Research on Chinese Patent Candidate Term Selection Based on Dependency Parsing (Postprint)

Authors: Yu Yan, Chen Lei, Jiang Jinde, Zhao Naixuan

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] To address the issues in Chinese patent candidate term selection methods, such as the need to develop different pattern matching rules for different datasets and low accuracy in patent term extraction, this paper proposes a Chinese patent term selection method based on dependency parsing to improve the accuracy of Chinese patent term extraction. [Method/Process] The method mainly consists of three steps: dependency parsing, pruning, and generating dependency subtrees. First, dependency parsing is performed on Chinese patents to obtain dependency trees. The dependency trees are then pruned to remove non-compliant dependency relations, generating dependency subtrees. Consecutive word strings are selected from these subtrees as candidate terms to extract Chinese patent terms. [Results/Conclusion] Experimental results demonstrate that compared with existing Chinese patent candidate term selection methods, the proposed Chinese candidate term selection method based on dependency parsing can effectively improve the accuracy of Chinese patent term extraction.

Full Text

Preamble

Research on the Selection of Chinese Patent Candidate Terms Based on Dependency Syntax Parsing

Yu Yan[1,2], *Chen Lei*[1], *Jiang Jinde*[3], *Zhao Naixuan*[1]
[1]Information Service Department, Nanjing Tech University, Nanjing 210009
[2]Department of Computer Engineering, Chengxian College, Southeast University, Nanjing 211816
[3]School of Business, Nanjing Xiaozhuang University, Nanjing 211171

Abstract

[**Purpose/Significance**] Existing methods for selecting Chinese patent candidate terms suffer from two major limitations: the need to manually define different pattern-matching rules for different datasets, and low accuracy in patent term extraction. To address these issues, this paper proposes a Chinese patent candidate term selection method based on dependency syntax parsing to improve the accuracy of Chinese patent term extraction. [**Method/Process**] The method consists of three main steps: dependency syntax parsing, pruning, and dependency subtree generation. First, dependency syntax analysis is performed on Chinese patents to obtain dependency trees. These trees are then pruned to remove non-compliant dependency relations, generating dependency subtrees from which continuous word strings are selected as candidate terms for Chinese patent term extraction. [**Result/Conclusion**] Experimental results demonstrate that compared with existing Chinese patent candidate term selection methods, the proposed dependency syntax parsing-based approach effectively improves the accuracy of Chinese patent term extraction.

Keywords: term extraction; dependency syntax parsing; Chinese candidate term selection
Classification Number: G202
DOI: 10.13266/j.issn.0252-3116.2019.18.013

1. Introduction

Patent documents contain rich solutions to problems across various domains. Effective patent document analysis can identify technological hotspots, recognize core technologies, and predict technological development trends, helping researchers gain inspiration and reducing innovation time and costs. Terms in patent documents provide structured knowledge units for analysis, embodying and carrying technical information, making them critical components of patent document analysis. Therefore, automatic term extraction from patent documents is a key challenge.

Current patent term extraction methods typically follow a two-step process: candidate term selection and candidate term ranking. Candidate terms are first extracted from corpora, then statistical information is used to calculate the likelihood of candidates being genuine terms, which are sorted accordingly. For Chinese patent term extraction, part-of-speech (POS) pattern matching is commonly used for candidate selection (e.g., "adjective + noun," "verb + noun" patterns). However, this approach has two main problems: (1) different Chinese patent text collections require manually defined matching rules, which is difficult to implement; and (2) while correctly selecting candidate terms, it may also introduce many non-term strings, such as "添加粉末" (add powder), even though it can correctly identify candidates like "氧化石墨烯" (graphene oxide).

Dependency syntax parsing reveals semantic modification relationships between words through dependency relations within sentences, enabling semantic under-

standing. It effectively compensates for the limitation of POS-based methods that struggle to capture deep semantic relationships. Therefore, this paper introduces dependency syntax parsing to Chinese candidate term selection for the first time, proposing a dependency syntax analysis-based method to improve Chinese patent term extraction accuracy.

