

面向方位词的时空逻辑语义分析^{*}

于知衡¹, 辛欣^{1,2}

¹(北京理工大学 计算机学院, 北京 100081)

²(北京理工大学 北京市海量语言信息处理与云计算应用工程技术研究中心, 北京 100081)

通信作者: 辛欣, E-mail: xxin@bit.edu.cn



摘要: 时空逻辑分析是指用逻辑符号准确表达实体间的时空关系. 传统的时空逻辑分析分为封闭域与开放域两种形式. 封闭域方法预先定义了表示时空逻辑的符号体系, 然后将自然语言转换成逻辑语言. 此类方法的优点是时空关系的表达准确, 但是由于人工定义的局限性, 所定义的体系并不能覆盖复杂的时空关系. 开放域的方法使用自然语言表示时空关系, 也就是将关键词进行抽取. 此类方法的优点是能够覆盖复杂的时空关系, 但是由于自然语言本身存在歧义性, 所表示的逻辑并不精确. 为了将自然语言表达的时空关系转化为逻辑语言, 从而更准确地表达时空信息, 针对如上问题展开研究. 考虑时空关系在语言学范畴主要通过方位词表达, 如果能将方位词的语义用逻辑符号加以定义, 那么既可以解决覆盖不足的问题, 也可以解决表达不精确的问题. 为此, 设计方位词的时空逻辑体系, 定义标注规范, 总结方位词的逻辑表达范围, 给出详细的标注准则; 基于该规范, 在人民日报和 CTB 两个数据集上手工标注样本 6190 条, 形成该任务的语料库; 最后基于该语料库, 利用大语言模型对方位词触发的时空逻辑表达式进行推理, 准确率可达到 70% 以上.

关键词: 时空逻辑; 语义分析; 句法结构; 方位词; 标注体系

中图法分类号: TP18

中文引用格式: 于知衡, 辛欣. 面向方位词的时空逻辑语义分析. 软件学报. <http://www.jos.org.cn/1000-9825/7553.htm>

英文引用格式: Yu ZH, Xin X. Localizer-oriented Spatio-temporal Logical Semantic Analysis. Ruan Jian Xue Bao/Journal of Software (in Chinese). <http://www.jos.org.cn/1000-9825/7553.htm>

Localizer-oriented Spatio-temporal Logical Semantic Analysis

YU Zhi-Heng¹, XIN Xin^{1,2}

¹(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

²(Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Application, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Spatio-temporal logical analysis refers to accurately expressing spatio-temporal relationships between entities using logical symbols. Traditional spatio-temporal logical analysis adopts two paradigms: closed-domain and open-domain approaches. Closed-domain methods predefine symbolic systems for representing spatio-temporal logic and then translate natural language into logical expressions. While ensuring accurate representation of spatio-temporal relationships, such methods face limitations in handling complex relationships due to the constraints of artificial definitions. Open-domain approaches extract keywords to represent spatio-temporal relationships using natural language itself. Although capable of covering complex relationships, these methods suffer from the semantic ambiguity inherent in natural language, resulting in imprecise logical representations. The purpose of this study is to convert natural language expressions of spatio-temporal relationships into logical language, enabling more precise representation of spatio-temporal information. To address the forementioned issues, this study considered the linguistic observation that spatio-temporal relationships in language are primarily expressed

* 基金项目: 国家自然科学基金面上项目 (62172044)

第一作者与本文相关工作完成于北京理工大学本科期间

收稿时间: 2025-03-24; 修改时间: 2025-08-25; 采用时间: 2025-09-16; jos 在线出版时间: 2026-02-11

through localizers. By defining the semantics of localizers through logical symbols, the proposed framework aims to overcome both the insufficiency of coverage and the lack of precision. Accordingly, a spatio-temporal logical framework for localizers is established, including 1) the design of annotation specifications that define the logical expression scope of localizers and provide detailed annotation guidelines; 2) manual annotation of 6190 samples from the People's Daily and CTB datasets to construct a task-specific corpus based on the proposed specifications; 3) application of large language models to perform logical reasoning on localizer-triggered spatio-temporal expressions, achieving an accuracy exceeding 70% based on corpus-driven inference.

Key words: spatio-temporal logic; semantic analysis; syntactic structure; localizer; annotation system

时空逻辑分析是针对一段自然语言文本,使用设定的逻辑符号将语言进行拆解,转化为结构化的语义信息,从而准确表达实体间的时空关系^[1-3].其作用是要把有歧义的自然语言表示为没有歧义的逻辑语言,这样能够帮助大模型解析时空逻辑信息,为其对自然语言的理解任务和逻辑推理任务提供支撑^[4-7].

现有的时空逻辑分析工作主要分为封闭域和开放域两种形式.以RCC和SpatialML为代表的封闭域时空逻辑分析方法,预先设定一系列符号语言来表示空间逻辑,然后对自然语言文本进行标注,将标注结果链接来表示时空逻辑^[4,8,9],从而使自然语言文本转化为结构化的符号语言.由符号定义空间逻辑的思路可得到精确标注结果,但无法有效覆盖空间逻辑,也就是说它是不完备的.例如SpatialML无法处理“北京和天津之间”或“山洞顶部”等情况,覆盖范围存在局限性,使得某些空间关系不能直接映射到现有的符号语言体系内.在开放域进行时空逻辑分析的研究思路,以语言学知识驱动的研究方案^[10-13]为代表,从语言出发标注空间逻辑语义,所有空间语义角色的属性值都是自然语言的文本片段.用语言本身作为标注的属性值,标注范围可以很广,泛化能力强,能够分析绝大多数存在空间逻辑的语言文本,但是其属性值难以统一,存在标注结果不精确的问题.

现有工作要么侧重封闭域形式,存在一定局限性;要么侧重于开放域形式,语义相对模糊.本文的研究动机是为时空逻辑分析提出一种新的解决思路,作为对现有研究不足之处的补充.时空逻辑在语言学上的表示主要通过方位词来触发,因此通过考察方位词语义来定义逻辑符号,就可以建立起将时空关系转换为逻辑语言的映射结构,既能够解决符号语言系统覆盖面不足的问题,也可以解决属性标签值表达不精确的问题.目前,从语言学角度(如方位词短语)考虑设计符号语言的相关研究较不充分,缺乏相应标注数据集.方位词在自然语言中是一个相对封闭的集合,多数常见的时空逻辑都通过方位词进行表达,所以将方位词触发的时空逻辑定义为符号语言,既能具备封闭域形式精确性高的优势,又能兼顾开放域形式泛化能力强的优势.为此,本文将关注方位词短语,依据其蕴含的时空逻辑语义信息,确立标注规范.在此基础上,构建相应的标注语料库,并针对时空逻辑语义的分类以及方位词分别进行语义分析,建立起方位词与时空逻辑之间的映射关系.本文的核心思想是从句法结构出发,考察方位词(localizer, LC)的含义,确定锚点(Anchor)来描绘时空逻辑关系,从而更准确地表达时空信息.图1展示了本文研究思路与现有研究方法的不同,具备两种形式的优点,同时又避免了它们现有的问题,还可以进一步分析时间逻辑和空间逻辑在概念领域的引申逻辑.

本文主要完成了如下的工作.首先,构建时空逻辑语义分析标注规范.考虑以句法结构为起点分析语义表示,以方位词为中心分析时空逻辑语义信息的方法.核心是利用球坐标系思想对语义信息作拆解,建立起一种基于层级结构,以方位词为导向的标注规范(后文统称“该规范”).该规范共包含7个标签,分别是方位词<LC>、锚点<Anchor>、时空逻辑类型<Category>、层级作用形式大类<Heirarchy>、层级作用形式小类<Subsidiary>、方位指向<Orientation>和距离<Distance>.其中时空逻辑类型分为3大类,即空间定位逻辑<Locating>、时间定序逻辑<Sequencing>和引申概念逻辑<Specifying>;层级作用形式小类共有19类,分属于相离<Disjoint>、相交<Intersect>、包含<Contain>和变化态<Mutable>这4大类;方位指向的常见取值有27种;方位词、锚点和距离则都为原文文本片段.其次,建立了时空逻辑语义分析标注语料库.进一步刻画了数据集标注类别分布,并分别以方位词和标注类别为中心作语义分析,串联起句法结构和语义表示之间的映射关系.语料库包含CTB新闻数据集^[14]和人民日报数据集^[6,7]共6190条时空逻辑关系信息,数据公布地址为<https://git-hub.com/NP-NET-research/Localizer-Spatio-Temporal-Semantic-Annotated-corpus>,可供研究者免费下载使用.最后,通过大模型完成了评测任务.本文对大语言模型进行了LoRA微调,验证了自动转换的效果,通过实验证明了该标注规范的合理性,直观说明通过句法结构分析时空逻辑语义信息的意义.

