Research on Knowledge Ontology Construction Techniques for Pre-Qin Classics: Postprint

Authors: He Lin, Yaling Chen, Sun Kedi

Date: 2023-04-01T16:15:51+00:00

Abstract

[Purpose/Significance] The construction of semantic ontologies for classical texts can facilitate their mining and analysis. However, significant grammatical differences between classical and modern texts pose considerable challenges to the construction of semantic ontologies for classical texts. [Method/Process] This paper employs natural language processing techniques to explore ontology construction methods for pre-Qin classics. Adopting the CIDOC CRM framework, which is widely used in the international cultural heritage domain, we design an ontology model for pre-Qin classics. Addressing the content characteristics and syntactic features of classical texts, we combine rule-based extraction with conditional random field methods to propose a technique for automatic ontology instance acquisition, and evaluate it using the Zuo Zhuan as experimental corpus. [Results/Conclusion] Experimental results demonstrate that the ontology instance extraction techniques proposed in this paper can effectively improve the efficiency of ontology construction for classical texts. The F-score for rule-based ontology instance extraction experiments is approximately 93%, while the F-score for the optimal feature template in conditional random field-based ontology instance extraction reaches 82.51%. Part-of-speech and positional information are of significant importance in ontology instance acquisition.

Full Text

Research on Ontology Construction Techniques for Pre-Qin Chinese Classics

He Lin, Chen Yaling, Sun Kedi
Department of Information Management, Nanjing Agricultural University, Nanjing 210095

Abstract

[**Purpose/Significance**] Constructing semantic ontologies for classical Chinese texts can facilitate the mining and analysis of these texts. However, significant grammatical differences between classical and modern Chinese pose substantial challenges for ontology building. [**Method/Process**] This paper explores ontology construction methods for Pre-Qin classics using natural language processing techniques. We design an ontology model for Pre-Qin classics based on CIDOC CRM, an international standard framework for cultural heritage. Combining rule-based extraction with Conditional Random Fields (CRF), we propose a comprehensive technique for automatic ontology instance acquisition, tested using *Zuo Zhuan* as experimental corpus. [**Results/Conclusion**] Experiments demonstrate that our proposed instance extraction technique effectively improves ontology construction efficiency for classical texts. The rule-based method achieved an F-score of approximately 93%, while the CRF-based method reached 82.51% using the optimal feature template. Part-of-speech and positional information prove crucial for ontology instance acquisition.

Keywords: Pre-Qin classics; Zuo Zhuan; Ontology construction; Conditional Random Fields; Rule matching
Classification Number: G254
DOI: 10.13266/j.issn.0252-3116.2020.07.002

The Chinese civilization has maintained an unbroken lineage for five thousand years, largely due to the vast sea of classical texts preserved from various historical periods. Formalizing and modeling the knowledge embedded in these classics, and revealing their associative relationships, not only facilitates the identification, understanding, and sharing of traditional cultural knowledge but also promotes deep mining of classical texts and advances digital humanities research. In the first half of the 20th century, the Harvard-Yenching Institute compiled the printed concordance series *Hanxue Yinde Congkan*, advancing knowledge representation from document-level to lexical-level units. The development of semantic web technologies like ontologies has injected new vitality into knowledge organization for classical texts. Scholars have undertaken related research on genealogical resource description ontologies, the "Twenty-Four Histories" ontology, and Traditional Chinese Medicine ancient books ontology. However, compared to existing ontology construction methods, building ontologies for Pre-Qin classics faces two major challenges: designing the semantic framework and extracting semantic relationships.

2. Literature Review

Recent years have witnessed fruitful achievements in the digitization of classical texts, enabling exploratory research on automatic word segmentation, part-of-speech tagging, named entity recognition, and lexical semantics for classical Chinese, yielding significant results. For word segmentation, CRF-based methods have achieved excellent performance, with experiments on texts like *Mencius*,

Zuo Zhuan, *Book of Han*, and *Book of Poetry* reporting F-scores approaching 98%. For part-of-speech tagging, scholars have conducted experimental studies on texts including *Chu Ci*, *Ming History*, *Zuo Zhuan*, and *Analects*, with the best results achieving F-scores near 95%. Research on named entity recognition has been relatively limited compared to segmentation and POS tagging. Some scholars have investigated methods for extracting person names, place names, and temporal expressions from classics such as *Mencius*, *Zuo Zhuan*, *Twenty-Four Histories*, and *Records of the Three Kingdoms*, while others have studied the extraction of place names and products from local gazetteers. Lexical semantics research has built upon achievements in automatic segmentation and POS tagging, providing foundations for syntactic and semantic annotation. Researchers have employed various techniques for word sense disambiguation and shallow syntactic parsing in classical Chinese.