2. Related Research

2.1 Term Extraction

Terms refer to general (concrete or abstract) theoretical concepts within specialized knowledge domains and constitute important components of domain knowledge systems, conveying substantial domain knowledge. Term extraction is the process of automatically identifying terms from text.

Current term extraction methods fall into two categories: unsupervised and supervised. Unsupervised methods typically combine linguistic and statistical approaches to extract terms from text collections with minimal manual intervention, strong applicability, and consistency. Supervised methods employ machine learning techniques such as maximum entropy models and conditional random fields to learn features from training data for term extraction. While supervised methods can overcome the limitation of unsupervised methods in recognizing low-frequency terms and achieve higher precision and recall, they require large-scale manually annotated corpora for training and remain immature, necessitating further experimentation and validation. Currently, no comprehensive, large-scale annotated corpus exists specifically for patent documents. Moreover, with rapid technological development, numerous new terms continuously emerge. Unsupervised methods can extract terms with minimal manual intervention, offering an effective solution to the difficulty of obtaining annotated corpora. Therefore, this study focuses on unsupervised patent term extraction.

Although many unsupervised methods exist, they typically follow a "candidate first, rank later" workflow: candidate term selection and candidate term ranking.

2.1.1 Candidate Term Selection Candidate term selection methods generally fall into three categories: n-gram filtering, noun phrase chunking, and POS pattern matching. N-gram filtering typically removes stop words and semantically light words (e.g., particles, modal particles) or manually selects words with poor word-formation ability, then traverses the text to obtain all n-gram sequences, selecting qualified multi-word units according to certain rules. This approach is simple and flexible in setting n-gram length but introduces too many non-term strings, affecting extraction accuracy.

Since terms are typically noun phrases, noun phrase chunking identifies noun phrases from POS-tagged text sequences. Noun phrases usually follow specific POS patterns (e.g., "adjective + noun"). While simple and fast, this method is primarily applied to English term extraction because Chinese noun phrase modification rules are complex and not limited to adjectives and nouns.

POS pattern matching shares the same basic idea as noun phrase chunking, assuming term POS sequences follow specific patterns. The difference lies in defining more complex matching patterns. Its advantage is the ability to specify targeted rules for Chinese text characteristics, making it a mainstream method for Chinese candidate term selection. However, it requires manually defining different rules for different datasets and may introduce excessive non-term strings.

2.1.2 Candidate Term Ranking Candidate term ranking primarily uses termhood and unithood metrics to measure the likelihood of candidates being genuine terms.

Termhood measures a candidate's domain relevance from the perspective of term membership. Common statistics include word frequency and C-value (and its variants). Word frequency measures domain relevance based on candidate occurrence frequency but neglects low-frequency terms. C-value improves upon word frequency by considering phrase nestedness, incorporating candidate frequency, length, and the frequency and count of longer candidates containing the current candidate. C-value is simple, adaptable, and language/domain-independent but still relies mainly on frequency, failing to effectively filter high-frequency non-term strings or correctly extract low-frequency terms. Some studies have attempted C-value improvements, such as PCC-value (incorporating document frequency) and STC-value (utilizing similar candidate information sharing term components).

Unithood measures the structural stability of candidate terms—the binding strength between internal components. Mutual information (MI) is a common unithood metric that calculates co-occurrence frequency of term components to measure dependency strength. While MI effectively reflects binding strength between character strings, it overestimates the strength of low-frequency but consistently adjacent strings. Variants like PMIk and EMI have been proposed to address this issue.

2.2 Dependency Syntax Parsing

Dependency syntax parsing is based on the assumption that syntactic structure essentially contains associations between word pairs, where one word governs another—this governor-dependent relationship is called a dependency relation. Dependency syntax analysis considers the core verb as the central component governing other words while itself being governed by none. All governed words subordinate to the governor through dependency relations. Dependency parsing reveals syntactic structure and discovers grammatical features and semantic connections by analyzing dependency relations between words. According to dependency grammar axioms, in a complete sentence, no word can depend on two or more other words; all semantic connections interweave to hierarchically structure the linear sentence into a dependency tree, reflecting semantic modification relationships independent of physical position.