<p style="text-align: center;">现有的封闭域形式 (以SpatialML为例)</p> <pre>input:我居住在一个离中石家庄太行山药香谷以南大约9公里的镇子里。 output:<PLACE id=1 type="PPLA" country="CN" form="NAM">石家庄</PLACE> <PLACE id=2 type="MTS" country="CN" form="NAM">太行山</PLACE> <PLACE id=3 type="MTN">药香谷</PLACE> <SIGNAL id=4 以南</SIGNAL> <SIGNAL id=5 9公里</SIGNAL> <PLACE type="PPL" id=6 form="NOM" ctv="TOWN">镇子</PLACE> <PATH id=7 distance="9.km" direction=source=3 destination=6 signals="2 3"> <LINK id=8 source=1 target=6 linkType="IN"/></pre> <p style="text-align: center;">封闭域形式存在的问题</p> <ol style="list-style-type: none"> 1. 按照可能存在的语义确立属性值标签库,处理时空逻辑有局限性; 2. 抽象化处理属性值标签,碎片化标注文本关键词,通过链接来按照实际语义构建关系,不呈现中间分析过程,不易于理解; 3. 专注于针对性地处理空间逻辑或针对性地处理时间逻辑,不同时标注时间和空间,以及其引申的概念领域的逻辑关系。 	<p style="text-align: center;">现有的开放域形式 (以SpaCE2023为例)</p> <pre>input:我居住在一个离中石家庄太行山药香谷以南大约9公里的镇子里。 output:[{"role":"空间实体","text":"我"}, {"role":"事件","text":"居住"}, {"role":"处所","text":"镇子里"}, {"role":"空间实体","text":"药香谷"}, {"role":"处所","text":"石家庄太行山"}, {"role":"事件处所","text":"中"}, {"role":"空间实体","text":"镇子"}, {"role":"参照实体","text":"药香谷"}, {"role":"距离","text":"9公里"}]</pre> <p style="text-align: center;">开放域形式存在的问题</p> <ol style="list-style-type: none"> 1. 大多数属性的标注属性值都是原文文本片段,不进行精确的语义指向,逻辑计算时存在歧义; 2. 所有标注针对文本片段,需要从长片段中按语义定位标注对象; 3. 专注于针对性地处理空间逻辑或针对性地处理时间逻辑,不同时标注时间和空间,以及其引申的概念领域的逻辑关系。
<p style="text-align: center;">面向方位词进行时空逻辑语义分析例子和优势</p> <pre>input:我居住在一个离中石家庄太行山药香谷以南大约9公里的镇子里。 output:<Phrase id=1 LC="中" Anchor="["石家庄"太行山]" Category="locating"空间定位逻辑" Hierarchy="intersect相交" Subsidiary="overlap重叠" Orientation="inner[更内层]"> 西石家庄太行山</Phrase> <Phrase id=2 LC="以南" Anchor="药香谷" Category="locating"空间定位逻辑" Hierarchy="disjoin相离" Subsidiary="outside相离" Orientation="south[南]"> Distance="9公里">药香谷以南</Phrase> <Phrase id=3 LC="里" Anchor="镇子" Category="locating"空间定位逻辑" Hierarchy="contain包含" Subsidiary="in范围内">镇子里</Phrase></pre> <ol style="list-style-type: none"> 1. 属性值是规定的符号语言,语义含义精确,克服开放域模糊的缺点;方位词是一个重要的表示时空逻辑的完备集合,从方位词出发能包含绝大多数时空关系,具有完备性; 2. 从句法结构(方位词)出发,有效建立起时空逻辑语义及其语言学上表示的联系,呈现逻辑语义分析中间过程,增强可解释性; 3. 可以从方位词出发,同时考虑时间逻辑、空间逻辑以及通过方位词表达的时间和空间在概念领域的引申概念逻辑,将它们纳入同一个框架下进行逻辑语义分析。 	

图 1 本文研究思路与现有研究方法的不同

本文第 1 节简述现有的时空逻辑语义分析相关工作, 以及其存在的一些问题. 第 2 节阐明面向方位词进行时空逻辑语义分析任务时, 需要注意的一些基本概念. 第 3 节给出时空逻辑语义分析的标注规范. 第 4 节阐明标注流程并对标注数据集进行分析, 考察方位词体现时空逻辑语义信息的特点. 第 5 节进行实验介绍及结果分析. 第 6 节为本文总结.

1 时空逻辑语义分析相关工作

现有时空逻辑语义分析的研究方法主要分为封闭域形式和开放域形式, 针对某段自然语言文本, 直接标注其空间或时间语义信息. 以封闭域形式进行的研究主要以 TimeML 和 SpatialML 为代表, 聚焦于将时间或空间关系转化为预先设定好的一套符号语言, 建立符号语言的标签库. SpatialML (spatial markup language) 是一种空间标记语言, 为空间逻辑语义分析提供了标准化框架, 从文本中的空间关系标志词入手, 碎片化地进行标注, 通过链接的方式表达空间逻辑信息^[15], 以实现对复杂空间推理任务的支持. SpatialML 设定了一系列对应于原文片段的标注角色, <PLACE>用于标注空间实体, <SIGNAL>用于标注空间关系信号词, 如方位词或距离, <LINK>为已标注的空间实体赋予语义角色, 通过符号将空间实体联系起来, <PATH>负责明确复杂空间关系的发生路径. 同时 SpatialML 对所有元素的属性值都设定了封闭的符号标签库, 将标注内容与符号语言严格对应起来. 例如空间实体<PLACE>的类型可能是<国家><河流>或<山脉>等, 空间关系<LINK>则可以是<包含><附近>或<交叠>等. SpatialML 从语义角度出发将空间关系转化为符号语言, 然后将自然语言文本与这样的符号标注对应, 保证空间逻辑语义分析具有精确性和整洁性, 不易产生歧义. TimeML 也采用了相同的理念, 专门处理时间关系^[16]. 应用这样的语义分析思想的还有区域链接演算 (region connection calculus, RCC)^[17], 以区域作为空间推理的实体, 更重视定性空间关系的表达, 使用连接关系区分不同拓扑形状增强计算表达能力. 相关研究针对中文地理命名体识别也以类似 SpatialML 的方式构建了标注规范^[12], 并对事件时空信息标注进行了探讨研究, 着重考虑地理空间位置关系以及文本中事件的时间、地理空间位置等要素. 但是这一类研究在时空信息标注上侧重于事件实体^[18], 并简化了时间定序逻辑, 分开标注时间和空间逻辑. 封闭域的研究方法还包括如语义金字塔概念^[19], 以及地理空间语言标注工具 (TESLA)^[20].

封闭域形式处理时空逻辑存在不足. 一方面, 如 SpatialML 需要事先定义好尽可能完备的符号语言, 进行语义分析有时会出现理解偏差. 该研究的思路是按照可能存在的空间实体类型以及时空关系来定义属性值标签库, 比如认定潜在的空间实体是<城镇>或<国家>等, 并明确定义为哪个国家, 为每一个国家都设定相应符号. 也即 SpatialML 对空间实体<PLACE>的属性有精确限定, 但其涵盖面可能受限, 部分空间实体类型无法定义. 例如“山洞”或“地窖”就没有进行符号规定, 难以准确标注, 使得 SpatialML 的精确性优势无法充分体现. 类似地, 这一方法

对空间关系的规定也存在部分空白,如<之间>的概念没有准确的表达方式,且<SIGNAL>的符号化有所局限.解决这种局限的简单思路是穷尽所有可能存在的空间实体类型或空间逻辑语义,但这会导致标签系统过于复杂;如果仅标注相近的含义,又会使标注结果不精确.另一方面,封闭域定义的符号语言抽象.如 SpatialML 以碎片化形式标注文本中的关键词,通过链接方法来按照实际的语义构建关系,这种思路不呈现中间的联系和分析过程,削弱了逻辑语义的分析能力和可解释性.对于一段自然语言文本, SpatialML 按照经验理解从实际空间语义角度进行标注,没有考虑空间关系和语言表示的明确映射关系, TimeML 以及其他封闭域形式也存在类似特征^[16,19,20].

开放域形式处理时空逻辑的代表性工作 是语言学知识驱动的中文空间语义理解方面的研究,相关研究提出从知识到数据的研究思路,构建了标注规范及相应标注语料库^[10-13].该研究进一步设计了空间语义理解的多项任务,即中文空间语义理解评测 (spatial cognition evaluation, SpaCE),要求机器可以从文本中提取标注空间实体、方位和事件等信息,识别空间语义角色,判断不同语言表述是否描述了相同的空间场景等^[10-13].以 SpaCE2023 进行空间逻辑语义分析的标注语料库为例,它针对整条句子中的多条空间信息,以空间实体为核心分别标注语义角色.可供标注的语义角色共 15 个,如<处所><事件><部位><时间>和<路径>等,以具体属性值是原文内容的文本片段型角色为主,也存在部分限定取值的标签型角色如<事实性>等.由此给出的空间实体及空间方位信息,描述了“某空间实体在某时经某事件,满足某种空间方位关系,且这一命题的事实性”,可以囊括绝大多数空间关系,灵活性很高,所提取到的逻辑语义信息更为全面^[10,13].

开放域形式处理时空逻辑也存在不足.其一,如 SpaCE2023 着重于标注空间语义角色,标注结果的大多数属性值都是原文文本片段,像是<事件>的取值可能为<压>或<驻留>, <部位>取值可能是<眼睛>或<下巴>等抽象语言表达,含有一定艺术修饰成分,不适合直接作为具体标注值.这种现象使得开放域形式的标注结果语义指向相对模糊,不利于严谨的逻辑计算.其二,开放域方法考虑针对原文文本片段进行标注,属性取值是文本内容,不在语言学和逻辑语义之间建立起有效的联系,使标注结果可解释性弱.

总体而言,如 SpatialML 的封闭域方法通过符号语言使得逻辑语义分析结果具有精确性,但受限干预设标签库的覆盖能力,且基于语义设置标签,文本到标签的映射关系不够直接.如 SpaCE 的开放域方法相较而言更加灵活,能够提升语义表达的覆盖范围,但牺牲了逻辑计算的严谨性.同时两种方法都更多地单一考虑时间逻辑或空间逻辑,不考虑同时对时空逻辑语义进行标注.本文提出的新思路可以具备两种形式的优点,并对相关研究的空白领域进行补充.

2 面向方位词的时空逻辑语义分析任务基本概念

2.1 时空逻辑语义分析

时空逻辑可分为 3 类: 其一是空间逻辑,即空间实体定位关系; 其二是时间逻辑,即时间点段定序关系,涉及时间点和时间段之间的时间关系; 其三是时间逻辑和空间逻辑的引申概念逻辑,一般是条件或限定关系,例如<文中>或<时代背景下>.图 2 是一则时间序列逻辑语义分析的示例.

