Research of Semantic Role Labeling and Application in Patent knowledge Extraction

Ling'en Meng
Institute of Scientific and Technical Information of China, Beijing
mengle2013@istic.ac.cn

Yanqing He *
Institute of Scientific and Technical Information of China, Beijing
heyq@istic.ac.cn

Ying Li *
Institute of Scientific and Technical Information of China, Beijing
liying@istic.ac.cn

ABSTRACT

Semantic Role Labeling (SRL) is a leading task of identifying arguments for a predicate and assigning semantically meaningful labels to them. SRL is crucial to information extraction, question answering, and machine translation. When applied to patent text, existing tools for SRL have unsatisfying performance because of long sentences. To improve performance in patent SRL systems, this study separates each sentence in patent abstracts into a simpler structure, and then labels semantic roles for the simplified sentence. At last, semantic information and semantic framework for frequently used words are used to extract patent knowledge. Our work demonstrates that the method used in this article can improve the performance in SRL system and obtain beneficial knowledge from patents.

Categories and Subject Descriptors

I.2.7 [**Computing Methodologies**]: Language Constructs and Features –*Language parsing and understanding*, *Text analysis.*

General Terms

Algorithms, Experimentation, Languages

Keywords

Semantic role labeling, Patent text, Patent knowledge extraction

1. INTRODUCTION

Semantic Role Labeling is the process of annotating the predicate-argument structure in text with semantic labels. SRL includes two sub-tasks: the identification of syntactic constituents that are semantic roles probably, and the labeling of those constituents with the correct semantic role [1]. Most of current researches on SRL focus on using supervised learning method including generative model and discriminate model. The generative model is firstly used in the SRL classification model. This model has fast training rate and the dependence on the training corpus is not strong. But the poor description ability and strong assumption of features independence lead to unsatisfactory performance. Discriminate models directly estimate the final goal of optimization-- conditional probability. The process is usually performed by iterative methods to find some optimized coefficients. Discriminant models generally include linear interpolation, SVM[2], Perceptron[3], SNoW(Sparse Network of Winnows)[4], Boosting[5], Maximum Entropy, Decision tree, Random forest[6], etc. Combining the results produced by multiple classifiers is a development direction and can obtain better results than any one classifier. These supervised learning methods above are often dependent on the effect of syntactic parsing and accurate annotation of SRL. It is widely used in information extraction, question answering, and machine translation.

SRL has the vital significance in shallow semantic parsing for text information, especially patent texts. Patent texts contain useful information about technologies. Analyzing patent texts can master the present situation of patent texts, predict the hotspot timely and grasp the trend of the technology. The existing patent platforms Patsnap (http://cn.patsnap.com/), TechGlory (Patent risk controls and competitive intelligence analysis system. http://www.tek-glory.cn/), and Wang Xuefeng[7] use a manually annotated corpus, they have high cost and low speed. Researchers also adopt automatic extraction method to obtain key information from patent texts. Jiang Caihong[8] constructs an ontology and writes rules for patent knowledge extraction. Zhai Dongsheng[9] uses ontology knowledge and semantic inference measure to construct a reference network of patent.

This article introduces SRL information combined with a semantic framework rules to extract patent technical topic from patent abstract. As we all know, patent text usually has the characteristic of long sentences with complex structures. As SRL systems are ported into patent texts, they get poor results and affect the effectiveness of the semantic analysis and knowledge extraction. Compare the following examples:

Long sentence:

A plurality of resonance units are arranged [ARGM-TMP in the shell], wherein one end of each resonance unit is fixed on the inner wall at one side of the shell.

Simplified sentence:

A plurality of resonance units are arranged [ARGM-LOC in the shell]

one end of each resonance unit is fixed on the inner wall at o ne side of the shell.

It's obviously that the sematic tag ARGM-TMP (ARGM-TMP represents time, more details in 2.2) in long sentence is wrong. The correct tag is ARGM-LOC (ARGM-LOC represents location) in the simplified sentence. To resolve the above problem, our approach separates each long complicated sentence in patent abstracts into a simpler structure, then labels semantic roles for the simplified sentences, finally, synthesizes all the semantic labels and semantic framework to extract patent topic. Finally,

Published at Ceur-ws.org

Proceedings of the First International Workshop on Patent Mining and Its Applications (IPAMIN) 2014. Hildesheim. Oct. 7th. 2014.

At KONVENS ´14, October 8-10, 2014, Hildesheim, Germany.

SRL information is used to extract patent knowledge from patent abstract and obtains beneficial topic knowledge from patents.

