ICIC Express Letters
Volume **10**, Number **7**, July **2016**

ICIC International ©2016 ISSN 1881-803X
pp. 1725–1730

CONSTRUCTION OF SEMANTIC LEXICON OF DERMATOLOGY

Yang Zhou[1], Nan Xiang[1], Ruixiang Wang[1], Yao Liu[2]
Xingliang Qi[1] and Zhenguo Wang[1]

[1]Institute of Chinese Medical History and Literature
Shandong University of Traditional Chinese Medicine
No. 4655, Daxue Road, Jinan 250355, P. R. China
zhenguow@163.com

[2]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. *This paper introduces the construction of the Semantic Lexicon of Dermatology by using the theory and technology of Natural Language Processing (NLP) which can provide the database, such as automatic semantic analysis, word sense disambiguation, for NLP. This paper analyzes some problems of the Semantic Lexicon of Dermatology including system, standard and code of classification, range of vocabulary, etc. It also introduces the main function of the editing system of the Semantic Lexicon of Dermatology.*

Keywords: Semantic Lexicon, Dermatology, Natural Language Processing (NLP)

1. **Introduction.** Construction of the Semantic Lexicon is the basic work of Natural Language Processing (NLP). The Semantic Lexicon occupies a key position in the study of NLP. Concerned people around the world have developed many large-scale Semantic Dictionaries. Semantic Dictionaries in medical field have also been constructing gradually [1]. With the development of Precision Medicine, the study of NLP in medical field has become a new research hot topic. Accordingly, the construction of medical professional Semantic Lexicon is imperative.

Usually, there are two ways to construct the Semantic Lexicon. One way is based on the subjective judgment of experts in concerned fields. The other way is based on Automatic Clustering [2]. Construction of the Semantic Lexicon of Dermatology combines the two ways mentioned above. On one hand, construction of the professional Semantic Lexicon must take the existing research results in the field as a reference which means that we should make full use of the existing dermatology dictionaries [3]. On the other hand, we should introduce new terms of dermatology by using Dermatology Lexicon Editing System based on corpus of works and thesis of dermatology. The Semantic Lexicon of Dermatology will be used in NLP widely.

This paper introduces the principle for construction of the Semantic Lexicon of Dermatology, the main function and framework of the Editing System of the Semantic Lexicon of Dermatology.

2. Principles for the Construction of the Semantic Lexicon of Dermatology.

2.1. **Principles for collecting terms.** The terms have wide coverage. It contains not only dermatology, venereology, cosmetology, dermatological surgery, but also clinical dermatology and basic medical knowledge, such as molecular biology, immunology, biochemistry, pathology, microbiology, pharmacology, and biology.

We collect new words, including disease names, drug names and other professional terms.

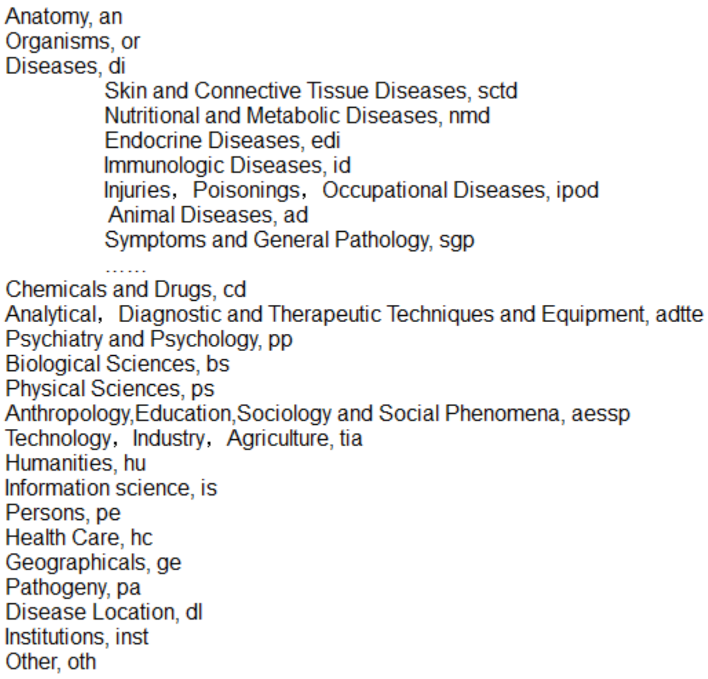
We choose the vocabulary with high frequency use for translating English to Chinese. If there are multiple Chinese words which express the same conceptual mean, we will get the word which is used most commonly.

