ontoNLP: A Mechanism for Transforming Textual Data into Structured Data

Jun WANG[1,2], Tong WEI[1,2,*], Xuemei TANG[1,2], Zhaoji WANG[3]

1. The Department of Information Management, Peking University
2. The Research Center of Digital Humanities, Peking University
3. School of International Chinese Language Education, Beijing Normal University

Abstract In the realm of digital humanities, the focus is on steering humanities research towards data-driven methodologies. Data-driven humanities research commonly relies on structured datasets. However, for humanities scholars, effectively converting textual data into structured data poses a significant challenge. This challenge is especially pronounced when dealing with large-scale textual data under unified standards. Annotation plays a pivotal role in dealing with this challenge. As a response, this article delves into the intersection of ontology and natural language processing (NLP) technology within the realm of annotation. It proposes an ontoNLP mechanism for automated annotation based on pre-defined ontology, aiming to facilitate the transformation of textual data into structured datasets. And, we developed an online system WuYuDian (吾与点) for automatic tagging ancient Chinese texts which is now openly accessed on Web.

1. Introduction

Ontology, a cornerstone in the field of knowledge representation, serves as a structured framework for capturing and organizing concepts and relationships within a specific domain. At its core, ontology provides a shared vocabulary and a set of defined rules that enhance communication and understanding of concepts among humans, and crucially, between humans and machines. Natural Language Processing (NLP) stands at the crossroads of linguistics and artificial intelligence, with the aim of bridging the gap between human language and computer comprehension.

Annotation is the core task of transforming textual data into structured data. OntoNLP is a mechanism designed to convert textual data into structured data. Its fundamental idea is to harness the strengths of both ontology and NLP technology through annotation, facilitating to process large-scale text data and organize knowledge in a unified annotation scheme. Based on ontoNLP mechanism, we developed an online automatic annotation tool —— 吾与点 (Wu-Yu-Dian). WYD is designed to facilitate the training data generation for machine learning, and the data extraction from textual materials for data-driven digital humanities works.

2. ontoNLP mechanism

ontoNLP mechanism is guided by the research questions, which is different from other annotation tools (Figure 1). In this mechanism, users first identify the research question, then collect textual data and define the domain ontology based on the research question. The concept layer in the domain ontology guides named entity annotation, while relationships guide relationship

[1] The corresponding author is Tong WEI, weitong1989@pku.edu.cn. This research is supported by the NSFC project "the Construction of the Knowledge Graph for the History of Chinese Confucianism" (Grant No. 72010107003).

annotation. After the annotation is completed, manual proofreading and modification are required, followed by the export of structured data. In this mechanism, ontology, as a formal and clear representation of knowledge and a conceptual framework for understanding the relationships between entities, is the cornerstone of automatic annotation. It provides a structured vocabulary and a set of rules to guide the extraction of information from unstructured text. NLP, when combined with ontology, helps extract contextual information from text and achieve automatic annotation of large-scale texts.