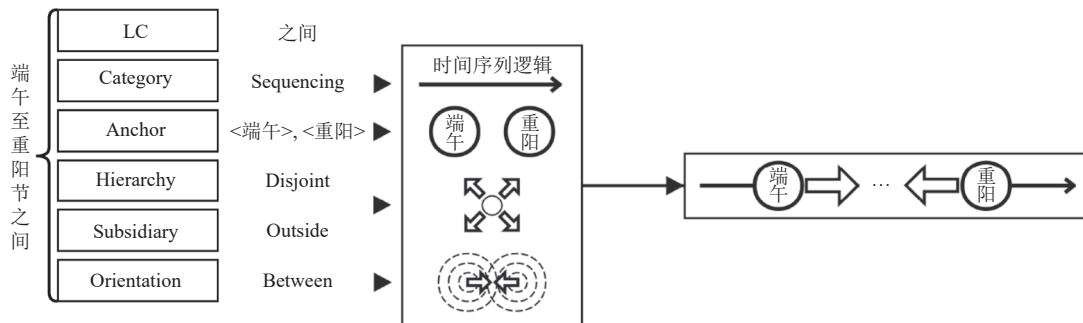


图 2 时空逻辑语义分析示例

2.2 面向方位词分析时空逻辑

时空逻辑信息主要体现在方位词短语中, 由方位词传递的时空逻辑语义具有广泛适用性、多样性和变化性. 广泛适用性说明方位词可以表达时间逻辑、空间逻辑和二者概念领域的引申逻辑, 并且能够传达的时空逻辑语义丰富, 是最基本的时空逻辑信息表达形式. 多样性即同样的时空逻辑信息可以通过不同的方位词表达, 不局限于某种特定句法形式. 变化性即某一方位词短语或句法形式, 在不同语言环境下会传递不一样的信息, 需要结合时空逻辑对象以及上下文语境理解, 这样的情况更为常见. 这些特点使方位词在时空逻辑信息表示中, 比动词传递信息准确, 比名词应用环境广泛, 比形容词存在更多二义性和模糊性而更加灵活. 需要说明的是, 本文讨论的是广义方位词范围, 包括空间方位词及时间方位词. 目的是在分析时空逻辑结构时, 能够保持分析框架的一致性. 时间方位词可视为空间方位词在时间领域的延伸, 主要用于说明事件的时间序列关系. 在《现代汉语八百词(增订本)》中有说明, “来”在表示“以来”时可归入方位词, 但属于“时间方位词”, 并非传统空间意义上的方位词. 本文的处理亦基于此类学理用法.

图3以空间逻辑为例, <Sentence>不同, 但<Phrase>都是<墙上>. 表1以3种时空逻辑类型为例, 展现了方位词的应用特征. 在本文采用的新闻数据集中, 方位词会触发全部3种时空逻辑类型, 可以全面地考虑方位词和时空逻辑的映射关系. 第3节将介绍面向方位词进行时空逻辑语义分析标注规范的设计思路及其具体内容, 通过标注规范将这类方位词短语的时空逻辑语义进行结构化处理, 呈现语言学表示和语义信息的联系.

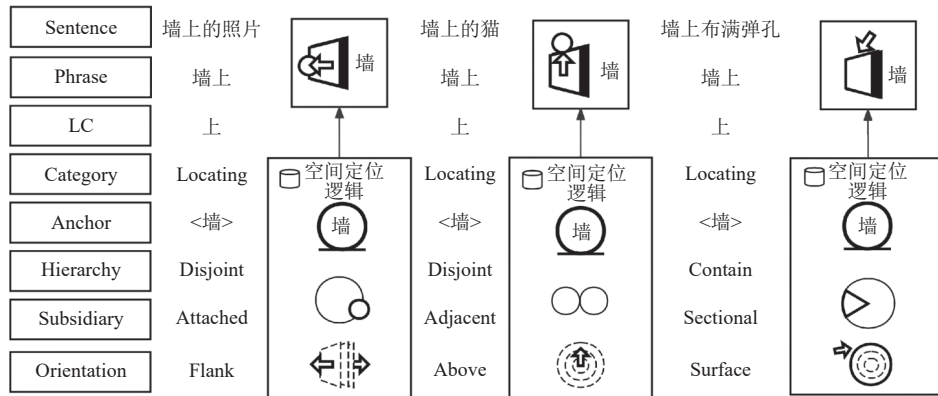


图3 方位词应用特征示例

表1 方位词短语中的时空逻辑信息示例

Example	Sentence or Phrase	Localizer	Explanation
1	车上(的货物)/车内(的货物)/车里(的货物)	上/内/里	The goods are inside the car
2	墙上(的涂鸦)/墙侧(的涂鸦)	上/侧	Graffiti painting on the wall surface
3	七月末/七月底/七月尾	末/底/尾	The last period of July
4	此前提下/基础上/此前提下/基础上	下/下/上	Based on

3 时空逻辑语义标注规范

3.1 层级导向标注规范基本架构

本文提出基于层级作用形式、以方位词为导向的标注规范, 该规范可以充分提取句式中方位词短语所蕴含的时间逻辑、空间逻辑或引申逻辑语义信息. 标注规范的基本架构见表2. 锚点<Anchor>一般是原文片段, 但在特殊语境下可取特定值如<文本时间>. 时空逻辑类型<Category>、层级作用形式大类<Hierarchy>、层级作用形式小类<Subsidiary>和方位指向<Orientation>存在确定的类别或取值集合, 相关规范见第3.3节内容. 距离<Distance>作为可附加对象是原文片段, 一般取空值, 在例子表示时会进行省略.

该标注规范是本文基于数据集和相关研究,提出的面向方位词短语的标注规范,利用了空间关系和球坐标系的思想.空间关系是时空逻辑的基础,通过3种空间方位的参考框架(外部、直接和内部参考框架)进行划分^[21,22],球坐标系考虑通过角度和距离来刻画相对位置关系.时空逻辑存在于特定的时空中,时空逻辑语义分析的首要任务就是找到锚点<Anchor>,以锚点作为原点建立起一个明确的时空范畴^[4].基于空间方位参考框架和球坐标系,将距离阶梯化定性处理为层次概念,转化为<Hierarchy>来刻画时空逻辑语义,确定时空对象相对于锚点的层级;将角度转化为方位指向<Orientation>,确定时空对象相对于锚点的方向.通过这种方法可以建立起该标注规范的基本架构,从而充分提取数据集中方位词短语所蕴含的时空逻辑语义信息.

表2 标注规范基本架构

项目	中文	解释
Phrase	短语	当前所处理的方位词短语
LC	方位词	时空逻辑对应的方位词
Category	类型	明确是时空逻辑类型
Anchor	锚点	当前时空逻辑的锚点
Hierarchy	层级作用形式大类	时空对象层级位置大类
Subsidiary	层级作用形式小类	时空对象层级位置小类
Orientation	方位指向	对象相对于锚点的方向
Distance	距离属性	附加的量化距离信息

3.2 时空逻辑语义分析形式化定义

对存在时空逻辑语义信息的句子<Sentence>及其中的方位词短语<Phrase>(如图3中的句子<墙上的照片>及其方位词短语<墙上>,表示为<[墙上]的照片>)进行分析时,按照以下流程逐步处理并生成结构化输出:第1步,结合句子信息判断方位词短语中的方位词<LC>及其时空逻辑语义类型<Category>,例如根据方位词<上>可判定其类型为空间定序逻辑<Locating>;第2步,确定该时空逻辑的锚点<Anchor>,如从方位词短语中可以提取出锚点<墙>;第3步,拆解逻辑语义,判明该时空逻辑层级作用形式类别<Hierarchy>和<Subsidiary>,如依据<上>并理解<照片>与<墙>的关系,可以解析出层级作用形式大类为<Disjoint>,小类为<Attached>;第4步,判断时空逻辑信息是否已经由层级作用形式类别充分提炼,若未充分提炼则需补充方位指向<Orientation>以及距离<Distance>,本例中<上>的时空逻辑语义未能被层级作用形式充分表达,结合上下文补充方位指向取值为<flank>,距离取值为<null>;最后,整理得到结构化输出如<“LC”:“上”,“Anchor”:“[墙]”,“Category”:“Locating”,“Hierarchy”:“Disjoint”,“Subsidiary”:“Attached”,“Orientation”:“flank”,“Distance”:null>,类似图3中所示.以下给出面向方位词的时空逻辑语义分析形式化定义.

形式化定义:给定非结构化输入:

$$x = (\text{Sentence}, \text{Phrase}) \quad (1)$$

识别时空逻辑语义,得到结构化输出:

$$y = (\text{LC}, \text{Category}, \text{Anchor}, \text{Hierarchy}, \text{Subsidiary}, \text{Orientation}, \text{Distance}) \quad (2)$$

3.3 标注规范

本节将依次给出时空逻辑类型<Category>、层级作用形式大类<Hierarchy>、层级作用形式小类<Subsidiary>和方位指向<Orientation>的标注规范,阐明其标注原则,并给出条件性语义类的处理方法.为保证标注严谨性和时空逻辑语义分析的一致性,标注规范中部分属性的类别或取值是特定集合的元素,结构化表达如公式(3).其中层级作用形式的小类的取值集合受相应大类限定,其可能的类别集合,是以相应数据条目选定的<Hierarchy>类别作为条件的集合:

$$\begin{cases}
 Y_{Category} \in M_{Category} \\
 Y_{Hierarchy} \in M_{Hierarchy} \\
 Y_{Subsidiary} | Y_{Hierarchy} \in H_{Subsidiary} | Y_{Hierarchy} \\
 Y_{Orientation} \in M_{Orientation}
 \end{cases}
 \quad (3)$$

时空逻辑类型<Category>明确语义逻辑的基本模式, 是提取方位词短语语义信息的第 1 步. 时空逻辑类型划分为以下 3 类: 空间位置对象的定位逻辑<Locating>、时间事件对象的定序逻辑<Sequencing>以及时空逻辑在概念领域的推广, 引申含义的限定逻辑<Specifying>. 空间位置对象定位逻辑反映三维空间内空间实体的位置关系, 例如“房屋后”或“城南”; 时间事件对象定序逻辑反映一维时间轴上事件的次序关系, 例如“三年前”或“春节以来”; 引申含义限定逻辑是空间和时间逻辑抽象化的延伸, 以辅助理解概念关系, 例如“法律上”或“文章之外”. 3 种时空逻辑类型满足边界清晰可分的特点, 对具体语句及相应方位词短语进行分析时, 不会产生分歧. 标注规范中层级作用的关系规范如图 4 所示.

