Postprint: Semantic Knowledge Extraction for Plant Species Diversity

Authors: Liu Jianhua, Wang Ying, Zhixiong Zhang, Li Chuanxi

Date: 2017-11-08T00:00:00+00:00

Abstract

[Objective] To expand the species-centric framework for plant species diversity extraction and explore semantic knowledge extraction methods. [Method] By integrating mainstream research in current biodiversity extraction, we adopt a species-centric approach to design a knowledge extraction framework encompassing multiple entities and their interrelationships, and leverage numerous existing professional databases to design and implement corresponding recognition methods. [Result] We design a species-centric knowledge extraction framework, explore semantic knowledge extraction methods for multiple entities and inter-entity relationships, and expand the content and approaches for extraction in the domain of plant species diversity. [Limitations] The completeness and accuracy of entity recognition are significantly influenced by the underlying knowledge base, and the types of inter-entity relationships are limited to co-occurrence, hierarchical, and grammatical relations, requiring further investigation. [Conclusion] This study expands the content and approaches of plant species diversity extraction, and can effectively support semantic retrieval and scientific computing.

Full Text

Extracting Semantic Knowledge from Plant Species Diversity Collections

Liu Jianhua[1,2], **Wang Ying**[1], **Zhang Zhixiong**[1], **Li Chuanxi**[3]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](China Great Wall Asset Management Co., Ltd, Beijing 100045, China)

Abstract

Objective: This study aims to expand the species-centered extraction framework for plant species diversity and explore methods for semantic knowledge extraction. **Methods:** Building on mainstream biodiversity extraction research, we designed a knowledge extraction framework encompassing multiple entities and their relationships, centered on species, and developed corresponding recognition methods by leveraging existing specialized databases. **Results:** We designed a species-oriented knowledge extraction framework and explored semantic knowledge extraction methods for various entities and their relationships, thereby expanding the scope and approach of plant species diversity extraction. **Limitations:** The completeness and accuracy of entity recognition are significantly influenced by the underlying knowledge base, and relationship types are currently limited to co-occurrence, hierarchical, and syntactic relationships, requiring further research. **Conclusion:** This research expands the content and methodology of plant species diversity extraction, effectively supporting semantic retrieval and scientific computation.

Keywords: Plant Species Diversity; Plant Species; Knowledge Extraction; Relation Extraction

2. Related Research Overview

Through the efforts of numerous researchers, many information extraction tools have been developed for the biodiversity domain. These tools employ either single methods such as natural language processing, dictionary matching, machine learning, rule-based templates, or shallow/deep syntactic parsing, or combine several of these approaches. The majority focus on recognizing various species names (scientific names, synonyms, common names, variety names, etc.), while some tools also address the identification of species traits. Thessen et al. [**?**] reviewed current research on species name recognition using natural language processing and machine learning algorithms in biodiversity science. Naderi et al. [**?**] introduced various extraction tools for the biomedical domain under the GATE framework. These publications provide comprehensive reviews of conventional biodiversity information extraction workflows, mainstream methods, and major tools at each stage. Rather than repeating these reviews, this paper examines current important biodiversity information extraction tools to discuss the extraction content in the plant species diversity domain, providing a reference foundation for our proposed knowledge extraction framework.

Current research on plant species diversity extraction can be summarized in the following areas:

2.1 Species Name Recognition and Normalization

Due to language differences and regional naming conventions, the same species name appears in scientific literature in various forms. Some follow the standard binomial (or trinomial) nomenclature in Latin, consisting of a genus name followed by a species epithet, with the genus capitalized and the species in lowercase, both in full form, often followed by the author' s surname [**?**]. Others use abbreviated forms with only the genus initial and full species epithet. Common names may appear in English or other languages, with the same species potentially having different common names across countries or regions [**?**]. These issues significantly increase the difficulty of species name recognition. Consequently, many researchers focus specifically on species name recognition, normalization, and organization, which represents the mainstream of current plant species diversity extraction research. Notable achievements include dictionaries for species name recognition and normalization such as NCBI Taxonomy [**?**], BioNames [**?**] (an online database linking animal names with source descriptions, taxonomy, and phylogenetic trees), and Species 2000 Global Species Catalog [**?**], as well as mature recognition tools like NetiNeti [**?**], OrganismTagger [**?**], Linnaeus [**?**], and TaxonGrab [**?**].