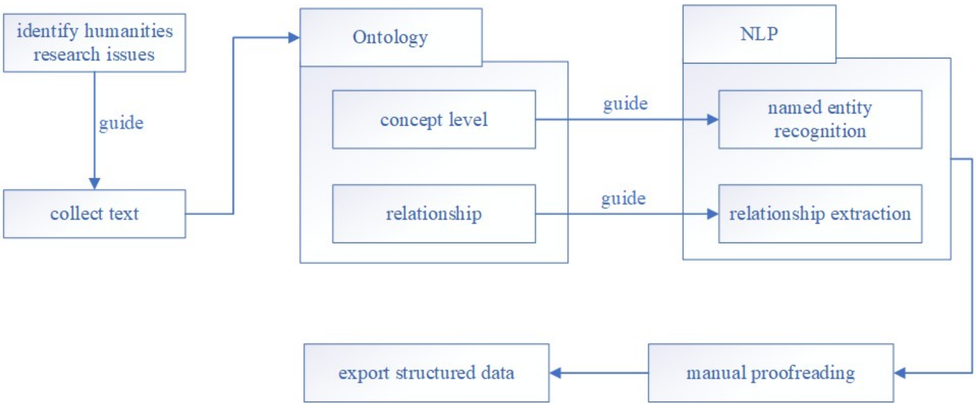
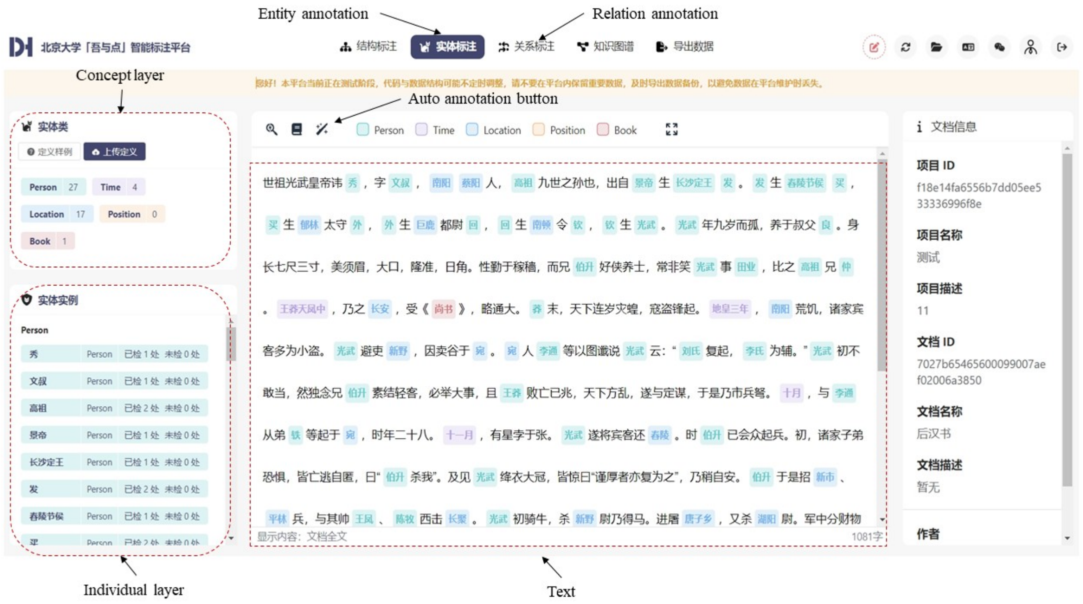


Figure 1. ontoNLP mechanism.

3. WuYuDian: An online annotation tool based on ontoNLP

Based on the ontoNLP mechanism, we developed WuYuDian that is an online automatic annotation tool (Figure 2) capable of annotating entities and extracting relationships. It is accessible at: https://www.wyd.pkudh.org. This tool can currently only handle Chinese ancient texts, and we will expand it into a multilingual processing tool in the future. The core functions of this tool include entity annotation (Figure 2) and relationship annotation (Figure 3). Entity annotation comprises concept layer, individual layer, and text layer. Users first define a conceptual model and upload it at the concept layer. Subsequently, they can initiate automatic annotation by clicking the auto-annotation button based on their conceptual model. Upon completing automatic annotation, users can manually proofread and modify in the text area. The automatic annotation function is facilitated through named entity recognition, achieved by transfer learning and incremental training on the BERT model. Currently, the entities that can be automatically annotated encompass five types: time, location, person, official, and book title, with an accuracy rate of approximately 92% in ancient Chinese texts. Furthermore, the platform offers users manual annotation and proofreading functions, with annotated entity results summarized and displayed at the individual level.



