从句子图到篇章图——基于抽象语义表示的篇章级共指标注体系设计

张艺璇 [1,2], **李斌** [1,2,3], **许智星** [1,2], **卢芃秀** [1,2]
1 南京师范大学文学院, 江苏南京 210097
2 南京师范大学语言大数据与计算人文研究中心, 江苏南京 210097
3 澳门大学人文学院, 中国澳门 999078
xiaoxian224@gmail.com, libin.njnu@gmail.com,
xzx0828@live.com, 230102030@njnu.edu.cn

摘要

篇章共指体现篇章概念的动态转移，成为近年研究热点。本文在梳理共指理论研究的基础上，综述了相关语料库及解析方法，发现共指语料库仍存在以下两个问题：共指关系标注粗疏与基本不考虑整句语义表示的融合。本文以句子级语义标注体系（中文抽象语义表示）为基础构建篇章共指体系，构建了 100 篇共指语料库。本体系涵盖 52 种句内语义关系和 8 种篇章共指关系，二者相结合构建的篇章共指语义图，为篇章级语义分析提供新的框架和数据资源。

关键词： 篇章共指 ；抽象语义表示 ；篇章语义结构 ；中文信息处理

From Sentence Graphs to Discourse Graphs: Designing a Discourse-Level Anaphora Annotation System Based on Abstract Semantic Representation

Yixuan Zhang[1,2], **Bin Li**[1,23], **Zhixing Xu**[1,2], **Pengxiu Lu**[1,2]
1 School of Chinese Language and Literature, Nanjing Normal University, China
2 Center for Language Big Data and Computational Humanities, Nanjing Normal University,China
3 Faculty of Arts and Humanities, University of Macau, China
xiaoxian224@gmail.com, libin.njnu@gmail.com,
xzx0828@live.com, 230102030@njnu.edu.cn

Abstract

Discourse coreference reflects the dynamic conceptual transitions within a text and has become a research focus in recent years. This paper, based on a review of coreference theory, summarizes related corpora and parsing methods, identifying two main issues in existing coreference corpora: coarse annotation of coreference relationships and a lack of integration with complete sentence semantics. Using a sentence-level semantic annotation system (Chinese Abstract Meaning Representation) as a foundation, this paper constructs a discourse coreference system and a corpus of 100 texts. This system encompasses 52 types of intra-sentence semantic relationships and 8 types of discourse coreference relationships. The combination of these elements constructs a discourse coreference semantic graph, providing a new framework and data resource for discourse-level semantic analysis.

Keywords: Discourse Anaphora , Abstract Semantic Representation , Discourse Semantic Structure , Chinese Information Processing

1 引言

指代（Reference）不仅表现为语法上的替代关系或语义上的同指关系，还充当着话题转换的衔接手段，对挖掘篇章概念转移、理解自然语言有重要意义，目前已成为理论语言学和计算语言学关注的热点问题之一。

鉴于研究目的与分析方法的差异，目前理论语言学和计算语言学在指代领域的术语使用有所区别。理论语言学常使用“回指（Anaphora）”来表示指代现象，广义的回指可分为直接回指和间接回指两类 (Ariel, 1990)。直接回指是指两个语言成分之间的关系，对其中一个成分的解释，取决于对另一个成分的解释 (Huang, 1984)，间接回指是指回指语和先行语之间没有明显的指代关系，而需要经过语用推理才能建立指称关系 (Erkü and Gundel, 1987)。在计算语言学中，与“直接回指”相对的是“共指（Coreference）”，与“间接回指”相对应的是“桥接关系（Bridging Relation）”。但间接回指（或桥接关系）由于其复杂性与模糊性，被学界研究长期排除在外，故相关研究发展缓慢且未成体系。尽管理论语言学界已有学者 (王军, 2004; 王军, 2013; Vieira and Poesio, 2000) 尝试总结常见间接共指类别，计算语言学界也构建了一系列桥接关系语料库，如 ISNotes (Markert et al., 2012)、BASHI (Rösiger, 2018)、ARRAU (Poesio et al., 2008; Uryupina et al., 2016) 和 SciCorp (Rösiger, 2016)。但并没有对该现象的本质上达成一致认同，故目前本文仅讨论具有概念同一性的共指现象。