2.2 Species Trait Recognition

For taxonomic researchers, various trait descriptions such as root, stem, and leaf color and length are crucial references for species classification. Therefore, some bioinformatics researchers have explored automatic recognition methods for species traits. Taylor [**?**] manually established rules and dictionaries based on textual syntactic features to identify species parts, characteristics, and states. Tang et al. [**?**] built upon related research and used predefined templates for supervised learning to generate rules for recognizing leaf shape, size, color, arrangement, and fruit shape characteristics. CharaParser employs heuristic methods and syntactic feature-based rule generation to effectively recognize multiple species traits [**?**]. Duan Yufeng et al. [**?**] continue to explore information extraction from Chinese plant species diversity description texts.

2.3 Biological Network Recognition

Various biological entities (species, molecules, genes, proteins, etc.) have multiple relationships that can be expressed as network graphs, enabling biological system analysis through graph analysis [**?**, **?**]. Proteins and genes are focal points in biomedicine, and research on their recognition extends beyond plant species diversity. In current plant species diversity literature, species genetic sequencing can identify phylogenetic relationships, and protein or gene technologies can influence or alter biological environments or characteristics for research purposes. Therefore, protein and gene recognition involves not just identifying named entities but also identifying biological networks formed by various entities through verbs (or verb phrases), prepositions (or prepositional phrases), possessives, and other connections.

Overall, current biodiversity information extraction research, particularly in plant species diversity, mostly focuses on exploring recognition methods for specific information types, aiming to structurally describe plant species diversity characteristics or assist in species identification. Systematic research and framework design for knowledge-based organization and semantic retrieval of scientific literature content are rare. Based on current research achievements, this paper systematically designs a semantic organization knowledge framework and explores rapid recognition methods for corresponding knowledge units from a practical application perspective.

3. Semantic Knowledge Framework Design

To conduct semantic knowledge extraction for plant species diversity, we must first identify what content to extract from target resources—that is, construct a reasonable semantic knowledge description framework. This framework serves as an important basis for describing semantic knowledge units and their relationships in this domain and supports subsequent knowledge organization and revelation. Therefore, based on analysis of existing research and the practical requirements of the National Science Library, Chinese Academy of Sciences' project to "construct a biodiversity domain ontology building and semantic organization application demonstration platform," we designed a semantic knowledge framework to support this demonstration platform. This section details the framework' s design process and content.

3.1 Hierarchy of the Semantic Knowledge Framework

In constructing this framework, we first used "Oryza sativa (rice species)" as a search term to retrieve 100 scientific articles from PubMed' s *Plant Physiology* and *The Plant Cell* journals. We then manually annotated these articles and consulted experts from the Institute of Botany, Chinese Academy of Sciences to confirm the annotated knowledge units. Manual annotation was conducted at three levels, as shown in Figure 1 [Figure 1: see original paper].

Figure 1. Example of Manual Annotation Hierarchy

(1) **Sentence Level:** The focus is achieving structured organization of scientific literature by identifying knowledge sentence groups for specific purposes. In scientific literature, much knowledge cannot be simply presented as individual knowledge units (phrases) or relationships between units—for example, a complete experimental condition (the combined effect of chemical element concentration and temperature control) or a complete experimental process may contain multiple knowledge units and relationships. For such information, we extract and identify multiple closely related phrases or short sentences to form knowledge sentence groups and determine their types, such as research methods, experimental processes, or research results, thereby reorganizing scientific literature knowledge.

(2) **Knowledge Unit Level:** Scientific literature frequently contains numerous knowledge units with clear semantic categories, often appearing as named entity names or phrases that reveal deep textual content. Extracting and identifying these knowledge units enables fine-grained revelation of literature content, which is significant for subsequent semantic retrieval.

(3) **Knowledge Unit Relationship Level:** Knowledge units do not exist independently and dispersedly in literature but often form various semantic associations through co-occurrence, subject-predicate-object expressions, etc. Combining these semantic associations can maximize deep text mining.

Among these three annotation levels, sentence-level extraction research is relatively independent and has been discussed in our previous work [**?**], so this paper will not elaborate further. The following sections focus on detailed discussion of knowledge units and their relationships.

