Postprint: A Scientific Named Entity Recognition Algorithm Based on Dependency Syntactic Features

Authors: Zhao Huaming, Qian Li, Yu Li

Date: 2023-04-01T16:15:55+00:00

Abstract

[Purpose/Significance] This study explores the recognition and extraction of scientific research named entities and their relationships, aiming to improve identification performance in complex scenarios such as long sentences, thereby providing references and insights for further applications.

[Method/Process] Based on dependency syntax feature analysis, this paper proposes a method for extracting relationships between scientific research named entities. The process includes: using the Standford Tagger tool to perform part-of-speech tagging on the target text; based on the tagging results, segmenting the target text into semantically coherent fragments with standardized structure around the core predicate and SAO structure; through dependency syntax analysis, identifying subjects and objects semantically related to the core predicate to form (entity, relation, entity) triplets.

[Results/Conclusion] Comparative tests with mainstream algorithms such as Ollie and Reverb demonstrate that the proposed method can effectively improve the accuracy of scientific research named entity recognition.

Full Text

A Research Entity Recognition Algorithm Based on Dependency Parsing Features

Zhao Huaming[1], **Qian Li**[1,2], **Yu Li**[1] [1] National Science Library, Chinese Academy of Sciences, Beijing 100190 [2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: **[Purpose/Significance]** This study explores the recognition and extraction of research entities and their relationships, aiming to improve recog-

nition performance in complex cases such as long sentences, thereby providing reference for further applications. [**Method/Process**] Based on dependency syntactic feature analysis, we propose a method for extracting research entity relationships. The process includes: using the Stanford Tagger tool for part-of-speech tagging of target text; based on the tagging results, segmenting the target text into semantically coherent fragments with standardized structure around core predicates and SAO structures; through dependency parsing, identifying subjects and objects semantically related to core predicates to form (entity, relation, entity) triples. [**Result/Conclusion**] Comparative tests with mainstream algorithms such as Ollie and ReVerb demonstrate that this method can effectively improve the accuracy of research entity recognition.

Keywords: dependency parsing; research named entity; entity recognition; relation extraction
Classification: G250
DOI: 10.13266/j.issn.0252-3116.2020.11.012

Introduction

In the era of big data, extracting useful information from massive datasets has become a challenging and hot issue in natural language processing and data mining. As a fundamental component of natural language processing, entity and entity relation extraction provides crucial technical support for lexical, syntactic, and semantic analysis, and is widely used in information extraction, information retrieval, recommendation systems, classification and clustering, automatic summarization, question answering, knowledge discovery, sentiment analysis, knowledge base construction, and many other natural language processing tasks.

For complex entity relations, Xu Fen et al. [1] proposed a feature vector-based method for entity and relation extraction that integrates features such as words, part-of-speech tags, entity attributes, and inter-entity relationships. Their research demonstrates that multi-level linguistic features can effectively improve entity relation extraction performance. N. Kambhatala [2] integrated entity words, entity types, entity reference patterns, overlaps, dependency trees, and parse trees, implementing entity and relation extraction based on maximum entropy models. Guo Xiyue et al. [3] proposed a method combining syntactic and semantic features for Chinese entity relation extraction, which primarily integrates dependency relations, distance between entities and core predicates, and semantic role labeling, effectively identifying multiple types of entity relations. Gan Lixin et al. [4] further incorporated dependency syntactic combination features and verb dependencies, significantly improving the variety of recognized relation types. H. Li et al. [5] proposed a location-semantic feature-based method that leverages the computability of location features and the interpretability of semantic features, integrating information gain from word positions with semantic calculations based on HowNet. Experimental results show that combining location and semantic features outperforms using either alone. Xi Bin et al. [6]

also improved entity relation extraction performance by effectively combining various lexical, syntactic, and semantic features internally and across feature sets.

