ICIC Express Letters
Part B: Applications
Volume **7**, Number **3**, March **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **693–699**

A STUDY ON THE EXTRACTION OF CHINESE PATENT KNOWLEDGE BASED ON QUALIA STRUCTURE AND SEMANTIC ROLES LABELING

Wei Chin, Yao Liu and Xiaodong Qiao

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

Received August 2015; accepted October 2015

Abstract. *This project aims to help users quickly grasp the essence of the patent documents and optimize the patent services. The previous researches on the extraction of the patent documents are mainly based on dictionaries and/or templates. However, such approaches have some problems in data processing. The present study adopts the theories of linguistics, specifically the qualia structure in generative lexicon theory, to the extraction of patent documents in Chinese. The patent verbs have been classified into 4 categories (invention type, form, constitution, function) and the semantic roles involved in important knowledge are divided into 15 categories (domain, part, material, etc.). The extraction based on such classification can overcome the above-mentioned difficulties and realize automatic reasoning in the extraction of patent documents.*

Keywords: Patent, Information extraction, Qualia structure, Semantic roles

1. **Introduction.** The patent literature is a type of information resources, which carries abundant of technology information. It helps to circulate the latest technology information and brings much economy profit. According to the statistics by the World Intellectual Property Organization, the patent literature covers around 90% to 95% of the new invention each year and the amount of patent applications increases each year by around 1000000 [1]. Facing these massive patent documents, it has been a key project to optimize the patent services and help users quickly grasp the essence of the documents.

The previous research on the extraction of patent literature is mainly dealing with the literature in mechanics. Zhao et al. make a research on 860 pieces of patent documents in the field of new energy vehicles with deep annotations. 973 cases of different verbs are retrieved and the top 10 verbs of high frequency is *baokuo* ‘include’, *lianjie* ‘connect’, *kongzhi* ‘control’, *juyou* ‘have’, *tigong* ‘supply’, *qudong* ‘drive’, *zaiyu* ‘lie in’, *chansheng* ‘produce’, *yongyu* ‘be used in’, *shezhi* ‘set’ [1]. Jiang et al. divide the abstract of patent documents into 5 parts: the overall judgment on the patent documents, the methods or technology used or abandoned, the principles involved, the components of the patent and the comments by the holder of the patent. In order to retrieve the information, they collect those verbs which express important information and the thesaurus reach around 50 pieces of papers. The extraction result is not very satisfactory as the f-value on action principle (31.19%) and the f-value on constitution and structure (47.54%) are not satisfying [2]. Based on verb sematic frame, Wang et al. distinguish three types of verbs: verbs of components, verbs of properties and verbs of functions [3].

From the above discussion, we know the extraction on patent literature is from the verb-based templates. In the above mentioned analysis, Wang et al. discuss the English patent literature [3], whose extraction is different from the one in Chinese. Zhao et al. have successfully found the 10 frequently used verbs [1]; however, these verbs have not been classified accordingly. Jiang et al. have made a comprehensive analysis on the verbs.

However, the extraction performance on action principle and constitution & structure is not satisfying. One reason may be related to the insufficiency of thesaurus and the failure of the extraction rules in covering all the language phenomena involved [2]. Sometimes it is hard to distinguish whether an element refers to constitution & structure or action principle. Take *lianjie* 'connect' as an example. In [*lianjie you*], it refers to the constitution & structure; while in [*jiang N1 he N2 lianjie, shide*...], it refers to action principle. If the sentence meaning or the clause meaning is not considered as a whole, the efficiency of extraction will not be high. As pointed out by Jiang et al., the extraction rules seem difficult to cover all the language phenomena. This is the weakness of the template-based extraction methods. What is more, the templates are always confined in a certain area and such methods could not help the computer to realize automatic reasoning. Given the above-mentioned problems, we propose to take a sematic structure analysis of the extraction of patent documents, instead of template-based methods. We adopt the qualia structure in generative lexicon theory and the semantic roles annotation, to the extraction of patent documents in Chinese. This analysis can overcome the above-mentioned difficulties and realize automatic reasoning.