3.2 Content of the Semantic Knowledge Framework

In this framework, knowledge units and their relationships are essential components. Based on manual annotation results, current focal points in biodiversity extraction, project requirements, and feasibility of subsequent extraction and recognition, we designed the plant species diversity semantic knowledge framework shown in Figure 2 [Figure 2: see original paper]. The knowledge units (shown as boxes in Figure 2) are species-centered, extending to various knowledge units related to species. For plant species attribute descriptions, we reused some concepts from the Plant Ontology (PO)[1], which essentially covers the main knowledge points in current plant species diversity literature. Beyond hierarchical relationships (shown as arrowed connections in Figure 2), different categories of knowledge units also have associative relationships (non-arrowed connections in Figure 2). Through co-occurrence, syntactic, and semantic analysis, we can construct factual triples between these knowledge units to support further text analysis.

[1] The Plant Ontology, funded by the U.S. National Science Foundation (NSF), is a controlled vocabulary for plant structure and growth stages.

Figure 2. Plant Species Diversity Semantic Knowledge Framework

4. Implementation of Semantic Knowledge Extraction

The construction of the plant species diversity semantic knowledge framework facilitates the collection, organization, and structuring of existing knowledge, storing records from current plant ontology databases as instances of various knowledge types in Figure 2. It also provides clear targets for further knowledge extraction. Based on this framework, we conducted knowledge extraction in the plant species diversity domain following the process described below.

4.1 Corpus Integration and Experimental Data Selection

Based on the semantic knowledge organization framework and through expert consultation and reference to relevant research from the Institute of Botany, Chinese Academy of Sciences [**?**], we compiled and integrated domain terminology and vocabulary including the G2000[1] plant ontology database, NCBI species database [**?**], address name gazetteers, and small compound names from Chemical Entities of Biological Interest[2]. We integrated instances according to the knowledge units defined in the semantic organization framework, ultimately forming nearly 170,000 instance entries. These domain resources can be used directly as lexicons for annotating knowledge unit instances and can also support semi-automatic construction of entity recognition rule bases for identifying new instances.

Additionally, we obtained 23,000 journal abstracts from *Plant Physiology* and *The Plant Cell* in PubMed, and 27,049 scientific abstracts from Web of Science based on a list of 20 core journals provided by the Institute of Botany, Chinese Academy of Sciences, constructing an experimental dataset of over 50,000 articles for knowledge extraction experiments.

[1] A plant species ontology database provided by the Institute of Botany, Chinese Academy of Sciences.
[2] Chemical Entities of Biological Interest (ChEBI) is a freely available ontology of biochemical entities focusing on small molecular compounds.

4.2 Knowledge Extraction Framework Design

To better recognize knowledge units and their relationships, we designed the knowledge extraction framework shown in Figure 3 [Figure 3: see original paper].

Figure 3. Semantic Knowledge Extraction Framework

(1) **Input Data Sources:** Include scientific literature to be processed and related domain resources (plant diversity ontologies, NCBI species database, etc.).

(2) **Extraction Tools and Methods:** Employ different natural language processing tools (including Stanford Parser, Berkeley Parser, etc.) to achieve part-of-speech tagging, syntactic dependency analysis, and semantic analysis of sentences.

(3) **Entity Extraction and Relation Extraction:** These are cross-iterative processes. Entity extraction itself is iterative—newly recognized named entities are added to user dictionaries for subsequent recognition rounds. Relation extraction results can also discover new entities, which are then used in the next round of relation discovery.

(4) **Information Extraction Result Storage:** Based on result types, we use both RDF storage and database storage for entities and relationships.

4.3 Knowledge Unit Instance and Relationship Extraction

We conducted extraction using dictionaries, rules, and syntactic analysis methods. Direct lexicon-based entity annotation is fundamental to all knowledge extraction research. This study primarily relied on dictionaries to extract some species names, geographical locations, chemical elements and compounds, and domain subject terms. The specific process is similar to existing research and will not be elaborated here. This section focuses on rule-based instance extraction and new instance recognition methods.

(1) Knowledge Unit Instance Annotation and Extraction To identify knowledge unit instances beyond those in dictionaries, we designed a dictionary-based approach combining rules and statistical methods. The process includes:

Rule-Based Knowledge Unit Recognition: Although knowledge unit instances in scientific literature have various forms, our analysis revealed common patterns in their constituent words regarding morphology, part-of-speech, and combination methods (e.g., for persons, institutions, numerical information, equipment). For such knowledge unit instances, we can effectively improve recognition accuracy through manually assisted rule writing. We explored the following general process for rapid rule construction:

1. Collect sample instances of a specific category and perform word segmentation, sentence splitting, and part-of-speech tagging.
2. For simply structured instances like years, dates, experimental data, and related numerical descriptions, construct patterns using morphological rules.
3. Remove prepositions, adverbs, and other non-meaningful words from the segmentation results and identify special proprietary vocabulary (category indicator words) from a frequency perspective.
4. For instances with category indicator words, represent each instance entry as a pattern of part-of-speech and word form after natural language processing, where feature words and non-noun words maintain their original string output. As shown in the examples in Figure 4 [Figure 4: see original paper] (where Token represents segmentation, Token.orth represents orthography, and Token.category represents part-of-speech), we statistically analyze feature word positions in sample patterns and classify sample entries into three categories based on feature word position: "head," "middle," and "tail." Within each category, we further classify by word form combination: entries without prepositions or possessives, entries with prepositions, entries with possessives, and entries with both. This secondary classification yields effective pattern combinations. For example, for universities, the final patterns include: "of NN/NNS" *(NN/NNS indicates capitalized or all-capital nouns or plural nouns,* represents multiple NN/NNS)," NN/NNS* "," (NN/NNS)(' s)NN/NNS* "," NN/NNS*

NN/NNS*," etc. These learned patterns are converted to finite state machines for instance recognition.

Figure 4. Example of Knowledge Unit Instance Sample Pattern Output

For instances without category indicator words, we collect sample sentences containing certain research element instances and manually label them as training samples. We perform word segmentation, part-of-speech tagging, and syntactic parsing to obtain linguistic features. Next, we extract n context words adjacent to research element instances in sample sentences (n is adjustable; following Jiang et al. [**?**], we set n=4), calculate their frequency, and select the top three most frequent adjacent words. If these high-frequency words appear in more than 50% of entries, they can be considered semantic features as pre-context or post-context words.

For these pre/post-context words, we obtain their synonym sets Synset[train] from WordNet. For sentences to be analyzed, we similarly extract n adjacent context words and obtain their synonym sets Synset[test] from WordNet. We calculate the sum of similarities between Synset[train] and Synset[test] for the n adjacent words using formulas (1) and (2). We use maximum similarity between synonym sets instead of direct word-to-word similarity to account for variations in word choice, morphology, and spelling in real texts.

Where Sim is the final overall similarity and Sim(nw) is the similarity of each adjacent word.

Dictionary Similarity-Based Knowledge Unit Instance Recognition: While dictionary-based recognition cannot identify new instances, dictionaries provide important support for new instance recognition. Based on features like word form, part-of-speech, frequency, and syntactic components, we can select candidate sets of new knowledge unit instances. We then calculate edit distance between candidates and dictionary instances to obtain similarity scores for recognizing unknown words.

Syntactic Analysis-Based Knowledge Unit Instance Recognition: Entity extraction and relation extraction are cross-iterative processes—relation extraction results can discover new instances. For candidate instances unrecognizable by rules or dictionary similarity, we can use syntactic dependency and grammatical relationships (coordinate sentence components) obtained from syntactic analysis, combined with statistical algorithms, to determine instance semantic types.

Specifically, parsers represent sentences as hierarchical syntax trees. For example, the sentence "Bell, based in Los Angeles, makes and distributes electronic, computer and building products." produces the syntax tree shown in Figure 5 [Figure 5: see original paper], using Penn Treebank [**?**] tags compatible with most part-of-speech tagging systems.

Figure 5. Syntax Tree Generated by Parser [?]

In addition to syntax trees, parsers provide dependency analysis results, as shown in Figure 6 [Figure 6: see original paper]. In the dependency analysis, instances are shown in parentheses, with keywords like nsubj and partmod indicating specific dependency relations.

Figure 6. Dependency Parsing Results and Explanation

As shown in Figure 6, syntactic analysis clearly identifies compound noun phrases (NP modules in syntax trees), with dependency features revealing relationships within phrases. For example, conj_{and}(electronic-11, computer-13) shows the coordinate relationship between "electronic" and "computer." If either' s semantic category is determined, the other' s category can be inferred, enabling instance type annotation.

(2) Relationship Extraction Knowledge unit relationships include many types: co-occurrence, appositive syntax, coordinate syntax, factual relationships, and semantic hierarchical relationships. Co-occurrence is simplest—two knowledge unit instances co-occurring within a specified window (full text, abstract, sentence) indicates a relationship, providing the most direct way to determine relevance. Since our texts are journal abstracts, we use sentences as co-occurrence windows. This simple relationship determination will not be detailed here; instead, we focus on syntax-based, factual, and semantic rule-based relationship identification.