These studies demonstrate that integrating dependency parsing and part-of-speech information can effectively improve entity relation extraction performance. Among common extraction tools, TextRunner [7], ReVerb [8], and R2A2 [9] utilize syntactic parsing algorithms for information extraction, while WOE [10], KrakeN [11], and Ollie [12] further integrate dependency parsing algorithms, achieving better extraction results. These tools are designed through careful linguistic analysis of relation phrase patterns in text to form pattern sets, which are then combined with regular expressions and pattern matching algorithms to achieve high-precision entity and relation extraction [13]. In recent years, research on deep learning-based entity relation extraction [14] has also yielded significant results. Tang Min et al. [15] enhanced deep learning models with an entity attention mechanism to distinguish semantic relations. Y. Lin et al. [16] proposed a relation extraction method from pure text that incorporates a multi-lingual neural relation extraction framework with attention mechanisms, effectively controlling noisy sentences.

2 Entity Relation Extraction Algorithm

2.1 Algorithm Design Approach

The overall design of the entity relation extraction algorithm based on part-of-speech tagging and dependency syntactic features is shown in Figure 1 [Figure 1: see original paper]. The algorithm principle is as follows:

(1) Using SAO structure as the basic sentence structure for long sentence segmentation and simplification. First, the Stanford NLP tool is used for part-of-speech tagging of input text. Then, based on SAO basic structure units and syntactic analysis, core predicates are identified and long sentences are segmented into finer-grained semantic structure units around these predicates and SAO structures.

(2) Through dependency syntactic feature analysis and modeling, subjects and objects semantically related to core predicates are identified.

(3) Integrating subjects, core predicates, and objects to form [entity, relation, entity] triples.

2.2 Long Sentence Segmentation Based on SAO Structure

This paper proposes a research entity relation extraction method that directly analyzes sentences through the fusion of part-of-speech tagging and dependency parsing to obtain final entity relation triples. Compared with the pattern matching approach mentioned above, this method improves upon two aspects: long sentence processing, where long sentences are segmented into standardized se-

mantic fragments around core predicates and SAO structures to facilitate accurate entity pair extraction and recognition; research entity relation identification, where the model is optimized through dependency analysis of core predicates and their auxiliary words, effectively improving recognition accuracy. For example, in the sentence "To efficiently handle high-dimensional data, we develop two deterministic algorithms that approximate the covariance matrices," tools like Ollie can only identify entity relations around the core verb "develop" (we; develop; two deterministic algorithms). However, from the sentence's meaning, the desired research entity candidate triple would be the relationship between research entities A and B in "developed A to handle B," i.e., (two deterministic algorithms; be developed to; high-dimensional data). This paper achieves extraction of important research entities and their relationships in scientific papers by adding syntactic dependency analysis algorithms for the verb "develop" and auxiliary word "to."

The SAO (Subject-Action-Object) structure theory originates from the Theory of Inventive Problem Solving (TIPS) and represents the basic functional unit for problem-solving methods [17]. From a grammatical perspective, SAO structure corresponds to SVO (Subject-Verb-Object) structure in sentences; from a semantic web RDF data model perspective, it corresponds to SPO (Subject-Predicate-Object) structure in triples. The introduction of SAO structure can effectively reveal component information and semantic relationships between components [18], forming a complete semantic understanding. In recent years, SAO structure has been widely applied in semantic analysis fields such as technology roadmap analysis [19] and technology evolution [20]. Compared with sentence-based analysis, SAO structure provides a more fine-grained semantic structure that enables deeper and more accurate mining and understanding of associations in text.

For entity extraction from long sentences, A. Gabor [21] first decomposes long sentences into a set of short sentences based on canonically structured patterns, then determines candidate triples from short sentences through natural logic inference. L. Corro [22] decomposes long sentences into a set of short sentences around seven basic sentence patterns, then determines candidate triples from short sentences through dependency analysis to improve extraction performance. Research shows that both canonical structures and basic sentence patterns contain core predicate components. Therefore, it is reasonable to segment long sentences into finer-grained semantic structures using core predicates as units and SAO structure as the basic structural unit and verification model, which facilitates accurate entity extraction in standardized semantic structures. This paper uses long sentence segmentation rather than decomposition into short sentences for two main considerations: syntactic dependency analysis in entity recognition can directly utilize the dependency parsing results of the long sentence, reducing intermediate steps and error rates; while satisfying SAO structure, the original sentence information is preserved as much as possible to minimize information loss.

The implementation process of long sentence segmentation based on SAO structure includes:

(1) Using the Stanford Tagger tool for part-of-speech tagging of long sentences. In the tagging results, nouns start with "NP" and verbs start with "VP," showing clear patterns.