2. **Theoretical Framework.** In this article, we mainly use the qualia structure and sematic roles analysis. We will briefly introduce them one by one.

2.1. **Qualia structure.** Qualia structure is a part of generative lexicon theory [4]. Generative lexicon deals with natural language semantics, in particular the semantics of words, i.e., the problem of compositionality. Pustejovsky characterizes a generative lexicon as a computational system involving at least four levels of representations: argument structure, event structure, qualia structure and lexical inheritance structure. Briefly, qualia structure is composed of four essential aspects of a word's meaning: FORMAL, CONSTITUTIVE, TELIC and AGENTIVE. FORMAL distinguishes it within a larger domain, which includes the quantity, shape, dimensions, color and positions. CONSTITUTIVE refers to the relation between an object and its constituent parts. TELIC refers to its purpose and function. AGENTIVE refers to the factors involved in its origin or "bring it about". Take 'book' as an example. Its FORMAL role is "to hold", which is used to carry information and knowledge; its TELIC role is to "to read"; and its AGENTIVE role is realized through writing [5]. Inspired by this classification, it is found these four aspects concern us most in analyzing patent documents. Specifically, it touches upon the following question: What are the characteristics of the patent (FORMAL)? What are the components of the patent (CONSTITUTIVE)? What could the patent be used to do or what improvement has been made (TELIC)? Who is the inventor of the patent or how is the invention created (AGENTIVE)? In the following part, we will classify the verbs in patent documents according to the invention type, characteristics, components and functions. Meanwhile, the argument structure and the related semantic roles are investigated.

2.2. **Semantic roles.** The extraction on patent documents could not go without verbs. By investigating the argument structure of the verbs, the semantic structure of the verbs involved in patent documents could be represented in a clear way, which helps the machine learning.

In line with HowNet [6] and Yuan [7], the semantic roles in patent documents have been classified into 15 types. The following are some examples.

(1) Main semantic roles

a. Principle semantic roles

A. Cause/Causer: the causer or initiator of an event

Ben faming de zhongjie zhuansu xiangbiyu tongchangde zhongjie zhuansu

this invention of n_mot we replacing.speed relatively normal n_mot we replacing.speed

mingxiande jiangdi, yinci shi ranliao xiaohao jiangdi.
obviously reduce so cause fuel take reduce
The n_mot we replacing speed of this invention is obviously lower than the usual speed; therefore, the fuel is reduced accordingly.

B. Relevant. It refers to the topic in the events of "relation" types except possession relation, e.g.:
ben faming shuyu huagong lingyu.
this invention belong.to chemical area
'This invention is related to the chemistry area.'

C. Whole. It refers to the entity which is the whole of its parts, e.g.:
gai xitong **baokuo** yi zhong...
this system include one type...
The system **includes** one type of... '

D. Possessor. It refers to the owner in possession relation, e.g.,
yi zhong cheng fang xing de dianhuaxue zujian, baokuo: **juyou**
one type present square shape of electrochemistry component include have
duanbu he wai biaomian de rongqi.
terminal.portion and outer surface of container
A square electrochemical element, which consists of a terminal portion and an outer surface.

One thing to note is that principle semantic roles usually locate the subject position, as the examples shown above.

b. Affected semantic roles

A. ResultEvent. It refers to the event which is caused by a certain action or event, e.g.,
Ben faming de zhongjie zhuansu xiangbiyu tongchangde zhongjie zhuansu mingxiande jiangdi, yinci shi ranliao xiaohao jiangdi.
The n_mot we replacing speed of this invention is obviously lower than the usual speed; therefore, the fuel is reduced accordingly.

B. Relative. The entity is relativized to the theme in the event.
ben faming sheji yi zhong lidianzi dianchi zhengji cailiao de zhibei gongyi.
this invention involve one type lithiumion batteries cathode material of make technique
'This invention involves a technology on the cathode material of lithiumion batteries.'

