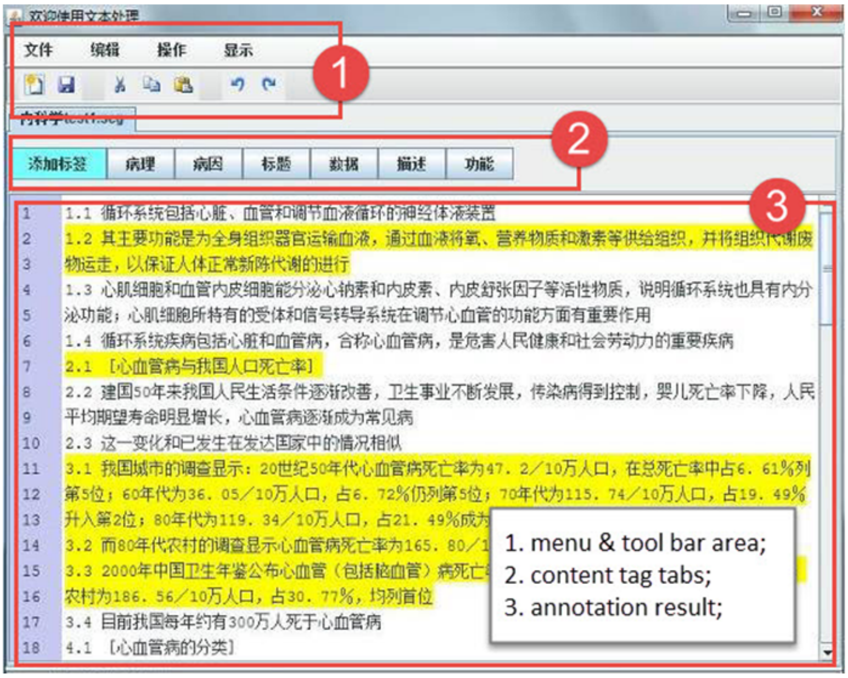
(2) Shallow syntactic parsing: adopting chunk definition, combined with IOB2 notation, the shallow parsing problem is transformed into a sequence labeling problem;

(3) Semantic role labeling: based on the identification sequence of arguments and arguments classification, SRL is conducted.

3.4. **Discourse annotation.** Discourse annotation aims to reveal the deep relations beyond the sentence level, and present the inner relations of the complicated discourses through normalized annotation process [11]. Taking consideration of the Chinese medical characteristics, we propose an easily operated discourse annotation process based on the traditional discourse theories and their corresponding corpus. We describe the deep

Table 1. Types of chunk in SRL

tag	meaning
ADJP	Adjective phrase
ADVP	Adverb phrase
CLP	CLP Classifier
DNP	Deg “的” phrase
DP	Determiner phrase
DNP	Deg “得” phrase
LOC	Location phrase
LST	List marker
NP	Noun phrase
PP	Prepositional phrase
QP	Quantifier phrase
VP	Verb phrase



1.1 循环系统包括心脏、血管和调节血液循环的神经体液装置
1.2 其主要功能是为全身组织器官运输血液，通过血液将氧、营养物质和激素等供给组织，并将组织代谢废
物运走，以保证人体正常新陈代谢的进行
1.3 心肌细胞和血管内皮细胞能分泌心钠素和内皮素、内皮舒张因子等活性物质，说明循环系统也具有内分
泌功能；心肌细胞所特有的受体和信号转导系统在调节心血管的功能方面有重要作用
1.4 循环系统疾病包括心脏和血管病，合称心血管病，是危害人民健康和社会劳动力的重要疾病
2.1 ［心血管病与我国人口死亡率］
2.2 建国50年来我国人民生活条件逐渐改善，卫生事业不断发展，传染病得到控制，婴儿死亡率下降，人民
平均期望寿命明显增长，心血管病逐渐成为常见病
2.3 这一变化和已发生在发达国家中的情况相似
3.1 我国城市的调查显示：20世纪50年代心血管病死亡率为47．2／10万人口，在总死亡率中占6．61%列
第5位；60年代为36．05／10万人口，占6．72%仍列第5位；70年代为115．74／10万人口，占19．49%
升入第2位；80年代为119．34／10万人口，占21．49%成为
3.2 而80年代农村的调查显示心血管病死亡率为165．80／1
3.3 2000年中国卫生年鉴公布心血管（包括脑血管）病死亡
农村为186．56／10万人口，占30．77%，均列首位
3.4 目前我国每年约有300万人死于心血管病
4.1 ［心血管病的分类］

1. menu & tool bar area;
2. content tag tabs;
3. annotation result;