2. SYSTEM ARCHITECTURE AND TECHNICAL DETAILS

In a patent text, an abstract contains its topic, effect, components and features; all of them are important information for the patent. The purpose of this article is to automatically extract the patent topic from the patent abstract. Patent topic mainly involves patent type and patent filed. An example is given in Figure 1. For the patent abstract, the phrase "*An electrically tunable filter*" indicates the patent type and the phrase "*technical field of electronic communication*" shows patent field. The two phrases: patent type phrase and patent field phrase, need to be extracted to form the patent technical topic.

Abstract: The embodiment of the invention provides **an electrically tunable filter**, relating to the **technical field of electronic communication**. The electrically tunable filter comprises a shell, a signal input end and a signal output end. A plurality of resonance units are arranged in the shell, wherein one end of each resonance unit is fixed on the inner wall at one side of the shell. Gaps are kept among adjacent resonance units. Two resonance units with the farthest distance from each other are respectively connected with the signal input end and the signal output end. Medium sheets which are used for adjusting the resonance frequencies of the corresponding resonance units by means of ascending and descending are arranged below the resonance units. The electrically tunable filter not only has a small number of tuning parameters, but also has a simple structure and can better realize the free movement of a center frequency point and the bandwidth of a passband.

Technical Topic: An electrically tunable filter, technical field of electronic communication

Figure 1. An example of patent abstract and its technical topic

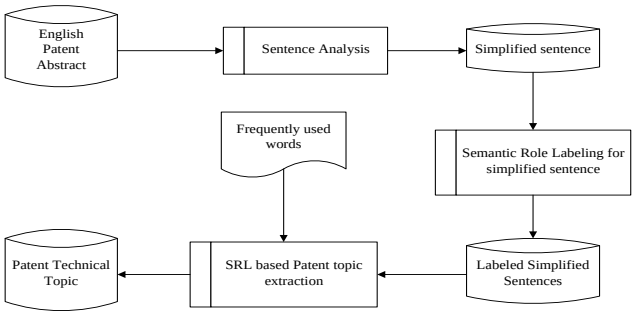

Figure2. The system processing pipeline

As shown in Figure2, our processing is divided into three steps. First, English patent abstracts are separated into simplified sentences by sentence analysis module. Next the simplified sentences are labeled with semantic roles. Finally, the frequently used words with semantic framework and labeled simplified sentences are input into a patent topic extraction module to obtain the patent technical topics.

2.1 Sentence Analysis

A patent abstract often contains long sentences, some of which may involve clauses, such as adverbial clause, object clause, attributive clause, etc. Clauses can generate inaccuracy in syntactic parsing. These errors even can transmit to SRL. For these reasons, we take out clauses in the long sentence, then, turn the long sentence into simplified sentences. Here we mainly separated attributive clause containing 'which' and 'wherein'.

Stanford Parser (http://cemantix.org/software.html) is introduced in order to support us to find clause boundaries. On account of the length of sentence over 70 words can't be parsed, the sentence over 70 words is divided at ';', 'wherein' before parsed. This practice can maintain the integrity of sentence structure. But there are still less than 7% sentences over 70 words, they are divided at the middle ',' by a simple iterative method. After parsing, If the long sentence contains 'wherein' clauses, we separate the long sentence at 'wherein' into two parts; if the long sentence contains 'which' clauses, we deal with them using a program, the pseudo-code is given in Figure 3.

```
Begin
Input :long sentence
  parsing the long sentence,we can get the syntactic tree —— parseLongSentence.
    if parseLongSentence contains guide word —— '(which)'
        find the guide word —— (which) in the syntactic tree, record the position as whichPosition.
        /*search from whichPosition，judge 'NP(…)' or 'VP(…)' which one come first,if NP(…)，record
        TRUE*/
            if search from whichPosition, 'NP' come first
                search from whichPosition，then take out the first S(…) close to
                whichPostition —— sentence1;
            else search from whichPosition, 'VP' come first
                search from whichPostition in the opposite direction,then take out the first
                NP(…) close to whichPosition;
                search from whichPosition，then take out the first S(…) close to
                whichPostition;
                combine NP(…) with S(…) as a new simplified sentence —— sentence2;
Output: sentence1,sentence2;
// print the sentence which removed the clause sentences.
Output:long sentence – sentence1 – sentence2;
End
```

Figure 3.The pseudo-code for sentence contains 'which' clauses

After sentence analysis, the patent abstract shown in Figure 1 is turned into some simplified sentences (bold fonts) shown in Figure 4.

Sub-sentences: The embodiment of the invention provides an electrically tunable filter, relating to the technical field of electronic communication.
The electrically tunable filter comprises a shell, a signal input end and a signal output end.
A plurality of resonance units are arranged in the shell.
One end of each resonance unit is fixed on the inner wall at one side of the shell.
Gaps are kept among adjacent resonance units.
Two resonance units with the farthest distance from each other are respectively connected with the signal input end and the signal output end.
Medium sheets are used for adjusting the resonance frequencies of the corresponding resonance units by means of ascending and descending
Medium sheets are arranged below the resonance units.
The electrically tunable filter not only has a small number of tuning parameters, but also has a simple structure and can better realize the free movement of a center frequency point and the bandwidth of a passband.