Dependency syntax parsing offers concise representation, small storage space, good computability, head-driven structure, minimal POS dependency, and universal dependency relations, making it suitable for flexible Chinese word order. While dependency parsing itself doesn't require relation classification, practical applications typically add labels to dependency tree edges to enrich syntactic information. The Harbin Institute of Technology's Language Technology Platform (LTP) dependency relation annotation system offers advantages including fewer relation types and easy comprehension.

Because dependency syntax parsing reveals semantic modification relationships through dependency relations, reflecting long-distance collocations independent of physical position, it has been widely applied in sentiment analysis, entity relation extraction, question answering, and trigger word recognition. For example, in sentiment analysis, researchers have proposed dependency rule and POS feature-based models for sentiment word identification. In entity relation extraction, methods use dependency relations to identify verb-predicate sentences. In question answering, dependency tree edge labels are modified to represent question decomposition information. In trigger word recognition, dependency syntax analysis improves trigger word extraction recall.

3. Chinese Patent Candidate Term Selection Based on Dependency Syntax Parsing

This paper introduces dependency syntax parsing to develop a Chinese candidate term selection method. The approach includes three main steps: dependency syntax parsing (Section 3.1), pruning (Section 3.2), and dependency subtree generation (Section 3.3). First, Chinese patent texts undergo dependency syntax analysis to obtain dependency trees. These trees are pruned to remove non-compliant relations, generating dependency subtrees from which continuous word strings are selected as candidate terms for Chinese patent term extraction.

3.1 Dependency Syntax Parsing

Dependency syntax parsing reveals syntactic structure by analyzing dependency relations between words, represented by directed arcs from governors to dependents, with core verbs as sentence governors. According to dependency grammar axioms, parsing hierarchically structures linear sentences into dependency trees.

Definition 1 (Dependency Tree): A dependency tree is denoted as T = (V, A, R), where V is the node set representing words, A is the set of directed arcs representing dependency relations (arcs originate from governors and point to dependents), and R is the root node representing the core verb. T satisfies: (1) R has indegree 0; (2) All nodes except R have indegree 1; (3) There is a directed path from R to any node.

[Figure 1: see original paper] shows dependency trees T1 and T2 obtained by parsing the sentences "本发明主要用于制备四氧化铁负载氮掺杂石墨烯复合材料" and "本发

明涉及一种共掺杂聚吡咯材料及其制备方法和应用" using HIT's LTP parser. Root points to core verbs "用于" and "涉及", with arc labels indicating dependency relation types (see). Letters under nodes represent POS tags (see). As shown, dependency parsing provides word relationships and shallow syntactic structures, offering a basis for Chinese patent candidate term selection.

3.2 Pruning

Since Chinese patent terms are generally noun phrases, certain dependency relations in dependency trees rarely appear within noun phrases and introduce noise affecting candidate selection. Therefore, we propose pruning dependency trees to reduce useless information before candidate selection. Analysis of relations in reveals that Chinese patent term internal relations are primarily attributive (ATT), coordinate (COO), left adjunct (LAD), and right adjunct (RAD) relations. Additionally, patent terms consist of domain-rich words and typically exclude stop words. We use HIT's stop word list plus manually selected patent-specific stop words like "发明" (invention) and "方法" (method).

Pruning Rule 1: Remove dependency relations other than ATT, COO, LAD, and RAD.
Pruning Rule 2: Remove dependency relations containing stop words.

[Figure 2: see original paper] shows pruned trees T1 and T2 from [Figure 1: see original paper], with gray arcs indicating removed relations and gray words indicating stop words.

3.3 Generating Dependency Subtrees

Chinese terms are typically phrases with nouns or verbs as cores modified by other components. We propose the dependency subtree concept for Chinese candidate term selection.

Definition 2 (Dependency Subtree): Given a dependency tree T = (V, A, R), a dependency subtree T' = (V', A', R') satisfies: (1) V' V, A' A, R' V'; (2) R' has indegree 0; (3) All nodes except R' have indegree 1; (4) There is a directed path from R' to any node; (5) R' is a content word (noun or verb).

