Guideline for Entity Annotation

Yanqi Ren, Chengzhi Zhang

Nanjing University of Science and Technology, Nanjing 210094, China

The following section will sequentially introduce the tasks involved in annotating method entity sentences, including the design of the annotation process, the use of annotation tools, and specific annotation examples. This part aims to demonstrate examples and the process of annotating method entity sentences.

A. Method Entity Annotation Task Description

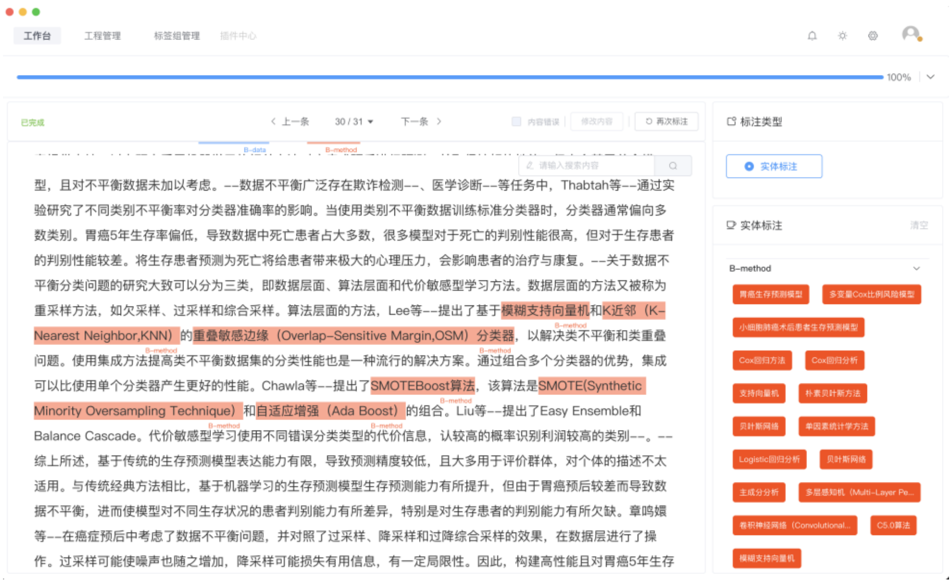
The current data annotation task focuses on labeling five types of method entities—method entities, data entities, index entities, tool entities, and theory entities—within the full-text content of academic papers in the field of "Information Resource Management" The original corpus of papers used in the study was sourced from the China National Knowledge Infrastructure (CNKI, https://www.cnki.net/), covering a time span from 2000 to 2022 and encompassing 21 journals. These journals include *Journal of Academic Libraries*, *Data Analysis and Knowledge Discovery*, *Archives Science Bulletin*, *Library Development*, *Archives Science Study*, *Library Tribune*, *Journal of the National Library of China*, *Researches on Library Science*, *Information Science*, *Library Journal*, *Information studies: Theory& Application*, *Library and Information Service*, *Journal of the China Society for Scientific and Technical Information*, *Document Information & Knowledge*, *Journal of Intelligence*, *Library and Information*, *Information and Documentation Services*, *Journal of Modern Information*, *Journal of Library Science in China*, and *Journal of Information Resources Management*, resulting in a total retrieval of 100,021 articles.

In the subsequent phase of topic modeling, abstract data corresponding to these articles were also obtained from the China National Knowledge Infrastructure (CNKI, https://www.cnki.net/). Subsequently, after excluding articles lacking a publication date, references, abstracts, full-text content, as well as those classified as cover articles, the effective dataset for this study comprised 59,084 articles. The study involved two rounds of sampling article data to construct the training set for the automatic extraction of method entities. The sampling method is outlined in the following Table a.

Table a: Method Entity Sampling Method

Sampling Count	Sampling Method
First round of sampling	The number of articles to be sampled from each journal for the first round of sampling was determined based on the smallest proportion of articles per journal, resulting in 83 articles.
Second round of sampling	The second round of sampling involved categorizing the already selected articles by year. Based on the standards of the first round of sampling, an additional 166 articles were added to ensure a balanced number of articles for each decade within the data collection timeframe.

A total of 249 sampled articles were obtained as the training set. To achieve the annotation objectives, the open-source annotation software Mark Studio was utilized to manually annotate method entities found in these 249 articles. The annotation interface of Mark Studio is shown in Figure a.