2.2. **Classification of the Semantic Lexicon of Dermatology.** Skin disorders are not only the diseases of skin, but also involve many human organs which are closely related with social and natural environment. Accordingly, the terms of dermatology not only contains dermatosis itself, but also involves many related medical specialties, such as internal medicine, surgery, pharmacy, physiology and pathology. Therefore, the classification of the Semantic Lexicon of Dermatology actually covers the whole medical system. Moreover, it is also related to the non-medical factors, such as natural science, and geography.

The classification of the Semantic Lexicon of Dermatology uses MESH classification system for reference and it has been extended and modified [4-6]. In order to collect the terms maximally, we use the existing dermatology dictionary, the latest clinical knowledge and the corpus of dermatology. We collect 300,000 terms and classify them into twenty-four first-rate categories.

First-rate category contains Anatomy, Organisms, Diseases, Pathogeny, Disease Location, Chemicals and Drugs, Analytical, Diagnostic and Therapeutic Techniques and Equipment, Psychiatry and Psychology, Biological Sciences, Physical Sciences, Anthropology, Education, Sociology and Social Phenomena, Technology, Industry, Agriculture, Humanities, Information Science, Persons, Health Care, Geographicals, Institutions, and Others (Figure 1).

Each category is divided into the lower categories; for example, the 'Pathogeny' category is divided into categories of Acute Geographical Factors, Neural Factors, Aggravated Factors, etc. And the category 'Aggravated Factors' is divided into categories of Insolation Weathering, Improper Diet, Improper Use of Drugs, etc. The lowest category is the fourth level directory.



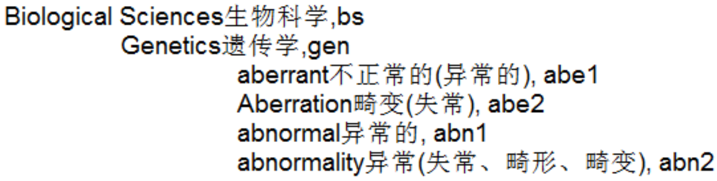
Anatomy, an
Organisms, or
Diseases, di
    Skin and Connective Tissue Diseases, sctd
    Nutritional and Metabolic Diseases, nmd
    Endocrine Diseases, edi
    Immunologic Diseases, id
    Injuries，Poisonings，Occupational Diseases, ipod
    Animal Diseases, ad
    Symptoms and General Pathology, sgp
    ……
Chemicals and Drugs, cd
Analytical，Diagnostic and Therapeutic Techniques and Equipment, adtte
Psychiatry and Psychology, pp
Biological Sciences, bs
Physical Sciences, ps
Anthropology,Education,Sociology and Social Phenomena, aessp
Technology，Industry，Agriculture, tia
Humanities, hu
Information science, is
Persons, pe
Health Care, hc
Geographicals, ge
Pathogeny, pa
Disease Location, dl
Institutions, inst
Other, oth

Figure 1. Classification of the Semantic Lexicon of Dermatology

2.3. Principles for marking terms. We use the first letter of English words as the tab of the term. For example, we mark 'biological sciences' as 'bs'.

If there is only one word, we use the first two letters, such as marking 'diseases' as 'di'.