Figure 3. GUI of discourse annotation platform

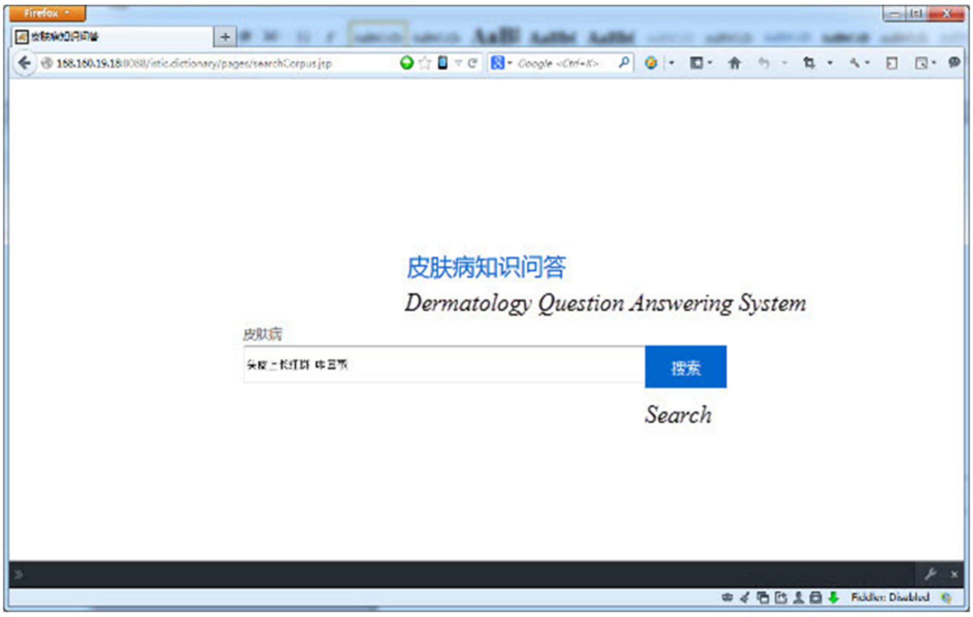
structure and meaning of the discourse in three aspects: content, relation and reference. Content tagging is to mark the fragments with clear meaning. General and domain specific tags are introduced to ensure the flexibility of the content tagging: general tags are regarded as the default tags, and domain tags can be expanded as needed within the medical fields. Relation annotation is to describe the logical relation between the adjacent components in a discourse. Reference tagging is to mark pronouns, nouns, etc. By cutting the medical unstructured text into paragraphs, sentences and clauses, the annotating process adopts a bottom-up way, from small to large linguistic unit, to tag the relations, and marks the core and subsidiary ones, until the formation of a discourse annotation tree is comparatively complete [12].

4. **Application.**

QA system allows users to use casual language to organize questions, unlike traditional search engines that require users to input the keywords in a rather compelled way. With the information of semantic annotation, semantic role labeling, and discourse annotation, we are able to match query words to a deep semantic structure and relation, so as to obtain a much more precise answer to present to users.

"*Dermatology and Venereology (Fifth Edition)*" is chosen as raw material of the corpus. With all the tools and platforms above, the corpus is being processed accordingly. From the Good Doctor Dermatology Online and Baidu Knowledge Dermatology Category, we select 100 typical questions in medical field. In this QA system application, the results are shown as Figures 4 and 5.

As can be seen from above, when a user inputs a question in the search box, the system analyzes the query text, decomposes the text into triples, conducts a search in the database, and presents the user an interactive dialog with choices to help further filter the results so as to provide the precise answer to the question. By clicking word links relevant to the question in the result page, a user can also browse related information in