C. OfPart: the part of an entity in the events of "include" type and "exclude" type, e.g.,
gai xitong baokuo yi zhong chuandongshi chuansong cheliang...
this system include one type transmission transfer car
'This system includes a type of transmission car.'

D. Possession: the entity is owned in possession relation, e.g.,
...**juyou** duan bu...
have terminal part
'...have/has a terminal part and...'

E. Coagent: the entity that should act equally together with another.
nei zhuanzi tiexin bei guding zai shuru zhou shang bing yu zhi yiqi xuanzhuan.
that internal.rotor iron.core passive fix on input axis and with it together spin
'The iron core of internal rotor is fixed on the input axis and it spins with the axis.'

Usually, affected semantic roles collocate with principle semantic roles, as ResultEvent usually collates with Causer, and Relevant with Theme.

(2) periphery semantic roles

a. Means

A. Instrument: the entity which is used as a tool in an event, e.g.,
tongguo lianghuade dianchi xingnengshuju ..., yuce dianchi xingneng.

by quantified battery SOF predict batteries performance
'The performance of the batteries could be predicted through the quantified SOF of batteries.'

B. Material: the entity out of which something is made.
Shuangjiban you buxiugang boban jicai ji liang ce biaomian de tan luo bomo goucheng.
bipolar.plate is stainless steel sheet and two side surface of carbon chromium thin film consist
'The bipolar plates consist of stainless steel sheet and carbon chromium thin film at the either side.'

C. Manner: the way in which the event happens
Jiang PVDF yong jingdianfangsi fangfa zhicheng nami xianwei.
use PVDF with electrospinning method make nano fibers
'Use the electrospinning method into PVDF to make Nanofibers.'
As shown above, the means arguments usually co-occur with verbs.

b. Circumstance

A. Location: in which an action takes place, e.g.,
suo shu xingxingchilun jigou zhiyu fadongji he fadianji zhijian .
SUO say planetary.gears structure lie engine and alternator between
'The planetary gears lie between the engine and the alternator.'

B. Range: the period of time which covers from the present till an event occurs in the future [6], e.g.,
yi zhong 12 zhen wenban zidong chongkongji shiyongyu zhizuo
one type 12 needle plate automatic needle-punching machine is.fit.for make
gongchen 600 zhen he 900 zhen wenban.
nominal 600 needle and 900 needle plate
'A needle-punching machine with 12 needle is fit for making a plate with nominal 600 needles or 900 needles.'

C. Domain: the area an event takes place in, e.g.,
ben faming ke yongyu xiandai chun diandong qiche shang . . .
this invention can apply.in modern pure electrical car upon
'This invention can be applied into electrical cars.'
The circumstance arguments are mainly used to express Location, Quantity, Time and Range involved in an invention.

3. **Solution.** In line with qualia structure, the verbs with important knowledge in patent documents have been sorted out. After studying the argument structure of these verbs, the rules on extraction could be made. The following are the detailed processes involved.

3.1. **Overall frame.** The data is pre-processed with segmentation, word tagging and syntactic analysis. The syntactic analysis involved directly influences the analysis of semantic roles. With the sematic roles annotation, the argument structure of those frequently used verbs are extracted, and the verbs are matched with the event patterns for further extraction. The flow chart is as Figure 1.

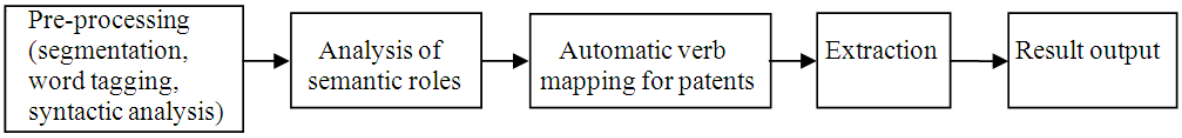
Pre-processing (segmentation, word tagging, syntactic analysis)
Analysis of semantic roles
Automatic verb mapping for patents
Extraction
Result output

FIGURE 1. The flow chart of knowledge extraction

3.2. **Extraction rules.** In line with the qualia structure, the frequently used verbs in patent documents are classified. In studying the argument structure of these verbs, some clue words have been singled out accordingly.