Appositive and Coordinate Syntax Relationship Extraction: As described in "Syntactic Analysis-Based Knowledge Unit Instance Recognition," some new instance recognition relies on appositive and coordinate relationships. Similarly, after confirming two instances' types, coordinate relationships identified through "and," "or," etc., in syntactic parsing can confirm appositive and coordinate syntactic associations.

Factual Relationship Identification: Factual relationships discussed here primarily refer to subject-predicate-object relationships in text—<S, P, O> (subject, predicate, object) facts that provide important support for subsequent reasoning. We designed the following process:

1. Input texts that have been segmented, split into sentences, and had knowledge unit instances recognized. Process each sentence iteratively. Create two empty relationship triple lists for non-verbal and verbal predicates.
2. Determine if each sentence contains one or more knowledge unit instances from Figure 2. If not, end analysis and continue to the next sentence. If containing two or more, proceed to step 3.
3. According to parser results, construct minimal simple sentences (containing only one subject-verb-object structure without any clauses) from the bottom of the syntax tree upward, building a simple sentence group. This becomes the second iteration point.

4. Determine if each simple sentence contains a knowledge unit instance from Figure 2. If not, end analysis and continue to the next sentence. If yes, proceed to step 5.
5. Use syntactic dependencies to extract subject-predicate-object relationships and construct (subject phrase, predicate verb, object phrase) triples.
6. Analyze these triples to determine if both subject and object phrases contain at least one knowledge unit instance from Figure 2. If yes, proceed to step 7. If all instances are in the same phrase, jump to step 8.
7. If both phrases contain only one instance each, check for semantic shift issues. If none, construct the relationship triple and add it to the verb relationship list. If shift exists, decide whether to discard based on shifted semantics. If phrases contain multiple instances, process them using permutations and combinations while noting ambiguities caused by coordination.
8. Analyze research element instances in relevant phrases and determine semantic relationships using their annotated types.
9. Output the verb relationship triple list.

Semantic Hierarchical Relationship Discovery: These relationships primarily involve possessives, fixed sentence patterns, and common expressions (such as "such as," "for example," "as well as"). We mainly extended Hearst patterns [**?**] and manually constructed over 20 relational rules to identify hierarchical relationships between instances.

4.4 Application of Knowledge Extraction Results

Using the above methods, we extracted 273,668 knowledge unit instances from titles and abstracts of over 50,000 relevant documents. The main extraction type distributions are shown in Table 1 (only showing results with >100 extracted instances).

Table 1. Distribution of Main Knowledge Units and Species Attribute Instances Extracted from Experimental Data

In addition to the knowledge unit instances shown in Table 1, we extracted 133,922 SPO syntactic relationships and 35,903 appositive relationships. Figure 7 [Figure 7: see original paper] shows partial SPO extraction results.

Figure 7. Partial SPO Syntactic Relationship Extraction Results in Plant Species Diversity Domain

Based on these extraction results and integrated domain knowledge bases and third-party resources, we constructed a semantic retrieval application demonstration platform for plant species diversity, providing users with domain knowledge revelation, semantic annotation, and ontology navigation. This validated the usefulness and effectiveness of our research. Figures 8 [Figure 8: see original

paper] and 9 [Figure 9: see original paper] show partial application demonstrations.

Figure 8. Knowledge Browsing, Retrieval, and Statistical Analysis Functions Based on Ontology Concepts or Entities

Figure 9. Co-occurrence Relationship Knowledge Graph for Single Articles Based on Semantic Knowledge Extraction

Compared with general biological knowledge extraction, the plant species diversity domain involves more complex knowledge unit types and relationships, such as ecological environments, species characteristics, and influencing factors. Therefore, designing the semantic knowledge framework requires considering more knowledge units from an application perspective, and specific recognition needs to comprehensively employ more domain-independent extraction methods to accommodate diverse knowledge unit instance extraction.

Based on analysis of current biodiversity information extraction research and the practical requirements of constructing a biodiversity domain ontology building and semantic organization application demonstration platform at the National Science Library, Chinese Academy of Sciences, this paper designed a plant species diversity semantic knowledge extraction framework and explored corresponding semantic knowledge extraction methods. This research primarily explored engineering-applicable knowledge organization frameworks and recognition methods from a practical application perspective. Therefore, dictionaries and manually written rules were essential components of our extraction approach. Consequently, the inherent limitations of dictionaries and manual rules somewhat restricted recognition completeness and accuracy. Future research will focus on refined recognition of various knowledge unit types.