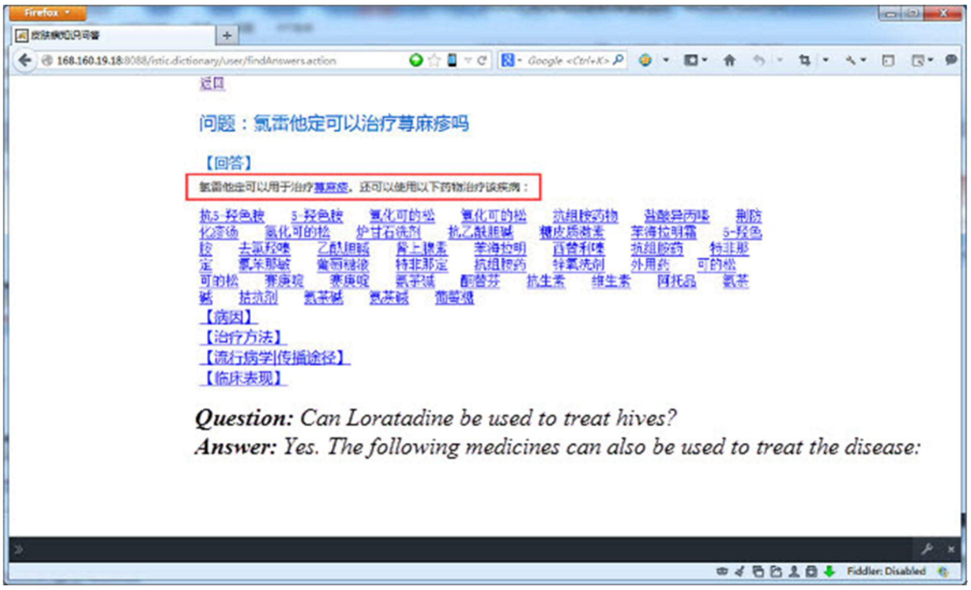
皮肤病知识问答

Dermatology Question Answering System

皮肤病

头皮上长红斑 伴三天

FIGURE 4. Homepage of QA system



问题：氯雷他定可以治疗荨麻疹吗

【回答】

氯雷他定可以用于治疗荨麻疹。还可以使用以下药物治疗该疾病：

抗5-羟色胺 5-羟色胺 氢化可的松 氢化可的松 抗组胺药物 盐酸异丙嗪 荆防化疹汤 氯化可的松 炉甘石洗剂 抗乙酰胆碱 糖皮质激素 苯海拉明霜 5-羟色胺 去氯羟嗪 乙酰胆碱 肾上腺素 苯海拉明 西替利嗪 抗组胺药 特非那定 氯苯那敏 葡萄糖酸 特非那定 抗组胺药 锌氧洗剂 外用药 可的松 可的松 赛庚啶 赛庚啶 氨茶碱 酮替芬 抗生素 维生素 阿托品 氨茶碱 拮抗剂 氨茶碱 氨茶碱 葡萄糖

【病因】
【治疗方法】
【流行病学|传播途径】
【临床表现】

Question: *Can Loratadine be used to treat hives?*
Answer: *Yes. The following medicines can also be used to treat the disease:*

FIGURE 5. Query result of QA system

his interests. It can be concluded that our methods can effectively process semantization of medical unstructured text.

5. Conclusions.

Based on the characteristics of Chinese medicine unstructured text, the paper proposes to process semantization of unstructured text in medical field by segmentation, POS tagging, semantic annotation, semantic role labeling, and discourse annotation to achieve semantization of unstructured text in medical field. The experimental results show that the proposed methods can effectively process semantization of medical unstructured text. In the future we intend to proceed along two lines in parallel: on one hand, to broaden the scope of semantization by adopting some other technologies and tools; on the other hand, to improve some key algorism to meet the needs of semantization in other fields other than medical field.

Acknowledgment. This work is partially supported by National Key Project of Scientific and Technical Supporting Program No. 2013BAH21B02. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] V. Nebot and R. Berlanga, Exploiting semantic annotations for open information extraction: An experience in the biomedical domain, *Knowledge and Information Systems*, vol.38, no.2, pp.365-389, 2014.

[2] S. Ou, Semantic representation of Chinese thesauri, *Library and Information Service*, vol.59, no.16, pp.110-118, 2015.

[3] Y. Liu, X.-F. Chen, Z. Sui, Y. Hu and Q. Zhao, Research on semantic method of library resources' organizing, *ICIC Express Letters*, vol.5, no.4(A), pp.1011-1017, 2011.

[4] H. Bai and X. Qiao, Study of semantic bibliography base on ontology and linked data, *New Technology of Library and Information Service*, vol.26, no.9, pp.18-27, 2010.

[5] J. Zhang and Y. Liu, Medical literature's semantic status and solution, *International Journal of Knowledge and Language Processing*, 2013.

[6] Y. Liu, H. Shi, D. 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.

[7] L. Qiu, L. Wu, K. Zhao and C. Hu, Improving Chinese dependency parsing with auto-extracted dependency triples, *Int. J. of Asian Lang. Proc.*, vol.22, no.2, pp.75-84, 2012.

[8] S. Yu, H. Duan, X. Zhu, B. Swen and B. Chang, Specification for corpus processing at Peking University: Word segmentation, POS tagging and phonetic notation, *Journal of Chinese Language and Computing*, vol.13, no.2, pp.121-158, 2003.

[9] X. Wang, *The Research and Implementation on Chinese Semantic Role Labeling Based on Lightweight Syntactic Information (in Chinese)*, Ph.D. Thesis, Peking University, 2012.

[10] W. Chen, Y. Zhang and H. Isahara, An empirical study of Chinese chunking, *Proc. of the COLING/ACL on Main Conference Poster Sessions*, 2006.

[11] Y. Zhou and N. Xue, PDTB-style discourse annotation of Chinese text, *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, vol.1, 2012.

[12] Y. Wang, *The Development and Application of Discourse Annotation Platform*, Ph.D. Thesis, Peking University, 2014.