Information Extraction from Chinese Plant Species Diversity Description Texts (Postprint)

Authors: Duan Yufeng, Huang Sisi

Date: 2017-10-11T00:00:00+00:00

Abstract

【Objective】 To realize information extraction from descriptive texts of Chinese plant species diversity. 【Method】 Supported by a Chinese plant species diversity ontology, a strategy of hierarchical screening and annotation at the paragraph, sentence, and concept levels was employed to extract information from descriptive texts based on rules. 【Results】 Tested on a sample containing 4,734 information points, the accuracy, recall, and F-measure of information extraction were 0.86, 0.85, and 0.85, respectively. 【Limitations】 The rule set needs to be further refined for expressions that cannot currently be extracted accurately. 【Conclusion】 The research scheme can effectively realize information extraction from descriptive texts of Chinese plant species diversity.

Full Text

Information Extraction from Chinese Plant Species Diversity Description Text

Duan Yufeng, Huang Sisi
Business School, East China Normal University, Shanghai 200241, China
Email: yfduan@infor.ecnu.edu.cn

Abstract

[**Objective**] To extract information from Chinese plant species diversity description text.
[**Methods**] Taking the plant species diversity domain ontology as the foundation, we adopt a strategy of stepwise selection and annotation at the paragraph, sentence, and concept levels.
[**Results**] A sample containing 4,734 information points was used for testing. The extraction achieved precision, recall, and F-measure values of 0.86, 0.85, and 0.85 respectively.

[**Limitations**] To address remaining challenges in extracting information from description text, the rule set requires further improvement.
[**Conclusions**] The research scheme effectively fulfills information extraction from Chinese plant species diversity description text.

Keywords: Information extraction; Plant species diversity description text; Chinese information processing; Ontology

1. Introduction

With the digitization of biodiversity heritage libraries such as the Biodiversity Heritage Library (BHL), natural language processing techniques have become increasingly important for extracting structured information from botanical literature. Key tasks include digitization, annotation, name recognition and discovery, and morphological character extraction. This study focuses on developing an effective framework for extracting information from Chinese plant species diversity descriptions using domain ontology and machine learning approaches.

2. Methodology

2.1 Ontology Foundation

The system builds upon the Plant Ontology (PO), which provides a structured vocabulary for plant morphological and anatomical structures. Key object properties include:

- `has_shape`, `has_arrangement`, `has_texture`
- `has_color`, `has_growth_form`, `has_length`
- `has_participant`, `located_in`, `part_of`
- `derives_from`, `preceded_by`, `develops_from`

The ontology enables semantic annotation of botanical characters at multiple levels of granularity.

2.2 Text Processing Pipeline

The processing workflow employs several core components:

Word Segmentation: ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) is used for Chinese word segmentation, with part-of-speech tagging and named entity recognition capabilities.

Information Extraction: Conditional Random Fields (CRFs) implemented in CRF++ 0.58 serve as the primary extraction model. The system uses a bootstrapping approach to iteratively improve annotation quality.

Feature Engineering: Features include lexical items, part-of-speech tags, ontology class memberships, and contextual windows. The Vector Space Model (VSM) represents text for similarity calculations.

2.3 Annotation Strategy

A stepwise selection and annotation approach operates at three levels:

1. **Paragraph Level:** Identify relevant description sections (e.g., morphology, phenology)
2. **Sentence Level:** Extract sentences containing target information
3. **Concept Level:** Annotate fine-grained morphological characters using ontology terms

The system processes TXT and XML formats, handling DOM structures and web documents.

---

3. Experiments

3.1 Dataset

The evaluation uses 4,734 information points extracted from the Flora of China (http://frps.eflora.cn). The dataset includes 17 morphological character types across multiple plant families and genera.

3.2 Evaluation Metrics

Performance is measured using standard information extraction metrics:

$$\text{Precision}(P) = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall}(R) = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$F\text{-measure} = \frac{2 \times P \times R}{P + R}$$

Contingency table calculations follow the standard n_{ij} notation where n_{11} represents true positives, n_{12} false negatives, n_{21} false positives, and n_{22} true negatives.