If two tabs are the same, we use the first three letters, such as marking "genetics" as "gen". If there is still a repetition, then we use the Arabia number to distinguish (Figure 2).



Biological Sciences生物科学,bs
Genetics遗传学,gen
aberrant不正常的(异常的), abe1
Aberration畸变(失常), abe2
abnormal异常的, abn1
abnormality异常(失常、畸形、畸变), abn2

Figure 2. An example of marking terms

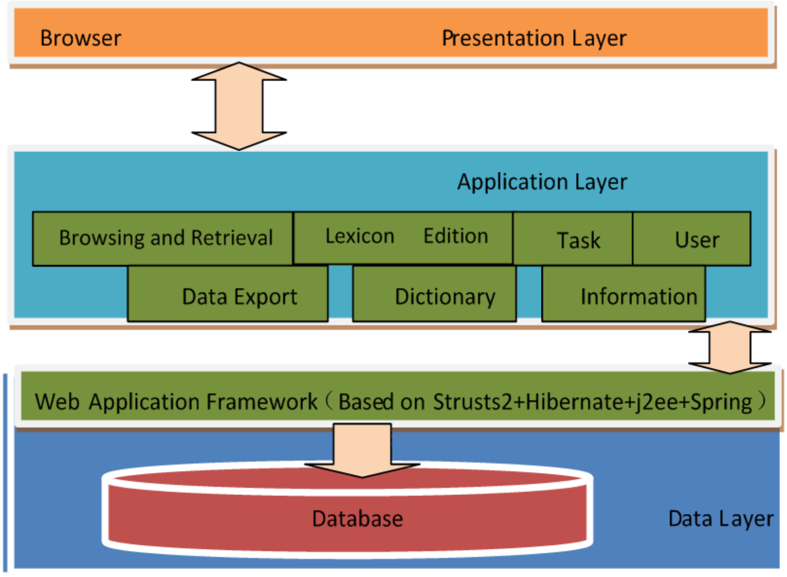


Figure 3. System framework

3. Editing System of the Lexicon of Dermatology. The system is based on the B/S architecture of the Internet. It can be accessed by the browser without the client. Professionals can use the system to modify and check the lexicon according to the reference statements and their own experience, and it supports multi-collaborative editing. It can also derive lexicon and word attributes.

There are three modules in the system structure. They are presentation layer, application layer and data layer. This system based on database is developed by using web application framework (based on Strusts2 + Hibernate + j2ee + Spring) (Figures 3 and 4). The function of 'User Management' module includes registration, login, modifying password. This part will not be detailedly described.

3.1. Proofreading and editing terms. It can add, delete, modify, query and upload the new domain lexicon. We can get the most relevant information from Internet sites in accordance with the data itself by using the system. There are four functions in the system of proofreading and editing terms, Batch Input, Terms Retrieval, Terms Editing, Terms Outputting (Figure 5).