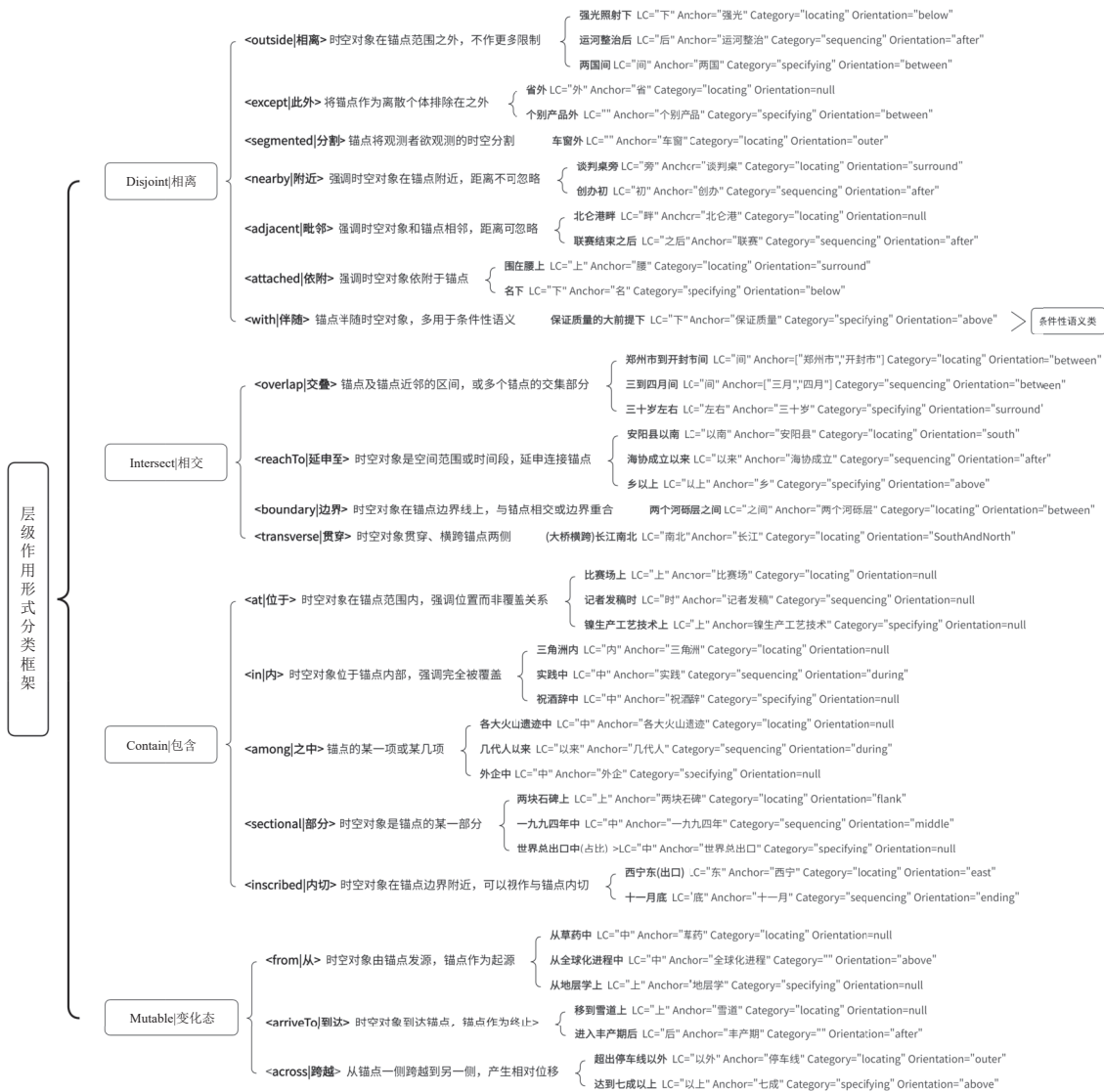


图 4 层级作用形式分类框架

层级作用形式作为该规范的核心部分,高度提炼了方位词短语所蕴含的时空逻辑语义信息,是将非结构化的时空逻辑语义转化为结构化计算机语言的关键步骤.层级作用形式分为两层,第1层是层级大类<Hierarchy>,第2层是细分小类<Subsidiary>,以树结构建立.层级小类受到层级大类类别的限制,某项时空逻辑属于特定层级大类类别时,细分小类只能是该大类下的类别.

层级作用形式大类共有4类,分别是相离<Disjoint>、相交<Intersect>、包含<Contain>和变化态<Mutable>.相离意味着时空对象和锚点“几乎是不重合的”,锚点作为时空对象的原点或基本参照物.此处相离与拓扑关系中相离不同,例如细分小类相邻<Adjacent>和依附<Attached>也会被纳入该大类中.相交强调锚点和时空对象相互联系或重叠,锚点作为时空对象的限定边界如<ReachTo>和<Boundary>,或中心点如<Overlap>和<Transverse>来确定时空对象.包含强调锚点和时空对象是隶属或嵌套关系,其中细分小类区分不同的包含状态或模式.例如<at>即锚点作为时空对象的位置,<Among>即锚点作为时空对象所隶属的集合,<Inscribed>即时空对象与锚点是内切的.变化态是对前3个类别的补充,其设置是基于对上下文语境可能附加提供逻辑语义的考虑,强调时空对象处于相对于锚点参照物的移动状态,比如正在到达<ArriveTo>或跨越<ArriveTo>锚点.另外,很多层级作用形式类别不对3种类型都适用,如一维线性时间定序逻辑中不存在横贯<Transverse>关系,分割<Segmented>关系也不会和时间定序逻辑中通过方位词表达.

方位指向<Orientation>是对层级作用形式的补充,在后者无法充分提炼时空逻辑语义时,以附加说明丰富结构化语言的语义信息.方位指向取值规范如图5所示分为4类,对应空间逻辑取值的空间方位、对应时间逻辑取值的时间方位、空间和时间逻辑都可以取值的时空方位.引申逻辑关系在不同情境下的方位指向取值规范没有限定,仅有条件性语义类需要作特殊说明.



图5 方位指向取值规范

引申逻辑关系中的条件性语义,是一种特殊的空间隐喻方法,对这种情况的方位指向取值的规范需要特别注意.因为自然语言中,常利用具象方位(如<上><下>)辅助表达抽象概念(如<基于某些原则>),通过具体的空间隐

喻,清晰传达复杂的含义. 对这类条件性语义的分析,不应舍弃其方位词原含义,而是以空间逻辑进行类比. 图6展示了该类方位指向的取值原则. 如<前提下>和<基础上>表明起到基石性质作用的条件,也即锚点是对对象概念对象的根基或基础,以<above>进行语义拆解;而如<支持下>作为附加的背景环境,在语义上不表明起到决定性的支撑作用的概念,以<below>进行语义拆解;对于<情况下>表明持续状态中的条件,通过<during>进行语义拆解. 条件性语义类是引申逻辑关系中的一大模块,方位词短语中的锚点形式多样且含义复杂,对条件性语义类的标注质量的控制是标注过程中的一项重要任务.

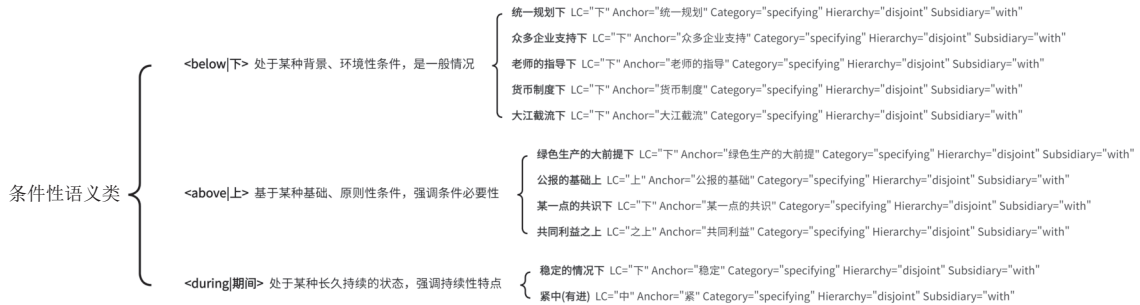


图6 条件性语义类标注规范

4 数据标注与分析

4.1 数据标注方法

本研究采用单人标注、多人校验的方式,分多轮迭代完成数据标注. 在第1轮标注中,基于初始版本的标注规范,由一名标注人员对小规模语料样本进行试标注. 完成后,与另一名抽检人员对标注结果进行抽检与讨论,重点关注属性是否冗余、取值是否明确、标注流程是否具备可操作性等问题^[6,7]. 根据讨论结果,删除了多余的属性和取值,优化了标注规范的结构与表述.

第2轮标注改为按语料文本顺序对整个数据集进行标注. 该轮标注过程中,阶段性抽检发现标注一致性较差、错误率较高. 分析讨论认为,单人按语料顺序标注易受上下文干扰和主观判断影响,导致同一语义产生不同标注结果. 此外,这种标注方法在错误校对时难以批量修正,检验矫正的成本很高.

针对以上问题,第3轮调整标注策略,以方位词为基本单元组织标注,具体步骤如下:首先,针对某一个方位词所有语料进行语义初筛,依据标注规范确定其类别及属性取值,形成该方位词的候选标注集合;然后系统标注该方位词下的全部语料,从已建立集合中选择合适的标注项. 标注过程中如果发现语义相近、可合并的项,则回溯调整来尽量保证一致性;如果遇到未覆盖的语义,则及时扩充候选集. 对语义相同的不同方位词,如先前已标注过<年末>,现在要标注<年底>,则复用先前已有标注项,确保跨词标注的一致性. 全部标注完成后,由标注人员进行人工抽检,发现错误时,集中修正当前方位词以及相关复用项的标注.

语义边界的确立是标注过程中的主要难点. 针对两种非常相近的语义,需决定是区分不同标注项(即视为不同语义),还是合并采用同一标注项(即视为相同语义). 这种取舍会影响标注质量. 本文遵循以下原则:若在日常语言使用中不易引起歧义,则合并这两种语义;若易导致歧义,则进行区分,这样可以确保标注结果既精确又具备足够数目的样本支持.