Overall, while certain progress has been made in automatic segmentation and POS tagging for ancient texts, further research is needed to better apply these results for extracting named entities and relationships, and to construct domain ontologies and other semantic tools that can drive deeper mining of classical texts. Currently, establishing ontology semantic frameworks requires manual participation. Automatically extracting relevant terms and attribute relationships from corpora using natural language processing and machine learning technologies represents a crucial pathway for ontology application. Recent years have seen numerous achievements in automatic ontology construction. Scholars have built large general-purpose ontologies like DBpedia Ontology and YAGO based on Wikipedia and WordNet. Natural science domains such as life sciences and geoscience have developed relatively large practical domain ontologies, including GeoNames Ontology, The Drug Ontology, UMLS SemNet, Gene Ontology, and SNOMED. In humanities and social sciences, researchers have attempted ontology construction in history and philosophy, developing ontologies for the Three Kingdoms period, KMT-CPC cooperation, national history, Twenty-Four Histories, and philosophy. The Shanghai Library designed an ontology model for its genealogical collections, Zhonghua Book Company developed the "Twenty-Four Histories" ontology for semantic organization of persons, places, and times, and the China Academy of Chinese Medical Sciences built a knowledge base of traditional Chinese medicine classics by extracting knowledge units and establishing attribute relationships. Scholars have also applied metadata and ontology technologies to information resource description and organization in drama, folklore, and other fields.

However, ontology construction for classical texts still faces numerous difficulties. On one hand, there remains a lack of top-level semantic description frameworks specifically for classical texts. On the other hand, the grammar and syntax of classical Chinese differ substantially from modern language, requiring further investigation into methods for mining concepts and conceptual relationships from classical texts. Against this backdrop, this paper attempts to design a universally applicable semantic framework by reusing existing ontologies, focusing on researching automatic methods and techniques for ontology instance

acquisition.

3. Ontology Model Construction for Pre-Qin Classics

3.1 Challenges in Pre-Qin Classics Ontology Construction

3.1.1 Semantic Framework Design The Pre-Qin period represents a crucial era for the origins of traditional Chinese thought and culture. Pre-Qin classics document the nation's philosophical ideas, traditional virtues, and humanistic spirit, resulting in broad categories of knowledge points with complex semantic relationships. Designing an ontology semantic framework that can formally and modelically describe this knowledge and reveal associative relationships constitutes the first major challenge. For the framework design, to enable understanding and sharing of knowledge embedded in Pre-Qin classics, we referenced numerous existing ontology projects. The CIDOC CRM (Conceptual Reference Model), proposed by the International Committee for Documentation, provides an object-oriented ontology describing the conceptual system and relationship definitions needed in cultural heritage work. Widely applied in tangible and intangible cultural heritage domains, this model offers excellent universality. Based on CIDOC CRM and Pre-Qin classics research literature, we organized the content of classics and conducted cluster analysis (see the companion paper "Theme Mining and Evolution Analysis of Social Development in the Spring and Autumn Period"), identifying five core categories: military, marriage, diplomacy, politics, and livelihood. Under this framework, we systematically categorized attribute relationship hierarchies corresponding to physical objects, symbolic objects, and conceptual objects embedded in classics, constructing an ontology model for classical texts.

3.1.2 Semantic Relationship Extraction Technology Scholars have explored various methods for ontology semantic relationship extraction using natural language processing and machine learning technologies. However, the grammar and syntax of Pre-Qin classics differ significantly from modern Chinese, preventing direct application of current NLP methods to relationship extraction. Using *Zuo Zhuan* as an example, Pre-Qin classics feature short sentence lengths and dispersed semantic themes across chapters. Corpus resources are also critical for semantic relationship extraction. While modern Chinese has accumulated large amounts of annotated corpora, publicly available annotated classical Chinese corpora are extremely limited due to the difficulty of classical grammar and limited digital resources. This poses tremendous challenges for relationship extraction. Our data source is the *Zuo Zhuan* corpus annotated by Chen Xiaohe's research team at Nanjing Normal University, which provides manual word segmentation and POS tagging—currently one of the few high-quality Pre-Qin corpora available. However, to extract semantic relationships from *Zuo Zhuan*, we still needed to add semantic annotations, a task complicated by the characteristics of classical Chinese grammar.