Figure 4. Simplified sentences of patent abstract shown in Figure 1

2.2 SRL System for Simplified Sentences

After obtaining the simplified sentences, we use the tool -- Automatic Statistical SEmantic Role Tagger (ASSERT) (about this tool, you can find more information by visiting http://cemantix.org/publications.html) to label them. A sentence is annotated with tags such as TARGET, ARG 0~5, ARGM. Each predicate verb of the sentence is marked with TARGET. ARG0、ARG1 respectively represents agent, patient. ARG2 - ARG5 have

different meanings in different situations. As to ARGM, it has thirteen subtypes, they are shown in Table 1.

Table 1. Subtypes of the ARGM modifier tag

ARGM-LOC	location	ARGM-CAU	cause
ARGM-EXT	extent	ARGM-TMP	time
ARGM-DIS	discourse connectives	ARGM-PNC	purpose
ARGM-ADV	general purpose	ARGM-MNR	manner
ARGM-NEG	negation marker	ARGM-DIR	direction
ARGM-MOD	modal verb		

More information about semantic roles please refer to Martha Palmer[10]. Table 2 shows the difference of SRL for patent abstract shown in Figure 1 and Figure 4.

Table 2 Difference of SRL between Long Sentence and Simplified Sentence

SRL errors in long sentences	Correct SRL results in simplified sentences
A plurality of resonance units ar e arranged*[ARGM-TMP in the shell]*, wherein one end of each resonance unit is fix ed on the inner wall at one side of the shell.	A plurality of resonance units ar e arranged *[ARGM-LOC in the shell]*
A plurality of resonance units ar e arranged in the shell, *[TARGE T wherein]* one end of each resonance unit is fixed on the in ner wall at one side of the shell.	In the process of separating the long sentence, word—'wherein' is removed. This error can be no more arise in simplified sentence.
[ARG1 Medium sheets which are used for adjusting the resonance frequencies of the corresponding resonance units by means of ascending and descending]	*[ARG1 Medium sheets]* are used for adj usting the resonance frequencie s of the corresponding resonance units by means of asc ending and descending

2.3 Patent Topic Extraction Based on SRL

As stated in the above, since patent topic includes two parts: type-phrase and field-phrase, we extract type phrase and field phrase separately. First, we build a frequently-used-words list for patent's topic. In this step, we manually annotated the patent abstracts in small-scale, and then the predicates appear frequently in the sentence that contains patent topic is collected to form this list. Next, we analyze every frequently-used-word to obtain its linguistic features and assign a framework of SRL information for each of them. The semantic framework can help us to decide which semantic role should be extracted as the patent topic. Two examples for the semantic framework of frequently-used-words is shown in Table 3. If a sentence contains 'provide' as the TARGET(the predicate tag of the sentence), ARG1 is taken out from the sentence as the type-phrase.

Table 3. TARGET semantic framework of frequently-used-words

Frequent Word	Semantic Framework	Example
relate	relate [to ARG2]	[ARG1 The invention] [TARGET relates][ARG2 to a double-shielded mineral-insulated cable]
provide	provide [ARG1]	[ARG0 The embodiment of the invention] [TARGET provides] [ARG1 an electrically tunable filter relating to the technical field of electronic communication]

Next, we match the word from the list with TARGET of each simplified sentence in the abstract. If matched, the phrase for semantic role ARG0~ARG5 of TARGET is extracted from this sentence according to its framework.

For the field-phrase, we firstly choose the labeled sentence that contains phrase with "field" between "[" and "]". If the semantic role for the phrase is ARGM, we extract the corresponding phrase as the field-phrase. Otherwise, we locate TARGET in the sentence containing "field", and then judge TARGET semantic framework to determine which semantic role should be extracted from ARG0 to ARG5.

In fact, in order to promote performance of extraction, post-processing methods are used, such as getting rid of the preposition at the beginning or removing some gerundial phrases.

3. EXPERIMENT

In this section, we perform an experiment to evaluate our patent topic extraction based on SRL. The evaluation standard - 'Precision', 'Recall', 'F1' are used to evaluate the system effect. We choose 50 patent abstracts relating to communication field as our experiment data. Detailed statistics of corpus is shown in Table 4. We take out the clauses from the long sentence by using described method in section 2.1. The experimental results are shown in Table 5. From the table, the precision of "which" clause is 73.61% and "wherein" clause reach a higher precision 96.07%. When putting them together, the precision is 79.61% and error analysis shows that the error mainly due to inaccuracy syntactic analysis even syntactic errors. Of course, the syntactic structure is lost for less than 7% of the sentences. This probably contributes to the small performance loss.