为推动共指消解（Coreference Resolution）任务的发展，计算语言学界开始标注共指语料库。尽管现有语料库的构建体系和标注方法已有一定发展，但仍存在以下两个问题，即共指关系标注粗疏以及基本不考虑整句语义表示的融合，导致共指消解仍有较大发展空间。因此，本体系将在中文抽象语义表示（Chinese Abstract Meaning Representation；CAMR）基础之上构建篇章共指标注体系。CAMR 是句子级标注体系，有 5 种核心语义角色关系和 47 种非核心语义角色关系，可有效挖掘共指词所担任的语义角色，提供语义关系。本体系以“概念同一性”为基本原则，从词形、词长和词类等角度区分 8 种共指关系，提供共指链内部、共指词之间的共指关系。先行词、共指词以及句内语义关系与篇章共指关系相结合可构建出篇章共指语义图，从语义角度归纳共指规律。基于本体系所构建的 100 篇共指语料库，可为共指解析提供数据支持。

2 直接回指的理论研究

已有大量学者从句法、认知、功能、篇章等角度对回指理论作出探索，其中影响深远的是认知角度的可及性理论（Accessibility Theory) (Ariel, 1990) 以及篇章角度的话题延续性模式（Topic Continuity Model) (Givón, 1983)。在两大理论提出的“可及性标示阶”与“话题延续性量表”的基础之上，学界逐步确定了频次、位置、间隔距离、语篇结构、干扰数量和句法成分等回指计量指标，但这些指标大多从表层分析回指，较少有学者构建回指的语义语料库，从语义角度归纳回指规律。纵观回指计量发展历程，研究焦点主要有三大趋势：其一，逐渐从名词和代词等具体实体的标注与计量 (徐赳赳, 2003; 许余龙, 2004; 宋宏, 2010; 石艳华, 2014) 转向概述回指等抽象类别 (刘东虹, 2014; 王秀丽, 2012; 吴丹苹, 2021; 寇鑫 and 徐坤宇, 2023)；其二，近年来，学界已不再局限于单纯地专研某一回指类型，而是将其放入句法、句式等宏观视野中探索规律。如蒋平 (2021) 基于认知语言学总结存现宾语的零形回指演进，冉晨 (2024) 分析数量名回指语的指称性质与回指确认方式；其三，回指研究视角逐渐从本体研究扩展到语言教学与应用中。如杨永生 (2020) 基于回指确认考察韩国学生汉语“这/那”句的习得情况，李榕等 (2021)、曹娜等 (2024) 借鉴徐赳赳 (2003) 等学者提出的计量指标挖掘韩国留学生的回指偏误、归纳二语学习者使用主语回指语的显隐现象。

3 共指的计算资源构建与解析研究

计算语言学借鉴现有理论标注了名词、代词、零形式等常见体词共指，并进一步探索动词、形容词甚至副词等谓词共指，一定程度上推动了共指研究。但谓词共指存在意义实虚难以界定等问题，仍处于探索阶段，体词共指的标注仍为学界主流。但这些语料库的标注信息相对粗疏，不能考察共指规律，也不能为现有计算模型提供充足的共指语义或知识信息。

3.1 共指相关语料库构建

面向计算建模的共指语料库可分为两类：一类是以 OntoNotes 为代表的共指语料库，一类是以 MS-AMR 为代表的包含共指信息的语料库。前者细致标注共指现象，却忽略了整句的语义

表示；后者全面表示整句语义表示，对共指现象的标注却仍待深化。但这两类语料库共有的不足在于均仅停留在了共指的同一性标注中，没有对共指链中共指词与先行词之间的关系作进一步分析。