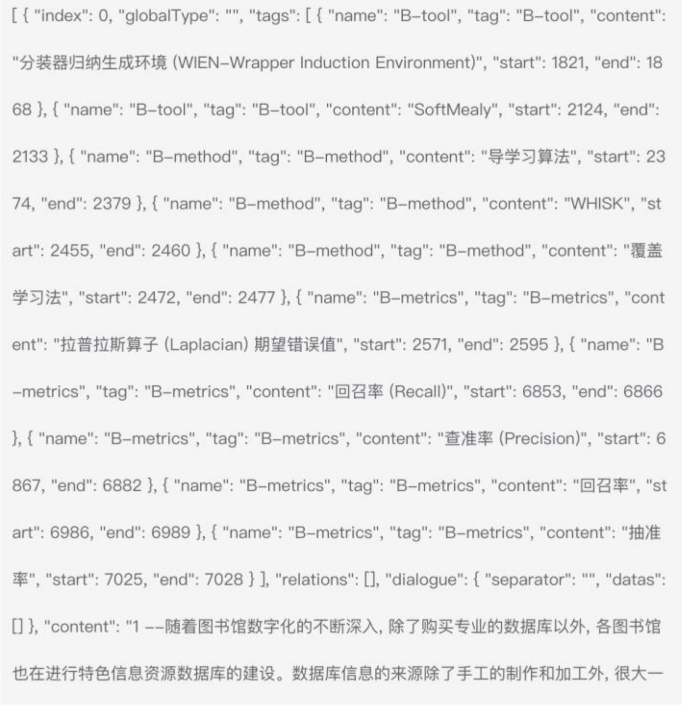
型，且对不平衡数据未加以考虑。––数据不平衡广泛存在欺诈检测––、医学诊断––等任务中，Thabtah等––通过实验研究了不同类别不平衡率对分类器准确率的影响。当使用类别不平衡数据训练标准分类器时，分类器通常偏向多数类别。胃癌5年生存率偏低，导致数据中死亡患者占大多数，很多模型对于死亡的判别性能很高，但对于生存患者的判别性能较差。将生存患者预测为死亡将给患者带来极大的心理压力，会影响患者的治疗与康复。––关于数据不平衡分类问题的研究大致可以分为三类，即数据层面、算法层面和代价敏感型学习方法。数据层面的方法又被称为重采样方法，如欠采样、过采样和综合采样。算法层面的方法，Lee等––提出了基于模糊支持向量机和K近邻（K–Nearest Neighbor,KNN）的重叠敏感边缘（Overlap–Sensitive Margin,OSM）分类器，以解决类不平衡和类重叠问题。使用集成方法提高类不平衡数据集的分类性能也是一种流行的解决方案。通过组合多个分类器的优势，集成可以比使用单个分类器产生更好的性能。Chawla等––提出了SMOTEBoost算法，该算法是SMOTE(Synthetic Minority Oversampling Technique）和自适应增强（Ada Boost）的组合。Liu等––提出了Easy Ensemble和Balance Cascade。代价敏感型学习使用不同错误分类类型的代价信息，认较高的概率识别利润较高的类别––。––综上所述，基于传统的生存预测模型表达能力有限，导致预测精度较低，且大多用于评价群体，对个体的描述不太适用。与传统经典方法相比，基于机器学习的生存预测模型生存预测能力有所提升，但由于胃癌预后较差而导致数据不平衡，进而使模型对不同生存状况的患者判别能力有所差异，特别是对生存患者的判别能力有所欠缺。章鸣嬛等––在癌症预后中考虑了数据不平衡问题，并对照了过采样、降采样和过降综合采样的效果，在数据层进行了操作。过采样可能使噪声也随之增加，降采样可能损失有用信息，有一定局限性。因此，构建高性能且对胃癌5年生存

Figure a: Annotation tool Mark Studio's annotation interface example

After completing the annotation, export the annotation data and modify the format according to the training requirements.