3.2 Annotation Methods for Pre-Qin Classics

Ontology semantic relationship extraction essentially involves extracting "subject-predicate-object" triples from text. Based on this principle, we combined syntactic and role information, using role-based BIO annotation to label classical texts. Table 1 shows the label meanings, including Agent (施事者), Patient (受事者), Instrument (工具), Location (处所), and Time (时间). Following this method, we annotated *Zuo Zhuan* according to the Pre-Qin ontology semantic framework. For example, the sentence "仲庆父请伐齐师" (Zhong Qingfu requested to attack the Qi army) yields the annotation shown in Table 2 . In the results, predicate V corresponds to property types in the ontology model, the agent E1 corresponds to the domain of object properties, and the patient E2 corresponds to the range. Therefore, obtaining large quantities of BIO-annotated results is essential for ontology instance extraction.

4. Research on Ontology Instance Acquisition Technology for Pre-Qin Classics

Addressing the syntactic characteristics of Pre-Qin classics, this paper investigates a hybrid method combining rule-based and CRF approaches to maximize utilization of inherent expression patterns in classical Chinese for ontology instance extraction, tested on *Zuo Zhuan*.

4.1 CRF-Based Object Property Relationship Acquisition

Object property relationships primarily represent attributes between two classes, corresponding to agent and patient categories in BIO annotation, with property types determined by trigger verb semantics. To acquire numerous BIO-annotated results, we treat role acquisition as a sequence labeling problem, using Conditional Random Fields (CRF) to identify relevant roles. CRF, proposed by Lafferty et al. in 2001, is a conditional probability distribution model that fits the target P(Y|X). Based on input random variable X, it constructs feature functions from feature templates as statistical data for training, predicts the joint probability distribution of output random variable Y, and ultimately finds the optimal output sequence with highest probability. Feature selection significantly impacts training results, with features divided into state transition matrices and observation sequence features. Table 3 shows a sample of training data features and target labels. Table 4 illustrates five features from the feature template when "O-E1" is set as the current word label and "师" (army) as the current word. As CRF algorithms are well-established, we omit implementation details here, with recognition effects of different feature templates discussed in the experimental evaluation section. In the Pre-Qin ontology, different category property relationships are determined by trigger verb semantics. After acquiring relevant roles, we identify the corresponding ontology property types through the semantic types of trigger verbs in the sentences.

4.2 Trigger Verb Identification

Analysis of classical texts reveals that trigger verbs often determine sentence semantic types. These verbs are crucial for identifying property relationships in the Pre-Qin knowledge ontology. By identifying trigger verb semantic types and acquiring their corresponding BIO roles, we complete category property relationship extraction. To build trigger verb collections for ontology property relationships, we first 统计动词词频 (statistically analyzed verb frequencies) in Pre-Qin texts, manually determined initial trigger verb sets based on category property relationships in the ontology model, then used Bootstrapping self-expansion algorithms to augment and refine the trigger verb collection.

Bootstrapping is an unsupervised machine learning algorithm based on statistical knowledge. It begins with an initial core word set, calculates co-occurrence degrees between words to determine correlation, sets a threshold, and adds words exceeding this threshold to the core set through iterative training until the set stabilizes. Specifically, we use tf-idf values as word weights to select top-k new words or manually confirm initial theme sets, then evaluate each candidate word using function T to calculate scores, adding top-ranked words to generate new theme sets iteratively. The iteration can be controlled by manually set counts or theme set sizes. Formula (1) shows the calculation, where s and w represent the core word set and a candidate word respectively, and F(w,s) represents the total co-occurrence frequency between the candidate word and all core words.

4.3 Rule-Based Data Property Relationship Acquisition

Data property relationships (DataProperty) do not involve relationships between two classes but rather basic characteristics of entities, such as a person's name, gender, nationality, and official position. Analysis of Pre-Qin texts reveals that these basic attribute relationships exhibit certain patterns due to classical Chinese syntactic characteristics. Therefore, we can extract attributes with inherent expression patterns using regular expressions. The identification process involves: (1) analyzing words directly tagged as nr (person name) by composition and position; (2) statistically analyzing POS tags and words surrounding person names to identify other vocabulary containing basic attribute relationships; (3) constructing rules based on this analysis; and (4) building regular expressions to identify all instances meeting requirements.