3.1.1 共指语料库

最为典型的共指语料库为 OntoNotes (Weischedel et al., 2013) 语料库，该语料库也是目前共指评测认可度最高的数据集。其标注规范成为众多共指标注体系的借鉴样本，使共指语料库迈向多语言时代。尽管语料库数量在不断增加，规模不断扩大，语种不断丰富，但标注信息和标注方法仍并未发生明显更新。除了该系列语料库，共指语料库还包括 MUC (Hirschman and Chinchor, 1998)、ACE (Doddington et al., 2004) 等早期语料库，另外还有一些标注单一共指信息的语料库，如汉语零指代语料库 (孔芳 et al., 2021)、法汉指称链条平行语料库 (胡霄钦 and 王秀丽, 2021)。这些语料库将共指标注从体词扩展到谓词，为探索篇章共指提供新的渠道。但这些语料库几乎不涉及共指的语义信息，也并未挖掘共指链内部的共指关系。MUC6（Message Understanding Conference）开创了共指标注体系的先河。尽管其标签分类较为简单，但研究学者已逐渐认识到共指的复杂性，并有了灵活表示共指词的倾向性（如设置“MIN”标签以表示包含在整词之内的部分字符串）。但其细则也存在值得商榷之处，如将人物与其头衔等职位标注为共指关系、不补充标注零形式等问题。ACE（Automatic Content Extraction）在 MUC6 的基础之上，对共指进一步标注。其语料不再局限于英语，而是迈向了多语言资源的构建，ACE 也是最早针对中文指代消解的国际评测语料资源。MUC6 和 ACE 对共指的标注较为简单，直到 OntoNotes 之后共指标注迎来飞跃。OntoNotes 取消对共指标注的限制，除名词、代词之外，还可标注动词、形容词等信息，对指代消解任务有新的推进。在共指层主要标注两类信息，即同一性（IDENT）和同位语（APPOS）。同一性共指（IDENT）用于指代共指，意味着代词、名词性和特定所指对象的命名提及之间的联系，不包括一般的、未指定的或抽象实体。同位语（APPOS）逻辑上代表属性，因此被单独处理。自 OntoNotes 问世之后，主要有两大研究现状与趋势：其一，在 OntoNotes 标注规范的基础之上构建了许多大型语料库。Ghaddar 等 (2016) 继承 OntoNotes 标注方法，以维基百科为语料构建了 WikiCoref 语料库；Chen 等 (2018) 在 OntoNotes 规范的基础之上改良并构建了目前最大规模的 PreCo 英语共指语料库；Poesio 等 (2019) 基于游戏化方法构建了众包共指语料库，是最大的应用共指语料库之一。其二，近三年的共指资源主要集中于非英语语言以实现跨语言共指解析。这些语料库包括俄语共指语料库 RuCoCo (Dobrovolskii et al., 2022)、涵盖英法德葡四种语言的多语言共指语料库 ParCorFull2.0 (Lapshinova-Koltunski and Ferreira, 2022)、基于 UD（Universal Dependencies）构建的多语言语料库集合和共指消解的标准化格式 CorefUD (Nedoluzhko et al., 2022) 以及荷兰语跨文档事件共指解析大型数据集 (De Langhe et al., 2023) 等数据资源。

3.1.2 融合共指信息的整句语义标注语料库

以 PDT (Mikulová et al., 2006)、UMR (Uryupina et al., 2016) 和 MS-AMR (O’ Gorman et al., 2018) 为代表的语料库实际是句子级语义标注语料库，并非专为共指而设计。尽管一定程度弥补了共指语料库缺乏整句语义信息的不足，但在共指规范上仍待完善。

PDT 的共指信息有三种，即语法共指、文本共指和特殊类型共指，主要标注了代词、动词、省略、概述共指等共指信息。UMR 是建立在 AMR 基础之上、具有语言包容性的语料库，除了继承了 AMR 的句子级表示之外，在共指方面采用关系三元组的形式标注实体共指和事件共指。MS-AMR 参考了 OntoNotes 的标注方式在句子级 AMR 基础之上扩展句间共指信息。

3.2 共指关系的计算

从数学角度而言，共指关系实质是一种等价关系，因此消解过程就是等价类划分的过程 (宋洋 and 王厚峰, 2015)。在过去的十年中，共指解析已经从基于规则的方法发展到基于学习的方法 (周炫余 et al., 2014)。基于规则特征的泛化能力较差，理解和实现比较简单 (Lang et al., 2007)，基于学习的方法开始引入开放知识作为额外特征，改善了模型效果高度依赖特征工程的弊端。近年来采用深度学习算法，如多层感知器/循环神经网络方法、基于知识的方法和基于变压器的方法 (Liu et al., 2023)，逐渐突破大规模语料的限制，大大增强了模型的深层语义学习能力和泛化性能 (陈远哲 et al., 2019)。Liu 等 (2023) 总结出五点共指解析的挑战及发展方向，即：下游任务缺乏带有共指解析标签的数据集；缺乏符号特征与子符号特征的组合；需结合现有语

言研究及认知直觉；当前模型需压缩所用资源以实现多任务学习；仍需超大规模语言模型的出现。Liu 指出，尽管已有大量研究成果为数据注释和模型设计奠定理论基础，但这些语言和认知发现却很少被纳入基于深度学习的共指解析模型中，未来可结合额外的符号特征（如语义特征和知识表示）和子符号方法（如词嵌入） (Mao et al., 2018; Cambria et al., 2022; Mao et al., 2022)，以解决共指数据稀疏问题，提高共指消解的精度。