3.3 Results

The system achieved the following performance on the test set:

- **Precision:** 0.86
- **Recall:** 0.85
- **F-measure:** 0.85

These results demonstrate effective extraction capability for Chinese plant species diversity descriptions. The bootstrapping approach improved recall from an initial 0.7138 to 0.85 while maintaining high precision.

---

4. Discussion

The stepwise ontology-driven approach proves effective for domain-specific information extraction. Current limitations include handling out-of-vocabulary terms and complex nested descriptions. Future work will focus on expanding the rule set and integrating deep learning methods to improve coverage of rare morphological characters.

References

[1] BHL. Biodiversity Heritage Library [EB/OL]. [2015-09-27]. http://www.biodiversitylibrary.org/

[2] Thessen A E, Cui H, Mozzherin D. Applications of Natural Language Processing in Biodiversity Science [J]. Advances in Bioinformatics, 2012: Article ID 391574. doi: 10.1155/2012/

[3] Vanel J M. Worldwide Botanical Knowledge Base [EB/OL]. [2011-10-11]. http://wwbota.free.fr/

[4] Cui H, Heidorn P. The Reusability of Induced Knowledge for Automatic Semantic Markup of Taxonomic Descriptions [J]. Journal of the American Society for Information Science and Technology, 2007, 58(1): 133-149.

[5] Duan Yufeng, Hei Zhenzhen, Ju Fei, et al. Study on Semantic Markup of Species Description Text in Chinese Based on Auto-learning Rules [J]. New Technology of Library and Information Service, 2012(5): 41-47.

[6] Duan Yufeng, Hei Zhenzhen, Ju Fei, et al. Semantic Annotation of Species Description Text in Chinese Literature by Naïve Bayes Classifier [J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(8): 805-812.

[7] Taylor A. Extracting Knowledge from Biological Descriptions [C]. In: Proceedings of the 2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases. 1995: 114-119.

[8] Wood M M, Lydon S J, Tablan V, et al. Using Parallel Texts to Improve Recall in IE [C]. In: Proceedings of Recent Advances in Natural Language Processing (RANLP' 03). 2003: 505-512.

[9] Tang X, Heidorn P B. Using Automatically Extracted Information in Species Page Retrieval [OL]. [2011-08-10]. http://www.tdwg.org/proceedings/article/view/195/

[10] Gruber T R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing [J]. International Journal of Human-Computer Studies, 1995, 43(5-6): 907-928.

[11] Soderland S. Learning Information Extraction Rules for Semi-Structured and Free Text [J]. Machine Learning, 1999, 34(1-3): 233-272.

[12] Xiang Yang, Wang Min, Ma Qiang. Research on Jena-based Ontology Building [J]. Computer Engineering, 2007, 33(14): 59-61.

[13] Abascal R, Sanchez J A. X-tract: Structure Extraction from Botanical Textual Descriptions [C]. In: Proceeding of the String Processing & Information Retrieval Symposium & International Workshop on Groupware. 1999: 2-7.

[14] Diederich J, Frotuner R, Milton J. Computer-assisted Data Extraction from the Taxonomical Literature [OL]. [2011-08-15]. http://math.ucdavis.edu/~milton/genisys.html

[15] Cui H. CharaParser for Fine-grained Semantic Annotation of Organism Morphological Descriptions [J]. Journal of the American Society for Information Science and Technology, 2012, 63(4): 738-754.

[16] Cui H, Singaram S, Janning A. Combine Unsupervised Learning and Heuristic Rules to Annotate Morphological Characters [J]. Proceedings of the American Society for Information Science and Technology, 2011, 48(1): 1-9.

[17] Sha Lihua. Research on Semantic Annotation for Domain Documents [D]. Changchun: Jilin University, 2009.

[18] Shi Jing. Information Extraction and Analysis Based on Plant Ontology [D]. Yangling: Northwest Agriculture and Foresty University, 2010.

[19] Cui H. CharaParser for Fine-grained Semantic Annotation of Organism Morphological Descriptions [J]. Journal of the American Society for Information Science and Technology, 2012, 63(4): 738-754.

[20] Apache Jena [CP/OL]. https://jena.apache.org/

[21] CRF++: Yet Another CRF Toolkit [CP/OL]. https://taku910.github.io/crfpp/

[22] Flora of China Editorial Committee. Flora of China [DB/OL]. [2007-09-28]. http://frps.eflora.cn/

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