(2) Preprocessing the part-of-speech tagging results, mainly marking non-finite verbs to distinguish them from core verbs, as their part-of-speech tagging forms are similar (both starting with "VP"). For example, gerunds like "doing" are tagged as "(VP (VBG doing))" and infinitives as "(VP (TO to))." This preprocessing reduces noise and improves accuracy.

(3) Using symbols such as "SBAR" (clause marker), ",", and "CC" (coordinating conjunction marker) as feature identifiers for preliminary segmentation of long sentences.

(4) Based on SAO structure, verifying and merging preliminary segmentation results to ensure each segment contains only one core predicate, and outputting final segmentation results.

For example, the sentence "Then, two models of damping in a tall building, the artificial neural network (ANN) model and the auto-regressive (AR) model, are established by employing ANN and AR methods, and used to predict the damping values at high amplitude level, which are difficult to obtain from field measurements." (from the abstract of a paper titled "Damping in buildings: its neural network model and AR model") is initially segmented into 8 fragments, which after verification and merging results in 3 fragments, with the middle fragment being "and used to predict the damping values at high amplitude level."

2.3 Entity and Relation Extraction Based on Dependency Syntactic Features

Based on the Stanford Typed Dependency dependency relation functions, we analyze SAO structure-based semantic fragments to identify core predicates and their semantically related subjects and objects. Using the first two fragments of the example sentence above, we illustrate the research entity relation extraction process, which mainly includes entity extraction, entity relation recognition, and dependency syntactic feature analysis. The corresponding dependency analysis, chunk analysis, and part-of-speech tagging examples are shown in Figure 2 [Figure 2: see original paper].

2.3.1 Entity Extraction Based on part-of-speech tagging results and basic sentence patterns, entity extraction rules are organized around predicates for entity extraction. The implementation process includes two main steps: Using pattern matching tools (Tregex) [23] to identify minimal NP chunks by executing the pattern "NP ! « NP." In syntax trees, the minimal NP chunk (noun

phrase) is considered the smallest unit for semantic processing. Processing results: NP chunks such as "two models of damping in a tall building," "ANN and AR methods," and "the damping values" are each treated as independent chunks. Using the rule "A established by B" to extract entity objects A and B. The candidate triple result for the first semantic fragment is: (two models of damping in a tall building; be established by; ANN and AR methods), with entity types "question" and "method" respectively. Using the rule "A used to B," entity objects A and B are extracted. The candidate triple result for the second semantic fragment is: (two models of damping in a tall building; be used to; the damping values), with entity types "method" and "question" respectively.

2.3.2 Entity Relation Recognition Entity relation extraction primarily identifies part-whole, composition, agency, and causal relationships between entity objects. This paper draws on Jiang Ting's [24] classification of common entity relation types in academic literature and uses WordNet [25] to supplement and expand verbs (predicates) in major relation types that have term category dependencies. Considering the characteristics of scientific literature, auxiliary words ("to," "for," "with," "as," "in," etc.) dependency rules and extraction rules are added to the SVOA model [22] to improve research entity recognition capability, such as: useMethod for Question.

2.3.3 Precision Recognition Based on Dependency Parsing Analysis Notably, during entity object extraction of the second semantic fragment, entity object A does not explicitly exist. This paper uses dependency relation analysis and relation chain calculation for associative recognition. As shown in Figure 2, the dependency relation chain is "used->established->models." The dependency relation identification pattern is: "({}=object> conj:and {lemma:used}={})> nsubjpass{}=subject." Through pattern matching, the associated entity object A for "used" is found to be "two models of damping in a tall building." Similarly, based on dependency relation chain analysis, entity coreference can be resolved, semantically related entities can be identified, and entity clustering and merging can be achieved.