3.2.1. *The classification of verbs in patent documents.* In line with the qualia structure of arguments, 4 types of verbs could be classified in patent documents: they are invention type (cf. AGENTIVE role), formal type (cf. FORMAL role), constitutive type (cf. CONSTITUTIVE role) and function type (cf. Telic role). The following are some examples:

(1) Invention type: the area or type of an invention, product or technology. e.g.,
ben faming **shuyu**...
this invention belong.to
'This invention belongs to...'

(2) Formal type: the traits which distinguish the current patent from others.
qi tezheng **zaiyu**...
it traits lie.in
'The traits lie in...'

(3) Constitutive type: the relations between the component involved in the patent. e.g.,
chongkongmo **baokuo**...
punching.die include
'The punching die includes...'

(4) Function type: the function of the patent.
...jinger **shixian**...
then realize
'Then ... is realized.'

As exemplified as above, *shuyu* in (1) tells us the invention type of a patent, *zaiyu* 'lie in' in (2) brings out the traits, *baokuo* 'include' in (3) tells us the component information and *shixian* 'realize' in (4) expresses the function of the patent. In this way, the verbs in patent documents on machinery is summarized as below.

Table 1. Verbs extracted in patent documents

Knowledge	verbs
Invention type	*shuyu* 'belong to', *sheji* 'involve', *tigong* 'supply', *gongkai* 'publicize', *shihe* 'fit', *wei* 'for'
Formal type	*you/juyou* 'have', *zai/zaiyu* 'lie in'
Constitutive type	*caiyong* 'use', *chuandi* 'transfer', *zhicheng* 'make', *goucheng* 'constitute', *zucheng* 'consist of', *shezhi/zhuangzhi* 'set', *anzhuang* 'equip', *lianjie* 'connect', *peihe* 'cooperate', *jiechu* 'touch', *sheyu/anzhuang yu/zhuangzhiyu/zhuangsheyu* 'set', *sheyou/anzhuangyou/zhuangzhiyou/zhuangsheyou/shezhiyou* 'equipped with'
Function type	*shixian* 'realize', *jiejue* 'solve', *youhua* 'optimize', *tigao* 'raise', *zengjia* 'increase', *cujin* 'promote', *jianshao/jiangdi* 'reduce', *kefu* 'overcome', *gaishan* 'improve', *dadao/dacheng* 'achieve', *baozheng* 'ensure', *xiaochu* 'eliminate', *fangzhi* 'prevent', *bimian* 'avoid', *zhidao* 'guide', *bianyu/youzhuyu/youliyu* 'be helpful in', *shiheyu* 'be fit for', *yongyu/yingyongyu* 'apply in', *ke/keyi* 'may', *neng/nenggou* 'can', *deyi* 'can', *shi/shide* 'cause'

3.2.2. *The interaction between argument structure and verbs.* Based on the types of the verbs and the argument structure involved, the clue words in patent documents on machinery is depicted as Table 2.

TABLE 2. The interaction between verbs and their argument structure in patent documents of machinery