4.2 标注数据描述统计

分别从CTB5数据集^[14]中随机抽取了1600条数据,从人民日报1998-01-2003版数据集^[6,7]中抽取了4590条数据,共6190条数据进行标注,数据公布地址为<https://github.com/NP-NET-research/Localizer-Spatio-Temporal-Semantic-Annotated-corpus>. 表3展示了标注数据集的基本情况. 表4列出了人民日报数据集中的高频次方位词,对这些方位词,按比例随机抽取标注,其中的百分数表示时空逻辑类型的样本在数据集中的占比. 对于其余方位

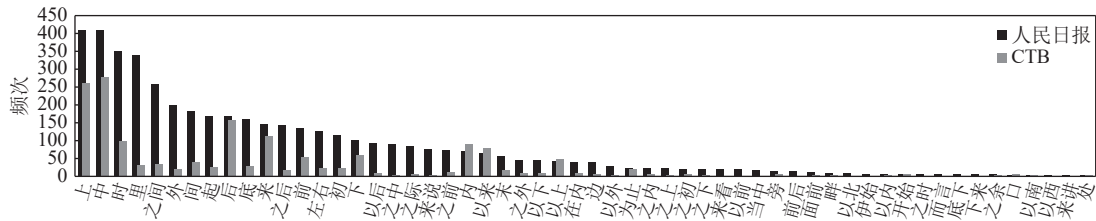
词,全部进行标注.方位词频次、层级作用形式分类和方位指向取值分布见图7.

表3 标注数据集基本情况

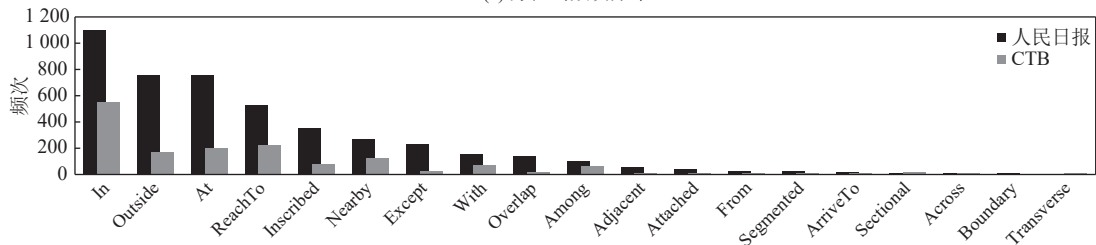
数据集	Number	Locating (空间定位逻辑)	Sequencing (时间定序逻辑)	Specifying (引申限定逻辑)
人民日报	4590	761 (16.6%)	2300 (50.1%)	1527 (33.3%)
CTB	1600	172 (10.8%)	917 (57.3%)	511 (31.9%)

表4 人民日报初始数据集高频方位词

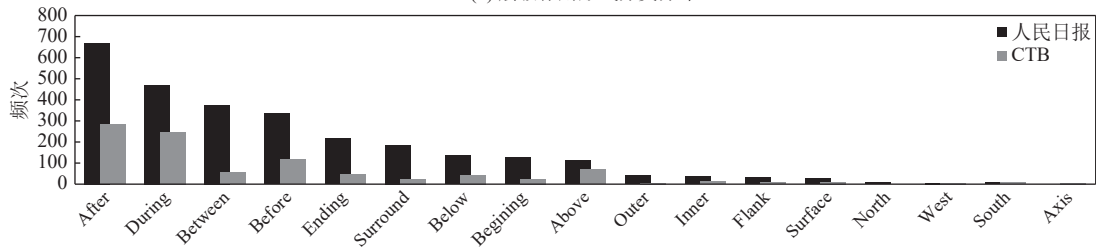
Localizer	中	上	后	时	里	下	来	内	前	以来	以上
Frequency	3130	3032	1332	1062	902	821	652	526	512	451	363



(a) 方位词频次分布



(b) 层级作用形式分类分布



(c) 方位指向取值分布

图7 数据集中方位词频次、层级分类和方位指向取值的分布

4.3 按照类别频次作语义分析

按层级作用形式类别进行语义分析,有助于归纳可以表达同一层级作用形式语义(简称层级语义)的方位词集合,理解标注规范架构语义拆解中,层级语义和方位指向语义(简称方向语义)所承担的不同职能,以及两种语义属性的联系.对数据集中出现频次最高的层级类别进行语义分析,如图8所示,例子结构如下:

$$\langle \text{Hierarchy-Subsidiary} \rangle - \langle \text{LC} \rangle - \langle \text{Orientation} \rangle - \text{Phrase} \tag{4}$$

具有分散性特征的时空逻辑语义可通过多种方位词表达,且这些方位词在语料中出现的频率彼此接近,均处于相对较高的水平.这使其表达形式丰富,方位词与逻辑语义的映射关系更复杂.以<in>为例,其直觉上主要由<里><内><中>体现,但<上><间><底><来>等方位词也以不可忽视的频率具有<in>的语义,使得<in>并不倾向于仅通过某几个词表达,而是分散于许多方位词的含义中.<Adjacent>也具有相同特征,其语义表达方式灵活性也很高,除高频词<上><后>等外,还可通过<边><旁><畔>等表达.在实验中,模型容易混淆具有分散性特征的时空逻辑语

义,需严格把控这类语义的数据标注质量.

具有集中性特征的时空逻辑语义主要通过某几个方位词表达,可能存在更多方位词体现该语义,但占比不高.如<at>的语义集中由<上><时>等表达,其他方位词体现<at>的频次很低,也即<at>语义总是倾向于由某几个方位词体现,而不像<ReachTo>等有分散性特征的语义相对均匀地由许多方位词表达.类似具有集中性特征的时空逻辑语义还包括<Among><Inscribed>等词,统一考虑可规范标注质量,增强模型对时空逻辑语义信息的分析理解能力.

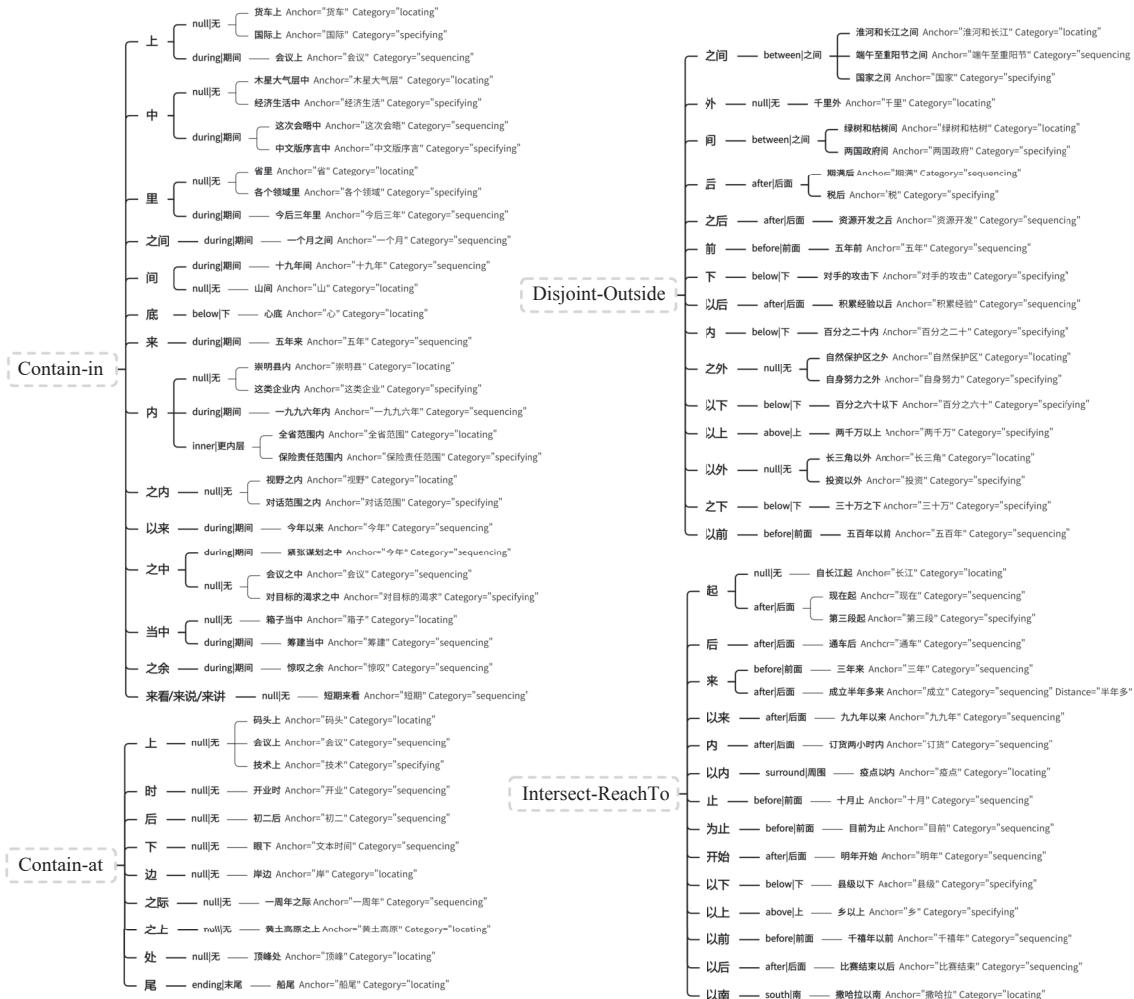


图 8 频次最高的 4 个类别语义分析

诠释完备的层级语义的信息丰富,能够充分体现方位词的时空逻辑语义,此时相应的方向语义为空值或映射元素简单.例如<in>仅有几个方向语义,且以空值和“during|期间”居多;<at>的方向语义则基本都是空值.诠释完备的层级语义本身已经表达清楚相应的时空逻辑语义,只需通过方位指向做有限度的补充说明.另一方面,这样的层级语义已经提取了方位词的绝大多数部分信息,实际上也方位指向施加了更严格的限定,使语义分析有迹可循.类似情况还包括<from><except>和<nearby>等.