B. Description of Exported Data

Field Descriptions: index: data index (starting from 0). content: original text (first 30 characters, can be hidden in export settings). tags: array of annotations. tags.name: annotation name. tags.tag: annotation label. tags.content: annotation content. tags.start: start of annotation (starting from 0). tags.end: end of annotation (tags.end - tags.start = length of annotation). As shown in Figure b.



```
[ { "index": 0, "globalType": "", "tags": [ { "name": "B–tool", "tag": "B–tool", "content":
"分装器归纳生成环境 (WIEN–Wrapper Induction Environment)", "start": 1821, "end": 18
68 }, { "name": "B–tool", "tag": "B–tool", "content": "SoftMealy", "start": 2124, "end":
2133 }, { "name": "B–method", "tag": "B–method", "content": "导学习算法", "start": 23
74, "end": 2379 }, { "name": "B–method", "tag": "B–method", "content": "WHISK", "st
art": 2455, "end": 2460 }, { "name": "B–method", "tag": "B–method", "content": "覆盖
学习法", "start": 2472, "end": 2477 }, { "name": "B–metrics", "tag": "B–metrics", "cont
ent": "拉普拉斯算子 (Laplacian) 期望错误值", "start": 2571, "end": 2595 }, { "name": "B
–metrics", "tag": "B–metrics", "content": "回召率 (Recall)", "start": 6853, "end": 6866
}, { "name": "B–metrics", "tag": "B–metrics", "content": "查准率 (Precision)", "start": 6
867, "end": 6882 }, { "name": "B–metrics", "tag": "B–metrics", "content": "回召率", "st
art": 6986, "end": 6989 }, { "name": "B–metrics", "tag": "B–metrics", "content": "抽准
率", "start": 7025, "end": 7028 } ], "relations": [], "dialogue": { "separator": "", "datas":
[] }, "content": "1 ––随着图书馆数字化的不断深入, 除了购买专业的数据库以外, 各图书馆
也在进行特色信息资源数据库的建设。数据库信息的来源除了手工的制作和加工外, 很大一
```

Figure b: Example of exported data

C. Method Entity Annotation Stage Division

To minimize subjective bias in manual annotation as much as possible and to enhance consistency, the annotation process of the randomly selected 249 articles was divided into three phases.

（1） The first phase was **the pre-annotation phase**, during which a subset of articles was randomly selected and independently annotated by two individuals according to the initial annotation guidelines. After annotation was completed, the two individuals compared the annotations article by article, discussed the differences in annotation sections, and then revised the annotation guidelines accordingly.

（2） The second phase was **the consistency calculation annotation phase**. Excluding the papers from the pre-annotation phase, two sets of paper data were randomly selected from the remaining papers and were independently annotated by three additional annotators according to the revised guidelines, ensuring that each set of paper data was annotated by two annotators. After the annotation was completed, the consistency of each set of paper data was calculated, resulting in values of 0.69 and 0.73, which met the consistency requirements.

（3） The third phase was the formal annotation phase. After the consistency calculation was completed, one of the annotators from the pre-annotation phase reviewed and verified the annotation results from the consistency calculation phase to establish a unified outcome, and this annotator completed the annotation of the remaining papers.

C.1 Detailed Description of Method Entity Annotation

The main content of this section is to provide clear definitions of five fine-grained research method entities, along with detailed explanations of their related annotation guidelines and specific examples.

C.1.1 Method Entity Type Classification

After conducting a literature review and extensive reading on the classification of research methods in the field of Information Resources Management (IRM), this study classified method entities in the IRM field into five types: “Method” “Tool” “Data” “Metrics” and “Theory”. The specific classification standards for method entities are provided in Table b.

Table b: Five Types of Fine-Grained Knowledge Entities & Definitions in the IRM Field

Type	Definition	Sample
Theory	Theoretical frameworks, laws, regulations, or academic theories, etc.	*文件运动理论(Record movement Theory)、文件生命周期理论(Theory of Records’ life Cycle)、赖普斯定律(Price Law)*
Method	Algorithms, models, and methods, etc.	*LDA、SVM、CNN、主成分分析法(Method of Principal Component Analysis)*
Data	Datasets, lexicons, dictionaries, literature, catalogs, etc.	*NTU、WordNet、Hot Net、DBpedia*
Tool	Open-source tools, software, programming languages, or platforms, etc.	*SPSS、stata、JAVA*
Metrics	Metrics, evaluation criteria, etc.	*召回率(Recall), F_1、精确率(Precision)*

The annotation was performed at the sentence level for research methods mentioned in the entire text of the paper, and if a sentence contained method entities (Method, Tool, Data, Metrics and Theory), all method entities within that sentence were annotated.

C.1.2 Method Entity Annotation Example

The main purpose of this section is to provide concrete examples of the final annotation guidelines, to aid in better understanding and application of these annotation rules. These examples will thoroughly demonstrate the annotation methods in various scenarios, ensuring consistency and accuracy in the annotation process.