Another notable issue is that after minimal NP chunk merging and pattern matching recognition, entity elements in candidate triples may contain multiple entities, requiring identification of their corresponding reasonable entity relations based on proximity principles or dependency parsing results. For example, in the sentence "Feed-Forward Back-Propagation Artificial Neural Network (FFBP-ANN) trained with Levenberg-Marquardt algorithm is used for estimation of different performance parameters of CMPA," the initial method entity extracted for the "estimation of different performance parameters of CMPA" problem is "Feed-Forward Back-Propagation Artificial Neural Network (FFBP-ANN) trained with Levenberg-Marquardt algorithm." However, this contains two entities: "Feed-Forward Back-Propagation Artificial Neural Network" and "Levenberg-Marquardt algorithm." Based on the dependency relation chain, the final method entity should be "Feed-Forward Back-Propagation

Artificial Neural Network" rather than "Levenberg-Marquardt algorithm." Partial dependency calculation results are as follows: [..., nsubjpass(used-14, Network-5), appos(Network-5, FFBP-ANN-7), acl(Network-5, trained-9), ..., nmod:with(trained-9, algorithm-12), auxpass(used-14, is-13), root(ROOT-0, used-14)...]. The analysis shows that "root(ROOT-0, used-14)" indicates the core word of the sentence is "used"; "nsubjpass(used-14, Network-5)" indicates "Network" is the subject of "used" ("nsubjpass" denotes passive nominal subject); "nmod:with(trained-9, algorithm-12)" indicates "algorithm" forms a compound noun with "trained" through "with" ("nmod" denotes compound noun modifier); "acl(Network-5, trained-9)" indicates "trained" modifies "Network." To reuse this analysis result in subsequent entity and relation recognition, this paper defines the dependency relation chain between "Network" and "used" as "ner.dep_{{{nsubjpass}}_{{identifier}}}()." Similarly, dependency relation chains between core predicates and auxiliary words, as well as coreference words, are summarized and modeled for common dependency chains to form a dependency relation chain discrimination model, enabling interface reuse and precise recognition of research entities.

3 Empirical Research

3.1 Experimental Design and Procedure

This paper extracts abstracts from the top 10 most cited papers published in the *Artificial Intelligence* journal in 2016 from the Microsoft Academic Database as experimental data. The main development tools used are Ollie-app-latest.jar, Reverb-latest.jar, Stanford-corenlp-3.9.2.jar, and Stanford-tregex-3.9.2.jar, with JDK 1.8 as the development environment. Using syntactic tagging and dependency relation chain analysis, we construct rule models and dependency models for research entity extraction to identify and reveal important terms and their relations in scientific texts. The effectiveness of our algorithm is then verified by comparing results with manually annotated baseline data and Ollie/ReVerb algorithms.

The experimental procedure mainly includes: (1) designing a research entity recognition algorithm using basic NLP tools, summarizing common syntactic and dependency models, and constructing a prototype system; (2) manually annotating important terms, entities, and their relations in experimental data as baseline data; (3) using Ollie and ReVerb open information extraction (IE) tools as comparison algorithms to obtain recognition results.

3.2 Open Information Extraction Tools

ReVerb [9] and Ollie [12] are open information extraction tools developed by the University of Washington that identify entity relations in arbitrary sentences to complete entity and relation extraction. ReVerb is an early work that primarily extracts verb-based entity relations—i.e., extracting S and O through A in SAO structure. Ollie is an upgraded version of ReVerb with significant improvements

in relation identification patterns and contextual information-assisted discrimination, representing a new generation of information extraction tools.

In terms of relation identification patterns, Ollie added relation discrimination patterns using nouns and adjectives as association media. For example, the extraction result for "Microsoft co-founder Bill Gates spoke at…" is (Bill Gates; be co-founder of; Microsoft), where "co-founder" serves as the relation medium as a noun. For contextual information-assisted discrimination, Ollie uses attributes and clause modifiers to improve extraction quality. For example, the extraction result for "Early astronomers believed that the earth is the center of the universe." is ((the earth; be the center of; the universe), AttributedTo=believe; Early astronomers), where attribute information indicates the conclusion contradicts simple extraction. Another example: "If he wins five key states, Romney will be elected President." yields ((Romney; will be elected; President) ClausalModifier=if; he wins five key states), where clause modifiers provide additional information.

3.3 Experimental Results Analysis

Following the experimental procedure, our algorithm, manual annotation, ReVerb, and Ollie algorithms are applied to the experimental data. Analysis is conducted in two parts: overall experiment analysis and single abstract instance analysis.