诠释不完备的层级语义信息简单,相应的方向语义映射元素多,不能充分体现方位词的时空逻辑语义,但也对方向语义进行了规范.如<Outside>和<ReachTo>的方向语义分散,说明这类层级语义对方位词拆解得到的语义信息少,需要方位语义进行更细致的限定和说明.类似的层级语义还有<Inscribed><Adjacent>和<Sectional>等.部分诠释不完备的层级语义在分析时不易区分和处理,其对应标注内容在不同的标注方法下语义边界模糊,容易造成

混淆. 这一方面的问题可以通过增大相应层级语义、时空逻辑语义的数据量来解决, 或在标注时增强标注的一致性, 对该层级语义的边界明确界定、统一标注.

4.4 按照方位词频次作语义分析

按方位词频次作语义分析有助于理解每个方位词的含义, 筛选出语义复杂的方位词进行针对性处理和探讨、语义单一的方位词简化标注和分析方法, 并构建不同方位词的联系. CTB 数据集中频次最高的方位词为<中><上><后><来><时>, 除此之外, 人民日报中频次最高的方位词为<里><之间><外><间><起><底>. 对它们在全部数据集出现情况进行语义分析, 如图 9、图 10 所示. 给出的例子结构如下.

$$\langle LC \rangle - \langle Category \rangle - \langle Orientation \rangle - Phrase \quad (5)$$

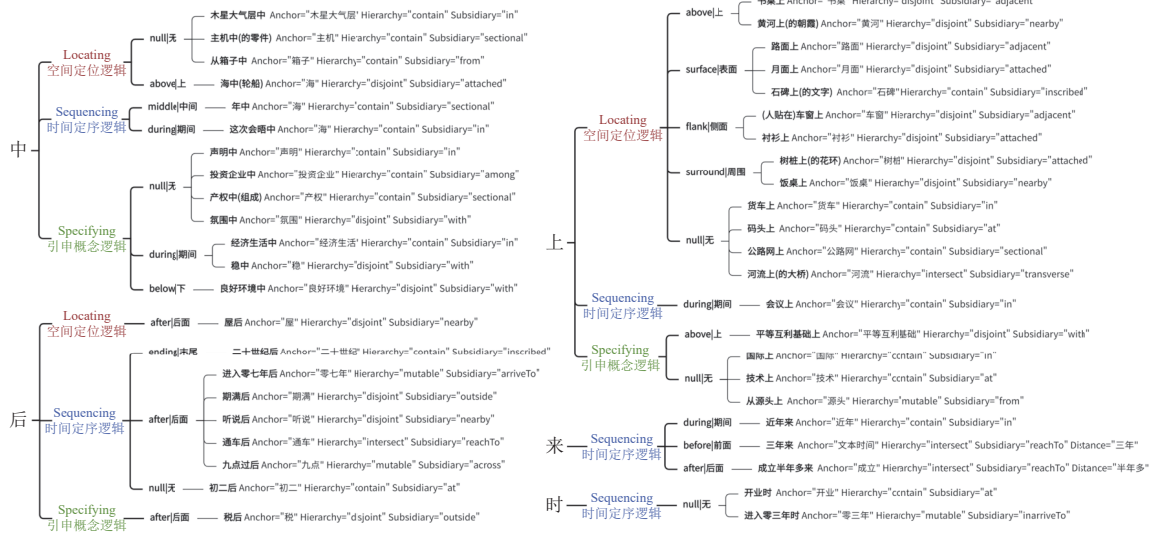


图 9 CTB 数据集中频次最高的 5 个方位词整体语义分析

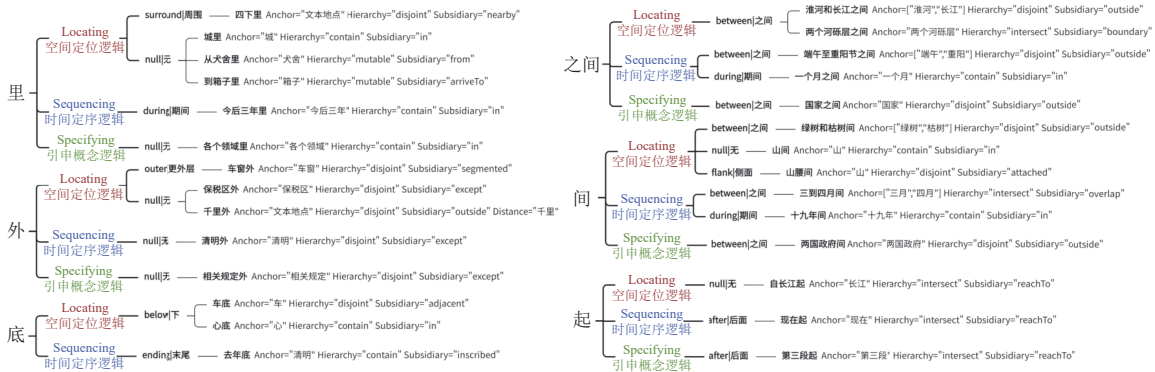


图 10 除图 9 中已给出的方位词以外, 人民日报数据集中频次最高的 5 个方位词整体语义分析

语义复杂的方位词层级作用形式分类显著多于其他方位词, 在不同情境有多种方向或层级含义. 如<上>在方向语义除<above>本义外还可以表示“锚点表面”“锚点周围”“期间内”或“锚点中”等, <后>在层级语义除<Outside>外还有可能是“在锚点附近”“锚点末尾部分”或“自锚点以后”等, 对模型辨别同一方位词不同情境层级语义或方向语义的能力提出了挑战.

语义明确的方位词类别单一、歧义较少, 只有单个或几个层级或方向语义, 标注方便、模型进行语义分析简单. 如<时>仅在时间逻辑中有<at>和<ArriveTo>两种层级语义, 这里<ArriveTo>还是<at>的动态特殊情况; <来>

则在语义上仅表示“到达某个时间点的一段时间”或“某个时间段内”, <起>的语义则只有前者, 标注歧义少; <外>除 <except>的“除锚点之外”和 <Outside>的“距锚点一段距离的位置”的两种本义外, 其他语义如 <Segmented>的“锚点更远处”的频次低, 语义构成简单. 多数方位词的语义明确, 仅少数方位词语义复杂. 因此在标注以及语义分析过程中筛选出语义明确的方位词是一项必要任务, 能够大幅简化标注流程、增强标注一致性, 提高模型语义分析能力.

有些方位词结构不同, 但语义表达类似、类别含义相近, 多出现在<以>和<之>作为首字的方位词中. 如<间>和<之间>都主要表示<between>或<during>的含义, 在句式结构上表现相近; 但是<间>的自由度更高, 还可以表示“在事物侧面”即<flank>等. <后>、<以后>和<之后>在部分含义上也相近, 但是后两者局限性强; <之后>多表示时间或引申逻辑中“与锚点相邻的次序事物”, 标注为<Adjacent>; <以后>仅在时间逻辑中表明“某个时间点更迟的时间”, 标注为<ReachTo>. 针对这样的情况需要配对处理, 关注相同的语义时, 注意句子结构上具体用法的不同, 以及微小的语义差异, 在数据标注中显著区别开来.

另外, 部分方位词语义相对单一, 但也存在一定复杂性. 例如<底>多数时候表示“时间段末尾”或“事物下侧”, 后者有时指事物内侧最下部, 标注语义为<contain><in>, 有时指事物外侧贴合的下部, 标注语义为<Disjoint><Adjacent>. 这类情况需要注意区分.

5 实验

本文的实验旨在探讨如下 3 个问题.

第一, 本文提出的标注规范体系是否能有效支撑时空逻辑语义分析任务? 不同的数据集上该标注规范表现如何?

第二, 在基于标注数据集进行 LoRA 微调时, 何种参数设置可以取得更优性能? 怎样的参数设置鲁棒性更强?

第三, 模型对该标注规范中哪些标注属性及类别更加敏感或更不敏感? 也即哪些标注属性及类别在测试中准确率表现更佳?

第 5.1 节介绍了基础实验准备, 以及本文如何设计实验解决上述问题. 第 5.2 节将呈现量化的实验结果并针对提出的问题探讨. 第 5.3 节会对错误典例进行分析, 并给出可能的优化方向.

5.1 实验设置

基于本文提出的标注体系构建两个数据集: CTB 数据集^[14] (包含 1600 条样本) 和人民日报数据集^[6,7] (包含 4690 条样本), 采用 glm3-6b 模型进行实验. 人民日报数据集数据量较大, 按照 12:3:7 的比例划分训练集、验证集和测试集, CTB5 数据集较小, 划分比例为 7:1:1.

为回答第 1 个问题, 验证该规范体系的有效性, 本文分别在两个不同体量的数据集上进行了 Few-shot 和 LoRA 微调实验, 从整体性能评估该规范在时空逻辑语义分析任务上的适用性, 直观呈现其在不同数据集上增强模型时空逻辑语义分析能力的效果. 为回答第 2 个问题, 分别针对两个数据集在多个参数组合下进行了微调实验, 具体参数设置如表 5 所示, 统一配置批处理量大小为 8、秩参数为 256、最大步数为 6000, 以控制比较不同参数对模型效果的影响, 并比对找到更具鲁棒性的参数组合. 为回答第 3 个问题, 判断模型对标注类别及属性的敏感性, 设置了两个综合评估指标, 包括除 Distance 属性外的一致准确率 (myA)、所有标注属性完全正确准确率 (AbsA); 另外设置多个单项属性或类别评估指标, 即仅针对方位词、时空逻辑类型、锚点等不同标注属性各自的准确率 (LCA、CatA 和 AncA 等), 以及空间逻辑、时间逻辑和引申逻辑每个类型自身的完全一致准确率 (LocA、SeqA 和 SpeA), 以呈现模型对该规范中不同时空逻辑类型的敏感程度 (指标定义见表 6, 准确率即正确数目除以判断的样本数目).