- **Principle of minimum scope：** Annotation personnel are only required to record the core concepts of method, tool, data, metrics, and theory entities, without the inclusion of adjectives or other modifiers, to reduce redundant information and ensure the conciseness and accuracy of the annotations.

【Example】

Sentence	Annotated Content	Type
不难发现,在重视档案服务社会化基本理论综合研究及高校档案服务社会化研究的同时...... (It is not difficult to find that while emphasizing the comprehensive theoretical research on the socialization of archival services and the research on the socialization of archival services in universities……)	档案服务社会化(The socialization of archival services)、高校档案服务社会化(The socialization of archival services in universities)	Theory

- **Principle of minimum scope (continued):** If there is a comparison between a method and its renamed version after improvement, it is necessary to determine whether to retain modifiers based on the contextual situation.

【Example a】

Sentence	Annotated Content	Type
论文以 h 指数为例,探索如何开展基于动态引用数据的 h 指数趋势分析方法和步骤。 (The article takes the h-index as an example to explore the methods and steps for conducting trend analysis of the h-index based on dynamic citation data.)	h 指数(h-index)、基于动态引用数据的 h 指数(The h-index based on dynamic citation data)	Metrics

For the metric research method entity “h 指数（h-index）”,if“基于动态引用数据的（Based on dynamic citation data）” is removed from “基于动态引用数据的 h 指数（The h-index based on dynamic citation data）”, it becomes indistinguishable from the previous entity “h 指数（h-index）” .However, based on the contextual information, this is intended to compare the improved metric research method entity with the existing metric entity, so “基于动态引用数据的(Based on dynamic citation data)” should be retained here.

【Example b】

Sentence	Annotated Content	Type
电子文件和电子文件管理呼唤有科学的理论作指导,而传统的文件(或档案)生命周期理论都难扛此重任;历史的重任便落到广义文件观指导下的新文件生命周期理论的出台上。 (Electronic records and their management call for scientific theories for guidance, yet traditional records (or archival) lifecycle theories are hard-pressed to fulfill this role. The historical responsibility thus falls on the new records lifecycle theory under the guidance of the broad view of records.)	传统的文件(或档案)生命周期理论(Traditional records (or archival) lifecycle theories), 新文件生命周期理论(、New records lifecycle theory)	Theory

Here, a comparison is conducted between the two entities, old and new, regarding the “生命周期理论(Lifecycle theory)”. If the modifiers “传统的文件（或档案）(Traditional records (or archival))” and “新文件(New records)” are not added, the two entities become indistinguishable. However, based on the context, it is clear that the intention is to compare the two entities, old and new, so the modifiers should be retained.

- **The Collective Entity:** The general terms for a category of method entities do not require annotation; only instances with specific names and defined meanings should be annotated. General terms such as “经典算法(Classical algorithm)” “机器学习(Machine learning)” and “深度学习(Deep learning)” do not require annotation; only specific algorithmic entities such as “CNN,” “RNN,” “SVM,” and “LSTM” should be annotated.

【Example a】

Sentence	Annotated Content	Type
本文重点采用定量研究方法,借助stata12.0统计软件分析2013年的流动人口动态监测数据。 (This article focuses on quantitative research methods and uses Stata12.0 statistical software to analyze the dynamic monitoring data of the floating population in 2013.)	stata12.0	Tool

“定量研究方法(quantitative research methods)” is a general term for method entities and does not have a specific meaning, so it does not require annotation.

【Example b】

Sentence	Annotated Content	Type
罗斯韦尔的5代技术创新理论是创新理论中对于知识的作用论述得较为深刻的理论。--对知识经济理论的形成与发展作出过贡献的其他重要技术创新理论还包括以下几个方面。 (Roswell’s fifth-generation technological innovation theory is one of the innovation theories that elaborate on the role of knowledge more profoundly. --Other important technological innovation theories that have contributed to the formation and development of knowledge economy theory include the following aspects.)	5代技术创新理论 (Fifth-generation technological innovation theory)	Theory

In this case, “知识经济理论(Knowledge economy theory)” and “技术创新理论(Technological innovation theories)” are general terms for theoretical entities and are not instances with specific meanings, so they do not require annotation.

- **Abbreviation:** If a sentence contains both a full name and its abbreviation, annotate both the abbreviation and its corresponding full name together.