Based on generated dependency subtrees, continuous word strings are selected as Chinese patent candidate terms. For example, [Figure 3: see original paper] shows subtrees T1,1–T1,8 and T2,1–T2,3 generated from pruned trees T1 and T2. Since compared methods select phrases with >1 word, we only select subtrees with $|V| > 1$ (containing at least 2 words). In Figure 3: see original paper, subtree T1,1 is discontinuous and discarded, while T1,2–T1,8 are continuous, generating 7 candidate terms. In Figure 3: see original paper, T2,1 and T2,2 are continuous, while T2,3 is discontinuous and discarded, generating 2 candidate terms.

and compare candidate terms extracted by n-gram, noun phrase chunking, POS pattern matching, and our method for the two example sentences. N-

gram includes all correct terms but introduces excessive incorrect candidates. Noun phrase chunking causes missed terms. POS pattern matching both misses terms and introduces many incorrect candidates. Our dependency syntax-based method identifies more correct terms with relatively fewer errors, providing a solid foundation for subsequent ranking.

4. Experiments

4.1 Dataset

To validate feasibility and effectiveness, we selected graphene patent literature for experiments. Graphene is one of the thinnest known materials with unique structure and excellent optical, chemical, electrical, and mechanical properties, becoming a hot research topic in physics, chemistry, and materials science with promising industrial applications. We retrieved 6,445 valid Chinese invention patents from the China National Intellectual Property Administration database (2014–2018) using "graphene" as keyword (search date: November 15, 2018), using titles and abstracts as the patent text dataset.

4.2 Evaluation Metrics

Given the large patent text volume, we use precision as the evaluation metric, assessing correctness of top-N extracted terms:

Precision = (Number of correctly extracted terms / Number of extracted terms) $\times$ 100% (Equation 1)

We evaluate N = 200–2000, manually judging results. To avoid subjectivity and domain knowledge limitations, we use Baidu Baike, Wikipedia, and other knowledge websites combined with expert evaluation to determine term correctness.

4.3 Experimental Results

4.3.1 Comparison of Candidate Term Selection Methods' Impact on Extraction Effectiveness We first compare typical candidate term selection methods with our proposed method:

(1) **n-gram**: Remove stop words and semantically light words, then traverse to obtain all n-gram sequences (n = 2–6).

(2) **NP**: Noun phrase chunking using regex (a|n)+ (n|vn).

(3) **pos1**: POS pattern matching rules in .

(4) **pos2**: POS pattern matching rules in .

(5) **dep**: Our dependency syntax-based method from Section 3.

For evaluation, we use two typical ranking algorithms: C-value and PMI. Table 7 shows the compared methods.

[Figure 4: see original paper] and [Figure 5: see original paper] show results. In C-value-based comparison ([Figure 4: see original paper]), n-gram+C-value achieves lowest precision (35.75% at N=1000). NP+C-value improves but remains lower than pos1 and pos2. pos2+C-value outperforms pos1+C-value, achieving 13.60% and 17.58% improvement over n-gram+C-value at N=1000. Our dependency-based method achieves highest precision, improving 24.37% over n-gram+C-value at N=1000. PMI-based comparison ([Figure 5: see original paper]) shows similar results, with dep+PMI achieving best performance.

4.3.2 Comparison of Candidate Selection vs. Ranking Impact on Extraction Effectiveness Unsupervised term extraction includes both steps, but research has focused mainly on ranking algorithm improvements. This section explores the relative impact of each stage. We use pos2 (best traditional selection method) with C-value and PMI as baselines, comparing against improved selection (dep) and improved ranking methods (PCC-value, STC-value, PMIk, EMI) as shown in .

Results ([Figure 6: see original paper], [Figure 7: see original paper]) show that improved ranking methods (pos2+PCC-value, pos2+STC-value) slightly outperform pos2+C-value (3.73% and 4.83% improvement at N=1000), while dep+C-value shows highest precision (6.77% improvement). Similarly, PMI variants show modest gains (1.35% and 1.93%), while dep+PMI achieves 3.76% improvement. These results indicate that improvements in the first stage (candidate selection) have greater impact on final extraction accuracy than ranking method refinements, suggesting candidate selection deserves more research attention.