4 篇章级共指标注体系

综上所述，理论语言学界用于计量分析的语料库，大多只能从表层信息（如回指分布、句法成分等）计量并分析回指规律，缺乏以语义为研究导向的语料库；计算语言学界用于共指消解的语料库，也缺乏兼顾语义关系与共指关系的语料库，以提高共指消解的精度。因此，本文基于句子级语义标注体系（中文抽象语义表示）构建篇章共指标注体系，涵盖语义信息与共指信息，为共指消解提供新思路。

4.1 句子级 CAMR 的标注体系

Abstract Meaning Representation（AMR）是一种抽象语义标注框架，采用单根有向无环图的表示方法，图中节点表示概念，边表示概念之间的关系 (Banarescu et al., 2013)。AMR 忽略语义较虚的成分（如冠词、单复数、时态等等）和形态变化，从原句中抽象出概念，不拘泥于原句词语，允许对其进行增添、删减和改动等操作以便更好表示语义关系。李斌等 (Li et al., 2019) 在 AMR 的基础之上结合汉语的语言特点继承并发展出 CAMR（Chinese AMR）。其中，共包括 5 种核心语义角色关系（arg0-arg4）、47 种非核心语义角色关系。

句子级 CAMR 体系是篇章共指标注体系的构建基础，以词为基础单位，以树结构为展示形式，语义结构的层级性有利于在复杂的语言表述中确定共指单位，并提供共指单位的语义功能。表 1 为句子级 CAMR 标注体系和篇章级共指标注体系比较。

体系名称	句子级 CAMR 标注体系	篇章级共指标注体系
数据形式	单根有向无环图	单根有向无环图
标注层级	句内（句子级）	句内、句间（篇章级）
语义单位	单句内的概念（汉语常以词表示）	篇章内的共指概念（汉语常以词或短语表示）
揭示信息	句内各个成分的语义结构	篇章内部共指成分的指代关系
文本示例	男孩 [1] 希望 [2] 女孩 [3] 相信 [4] 他 [5] 。[6] x5/希望-01 :arg0() x1/男孩 :arg1() x5/相信 :arg0() x3/女孩 :arg1() x5/x1	p0222_r3_entity :root() s3_x55 / university（杭州大学） :alias() s3_x26_x27 / 该校 :homo() s9_x24 / university（杭州大学） :reduce() s13_x4 / 大学 :reduce() s13_x19 / 大学

表 1: 句子级 CAMR 标注体系与篇章级共指标注体系比较

备注：在文本示例中，句子级 CAMR 标注体系展示的是单句的语义结构，而篇章级共指标注体系的示例则展示了在同一个篇章中指代“杭州大学”的所有语言单位。左侧，形如“:arg0”的标签为语义关系，“x+ 数字”表示词编号；右侧，形如“:alias”的标签为共指标签，“s+ 数字”表示句编号，“x+ 数字”表示词编号。

4.2 共指标注原则

4.2.1 概念同一性

本体系遵守“概念同一性”原则，只标注相同指代意义的词语。如表 2 文本中与“福利院”有关的词语共出现 7 次，但有 6 次（位于 S3、S6、S7、S8、S11）指代为“杭州市儿童福利院（位于 S3）”，构成共指链 2；有一次（位于 S4），泛指国内的福利院，在本文中与“中国孤儿院（S1）”形成共指关系，构成共指链 1。

备注：表 2 为 CTB8.0 语料中编号为 0222 的文本，为便于理解，省略了与“福利院”“孤儿院”无关的句子。

S1：丁豪成为 [**中国孤儿院**]$_1$ 长大的第一个残疾人大学生 S3：自幼成为孤儿并在 [**杭州市儿童福利院**]$_2$ 长大的残疾学生丁豪，日前收到了杭州大学的录取通知书，成为该校哲学与社会科学系的学生。 S4：据了解，今年十九岁的丁豪是新中国第一个由 [**福利院**]$_1$ 收养成长的残疾孤儿大学生。 S6：三岁时一场高烧，使他患上了严重的小儿麻痹后遗症，这一年他被 [**福利院**]$_2$ 收养。 S7：丁豪在 [**儿童福利院**]$_2$ 读完小学，随后进入附近乡里一所学校上初中。 S8：一开始由 [**福利院**]$_2$ 的老师接送，初二后住校，每个星期六回到 [**福利院**]$_2$ ，无论刮风下雨，从不缺课。 S11：[**儿童福利院**]$_2$ 十六年如一日地关心、爱护他，去年以来，仅为他交纳学费、买资料、买营养品，就花了三千多元人民币。