您好！本平台当前正在测试阶段，代码与数据结构可能不定时调整，请不要在平台内保留重要数据，及时导出数据备份，以避免数据在平台维护时丢失。

实体类

定义样例 上传定义

Person	27
Time	4
Location	17
Position	0
Book	1

实体实例

Person

秀	Person	已检 1 处 未检 0 处
文叔	Person	已检 1 处 未检 0 处
高祖	Person	已检 2 处 未检 0 处
景帝	Person	已检 1 处 未检 0 处
长沙定王	Person	已检 1 处 未检 0 处
发	Person	已检 2 处 未检 0 处
舂陵节侯	Person	已检 1 处 未检 0 处
买	Person	已检 2 处 未检 0 处

世祖光武皇帝讳秀，字文叔，南阳蔡阳人，高祖九世之孙也，出自景帝生长沙定王发。发生舂陵节侯买，买生郁林太守外，外生巨鹿都尉回，回生南顿令钦，钦生光武。光武年九岁而孤，养于叔父良。身长七尺三寸，美须眉，大口，隆准，日角。性勤于稼穑，而兄伯升好侠养士，常非笑光武事田业，比之高祖兄仲。王莽天凤中，乃之长安，受《尚书》，略通大。莽末，天下连岁灾蝗，寇盗锋起。地皇三年，南阳荒饥，诸家宾客多为小盗。光武避吏新野，因卖谷于宛。宛人李通等以图谶说光武云："刘氏复起，李氏为辅。"光武初不敢当，然独念兄伯升素结轻客，必举大事，且王莽败亡已兆，天下方乱，遂与定谋，于是乃市兵弩。十月，与李通从弟轶等起于宛，时年二十八。十一月，有星孛于张。光武遂将宾客还舂陵。时伯升已会众起兵。初，诸家子弟恐惧，皆亡逃自匿，曰"伯升杀我"。及见光武绛衣大冠，皆惊曰"谨厚者亦复为之"，乃稍自安。伯升于是招新市、平林兵，与其帅王凤、陈牧西击长聚。光武初骑牛，杀新野尉乃得马。进屠唐子乡，又杀湖阳尉。军中分财物

显示内容：文档全文 1081字

文档信息

项目 ID
f18e14fa6556b7dd05ee533336996f8e

项目名称
测试

项目描述
11

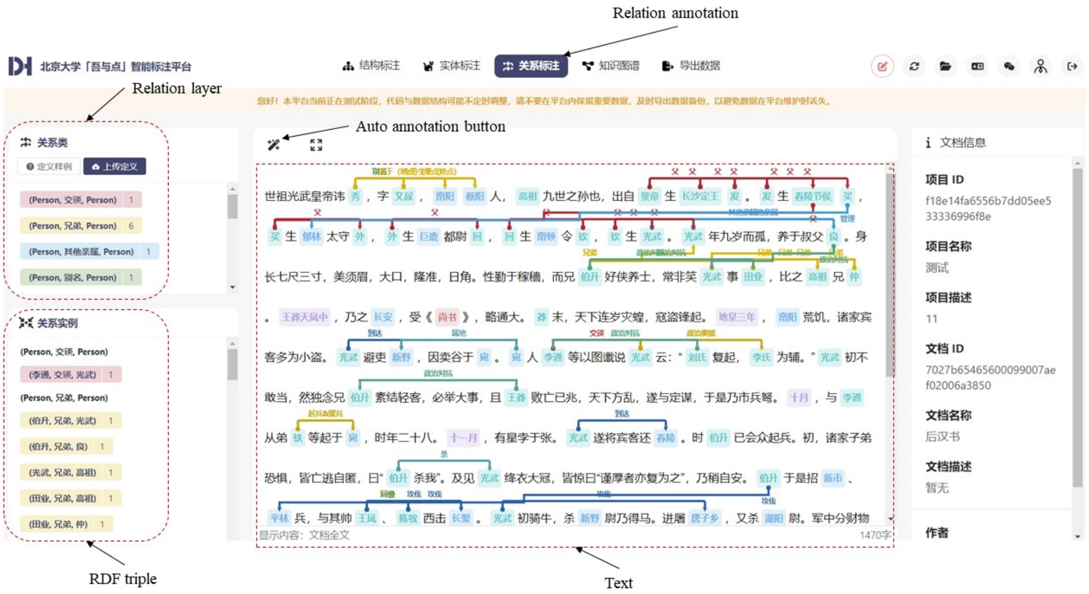
文档 ID
7027b65465600099007aef02006a3850

文档名称
后汉书

文档描述
暂无

作者

Figure 2. WuYuDian: an online tool of automatic annotation.