4.3.3 Detailed Analysis of Candidate Selection Methods lists top-10 highest-frequency candidate terms selected by n-gram, NP, pos1, pos2, and dep, with correct terms in bold. N-gram generates excessive noise. NP filters some noise but may miss correct terms (e.g., "氧化石墨烯" missed because "氧化" is a verb). POS methods (pos1, pos2) include richer POS patterns but introduce additional noise (e.g., "制备石墨烯" selected as "verb + noun"). Our dep method overcomes these issues without manual intervention, selecting more accurate candidate terms and reducing noise from verbs and other words, providing a solid foundation for ranking and achieving better extraction performance.

References

[1] FRANTZI K, ANANIADOU S, MIMA H. Automatic recognition of multi-word terms: the C-value/NC-value method[J]. International journal on digital libraries, 2000, 3(2): 115-130.
[2] ZHOU Lang, SHI Shumin, FENG Chong, et al. Chinese term extraction

method based on multi-strategy fusion[J]. Journal of the China Society for Scientific and Technical Information, 2010, 29(3): 460-467.
[3] WEI Xiaoli, SUN Yong, ZHANG Shukui, et al. Ontology concept acquisition based on maximum entropy model[J]. Computer Engineering, 2009, 35(24): 114-116.
[4] WANG Hao, WANG Miping, SU Xinning. Research on Chinese patent term extraction for ontology learning[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(6): 573-585.
[5] LI L, DANG Y, ZHANG J, et al. Domain term extraction based on conditional random fields combined with active learning strategy[C]//Proceedings of the North American chapter of the Association for Computational Linguistics. Stroudsburg PA: Association for Computational Linguistics, 2013: 16-23.
[6] CONRADO M, PARDO T, REZENDE S. A machine learning approach to automatic term extraction using a rich feature set[C]//The North American chapter of the Association for Computational Linguistics. Stroudsburg PA: Association for Computational Linguistics, 2013: 16-23.
[7] HU Apei, ZHANG Jing, LIU Junli. Chinese term extraction based on improved C-value method[J]. New Technology of Library and Information Service, 2013, 29(2): 24-29.
[8] DING Jie, LYU Xueqiang, LIU Kehui. Patent literature term extraction method based on boundary tag set[J]. Computer Engineering & Science, 2015, 37(8): 1591-1598.
[9] LIU Jian, TANG Huifeng, LIU Wuying. A Chinese term extraction method based on statistical techniques[J]. China Terminology, 2014, 16(5): 10-14.
[10] ZENG Zhen, LYU Xueqiang, LI Zhuo. A domain term extraction method for patent abstracts[J]. Computer Applications and Software, 2016, 33(3): 48-51.
[11] YANG Shuanglong, LYU Xueqiang, LI Zhuo, et al. Automatic recognition of terms in Chinese patent literature[J]. Journal of Chinese Information Processing, 2016, 30(3): 111-117.
[12] XU Chuan, SHI Shuicai, FANG Xiang, et al. Chinese patent literature term extraction[J]. Computer Engineering & Design, 2013, 34(6): 2175-2179.
[13] ZHANG Jie, ZHANG Haichao, ZHAI Dongsheng. Research on word segmentation method for Chinese patent claims[J]. New Technology of Library and Information Service, 2014, 30(9): 91-98.
[14] HU Wenmin, HE Tingting, ZHANG Yong. Chinese term extraction based on chi-square test[J]. Computer Applications, 2007, 27(12): 3019-3020.
[15] HAN Hongqi, ZHU Donghua, WANG Xuefeng. Method for extracting patent technology terms[J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(12): 1280-1285.
[16] YU Yan, ZHAO Naixuan. Patent term extraction based on general words and term components[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(7): 742-752.
[17] LIN Zifang, JIANG Xiufeng. New word recognition based on word internal patterns[J]. Computer and Modernization, 2010, 11(1): 162-164.
[18] PECINA P, SCHLESINGER P. Combining association measures for col-