表 2: “杭州市儿童福利院”共指示例

4.2.2 共指链可以嵌套

由于语言递归性，词语嵌套是语言表达的常用方式，因此极易出现共指词或共指链嵌套的情况。表 3 的例子共出现两条共指链，下标相同即为同一条共指链。S17 中的“其名”即存在共指词的嵌套现象：“其”指代“小海龟”，为共指链 2 的共指词；“其名”指代“卡洛塔”，为共指链 1 的共指词。当多个共指链的每一个共指词均存在嵌套现象时，则这些共指链存在嵌套现象。

S16：本届世界游泳锦标赛的吉祥物是一只名叫“ [**卡洛塔**]$_1$ ”的 [**小海龟**]$_2$ 。 S17： [**海龟**]$_2$ 的形象由意大利知名图案设计师瓦·隆巴多设计，[[**其**]$_2$ **名**]$_1$ 取自于一位天真可爱的意大利小女孩。 S18：别看 [**小海龟**]$_2$ 在沙滩上爬行往往显得笨拙，一旦到了海里，就变得灵活自如。 S19：在“[**小海龟**]$_2$ ”坚实的龟甲上嵌着由陆地、海洋组成的世界地图，寓意来自全球五大洲四大洋的游泳健儿会聚罗马，通过比赛，加强了解，增进友谊。

表 3: “卡洛塔”小海龟例句

备注：表 3 为 CTB8.0 语料中编号为 0308 的文本，为便于理解，省略了与“海龟”“卡洛塔”无关的句子。每行的“S+ 数字”为句子编号。

4.2.3 区分概念同一性和实体间关系

篇章内部存在大量指代不同、但关系密切的概念，极易混淆。但本体系严格区分这类现象，未来将总结该类现象设计普适性的实体间关系。如表 3 中存在的两条共指链，一条是实体链（“小海龟”链），一条是实体名称链（“卡洛塔”链），即为指代不同但关系密切的共指链。

4.3 共指关系类型

本体系标签综合词长、词形、词义和词类等角度，依据先行词和共指词的中心词和定语变化等指标设置标签类别，共 8 大类（见表 4）。由于“:encap”（概述共指）和“: illustrate”（分述共指）标签存在方向性，于表 5 中的“标签小类”分别作详细介绍。“示例”部分为当先行词为“两个美丽的女孩”且共指元素均指代这一女孩时，共指标签的适用情况。

5 共指语料库标注及计量分析

5.1 语料选择

我们从 CAMR v2.0 语料中选取了 100 篇文本进行共指标注，该文本在宾州中文树库（Chinese treebank；CTB）的编号为 chtb0222-chtb0323（chtb0319 因篇幅过短不予标注）。该语料库体裁涵盖经济、体育及生活，经数据清洗（去除记者、发表时间等新闻组成要素）后共有句子 1116 个，词次为 28303。由于该语料库同时拥有 CTB 句法信息和 CAMR 语义信息，可为共指标注与后期的计量检索提供极大帮助与便利。原因如下：其一，句子语义层级性以树结

编号	标签	含义	示例
1	:root	该词为先行词	两个美丽的女孩
2	:homo	共指词与先行词完全相同	两个美丽的女孩
3	:add	共指词在先行词基础之上添加某些词语	两个美丽善良的女孩
4	:reduce	共指词在先行词基础之上减少某些词语	两个女孩
5	:alias	共指词在先行词基础之上同时减少、添加某些词语（即改换其他表述）	两个贤惠的女孩
6	:pron	共指词是代词	她们
7	:illustrate	分述共指	——
8	:encap	概述共指	——

表 4: 共指元素标签类型及含义

标签	标签小类	含义
:illustrate	:illustrate	共指词对先行词分述
	:illustrate_of_alias	共指词对先行词总述，共指词是代词
	:illustrate_of_pron	共指词对先行词总述，共指词是短语
:encap	:encap_pron	先行词是小句，共指词是代词
	:encap_alias	先行词是小句，共指词是短语
	:encap_of_pron	先行词是代词，共指词是小句
	:encap_of_alias	先行词是短语，共指词是小句