5. Experimental Evaluation

5.1 Evaluation Methods

5.1.1 Experimental Corpus We selected the important Pre-Qin classic *Zuo Zhuan* as our experimental corpus, which includes word segmentation and POS tagging results. We performed BIO role annotation on this basis, with annotation samples shown in Table 5 . After manual annotation and consistency

checking, label frequency statistics are presented in Table 6 .

5.1.2 Evaluation Metrics We employed three evaluation metrics: Precision, Recall, and F-score: - Precision (P) = (Number of correctly predicted labels / Total number of machine-predicted labels) × 100% - Recall (R) = (Number of correctly predicted labels / Total number of actual labels) × 100% - F-score (F) = 2 × Precision × Recall / (Precision + Recall) × 100%

5.2 Trigger Verb Extraction Experiment

We conducted statistical analysis of all verbs in *Zuo Zhuan* using the Bootstrapping iteration method to acquire trigger verbs. Table 7 shows the extraction results for six property types: T2-Marriage, T3-Bear, T6-Garrison, T7-Attack, T8-Alliance, T9-Politics, E67-Birth, E69-Death, and T12-Career.

5.3 Evaluation of Rule-Based Data Property Relationship Recognition

Following the rule-based matching process in Section 4.3, we conducted rule extraction experiments on person-related corpora. The analysis proceeded in three steps: First, we analyzed words with nr (person name) POS tags. Table 8 shows that "子" (zi) and "公" (gong) are high-frequency characters. "子" appears frequently on the left due to Pre-Qin naming conventions and on the right because it was an honorific. "公" often represents a monarch when appearing as a single character, frequently appears as "公子" (prince) on the left, and denotes "Duke" when on the right. Besides these, left characters often include country names while right characters include generational names.

Second, we analyzed boundary words around person names. Table 9 shows that punctuation (w), verbs (v), and prepositions (p) are not valid components for person name extraction, while place names (ns) and common nouns (n) can combine with person names, serving as important clues.

Third, we analyzed left-boundary words with ns and n POS tags. Table 10 reveals that ns-tagged words are generally country names, while n-tagged words typically represent positions and family terms.

Based on this analysis, we identified the construction rule: [Country/Identity] + [Person Name]. Table 11 shows regular expression matching examples. The last word of each match represents the person name. If only one word exists and it's a country name, that becomes the person's nationality. If the suffix is "公" or "侯" or "伯", the person is a male monarch; if it's "氏", the person is female. Table 12 presents the rule-based matching results, showing gender identification achieved the highest F-score while official position information scored lowest. Error analysis revealed that complex person descriptions would make rules overly redundant and specific if all cases were included, creating a strong trade-off between recall and precision that warrants further investigation.

5.4 Evaluation of CRF-Based Object Property Relationship Recognition

Using CRF algorithms on texts filtered by trigger verb sets, we designed four templates for comparison: - Template 1: Window size of 1, using only word features - Template 2: Added POS features to Template 1 - Template 3: Added word-POS relationship features to Template 2 - Template 4: Added positional features to Template 3

Table 14 shows that Template 2 with POS features significantly improved results. Template 3 with word-POS relationships yielded modest improvements. Template 4 with positional information effectively enhanced recognition, making it our final template for BIO role identification.

Table 15 presents detailed results for Template 4, with F-scores above 80% for most labels including O-V, O-E1, O-E2, O-L, O-T, B-E1, and I-E1. Error analysis revealed high accuracy for single-character words but lower accuracy for multi-character words, likely due to the predominance of single-character words in classical Chinese and insufficient training samples. Expanding the training dataset could potentially improve results.

Conclusion

This paper, using *Zuo Zhuan* as a case study, explores automatic ontology construction methods and techniques for Pre-Qin classical texts. Based on an ontology model designed for classical text content, we investigated a hybrid approach combining rule-based and CRF methods for automatic ontology instance acquisition. Experiments demonstrate that our proposed method can effectively extract ontology instances from Pre-Qin classics. The constructed ontology can comprehensively describe persons and events in Spring and Autumn period society, helping researchers mine implicit knowledge from linear text information.