3.3.1 Overall Experimental Results Analysis Our algorithm's recognition results are compared with manually annotated baseline data and Ollie/ReVerb results through exact matching and approximate matching. Exact matching refers to one-to-one matching with baseline data. Approximate matching refers to semantic similarity-based matching, where terms with highly similar meanings to entities in baseline data are also considered correct recognition results. For example, in Table 2 , if the manually annotated baseline entity is "learning algorithms" and the algorithm's recognition result is "Conventional online learning algorithms," it is considered correct.

For evaluation metrics, this paper adopts precision and recall as indicators for entity and relation recognition effectiveness:

Precision = Ra Rh / Ra (Formula 1)
Recall = Ra Rh / Rh (Formula 2)

Where Precision denotes the precision metric, Recall denotes the recall metric, Ra is the number of entities in the algorithm-extracted entity set, Rh is the number of entities in the manually verified dataset, and Ra Rh represents the number of entities where extraction results match manual verification results.

Comparison results are shown in Table 1 . The results show that our proposed algorithm outperforms Ollie and ReVerb in entity recognition precision, entity

relation recognition precision, and recall (except for lower recall in entity recognition). Approximate matching achieves 76.6% precision for entity recognition and 78% precision for relation recognition with 75% recall. Dependency relation analysis and modeling of syntactic features play a key role in improving the accuracy of named entity recognition.

3.3.2 Single Abstract Instance Analysis Overall analysis proves the effectiveness of our algorithm to some extent. We further analyze a specific abstract instance from the experimental data and compare it with Ollie and ReVerb to demonstrate effectiveness. The experimental instance is titled "One-pass AUC Optimization." Partial recognition results are shown in Table 2 .

Compared with Ollie and ReVerb, our algorithm's advantages are:

(1) Better relation identification based on verb-auxiliary word combination models. For the original sentence: "To efficiently handle high-dimensional data, we develop two deterministic algorithms that approximate the covariance matrices," Ollie/ReVerb can only identify the relation based on the verb "develop" (we; develop; two deterministic algorithms). Our algorithm can additionally identify the relation based on the "develop to" combination model (two deterministic algorithms; be developed to; high-dimensional data). From the sentence's meaning, research literature entity extraction aims to identify the relationship between research entities A and B in "developed A to handle B," as shown in bold in Table 2. When multiple auxiliary words appear, such as "Their friendship developed through their shared interest in the Arts," our algorithm can also discriminate the dependency relations between "through" and "in" with the core predicate "developed" through dependency chain models to ensure recognition precision.

(2) Noise processing during SAO structure-based entity recognition effectively improves precision. Ollie and ReVerb have relatively high error rates. Table 2 shows Ollie results with confidence > 0.5. In raw results, "Conventional online learning algorithms…" was incorrectly recognized as 0.436: (Conventional online; be going only once through; training data) due to improper preprocessing of the gerund "learning," causing the noun phrase "Conventional online learning algorithms" to be incorrectly segmented.

(3) Better relation identification based on non-verb association media. For the sentence "We present a multilingual Named Entity Recognition approach based on a robust and general set of features across languages and datasets," both our algorithm and Ollie can identify the relation based on "based" (past participle as attributive modifier of "approach): (a multilingual Named Entity Recognition approach; be based on; a robust and general set of features). Our algorithm mainly adds present participle and past participle-based entity recognition models, along with dependency chain discrimination models for these relational connectives and auxiliary

words (”to,” “for,” “with,” “as,” “in,” etc.).

3.3.3 Error Analysis of Research Entity Recognition Algorithm Analysis reveals that errors in our algorithm mainly fall into four categories:

(1) Recognition errors caused by heavy reliance on part-of-speech taggers and dependency parsers account for about 46%. For example, in “The results display the potential of algorithm selection to achieve significant performance improvements across a broad range of problems and algorithms,” “display” is incorrectly tagged as “(VP (NN display)).” In “The optimization objective we study asks to minimize the expected total cost of reaching a state in the target set, while ensuring that the target set is reached almost surely,” “the target set is reached” is incorrectly tagged as “(S (NP (DT the) (NN target)) (VP (VBD set))) (VP (VBZ is) (VP (VBN reached)))”, where “set” is incorrectly tagged as a verb. The former error can be corrected later, while the latter is difficult to fix and depends on Stanford NLP tool upgrades.