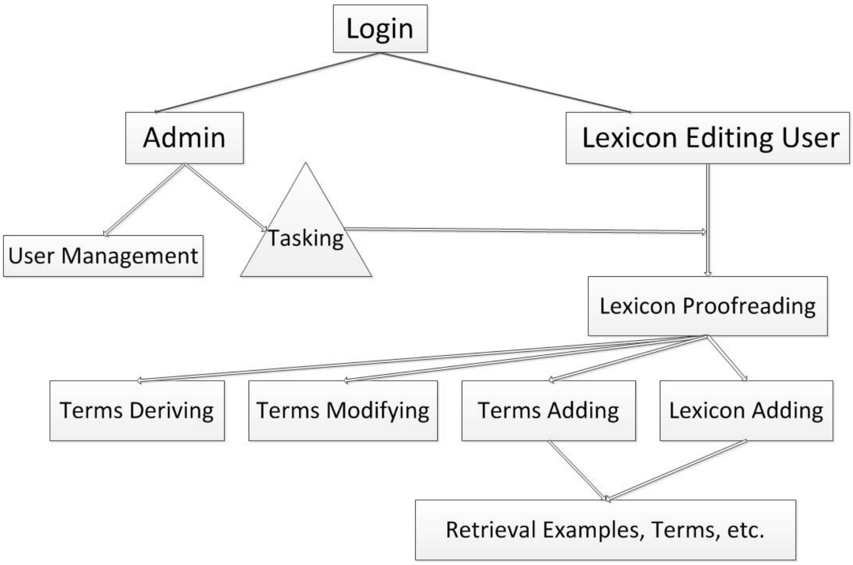
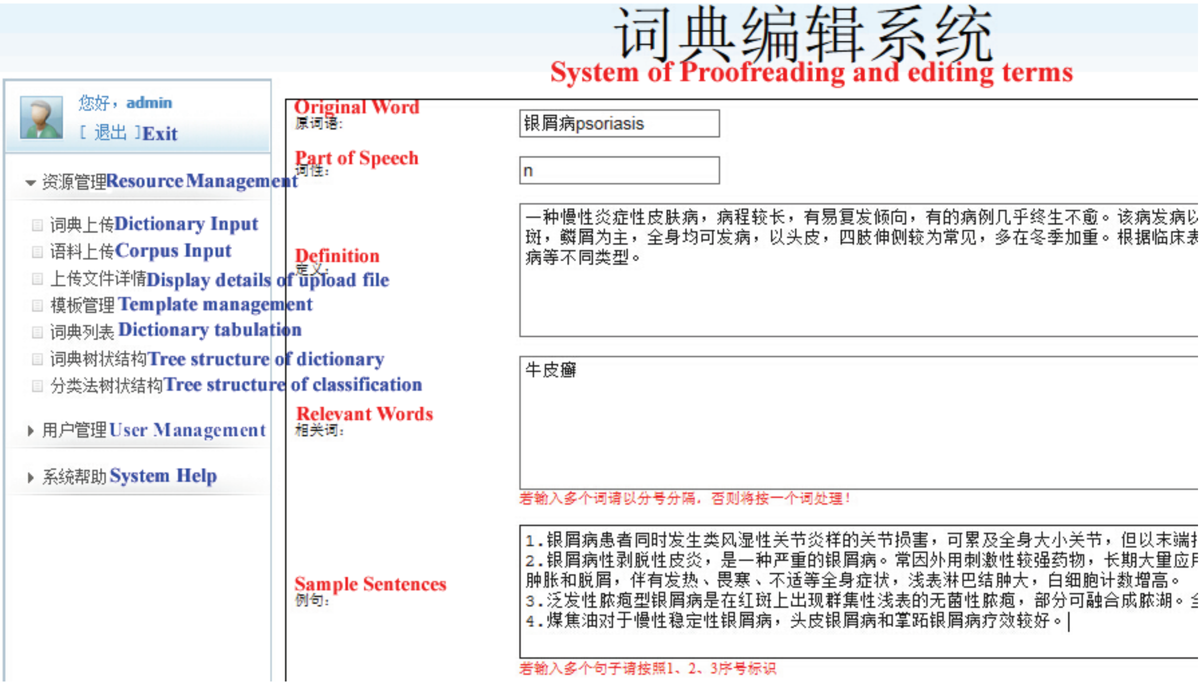


FIGURE 4. The procession of the system


词典编辑系统
System of Proofreading and editing terms
您好，admin
[退出]Exit
资源管理Resource Management
词典上传Dictionary Input
语料上传Corpus Input
上传文件详情Display details of upload file
模板管理 Template management
词典列表 Dictionary tabulation
词典树状结构Tree structure of dictionary
分类法树状结构Tree structure of classification
用户管理User Management
系统帮助 System Help