location extraction[C]//Proceedings of the COLING/ACL on main conference poster sessions. New York: ACM, 2016: 651-656.
[19] DU Liping, LI Xiaoge, YU Gen, et al. New word discovery based on improved mutual information algorithm for Chinese word segmentation system improvement[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(1): 35-40.
[20] ZHANG W, YOSHIDA T, TANG X, et al. Improving effectiveness of mutual information for substantival multiword expression extraction[J]. Expert systems with applications an international journal, 2009, 36(8): 10919-10930.
[21] ROBINSON J. Dependency structures and transformational rules[J]. Language, 1970, 46(2): 259-285.
[22] BAI Miaoqing, ZHENG Jiaheng. Research on verb-verb collocation methods[J]. Computer Engineering and Applications, 2004, 40(27): 70-72.
[23] LIU Huaijun, CHE Wanxiang, LIU Ting. Feature engineering for Chinese semantic role labeling[J]. Journal of Chinese Information Processing, 2007, 21(1): 79-84.
[24] WANG Huize, GONG Shengrong, LIU Chunping. Fisherfaces method combining global and local features[J]. Computer Engineering and Applications, 2008, 44(24): 194-196.
[25] CHE W, LI Z, LIU T. A Chinese language technology platform[C]//The 23rd international conference on computational linguistics. New York: ACM, 2010: 3-16.
[26] AGARWAL B, PORIA S, MITTAL N, et al. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach[J]. Cognitive computation, 2015, 7(4): 487-499.
[27] FENG Chong, LIAO Chun, LIU Zhirun, et al. Sentiment key sentence identification based on lexical semantics and syntactic dependency[J]. Acta Electronica Sinica, 2016, 44(10): 2472-2476.
[28] DENG Shuqing, LI Wanwei, XU Jian. Research on sentiment word identification based on syntactic dependency rules and POS features[J]. Information Studies: Theory & Application, 2018, 41(5): 137-142.
[29] QUAN C, WANG M, REN F. An unsupervised text mining method for relation extraction from biomedical literature[J]. Plos one, 2014, 9(7): 1-8.
[30] LI Mingyao, YANG Jing. Open Chinese entity relation extraction method based on dependency analysis[J]. Computer Engineering, 2016, 42(6): 201-207.
[31] GAN Lixin, WAN Changxuan, LIU Dexi, et al. Chinese entity relation extraction based on syntactic-semantic features[J]. Journal of Computer Research and Development, 2016, 53(2): 284-302.
[32] LI Chao, CHAI Yumei, GAO Minglei, et al. Research on answer extraction in Chinese question answering system using syntactic parsing and deep neural networks[J]. Small Microcomputer Systems, 2017(6): 1341-1346.
[33] LIU Xiong, ZHANG Yu, ZHANG Weinan, et al. Dependency parsing-based decomposition method for complex factual questions[J]. Journal of Chinese Information Processing, 2017, 31(3): 140-146.
[34] KLEIN S, MCCONLOGUE K, SIMMONS RF. Co-occurrence and dependency logic for answering English questions[J]. Journal of the American Society

for Information Science & Technology, 2014, 15(3): 196-204.
[35] WANG J, ZHANG J, AN Y, et al. Biomedical event trigger detection by dependency-based word embedding[J]. Bmc medical genomics, 2016, 9(2): 45-54.
[36] GAO Yuan, XI Yaoyi, LI Bicheng. Trigger word extraction method based on dependency parsing and classifier fusion[J]. Application Research of Computers, 2016, 33(5): 1407-1410.
[37] ZHANG Zhonghua, SU Fangfang, JI Donghong. Research on biomedical event trigger word recognition[J]. Application Research of Computers, 2017, 34(3): 661-670.
[38] ZHANG Leihan, LYU Xueqiang, LI Zhuo, et al. Research on domain ontology term extraction method[J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(2): 167-174.

Author Contributions: Yu Yan: proposed research idea, designed methodology, conducted experiments, wrote paper; Chen Lei: collected and cleaned data; Jiang Jinde: analyzed data; Zhao Naixuan: revised paper.

Note: The final promotional paragraph about "释文数字阅读服务平台" (Shiwen Digital Reading Service Platform) has been omitted as it is unrelated to the academic content of the paper.

Note: Figure translations are in progress. See original paper for figures.