(2) Recognition errors due to missing “exception rule restriction” templates account for about 20%. For example, in “Unfortunately, it is relatively easy to develop sophisticated models to help reduce the error of estimation by a few percent,” “to help reduce” could semantically be written as “to reduce” without major ambiguity, but the tagging result differs slightly from common infinitive structures and requires special handling. Another example: “During this research a prototype of a 3D cadastre was developed” commonly appears as “During this research, a prototype of a 3D cadastre was developed.” Adding a comma would clarify the sentence structure, requiring special rules to correct.

(3) Recognition errors caused by context coreference chain identification account for about 12%. Complex coreference chains not covered by the model cause errors. For example, in “Our results were satisfactory and were compared with those obtained by a learning system based on Self-Organizing Maps,” “those” is tagged as a determiner (DT) with coreference to “results.” In “We also take the opportunity to clarify some properties of the semidefinite relaxation, were it to be used for an actual nonconvex problem in this area,” “it” is tagged as a personal pronoun (PRP) with coreference to “properties.” In “In the present study, a time series neuro-fuzzy model is proposed that is capable of exploiting the strengths of traditional time series approaches,” “that” is tagged as a relative clause conjunction (SBARIN) with coreference to “model.” Coreference chain discrimination models have many types and rules, but this error can be reduced through further optimization.

(4) Other errors account for about 22%, such as special complex sentence patterns and special compound words.

Conclusion

Entity and relation extraction has proven useful in many natural language processing tasks [21]. This paper addresses the noise issues in long sentences and the particularities of research entity extraction, combining part-of-speech tagging and dependency parsing to improve entity extraction models that use pattern matching for triple extraction. The improvements are validated through examples. These new measures provide theoretical and practical reference value for further research on precise recognition and extraction of research entities and their relations. The main innovations and contributions include: improving long sentence processing to clarify semantic structures and enhance entity recognition precision; modeling dependency relation chains through analysis of core predicates and their auxiliary words to significantly improve research entity recognition and extraction; providing a basic approach to research entity recognition using research problems and their solutions as examples, facilitating research question answering applications.

Limitations include a small experimental dataset lacking large-scale application testing. Future work includes improving and expanding core predicate classification, accumulating predicate-based entity extraction templates, supplementing research entity recognition models using nouns/adjectives as association words, and developing confidence calculation methods.

References

[1] Xu Fen, Wang Ting, Chen Huowang. Chinese entity relation extraction based on SVM method[C]//Proceedings of the 9th National Conference on Computational Linguistics: Frontier Research and Applications of Content Computing. Dalian: Dalian University of Technology, Tsinghua University State Key Laboratory of Intelligent Technology and Systems, Chinese Information Processing Society of China, 2007: 497-502.

[2] Kambhatala N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. Stroudsburg: ACL, 2004: 1-4.

[3] Guo Xiyue, He Tingting, Hu Xiaohua, et al. Chinese entity relation extraction based on syntactic and semantic features[J]. Journal of Chinese Information Processing, 2014, 28(6): 183-189.

[4] Gan Lixin, Wanchang Xuan, Liu Dexi, et al. Chinese entity relation extraction based on syntactic and semantic features[J]. Journal of Computer Research and Development, 2016, 53(2): 284-302.

[5] LI H, WU X, LI Z, et al. A relation extraction method of Chinese named entities based on location and semantic features[J]. Applied intelligence, 2013, 38(1): 1-15.

[6] Xi Bin, Qian Longhua, Zhou Guodong, et al. Application of linguistic combination features in semantic relation extraction[J]. Journal of Chinese Information Processing, 2008, 22(3): 44-50.

[7] BANKO M, CAFARELLA M J, SODERLAND S, et al. Open information extraction from the Web[C]//Proceedings of the 20th international joint conference on artificial intelligence. Hyderabad: Morgan Kaufmann Publishers Inc., 2007: 2670-2676.

[8] FADER A, SODERLAND S, ETZIONI O. Identifying Relations for Open Information Extraction[C]//Proceedings of the 2011 conference on empirical methods in natural language processing. Stroudsburg: ACL, 2011: 1535-1545.