Table 4 Detailed statistics of experimental corpus

Data	Language	Number of sentences	vocabulary	Average sentence length
Long sentences	English	175	8195	47

Table 5 the Performance of sentence analysis

clauses	Precision(%)	Recall(%)	F1(%)
which	73.61	67.08	70.19

wherein	96.07	96.07	96.07
which+wherein	79.61	78.09	78.84

Using the SRL tool — ASSERT, we get the simplified sentences with semantic tags. Then patent topics are extracted from abstracts according the algorithm in section 2.3. In order to evaluate the performance of topic extraction, we let three experts label the topics in the 50 English patent abstracts, and then regard them as the golden standard. Three non-experts are asked to judge whether the extracted topics are correct. When more than two of them give a correct judgment for an extracted topic, we regard it is a right one.

The result shows that there are more than 35 patent abstracts which match the manual annotated results. This means our method has a 70% precision for topic extraction. After careful examination, we think the error results from two main reasons:

(1) The high-frequency words list has a small coverage of vocabulary. Their frameworks are not precise enough to get a correct patent type phrase or patent field phrase.

(2) If one sentence has predicates share same words, it is a challenge to decide which one is the best.

4. CONCLUSION

This research studied SRL and applied it to patent knowledge extraction. The patent abstract is separated into simplified sentences by sentence analysis, then labeled semantic role for them. Patent technical topic is generated by combing the patent type phrase and patent field phrase. The patent topics are automatically extracted from the simplified sentences with SRL. Our work demonstrates the method we used is effective.

Until now, the research only performed a simple preprocessing before SRL and our extraction rules of semantic framework are also far from comprehensive. In order to get more improvement, the following work needed to be done: (1) A high frequency vocabulary can be constructed in larger scale with deeper semantic information of patent context. (2) The pre-processing of SRL need to be further optimized. (3) This research only extracted patent technical topic and more information, such as patent components, patent characteristics and effect can be done. Our system will be modified to realize more patent information mining. We are supposed to further exploring in patent semantic level.

5. ACKNOWLEDGMENTS

This activity has been carried out within the China funded project, Natural Science Funds “context analysis on statistical machine translation for patent texts”(No.61303152).The work described in this paper could have not been possible without the collaboration of a number of people. We wish thank you our colleagues Jin WEI, Zhaofeng ZHANG, and Peng QU.

6. REFERENCES

[1] Sameer Pradhan, Wayne Ward, Daniel Jurafsky, Kadri Hacioglu and James H. Martin.2005. Semantic Role Labeling Using Different Syntactic Views. *ACL-05.Association for Computational Linguistics Annual Meeting*(Ann Arbor, MI(US),June 25-30,2005).2005,581-588.

[2] Sammer Pradhan, Kadrihacioglu, Valerie Krugler, Wayne Ward, Jamesh. Martin, and Daniel Jurafsky.2005. Support Vector Learning for Semantic Argument Classification. *Machine Learning Journal*. 60, 1/3(2005), 11-39.

[3] Hierarchical Recognition of Propositional Arguments with Perceptrons(2004). *In Proceedings of CoNLL 2004 Shared Task.*2004.

[4] P. Koomen, V. Punyakanok, D. Roth, and Wen-tau Yih. 2005.Generalized Inference with Multiple Semantic Role Labeling Systems. *Proceedings of CoNLL-2005*. (Ann Arbor, Michigan).2005,181-184.

[5] R. E. Schapire, and Y. Singer. Improved Boosting Algorithms Using Confidence-rated Predictions. 1998. *Proceedings of the Eleventh annual conference on Computational learning theory* .Madison,(WI(US);Madison, WI(US)). 1998,80-91.

[6] R. D. Nielsen, and S. Pradhan. 2004. Mixing Weak Learners in Semantic Parsing. *42nd Annual Meeting of the Association for Computational Linguistics* (Barcelona(ES)).2004,1-8.

[7] Wang Xuefeng, Wang Youguo, and Liu Yuqin. Construction of Patent Analysis System Based on Data Collaboration. *Library and Information Service*.57,14(2013),92-96.DOI=http://dx.doi.org/10.7536/j.issn.0252-3116.2013.14.01.

[8] Jiang Caihong, Qiao Xiaodong, and Zhu Lijun. 2009.Ontology-based Patent Abstracts’ Knowledge Extraction. *New Technology of Library and Information Service*. 2,(July.2009):23-28.DOI=http://dx.doi.org/10.3969/j.issn.1003-3513.2009.02.004

[9] Zhai Dongsheng, Zhang Xinqi, and Zhang Jie. 2013.Design and Implementation of Derwent Patent Ontology.*Information Science*. 31.12(2013):95-100.

[10] Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2004.The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*. 31,1(July,2004),71-105.DOI=http://doi.acm.org/10.1162/0891201053630264