Original Word 原词语: 银屑病psoriasis

Part of Speech 词性: n

Definition 定义: 一种慢性炎症性皮肤病，病程较长，有易复发倾向，有的病例几乎终生不愈。该病发病以
斑，鳞屑为主，全身均可发病，以头皮，四肢伸侧较为常见，多在冬季加重。根据临床表
病等不同类型。

Relevant Words 相关词: 牛皮癣

若输入多个词请以分号分隔，否则将按一个词处理！

Sample Sentences 例句:
1.银屑病患者同时发生类风湿性关节炎样的关节损害，可累及全身大小关节，但以末端
2.银屑病性剥脱性皮炎，是一种严重的银屑病。常因外用刺激性较强药物，长期大量应
肿胀和脱屑，伴有发热、畏寒、不适等全身症状，浅表淋巴结肿大，白细胞计数增高。
3.泛发性脓疱型银屑病是在红斑上出现群集性浅表的无菌性脓疱，部分可融合成脓湖。
4.煤焦油对于慢性稳定性银屑病，头皮银屑病和掌跖银屑病疗效较好。|

若输入多个句子请按照1、2、3序号标识

FIGURE 5. Proofreading and editing terms

3.1.1. *Batch input.* It can import the data into the database by using batch processing method. Select the TXT or Excel file, and click 'Upload' to import the data into the database. Then, it shows a word in a row in the TXT file, and it indicates word, English name, definition, another name in columns in the Excel file.

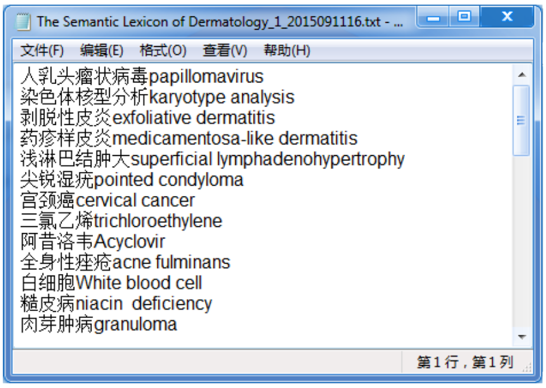
3.1.2. *Terms retrieval.* This system provides two search ways and search functions that base on combination of multiple choices. It is word characteristic retrieval, keywords searching and domain retrieval.

Users choose the characteristic of words from the word character tabulations in the search area. The system will provide the relative words tabulations of the word characters selected. We input keywords in the relative input box, and click 'search' button. We will get the word character selected and words information which contains the key words. The key words should be fuzzy matching which means the words contain the key words.

If we check the ‘yes’ choice of whether searching entry that has no match with upper and lower entry only, the query range will be limited in the entry that has no match with upper and lower entry.

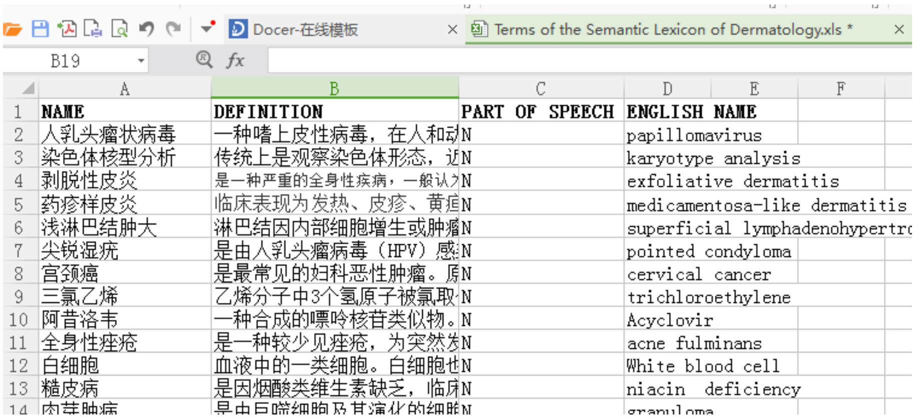
3.1.3. *Terms editing.* It can edit the related properties of terms, such as Chinese names. It can also add new terms with name, characteristic, related words, etc. At the same time, Chinese name, English name, another name and examples will be added into the corresponding properties by using retrieval module automatically.