References

[1] Thessen A E, Cui H, Mozzherin D. Applications of Natural Language Processing in Biodiversity Science [J]. Advances in Bioinformatics, 2012. DOI: 10.1155/2012/391574.
[2] Naderi N, Kappler T, Baker C J, et al. OrganismTagger: Detection, Normalization and Grounding of Organism Entities in Biomedical Documents [J]. Bioinformatics, 2011, 27(19): 2721-2729.
[3] Species [EB/OL]. [2016-04-12]. http://en.wikipedia.org/wiki/Species.
[4] Gerner M, Nenadic G, Bergman C M. LINNAEUS: A Species Name Identification System for Biomedical Literature [J]. BMC Bioinformatics, 2010. DOI: 10.1186/1471-2105-11-85.
[5] The NCBI Taxonomy Homepage [EB/OL]. [2016-04-12]. http://www.ncbi.nlm.nih.gov/Taxonomy/taxonom
[6] Page R D M. BioNames: Linking Taxonomy, Texts, and Trees [OL]. http://dx.doi.org/10.7717/peerj.190.
[7] Species 2000 [EB/OL]. [2016-04-12]. http://www.catalogueoflife.org/annual-checklist/2014/.
[8] Akella L M, Norton C N, Miller H. NetiNeti: Discovery of Scientific Names

from Text Using Machine Learning Methods [J]. BMC Bioinformatics, 2012. DOI: 10.1186/1471-2105-13-211.
[9] The OrganismTagger System [EB/OL]. [2016-04-12]. http://www.semanticsoftware.info/organism-tagger.
[10] Koning D, Sarlar I N, Moritz T. TaxonGrab: Extracting Taxonomic Names from Text [J]. Biodiversity Informatics, 2005, 2: 79-82.
[11] Taylor A. Extracting Knowledge from Biological Descriptions [C]//Proceedings of the 2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases. 1995: 114-119.
[12] Tang X, Heidorn P B. Using Automatically Extracted Information in Species Page Retrieval [C]//Proceedings of TDWG 2007. 2007.
[13] Cui H. CharaParser for Fine-grained Semantic Annotation of Organism Morphological Descriptions [J]. Journal of the Society for Information Science and Technology, 2012, 63(4): 738-754.
[14] Duan Yufeng, Huang Sisi. Information Extraction from Chinese Plant Species Diversity Description Text [J]. New Technology of Library and Information Service, 2016(1): 87-96.
[15] Li C, Liakata M, Rebholz-Schuhmann D. Biological Network Extraction from Scientific Literature: State of the Art and Challenges [J]. Briefings in Bioinformatics, 2013. DOI: 10.1093/bib/bbt006.
[16] Skusa A, Rüegg A, Köhler J. Extraction of Biological Interaction Networks from Scientific Literature [J]. Briefings in Bioinformatics, 2005, 6(3): 263-276.
[17] Bai Guangzu, He Yuanbiao, Ma Jianxia, et al. Application of Machine Learning with Limited Corpus to Identify Structure of Scientific Abstracts Automatically [J]. New Technology of Library and Information Service, 2014(7-8): 34-40.
[18] Xu Zheping, Cui Jinzhong, Qin Haining, et al. On the Architecture of Biodiversity e-Science Infrastructure in China [J]. Biodiversity Science, 2010, 18(5): 480-488.
[19] Jiang W, Guan Y, Wang X L. Improving Feature Extraction in Named Entity Recognition Based on Maximum Entropy Model [C]//Proceedings of the 5th International Conference on Machine Learning and Cybernetics. 2006: 2630-2635.
[20] De Marneffe M-C, Manning C D. Stanford Typed Dependencies Manual [OL]. http://nlp.stanford.edu/software/dependencies_{manual}.pdf.
[21] Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora [C]//Proceedings of the 14th International Conference on Computational Linguistics, 1992.

Author Contributions

Liu Jianhua: Proposed the overall framework, designed the semantic knowledge extraction framework, participated in implementation and development, wrote the main content, and proofread and revised the final version.
Wang Ying: Participated in framework design, corpus preparation, storage structure design, and data processing.

Zhang Zhixiong: Participated in framework design and provided revision suggestions.
Li Chuanxi: Responsible for implementation and development of extraction functions and provided development documentation.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is stored by the authors. Contact: liujh@mail.las.ac.cn.
[1] Liu Jianhua, Wang Ying, Li Chuanxi. Top40000 SPO Extraction Results.xls.

Received: 2016-04-14
Accepted: 2016-08-12

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