However, limitations remain. Trigger verb acquisition could be improved through external dictionary assistance to enhance accuracy and recall. Additionally, our instance extraction process only utilized word, POS, and positional information. Future work could improve extraction by enhancing syntactic parsing precision to obtain more features, and explore deep learning techniques to further boost instance extraction performance.

References

[1] Zong Fan. Let ancient books “come alive” [N]. Guangming Daily, 2017-11-30(14). [2] Xia Cuijuan, Zhang Lei. Application of linked data in digital humanities services for genealogy [J]. Library Journal, 2016, 35(10): 26-34. [3] Yu Tong, Cui Meng, Li Haiyan, et al. Application research of ISO technical specification “Semantic network framework of traditional Chinese medicine language system” [J]. China Medical Herald, 2016, 13(4): 89-92. [4] Dong Hui,

Xu Lei, Wang Fei, et al. Research on semantic analysis of Chinese historical records based on semantic systems [J]. Library Theory and Practice, 2015(4): 1-5, 46. [5] Chen Xiaohe. Information processing of Pre-Qin documents [M]. Beijing: World Book Publishing Company, 2013. [6] Ouyang Jian. Large-scale ancient text visualization analysis and mining for digital humanities research [J]. Journal of Library Science in China, 2016, 42(2): 66-80. [7] Zhu Xiao, Jin Li. Application effect exploration of CRF graphical models in POS tagging research of *Ming History* [J]. Fudan Journal (Natural Science), 2014, 53(3): 297-304. [8] Liu Liu, Li Bin, Qu Weiguang, et al. Automatic acquisition of temporal characteristics of Pre-Qin vocabulary and automatic determination of document era [J]. Journal of Chinese Information Processing, 2013, 27(5): 107-113. [9] Yu Lili, Ding Dexin, Qu Weiguang, et al. Research on classical Chinese word sense disambiguation based on conditional random fields [J]. Microelectronics & Computer, 2009, 26(10): 45-48. [10] Ren Feiliang, Shen Jikun, Sun Binbin, et al. Survey on domain ontology construction from text [J]. Chinese Journal of Computers, 2019, 42(3): 654-676. [11] Wimalasuriya DC, Dou D. Ontology-based information extraction: An introduction and survey of current approaches [J]. Journal of Information Science, 2010, 36(3): 306-323. [12] Wang Ying, Zhang Zhixiong, Sun Hui, et al. Research on semantic revelation and organization methods of national history knowledge [J]. Journal of Library Science in China, 2015, 41(4): 55-64. [13] Thakker D, Karanasiou S, Blanchard E, et al. Ontology for cultural variations in interpersonal communication: Building on theoretical models and crowdsourced knowledge [J]. Journal of the Association for Information Science and Technology, 2017, 68(6): 1411-1428. [14] Zhou Yaolin, Zhao Yue, Sun Jingqiong. Research on organization and retrieval of intangible cultural heritage information resources [J]. Information Science, 2017, 36(8): 166-174. [15] ISO technical committee 46 variations in interpersonal communication, subcommittee. Information and documentation: A reference ontology for the interchange of cultural heritage information [S]. ISO 21127:2014. Geneva: ISO, 2014. [16] Doerr M. The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata [J]. AI Magazine, 2003, 24(3): 75-92. [17] Gu Donggao. Chronological tables of major events in Spring and Autumn [M]. Beijing: Zhonghua Book Company, 1993. [18] Tong Shuye. Research on Zuo Zhuan of Spring and Autumn [M]. Shanghai: Shanghai People's Publishing House, 2019. [19] Chen Xiaojie. Research on constructing war knowledge maps of Zuo Zhuan based on ontology [D]. Nanjing: Nanjing Agricultural University, 2018. [20] Chen Yaling. Research on Pre-Qin personage knowledge ontology construction method based on CIDOC CRM [D]. Nanjing: Nanjing Agricultural University, 2019. [21] Chen XH, Li B, Feng MX, et al. Ancient Chinese corpus [M]. Philadelphia: Linguistic Data Consortium, 2017. [22] Lafferty JD, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]//Proceedings of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2001: 282-289. [23] Lü Yunyun, Li Yue, Wang Suge. Research on ensemble classifier for Chinese sentence sentiment classification based on Bootstrapping [J]. Journal of Chinese Information Processing, 2013, 27(5): 84-93.

Author Contributions:
He Lin: Topic selection, framework design, paper writing and revision;
Chen Yaling: Paper writing and algorithm implementation;
Sun Kedi: Data analysis and processing.

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