3.1.4. *Terms outputting.* This system provides three output formats of terms (Figures 6-8). One format is to export the words list of the current retrieval conditions to TXT. The export format is to display a word in line; the default file name is ‘user ID-current time’. Another way is to export the words list of the current retrieval conditions to an Excel table, the table head for the entry name, Chinese name, English name, etc. The default file name is ‘current time-user ID-property’. The last format is to export structured lexicon.


The Semantic Lexicon of Dermatology_1_2015091116.txt - ...
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
第1行，第1列

人乳头瘤状病毒papillomavirus
染色体核型分析karyotype analysis
剥脱性皮炎exfoliative dermatitis
药疹样皮炎medicamentosa-like dermatitis
浅淋巴结肿大superficial lymphadenohypertrophy
尖锐湿疣pointed condyloma
宫颈癌cervical cancer
三氯乙烯trichloroethylene
阿昔洛韦Acyclovir
全身性痤疮acne fulminans
白细胞White blood cell
糙皮病niacin deficiency
肉芽肿病granuloma

FIGURE 6. Format1 – TXT



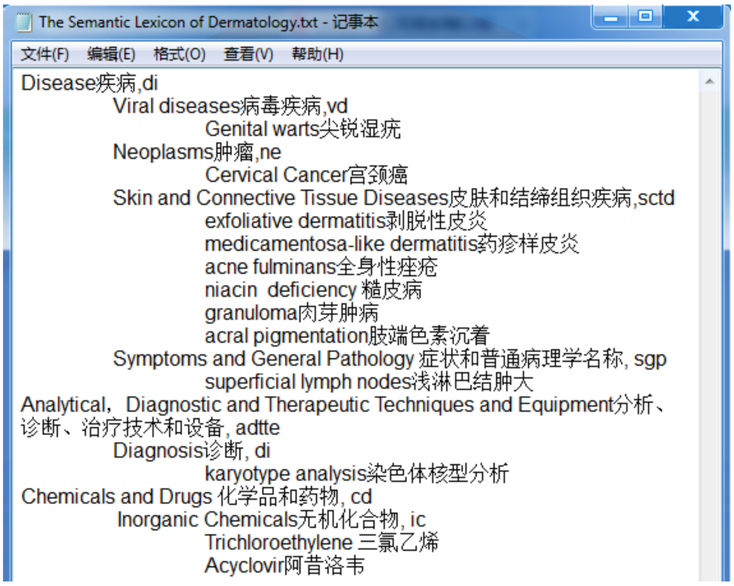
	A	B	C	D	E	F
1	NAME	DEFINITION	PART OF SPEECH	ENGLISH NAME		
2	人乳头瘤状病毒	一种嗜上皮性病毒，在人和动	N	papillomavirus		
3	染色体核型分析	传统上是观察染色体形态，近	N	karyotype analysis		
4	剥脱性皮炎	是一种严重的全身性疾病，一般认为	N	exfoliative dermatitis		
5	药疹样皮炎	临床表现为发热、皮疹、黄疸	N	medicamentosa-like dermatitis		
6	浅淋巴结肿大	淋巴结因内部细胞增生或肿瘤	N	superficial lymphadenohypertro		
7	尖锐湿疣	是由人乳头瘤病毒（HPV）感	N	pointed condyloma		
8	宫颈癌	是最常见的妇科恶性肿瘤。原	N	cervical cancer		
9	三氯乙烯	乙烯分子中3个氢原子被氯取	N	trichloroethylene		
10	阿昔洛韦	一种合成的嘌呤核苷类似物。	N	Acyclovir		
11	全身性痤疮	是一种较少见痤疮，为突然发	N	acne fulminans		
12	白细胞	血液中的一类细胞。白细胞也	N	White blood cell		
13	糙皮病	是因烟酸类维生素缺乏，临床	N	niacin deficiency		
14	肉芽肿病	是由巨噬细胞及其演化的细胞	N	granuloma		

FIGURE 7. Format2 – Excel table

3.2. **Displaying details of upload file.** The file of lexicon type can create hyponymy. The system can save it automatically and create hyponymy and synonyms.