您好！本平台当前正在测试阶段，代码与数据结构可能不定时调整，请不要在平台内保留重要数据，及时导出数据备份，以避免数据在平台维护时丢失。

关系类

定义样例 上传定义

(Person, 交谈, Person)	1
(Person, 兄弟, Person)	6
(Person, 其他亲属, Person)	1
(Person, 别名, Person)	1

关系实例

(Person, 交谈, Person)

(李通, 交谈, 光武)	1

(Person, 兄弟, Person)

(伯升, 兄弟, 光武)	1
(伯升, 兄弟, 良)	1
(光武, 兄弟, 高祖)	1
(田业, 兄弟, 高祖)	1
(田业, 兄弟, 仲)	1

世祖光武皇帝讳秀，字文叔，南阳蔡阳人，高祖九世之孙也，出自景帝生长沙定王发。发生舂陵节侯买，买生郁林太守外，外生巨鹿都尉回，回生南顿令钦，钦生光武。光武年九岁而孤，养于叔父良。身长七尺三寸，美须眉，大口，隆准，日角。性勤于稼穑，而兄伯升好侠养士，常非笑光武事田业，比之高祖兄仲。王莽天凤中，乃之长安，受《尚书》，略通大。莽末，天下连岁灾蝗，寇盗锋起。地皇三年，南阳荒饥，诸家宾客多为小盗。光武避吏新野，因卖谷于宛。宛人李通等以图谶说光武云："刘氏复起，李氏为辅。"光武初不敢当，然独念兄伯升素结轻客，必举大事，且王莽败亡已兆，天下方乱，遂与定谋，于是乃市兵弩。十月，与李通从弟轶等起于宛，时年二十八。十一月，有星孛于张。光武遂将宾客还舂陵。时伯升已会众起兵。初，诸家子弟恐惧，皆亡逃自匿，曰"伯升杀我"。及见光武绛衣大冠，皆惊曰"谨厚者亦复为之"，乃稍自安。伯升于是招新市、平林兵，与其帅王凤、陈牧西击长聚。光武初骑牛，杀新野尉乃得马。进屠唐子乡，又杀湖阳尉。军中分财物

显示内容：文档全文 1470字

文档信息

项目 ID
f18e14fa6556b7dd05ee533336996f8e

项目名称
测试

项目描述
11

文档 ID
7027b65465600099007aef02006a3850

文档名称
后汉书

文档描述
暂无

作者

Figure 3. The relationship annotation

Relationship annotation (Figure 3) also comprises three parts: the relationship layer, RDF triplet, and text layer. Users are required to upload the defined relationships in the relationship layer and then initiate the automatic annotation by clicking the auto-annotation button, which automatically annotates the relationships in the text layer. Relationship annotation is based on the entities identified during the entity annotation process. Once users finish the automatic annotation, they can manually proofread and modify in the text area. The relationship annotation currently supports nearly 30 types of relationships, including aliases, appointments, political alliances, conquests, management, familial relationships, etc. The model achieves an 80% accuracy in extracting relationships from historical texts.

4. Conclusion and Future work

This article primarily introduces the ontoNLP mechanism based on ontology and NLP technology for converting textual data into structured data. It then presents an online data annotation tool developed using the ontoNLP mechanism, focusing on entity annotation and relationship annotation functions. Annotation is a critical step in converting textual data into structured data. The WuYuDian online annotation tool developed in this article aims to assist humanities scholars in constructing datasets and conducting data-driven research. However, a limitation of this approach is its reliance on the NLP model, resulting in the inability to automatically recognize new entity types. To address this, our future work will involve training multi-domain and cross-language Bert models to enhance service provision or explore alternatives with large language models.