表 5: 共指部分标签小类及含义

构为展示形式，为共指词的确定提供精确定位；其二，CTB 涵盖分词、词性标注、句法等多种语言信息，涵盖学界现有的计量指标（如间隔距离、干扰数量、句法信息）；其三，CAMR 在 CTB 基础之上提供了语义信息，可揭示共指词变化的语义规律。

5.2 规范制定与标注过程

标注工作分为 3 个阶段。

（1）观察语料阶段。该阶段的主要工作是分析大量生语料，观察并标注出篇章中具有概念同一性的语言单位，在充分考虑语料质量与标注信息可操作性的基础之上，形成初步的标注规范。

（2）预标注阶段。该阶段将通过实践确认标注者对规范的理解，同时检验规范的通用性与可操作性，并在标注过程中及时调整规范，得到最终的标注规范。

（3）正式标注阶段。经过前期的标注工作，正式标注阶段采用了“程序粗提取 + 人工细标注”的方式，以词频为指标，在 CAMRv2.0 语料中提取潜在共指元素，建构粗糙的依存三元组，并在此基础调整修改，确保共指链标注无遗漏。最终形成完整的抽象语义表示篇章共指语料库。

5.3 语料计量分析

在这 100 篇文本中，共标注了 792 条实体共指链。每篇平均 7.92 条共指链，共指链长度处于 2-34 之间，平均每条共指链长度为 4.35，方差为 11.87，由此数据可以看出共指链长度变化极大，其中有 52% 的共指链长度处于 2-5 之间，其余共指链分布长度不等且分布不均；共指链跨句情况处于 0-34 之间，平均每条共指链跨 6.01 句，方差为 31.58；句内共指出现 509 次，句间共指出现 2937 次，平均每条链出现 0.64 次句内共指、3.71 次句间共指，方差分别为 1.73、7.11。经统计，各共指的出现频次见表 6（已按频次进行降序排列）。可以看到：首先，在共指关系中，共指词与先行词一致的情况占比最高，共出现 1495 次，但其方差也最高（6.88），说明这种共指关系出现次数最高，但在共指链中变化波动最大；其次，“:add”关系略低于“:reduce”关系，这可能与新闻语料的语言风格相关，即在论述一个概念时，新闻通常在首次提及时便提供最为详细的信息；最后，对于概述共指而言，使用代词指代先行词（小句），比使用短语指代先行词（小句）更为常用。

共指关系	总频次	平均频次/链	方差	比例（%）
:homo	1495.00	1.89	6.88	43.46
:root	792.00	1.00	0.00	23.02
:reduce	338.00	0.43	1.77	9.83
:add	309.00	0.39	1.09	8.98
:alias	278.00	0.35	0.75	8.08
:pro	200.00	0.25	0.91	5.81
:encap_pro	26.00	0.03	0.01	0.76
:encap_alias	22.00	0.03	0.02	0.64
:illustrate_of_alias	12.00	0.02	0.03	0.35
:illustrate	3.00	0.00	0.00	0.09
:illustrate_of_pro	1.00	0.13	0.13	0.03
:encap_of_alias	0.00	0.00	0.00	0.00
:encap_of_pro	0.00	0.00	0.00	0.00

表 6: 共指标签频次表

6 篇章共指语义图的构建

6.1 篇章共指语义图组成部分

本文所构建的篇章共指语义图为单根有向无环图，图中结点为共指元素，边表示共指元素之间的关系。其中，横向关系表示共指元素的句内语义关系，纵向关系表示共指元素的共指关系（见表 4、表 5）。上文中介绍到句子级 CAMR 共有 5 种核心语义角色关系和 47 种非核心语义角色关系，在篇章共指中常用的 CAMR 语义角色关系主要有以下几类。

6.1.1 核心语义角色关系

核心语义角色关系沿用 OntoNotes 体系的核心语义角色关系，标签及其代表关系如表 7 所示。

核心语义角色关系关系标签	关系
arg0	原型施事
arg1	原型受事
arg2	间接宾语、受益者、工具等
arg3	出发点、受益者等
arg4	终点

表 7: 核心语义角色关系汇总

6.1.2 非核心语义角色关系

CAMR 中，共有 47 种非核心语义关系，理论上均可以在共指标注中使用，但常见的语义角色关系有以下几种（见表 8）。

6.2 篇章共指语义图构建示例

篇章共指语义图为单根有向无环图，图中结点为篇章中所出现的共指概念，边表示共指概念之间的关系，即横向的句内语义关系和纵向的共指关系。本文将从横纵角度探讨篇章共指语义图结构，以证明基于 CAMR 体系标注和构建共指语义结构，有利于揭示篇章结构中共指链的动态变化，挖掘篇章共指语义信息。