The file of corpus type can analyze the corpus documents. The system resolves the corpus document to a corresponding lexicon file of Excel format. Then save it and create hyponymy automatically.

3.3. **Editing lexicon tree structure.** The lexicon tree structure can be modified by adding, deleting, modifying, sequencing, and checking. We can also drag the nodes to other nodes directly.



```
Disease疾病,di
        Viral diseases病毒疾病,vd
                Genital warts尖锐湿疣
        Neoplasms肿瘤,ne
                Cervical Cancer宫颈癌
        Skin and Connective Tissue Diseases皮肤和结缔组织疾病,sctd
                exfoliative dermatitis剥脱性皮炎
                medicamentosa-like dermatitis药疹样皮炎
                acne fulminans全身性痤疮
                niacin  deficiency 糙皮病
                granuloma肉芽肿病
                acral pigmentation肢端色素沉着
        Symptoms and General Pathology 症状和普通病理学名称, sgp
                superficial lymph nodes浅淋巴结肿大
Analytical，Diagnostic and Therapeutic Techniques and Equipment分析、
诊断、治疗技术和设备, adtte
        Diagnosis诊断, di
                karyotype analysis染色体核型分析
Chemicals and Drugs 化学品和药物, cd
        Inorganic Chemicals无机化合物, ic
                Trichloroethylene 三氯乙烯
                Acyclovir阿昔洛韦
```

Figure 8. Format3 – Structured lexicon

4. Application of the Semantic Lexicon of Dermatology.

Semantic Lexicon of Dermatology is a mechanical dictionary in the professional field which serves for the language information processing. We constructed lexical category system for information processing based on characteristics of skin diseases in the lexicon. It fulfills the words classification, and describes the different word attributes by classification to achieve the structuring in a way. It will excavate implicit knowledge from massive text. Semantic analysis plays an important role in the NLP. Semantic Lexicon of Dermatology will provide comprehensive semantic knowledge for the sentence meaning analysis, lexical ambiguity elimination which will improve accuracy of interpretation of documents.

5. Conclusion.

In this paper, firstly, we obtain the terms from the textbook of dermatology and the clinical literature. Secondly, we classify the terms of dermatology by using MESH classification system. Finally, we introduce the Editing System of the Lexicon of Dermatology.

With the development of dermatology, there will be many new professional vocabularies. In the future work, we will continue to collect vocabulary by using the Editing System of the Lexicon of Dermatology, so as to optimize the Dermatology Lexicon.

Acknowledgment. This work is partially supported by 973 Natural Research Program of China No. 2007CB512601. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] P. Ruch, R. H. Baud, A. M. Rassinoux, P. Bouillon and A. G. Robert, Medical document anonymization with a semantic lexicon, *Proc. of AMIA Symp.*, vol.7, no.1, pp.729-733, 2000.
[2] T. Chen and M. Sun, Automated construction of Chinese thesaurus based on self-organizing map, *Journal of the China Society for Scientific and Technical Information*, vol.26, no.1, pp.77-83, 2007.
[3] W. Xu, *English Chinese Dictionary of Dermatovenereology*, Phoenix Science Press, 2007.
[4] *https://www.nlm.nih.gov/class//terms_cond.html [EB]*, 2014.
[5] B. Zhao, *China Clinical Dermatology*, Phoenix Science Press, 2010.
[6] X. Zhang, *Textbook Series of New Century – Dermatovenereology*, People's Medical Publishing House, 2002.