[9] ETZIONI O, FADER A, CHRISTENSEN J, et al. Open information extraction: the second generation[C]//Proceedings of conference on artificial intelligence. Palo Alto: AAAI Press, 2011: 3-10.

[10] WU F, WELD D S. Open Information Extraction Using Wikipedia[C]//Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2010: 118-127.

[11] AKBIK A, LÖSER A. KrakeN: N-ary facts in open information extraction[C]//Proceedings of the joint workshop on automatic knowledge base construction and web-scale knowledge extraction. Stroudsburg: ACL, 2012: 52-56.

[12] SCHMITZ M, BART R, SODERL S, et al. Open language learning for information extraction[C]//Proceedings of the conference on empirical methods in natural language processing and computational natural language learning. Stroudsburg: ACL, 2012: 523-534.

[13] MAUSAM M. Open information extraction systems and downstream applications[C]//Proceedings of the twenty-fifth international joint conference on artificial intelligence. Palo Alto: AAAI Press, 2016: 4074-4077.

[14] Wu Wenya, Chen Yufeng, Xu Jin'an, et al. Survey of Chinese entity relation extraction research[J]. Computer and Modernization, 2018(8): 21-27.

[15] Tang Min. Research on Chinese entity relation extraction method based on deep learning[D]. Chengdu: Southwest Jiaotong University, 2018.

[16] LIN Y, LIU Z, SUN M. Neural relation extraction with multi-lingual attention[C]//Proceedings of the 55th annual meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017: 34-43.

[17] ILEV BAREIM, PROBERT D, PHAAL R. A review of TRIZ, and its benefits and challenges in practice[J]. Technovation, 2013, 33(2): 30-37.

[18] CHOI S, YOON J, KIM K, et al. SAO network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells[J]. Scientometrics, 2011, 88(3): 863-883.

[19] Guo Junfang, Wang Xuefeng, Qiu Pengjun, et al. Research on technology roadmap construction based on SAO analysis[J]. Studies in Science of Science, 2014(7): 976-981.

[20] Wang Xuefeng, Qiu Pengjun, Fu Yun. A new type of technology roadmap construction research—based on SAO structure information[J]. Studies in Science of Science, 2015(8): 1134-1140.

[21] ANGELI G, PREMKUMAR M J, MANNING C D. Leveraging linguistic structure for open domain information extraction[C]//Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing. Stroudsburg: ACL, 2015: 344-354.

[22] CORRO L D, GEMULLA R. ClausIE: Clause-based open information extraction[C]//Proceedings of the 22nd international conference on World Wide Web. New York: ACM, 2013: 355-366.

[23] Tregex, Tsurgeon and Semgrex[EB/OL].[2019-09-17]. https://nlp.stanford.edu/software/tregex.shtml.

[24] Jiang Ting, Sun Jianjun. Research on non-hierarchical relation extraction for academic resource ontology[J]. Library and Information Service, 2016, 60(20): 112-122.

[25] What is WordNet?[EB/OL].[2019-09-17]. https://wordnet.princeton.edu/.

Author Contributions

Zhao Huaming: Topic formulation, methodology design, algorithm testing, paper writing;
Qian Li: Research framework design, paper revision;
Yu Li: Corpus data collection and organization, paper revision.

A Research Entity Recognition Algorithm Based on Dependency Parsing

Zhao Huaming[1], Qian Li[1,2], Yu Li[1]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190
[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] To explore the recognition and extraction of research entities and their relationships, improve their recognition performance in complex situations such as long sentences, and provide reference for further applications. [**Method/process**] Based on the analysis of dependency syntactic features, a method for extracting research entity relationships is proposed, which includes: using the Stanford Tagger tool for part-of-speech tag-

ging of target text; based on annotation results, segmenting the target text into semantically coherent fragments with standardized structure around core predicates and SAO structures; through dependency parsing, identifying subjects and objects semantically related to core predicates to form (entity, relationship, entity) triples. [**Result/conclusion**] Comparative tests with mainstream algorithms such as Ollie and ReVerb show that this method can effectively improve the accuracy of research entity recognition.

Keywords: dependency parsing; research named entity; entity recognition; relation extraction

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