【Example】

Sentence	Annotated Content	Type
为了实现自适应,有两种神经网络模型可供利用,即:（MLP-Multi-Layer Perceptrons）多层感知器和（KFM-Kohonen Feature Map）科惑伦特征映射。 (To achieve adaptation, two neural network models can be utilized: Multi-Layer Perceptrons (MLP) and Kohonen Feature Map (KFM).)	(MLP-Multi-Layer Perceptrons)多层感知器(MLP-Multi-Layer Perceptrons),（KFM-Kohonen Feature Map)科惑伦特征映射(KFM-Kohonen Feature Map)	Method

- **Discrimination of entity types (especially theory):** Certain research methods, despite their wide application in the IRM field, may result in classification errors when entity types are determined solely by their names. Therefore, accurate identification of entity types necessitates a deep understanding of their intrinsic meanings and subsequent classification based on this understanding, to ensure classification accuracy and reliability.

【Example】

Sentence	Annotated Content	Type
其中最重要的经典模型就是英国经济学家哈罗德的经济增长模型和美国经济学家多马的经济增长模型,这两个模型虽是分别推演而出但却极为相似因而被人们合称为"哈罗德一多马模型"。 (Among them, the most important classic models are the economic growth model by British economist Harrod and the economic growth model by American economist Domar. Although these two models were developed separately, they are very similar and are thus collectively referred to as the "Harrod-Domar model.")	经济增长模型(Economic growth model), 哈罗德一多马模型(Harrod-Domar model)	Theory

Among these, methodology entities such as the "哈罗德一多马模型(Harrod-Domar model)" and the "经济增长模型(Economic growth model)" end with "模型(model)" , but it's essential to note that they represent economic theoretical framework models, classifying them as theoretical entities.

- **Terminal noun:** If words such as "算法(method)"、"模型(model)"、"框架(framework)"、"理论(theory)"、or "数据集(data)" are present at the end of a methodology entity, they must be tagged accordingly.

For example, the method entity "联想记忆神经网络模型(Associative Memory Neural Network Model)" may appear in a paper as either "联想记忆神经网络(Associative Memory Neural Network)" or "联想记忆神经网络模型(Associative Memory Neural Network Model)". In such cases, if the entity name is accompanied by the indicating term "模型(Model)", a full tag must be applied; if not, only "联想记忆神经网络(Associative Memory Neural Network)" should be tagged.

- **Conjunction：** If a conjunction (typically "和(and)" or "或(or)") connects nouns forming a single entity, the entire phrase must be tagged; otherwise, tag each entity individually.

【Example】

Sentence	Annotated Content	Type
Horta将AI和AAI指数称之为RCA和RCI指数来计算EU和U.S.的学科影响力。 (Horta refers to the AI and AAI indices as the RCA and RCI indices to calculate the disciplinary impact of the EU and the U.S.)	AI和AAI指数(The AI and AAI indices), RCA和RCI指数 (The RCA and RCI indices)	Metrics

• Conjunction (continued)： When method entities are connected by linking symbols such as “+”, “-”, or “=”, tag the entities individually without annotating the symbols.

【Example】

Sentence	Annotated Content	Type
硬件配置:Intel E5-2609v4 + NVIDIA TESLA P4*1 软件配置: Win10+Python3.6+Tensorfow1.5+Keras2.1+PaddlePaddle1.7 (Hardware configuration：Intel E5-2609v4 + NVIDIA TESLA P4*1 software configuration： Win10+Python3.6+Tensorfow1.5+Keras2.1+PaddlePaddle1.7)	IntelE5-2609v4、NVIDIA TESLA P4*1、 Win10、Python3.6、 Tensorfow1.5、Keras2.1、 PaddlePaddle1.7	Tool

• Quotation mark: When the method entities being annotated are enclosed in quotation marks, they should be tagged along with the quotation marks.

【Example】

Sentence	Annotated Content	Type
它是一种以“客户关系一对一理论”为基础,旨在改善企业与客户之间关系的新型管理机制，它包括一个组织机构判断、选择、争取、发展和保持客户所要实施的全部商业过程。 (It is a new management mechanism based on the “one-to-one customer relationship theory”, aimed at improving the relationship between a company and its customers. It encompasses all the business processes an organization undertakes to assess, select, acquire, develop, and retain customers.)	“客户关系一对一理论” (“one-to-one customer relationship theory”)	Theory