Knowledge	clue words
Invention type	a. (Relevant) *shuyu* ‘belong to’/*sheji* ‘involve’/*tigong* ‘supply’/*gongkai* ‘publicize’ (Relative) b. (Relevant) *wei* ‘for’ (Relative) c. (Relevant) *yong/yingyong/shiyongyu* ‘use’ NP de *lingyu* ‘scope’ /*fangfa* ‘method’ /*jishu* ‘technique’ (Relative or Domain) d. (Relevant) *keyi zuowei* ‘may use’ NP *de lingyu* ‘area’/*fangfa* ‘method’ /*jishu* ‘technique’ (Relative)
Formative type	a. *qi tezheng* ‘its characteristics’ *shi/you/zaiyu* ‘be/have/lie in’ (relative)
Constitutive type	a. *you/yong/caiyong* ‘be/use’ (Tool/Material/Manner/Range) *zhicheng/zucheng/goucheng* ‘make’ b. *zai* ‘at’ (Location) *shang* ‘up’ *sheyou/anzhuang/shezhi* ‘set up/equip with’ c. *yu* ‘with’ (Coagent) *lianjie/peihe/jiechu/shezhi* ‘connect/touch/set’ d. *tongguo/you* ‘by’ (Tool/Material/Manner/Range) *ba* ‘cause’ (Theme) *he* ‘and’ (Coagent) e. *liantong/lianjie/peihe/jiechu/shezhi* ‘connect/cooperate/set’ f. *tongguo* ‘through’ (Tool/Material/Manner/Domain) *chuandi/tigong* ‘tranfer/supply’ g. *tongguo/jingguo/caiyong* ‘by/through/by’ (Tool/Material/Manner/Range) *lai* ‘come’ (ResultEvt) h. *tongguo/jingguo/caiyong* ‘by/through/by’ (Tool/Material/Manner/Range) *shi/yishi/shide* ‘cause’ (ResultEvt)
Function type	a. (Theme) *ke yongyu* ‘can be used in’ (Domain/Range) b. *shi* ‘cause’ (Theme) *dedao* ‘get’ (ResultEvt) c. *dadao/dacheng* ‘achieve’ (ResultEvt) *de xiaoguo* ‘the effect’ d. *mudi/xiaogu/youdian zaiyu* ‘purpose/effects/strength lie in’ (ResultEvt)

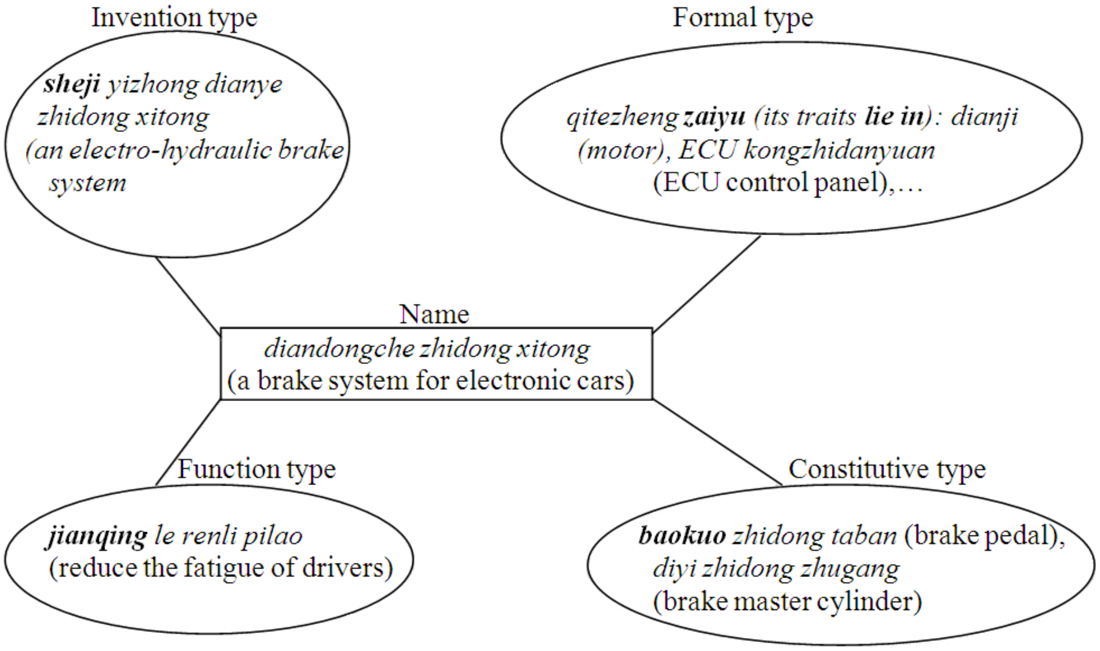
Invention type
sheji yizhong dianye zhidong xitong
(an electro-hydraulic brake system
Formal type
qitezheng **zaiyu** (its traits **lie in**): dianji (motor), ECU kongzhidanyuan
(ECU control panel),...
Name
diandongche zhidong xitong
(a brake system for electronic cars)
Function type
jianqing le renli pilao
(reduce the fatigue of drivers)
Constitutive type
baokuo zhidong taban (brake pedal), diyi zhidong zhugang
(brake master cylinder)