表 5 LoRA 微调的各参数设置

参数	1	2	3	4	5	6
LR	2E-5	2E-5	2E-5	2E-5	15E-6	23E-6
alpha	128	512	512	128	512	512
dropout	0.1	0.1	0.05	0.05	0.05	0.05

表 6 评估指标

Symbol	Name	Explanation
myA	Accuracy Except Distance	综合评判指标, 除Distance外一致的准确率
AbsA	Absolute Accuracy	综合评判指标, 完全一致的准确率
LCA	LC's Accuracy	LC识别准确率
CatA	Category's Accuracy	时空逻辑类型分类准确率
AncA	Anchor's Accuracy	锚点识别准确率
HieA	Hierarchy's Accuracy	层级作用形式大类准确率
SubA	Subsidiary's Accuracy	层级作用形式小类准确率
OriA	Orientation's Accuracy	方位指向准确率
DisA	Distance's Accuracy	距离准确率
LocA	Locating's Accuracy	仅针对空间逻辑的完全一致准确率
SeqA	Sequencing's Accuracy	仅针对时间逻辑的完全一致准确率
SpeA	Specifying's Accuracy	仅针对引申逻辑的完全一致准确率

5.2 实验结果定量分析

表 7 和表 8 展现了在人民日报数据集和 CTB 数据集上进行 Few-shot, 以及采用不同参数组合进行 LoRA 微调后的结果. 实验结果说明, 基于该标注数据集进行微调, 有助于提高模型时空逻辑语义分析能力, 多个单项指标准确率可以达到 85% 以上, 整体完全一致准确率 (AbsA) 接近 70%. 在人民日报数据集上采用学习率为 2E-5, alpha 值为 512, dropout 为 0.05 (组合 3), 延长训练步数到 10000 步得到实验结果记为 3*, 整体准确率突破 70%, 这表明该标注体系可以支撑起时空逻辑语义分析任务的要求.

表 7 人民日报数据集实验结果 (%)

实验	LCA	CatA	AncA	HieA	SubA	OriA	DisA	LocA	SeqA	SpeA	myA	AbsA
Few-shot	20.29	9.24	6.89	6.10	3.66	1.22	20.14	0	0	0	0	0
1	97.65	90.93	85.79	89.45	82.39	86.57	97.47	37.76	74.78	57.45	62.77	62.16
2	98.08	93.99	87.11	90.24	84.93	88.76	98.00	45.92	80.14	64.63	69.08	68.82
3	98.17	93.37	87.79	90.67	85.35	89.36	98.08	51.02	78.96	64.63	69.49	69.14
4	97.21	91.38	86.32	88.41	80.92	85.45	97.82	41.33	73.87	56.65	62.54	62.20
5	98.95	93.90	87.27	90.67	84.74	87.53	98.34	45.92	77.57	62.77	67.31	66.78
6	98.17	94.33	87.01	91.37	85.61	88.49	98.34	48.47	79.65	62.50	68.70	68.35
3*	98.26	94.68	87.62	91.46	85.88	89.97	98.26	53.06	78.96	65.69	70.18	69.83

注: 加粗数值表示在所有参数组合(不含3*)中取得的最高数值

表 8 CTB 数据集实验结果 (%)

实验	LCA	CatA	AncA	HieA	SubA	OriA	DisA	LocA	SeqA	SpeA	myA	AbsA
Few-shot	22.00	19.50	14.00	12.00	4.50	8.00	46.50	0	0	1.54	0.50	0.50
1	96.50	88.00	83.00	87.00	77.00	88.00	96.50	31.25	68.07	52.31	60.00	59.00
2	97.00	91.00	78.50	88.00	80.50	85.50	96.50	37.50	68.07	46.15	58.50	57.50
3	94.53	93.03	79.60	87.06	79.60	88.56	96.02	37.50	68.07	51.56	59.70	68.71
4	95.50	93.50	85.00	89.50	79.50	89.00	98.00	25.00	73.95	55.38	64.00	63.00
5	97.00	91.50	83.00	90.50	80.50	89.00	97.00	37.50	70.59	47.69	60.50	60.00
6	96.00	91.50	82.00	87.00	79.50	87.00	95.50	43.75	68.07	50.77	60.50	60.00

在人民日报数据集上, 当学习率为 2E-5, alpha 值为 512, dropout 为 0.05 (组合 3) 时, 模型表现最好, 其完全一致准确率 (AbsA) 接近 70%, 锚点识别能力 (AncA) 和方位识别能力 (OriA) 也均显著优于其他组合. 组合 2 相较于组合 3 仅更改 dropout 为 0.1, 性能却出现下降, 可见过高的 dropout 值限制了模型的表达能力. 使用较低或较高学习率时 (如组合 5 学习率为 15E-6, 组合 6 为 23E-6), 准确率明显降低, 说明 2E-5 是更合适本任务的学习率大小. 参数组合 1 和 4 的 alpha 值与秩参数比值较低, 效果相对最差. 在数据规模较小的 CTB 数据集中, 当学习率为

2E-5, alpha 值为 128, dropout 为 0.05 (即组合 4) 时, 模型表现最好, 其锚点、时空逻辑类型和方位指向的准确率与人民日报数据集上的模型水平接近, 该参数组合可能可以适用于小规模数据场景. 综合各项指标, 固定批处理大小为 8、秩参数为 256、最大步数为 6000, 较优的参数选择是设置学习率为 2E-5, alpha 值为 512, dropout 为 0.05 (组合 3).

观察表 8 中单项标注规范属性准确率, 其中方位词 (LCA)、时空逻辑类型 (CatA)、层级作用形式大类 (HicA)、方位指向 (OriA) 和距离 (DisA) 的准确率在 90% 以上, 层级作用形式小类 (SubA) 及锚点 (AncA) 的准确率则低于 90%. 所有参数组合中, 时间逻辑分析的准确率最高 (SeqA 接近 80%), 引申逻辑其次 (SpeA 接近 66%), 而空间逻辑的准确率 (LocA) 最低, 在 50% 左右. 由此可见, 模型对方位词属性 (LCA) 和时空逻辑语义类型 (CatA) 的敏感度相对最高, 对层级作用形式小类 (SubA) 的辨别力较弱, 而在空间逻辑语料上表现欠佳. 这可能是由于标注数据集中的空间逻辑数据较少: 人民日报数据集中占比 16.6%, CTB 中仅 10.8%, 样本量不足导致模型无法充分学习空间逻辑语义, 增加该类逻辑的标注语料可能有助于解决此问题.

5.3 错误典例分析

模型在处理某些复杂语义时存在判别错误, 表 9 展现了语义分析时的错误典例.

表 9 时空逻辑语义分析错误典例

#	Example	Summary
1	成为保税工厂后	<at>/<nearby>/<ReachTo> misclassified as <Outside>
2	(从)草药中	<from> misclassified as <at>/<in>
3	鞋底	Incorrect classification of <Attached> as <Inscribed>
4	新形势下	Unrecognized orientation <below>
5	迄今为止	Localizer positioning error, identified as <以来>
6	相互间	Anchor positioning error, identified as <相互>

实验中最常见的是就近语义识别错误, 模型会混淆层级作用形式的属性值, 常见于相近语义 (例 1/3), 或低频特殊语义 (例 2). 在例 1 中, 短语<成为保税工厂后>的方位词是<后>, 锚点是事件<成为保税工厂>, 时空逻辑类型是<sequencing>, 层级作用形式大类为<Disjoint>, 方位指向是<after>, 模型都能够正确分类; 但是其层级作用形式小类应为<nearby>或<ReachTo>, 体现上下文时间对象和锚点的关系, <ReachTo>表达语义<自从该事件以后>, <nearby>表达语义<该事件之后不久>, 都比模型错误分类的<Outside> (两事件仅有先后关系) 传递更丰富准确的语义. 例 2 是具有隐含移动语义的时空逻辑语义结构, 层次作用形式大类为变化态<mutable>, 语义难度较高, 模型可以准确识别空间逻辑中的变化态小类<from><ArriveTo>和<across>, 但对类似本例的引申逻辑<specifying>的变化态识别准确率较低, 容易简单理解为<at>或<in>. 这两种情况损失了部分原始语义信息, 不构成严重错误, 但反映出模型对细微语义差异的辨别力, 以及复杂环境下更丰富语义信息的提取能力不足. 而例 3 则是时空逻辑语义分析的完全错误, <鞋底>应归类为附着关系<Attached>, 但却被错误识别为内切关系<Inscribed>, 这种错误往往源于上下文语境不同, 例如<鞋底的纹路>应为<inscribed>, 而<鞋底的纸巾>则应该是<Attached>. 现有模型对空间实体关系的理解力不足, 从而产生混淆.

另外一大类是结构识别错误, 包括方位词语义遗漏或误判 (例 4/5)、锚点识别不完整 (例 6)、局部字符错误等. 例 4 是特殊条件类语义的识别错误, 模型无法识别其方位指向隐含语义<条件之下>, 对抽象概念的空间方位隐喻理解存在盲区. 例 5 的方位词完全误判为毫无关系的<以来>, 这种情况以及局部字符错误较少, 但反映出模型容易受较多训练语料的方位词影响, 在某些时候对较少训练语料的方位词存在误判. 例 6 暴露出模型在锚点识别中的边界错误, 不能准确抽取完整的语义单元, 对句子结构理解存在偏差, 形成错误的锚点识别逻辑; 类似地, 模型会正确提取锚点信息, 但无法去掉内部连词、整理为锚点数组形式 (如<甲和乙>应当整理为<[“甲”, “乙”]>). 此外, 模型有时判断基本正确, 但存在编码错误、锚点识别吞字 (如<甲市>识别为<甲>) 或多字问题, 这种情况占比约 10%, 需要予以关注.