FIGURE 2. The visualization of the extraction of patent knowledge

3.2.3. *Visual display.* Based on the rules, the knowledge in patent documents could be extracted and displayed visually. We will take the following paragraph as an example:

ben faming ***sheji*** *(involve) yizhong dianye zhidong xitong,* ***baokuo*** *(include) zhidong taban yu zhidong taban xianglian de diyi zhidong zhugang, qi tezheng* ***zaiyu*** *(lie in): hai baokuo dianji, tui gan, di'er zhidong zhugang, ECU kongzhi danyuan, suoshu dianji tongguo kongzhi tuigan yu di'er zhidong zhugang xianglian, ECU kongzhi danyuan jieshou zhidong taban de zhidong xinhao, kongzhi dianji dongzuo. ben faming caiyong zhixian dianji qudong zhidong zhugang huosai yundong yi dadao zhidong de mude, bao liu le chuantong zhenkong zhuli yeya zhidong xitong, quxiao zhenkong zhuliqi, jiang shuanggang zhidong zhugang gaiwei dangang,* ***shi*** *(cause) qi bu shou shifu zhenkong de xianzhi, ke* ***yongyu*** *(use) xiandai chundiandong qiche shang, suoxu renli hen xiao,* ***jianqing*** *(reduce) le renli pilao,* ***tigao*** *(raise) le jiashiyuan anquan.*

The words in bold are extracted and displayed, as shown in Figure 2. This could help users get to the point of the documents.

4. Conclusion.

In this article, we propose to study the verbs in patent documents in line with qualia structure and classify them into 4 types. Combining with the semantic roles involved, the verbs and the relevant argument structure have been constructed accordingly. The approach adopted here could help computer to realize automatic reasoning and extend to the patent documents in other areas.

Acknowledgment. This work is partially supported by National Key Project of Scientific and Technical Supporting Programs No. 2013BAH21B02, and China Postdoctoral Science Foundation No. 2014M550792. We gratefully acknowledge the helpful comments and suggestions of the reviewers, which have further improved this paper.

REFERENCES

[1] Y. Zhao, J. Gui, Y. Zhang, L. Zhu and C. Jiang, The study on text mining framework for patent document based on deep content indexing, *Digital Library Forum*, vol.54, no.11, pp.1-5, 2008.

[2] C. Jiang, X. Qiao, L. Zhu, J. Gui and Y. Zhang, GATE-based Chinese patent abstracts' extraction, *Digital Library Forum*, vol.54, no.11, pp.27-32, 2008.

[3] Z. Wang, Q. Qiu, P. Feng and S. Xie, Information extraction method of technical solution from mechanical product patent, *Journal of Mechanical Engineering*, vol.45, no.10, pp.198-206, 2009.

[4] J. Pustejovsky, *The Generative Lexicon*, MIT Press, Cambridge, MA, 1995.

[5] Z. Song, Logical metonymy, event coercion and noun-to-verb transformation, *Linguistic Sciences*, vol.12, no.2, pp.117-129, 2013.

[6] Z. Dong and Q. Dong, *HowNet and the Computation of Meaning*, World Scientic Publishing Company, 2006.

[7] Y. Yuan, A study of Chinese semantic knowledge system based on the theory of generative lexicon and argument structure, *Journal of Chinese Information Processing*, vol.27, no.6, pp.23-31, 2013.