针对上述问题,可以从以下方面进行改进:对就近语义识别错误,可以通过更精细的类别划分,或增加区分性强的语料进行训练;对于结构识别错误,需优化模型对锚点和方位词边界的敏感性,例如增加更明确的索引、更多的句法结构信息。另外,也需要针对性地增加低频方位词或类别的样本(通常其数目为个位数),以及特殊易混淆类别(例如<Inscribed><Adjacent>和<Attached>)的样本,从而增强模型的泛化能力。

6 总结

本文针对目前时空逻辑语义分析的问题,创新地提出了从方位词出发,以锚点为核心进行时空逻辑语义分析的方法,建立起基于层级作用形式的、以方位词为导向的标注规范架构。对标注数据分别以层级分类和方位词为中心作分析,串联起方位词和语义类型、句法结构和语义表示的关系。通过实验结果证明,面向方位词、建立锚点进行分析可以显著提高模型对时空逻辑信息的理解力。因此,考虑从句法结构进行时空逻辑语义分析以完善现有分析方法,存在可行性及必要性。

未来的主要研究方向如下。

(1) 基于当前数据集针对其他数据集进行标注验证,在此过程中对该标注规范架构和分类进行进一步优化。特别地,增大数据量以增强对空间逻辑的分析能力、丰富分类类别以更好理解引申逻辑语义信息。

(2) 考察动词的时空逻辑语义分析方法,并将动词和方位词的分析方法结合。

(3) 优化模型结构,探索专门处理文本时空逻辑语义分析任务的模型。

References

- [1] Zhang XY, Zhu SN, Zhang CJ. Annotation of geographical named entities in Chinese text. *Acta Geodaetica et Cartographica Sinica*, 2012, 41(1): 115–120 (in Chinese with English abstract).
- [2] Zhang XY, Zhang CJ, Zhu SN. Annotation for geographical spatial relations in Chinese text. *Acta Geodaetica et Cartographica Sinica*, 2012, 41(3): 468–474 (in Chinese with English abstract).
- [3] Leidner JL. Toponym resolution in text: Annotation, evaluation and applications of spatial grounding. *ACM SIGIR Forum*, 2007, 41(2): 124–126. [doi: [10.1145/1328964.1328989](https://doi.org/10.1145/1328964.1328989)]
- [4] Bateman JA, Hois J, Ross R, Tenbrink T. A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 2010, 174(14): 1027–1071. [doi: [10.1016/j.artint.2010.05.008](https://doi.org/10.1016/j.artint.2010.05.008)]
- [5] Ghosh S, Sengupta S, Mitra P. Spatio-temporal storytelling? Leveraging generative models for semantic trajectory analysis. In: *Proc. of the 2018 ACM Symp. on Neural Gaze Detection*. Woodstock: ACM, 2018. 8.
- [6] Yu SW, Zhu XF, Duan HM. The guideline for segmentation and part-of-speech tagging on very large scale corpus of contemporary Chinese. *Journal of Chinese Information Processing*, 2000, 14(6): 58–65 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2000.06.010](https://doi.org/10.3969/j.issn.1003-0077.2000.06.010)]
- [7] Yu SW, Duan HM, Zhu XF, Sun B. The basic processing of contemporary Chinese Corpus at Peking University SPECIFICATION. *Journal of Chinese Information Processing*, 2002, 16(5): 49–64 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2002.05.008](https://doi.org/10.3969/j.issn.1003-0077.2002.05.008)]
- [8] Schilder F, Versley Y, Habel C. Extracting spatial information: Grounding, classifying and linking spatial expressions. In: *Proc. of the Workshop on Geographic Information Retrieval at the 27th Annual Int'l ACM SIGIR Conf. Sheffield*, 2004. 3.
- [9] Randell DA, Cui Z, Cohn AG. A spatial logic based on regions and connection. In: *Proc. of the 3rd Int'l Conf. on Principles of Knowledge Representation and Reasoning*. Cambridge: Morgan Kaufmann, 1992. 165–176.
- [10] Zhan WD, Sun CH, Xiao LM. SpaCE: A linguistic knowledge-driven benchmark for spatial cognition evaluation. *Chinese Journal of Language Policy and Planning*, 2024, 9(5): 7–21 (in Chinese with English abstract). [doi: [10.19689/j.cnki.cn10-1361/h.20240501](https://doi.org/10.19689/j.cnki.cn10-1361/h.20240501)]
- [11] Zhan WD, Sun CH, Yue PX, Tang QT, Qin ZW. SpaCE2021: A new benchmark for Chinese spatial language understanding. *Applied Linguistics*, 2022(2): 99–110 (in Chinese with English abstract). [doi: [10.16499/j.cnki.1003-5397.2022.02.005](https://doi.org/10.16499/j.cnki.1003-5397.2022.02.005)]
- [12] Yue PX, Wang CW, Sun CH, Zhan WD, Sui ZF. Research on quality evaluation of Chinese spatial semantic understanding evaluation dataset. *Applied Linguistics*, 2023(1): 101–113 (in Chinese with English abstract). [doi: [10.16499/j.cnki.1003-5397.2023.01.006](https://doi.org/10.16499/j.cnki.1003-5397.2023.01.006)]
- [13] Xiao LM, Hu N, Zhan WD, Qin YH, Deng SR. Overview of CCL24-eval task 3: The fourth evaluation on Chinese spatial cognition. In: *Proc. of the 23rd China National Conf. on Computational Linguistics*. Taiyuan: ACL, 2024. 122–134.

- [14] Xue NW, Xia F, Chiou FD, Palmer M. The penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 2005, 11(2): 207–238. [doi: [10.1017/S135132490400364X](https://doi.org/10.1017/S135132490400364X)]
- [15] Mani I, Hitzeman J, Richer J, Harris D, Quimby R, Wellner B. SpatialML: Annotation scheme, corpora, and tools. In: *Proc. of the 6th Int'l Conf. on Language Resources and Evaluation*. Marrakech: European Language Resources Association (ELRA), 2008. 410–415.
- [16] Pustejovsky J, Castaño JM, Ingria R, Sauri R, Gaizauskas R, Setzer A, Katz G, Radev D. TimeML: Robust specification of event and temporal expressions in text. In: *Proc. of the 2003 AAAI Spring Symp. on New Directions in Question-answering*. Stanford: Stanford University, 2003. 28–34.
- [17] Cohn AG, Bennett B, Gooday J, Gotts NM. Qualitative spatial representation and reasoning with the region connection calculus. *GeoInformatica*, 1997, 1(3): 275–316. [doi: [10.1023/A:1009712514511](https://doi.org/10.1023/A:1009712514511)]
- [18] Zhang CJ, Zhang XY, Wang S, Liao JP, Chen XD. Annotation of spatial-temporal information of event in Chinese text. *Journal of Chinese Information Processing*, 2016, 30(3): 213–222 (in Chinese with English abstract).
- [19] Stock K, Jones CB, Tenbrink T. Speaking of location: Communicating about space with geospatial natural language. In: *Proc. of the Workshop on Speaking of Location 2019: Communicating About Space Co-located with the 14th Int'l Conf. on Spatial Information Theory (COSIT 2019)*. Regensburg: CEUR-WS.org, 2019.
- [20] Blaylock N, Swain B, Allen J. TESLA: A tool for annotating geospatial language corpora. In: *Proc. of the 2009 Annual Conf. of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Boulder: ACL, 2009. 45–48.
- [21] Kim JD, Ohta T, Tsujii J. Multilevel annotation for Information extraction: Introduction to the GENIA annotation. In: Witt A, Metzger D, eds. *Linguistic Modeling of Information and Markup Languages*. Dordrecht: Springer, 2010. 125–142. [doi: [10.1007/978-90-481-3331-4_7](https://doi.org/10.1007/978-90-481-3331-4_7)]
- [22] Fang JM. Cognitive structure of Chinese spatial orientation reference. *Chinese Teaching in the World*, 1999, (4): 32–38 (in Chinese with English).

附中文参考文献

- [1] 张雪英, 朱少楠, 张春菊. 中文文本的地理命名实体标注. *测绘学报*, 2012, 41(1): 115–120.
- [2] 张雪英, 张春菊, 朱少楠. 中文文本的地理空间关系标注. *测绘学报*, 2012, 41(3): 468–474.
- [6] 俞士汶, 朱学锋, 段慧明. 大规模现代汉语标注语料库的加工规范. *中文信息学报*, 2000, 14(6): 58–65. [doi: [10.3969/j.issn.1003-0077.2000.06.010](https://doi.org/10.3969/j.issn.1003-0077.2000.06.010)]
- [7] 俞士汶, 段慧明, 朱学锋, 孙斌. 北京大学现代汉语语料库基本加工规范. *中文信息学报*, 2002, 16(5): 49–64. [doi: [10.3969/j.issn.1003-0077.2002.05.008](https://doi.org/10.3969/j.issn.1003-0077.2002.05.008)]
- [10] 詹卫东, 孙春晖, 肖力铭. 语言学知识驱动的空间语义理解能力评测数据集研究. *语言战略研究*, 2024, 9(5): 7–21. [doi: [10.19689/j.cnki.cn10-1361/h.20240501](https://doi.org/10.19689/j.cnki.cn10-1361/h.20240501)]
- [11] 詹卫东, 孙春晖, 岳朋雪, 唐乾桐, 秦梓巍. 空间语义理解能力评测任务设计的新思路—SpaCE2021 数据集的研制. *语言文字应用*, 2022(2): 99–110. [doi: [10.16499/j.cnki.1003-5397.2022.02.005](https://doi.org/10.16499/j.cnki.1003-5397.2022.02.005)]
- [12] 岳朋雪, 王诚文, 孙春晖, 詹卫东, 穗志方. 中文空间语义理解评测数据集质量评估研究. *语言文字应用*, 2023(1): 101–113. [doi: [10.16499/j.cnki.1003-5397.2023.01.006](https://doi.org/10.16499/j.cnki.1003-5397.2023.01.006)]
- [18] 张春菊, 张雪英, 王曙, 廖建平, 陈晓丹. 中文文本的事件时空信息标注. *中文信息学报*, 2016, 30(3): 213–222.
- [22] 方经民. 汉语空间方位参照的认知结构. *世界汉语教学*, 1999, (4): 32–38.

作者简介

于知衡, 硕士生, 主要研究领域为自然语言处理.

辛欣, 博士, 副教授, CCF 专业会员, 主要研究领域为基于大语言模型的多模态信息检索与推理技术.