我们选用了 CTB8.0 中第 0226 篇新闻，该文本中含有实体共指链 13 条，共指链长度为 1-7 句不等，跨句情况为 1-9 句不等，共出现 7 次句内共指和 47 次句间共指。我们依据共指关系类

序号	非核心语义角色关系	中文说明	序号	非核心语义角色关系	中文说明
1	location	处所	6	poss	领属
2	manner	方式	7	subset	子集
3	medium	媒介	8	superset	父集
4	mod	修饰	9	topic	话题
5	part-of	部分			

表 8: 用于共指标注的常见非核心语义角色关系

型与句内语义关系，架构该文本的篇章共指语义图，如图 1 所示[1]。图 1 以句子为基本单元，每一个虚线框都代表一个句子平面，框的左上部为文本原句，虚线框内不同颜色的实线框为该文本中的共指元素，已用颜色高亮加以区别，且标注在文本原句中。图中的有向实线分为横纵两个方向，横向表示共指元素的句内语义关系（如“:arg0”标签），纵向表示共指元素之间的共指关系（如“:homo”标签）。下面将以该文本为示例，从横纵角度探讨篇章共指语义图结构。

（1）横向：揭示句子内部共指信息。

横向以句子为基本单元，可以揭示：（1）每句共指元素的数量及嵌套情况。如 S1 句共分布有 5 种共指现象，其中有 3 种存在层层嵌套关系。（2）每句的句内共指情况。如 S3 中“两军两国之间的合作”和“上述问题”具有概念同一性，为句内共指。（3）共指元素的句内语义关系。如在 S1 中，“迟浩田”的句内语义关系为“:arg0（施事）”，“坦桑”的句内语义关系为“:mod（修饰）”。

（2）纵向：揭示句子之间概念转移。

纵向以共指链为基本单元，可以揭示：（1）共指链的起止、长度及其分布。如“会谈”共指链，起于 S1，止于 S9，跨句情况为 9 句；共含有 3 个共指元素，分布于 S1、S3、S9，尽管出现次数较少，但几乎贯穿本篇始终。（2）共指元素之间的共指关系。如“会谈”共指链，在全文中并未出现变化；如“两国”共指链在篇章中也会使用“中国和坦桑”这类表达。

总之，语义关系将共指同一句、不同共指链之间的共指元素连结，共指关系将不同句、同一共指链之间的共指元素，构成了共指语义结构的基本骨架。该图在句子层揭示了共指元素的句内语义关系，在篇章级揭示了共指元素之间的共指关系，通过横纵关系综合观察，可总结篇章结构中共指变化的动态性。

7 结论与未来工作

共指是近年篇章研究热点，尽管计算语言学领域相关文章更新迭代较快，但其大多为指代消解模型的构建更新。且近年发布的多语言共指资源大多依据 2013 年发布的 OntoNotes 标注体系而构建，在标注体系的根本设定上并未有较大突破。因此，这些语料库均存在共指关系粗疏以及基本不考虑整句语义表示的融合等问题。本文在句子级 CAMR 体系基础之上探索共指标注体系，从整句语义结构探索共指的演变推进；区分了 8 种共指关系类型，对共指关系作出一定阐释；从 CAMR 中继承的句内语义关系，可结合篇章共指关系构建篇章共指语义图。基于该标注体系构建了 100 篇新闻共指语料库。该语料库的标注信息不仅涵盖了频次、位置、间隔距离、句法成分等经典计量指标，还拓展了基于语义结构的计量信息。在未来，我们将从以下三个方面开展工作：首先，在标注体系与标注规模上，我们将扩大新闻语料的标注规模，并尝试探索记叙文、小说等共指现象丰富的语料，完善共指标注方案；其次当标注体系稳定之后，将依据该规范扩大语料库规模。最后，基于共指标注体系发展实体间关系和篇章修辞结构，尝试建构整全的篇章语义结构图。

[1]清晰版可见网页：https://www.camrp.tech/DAMR/

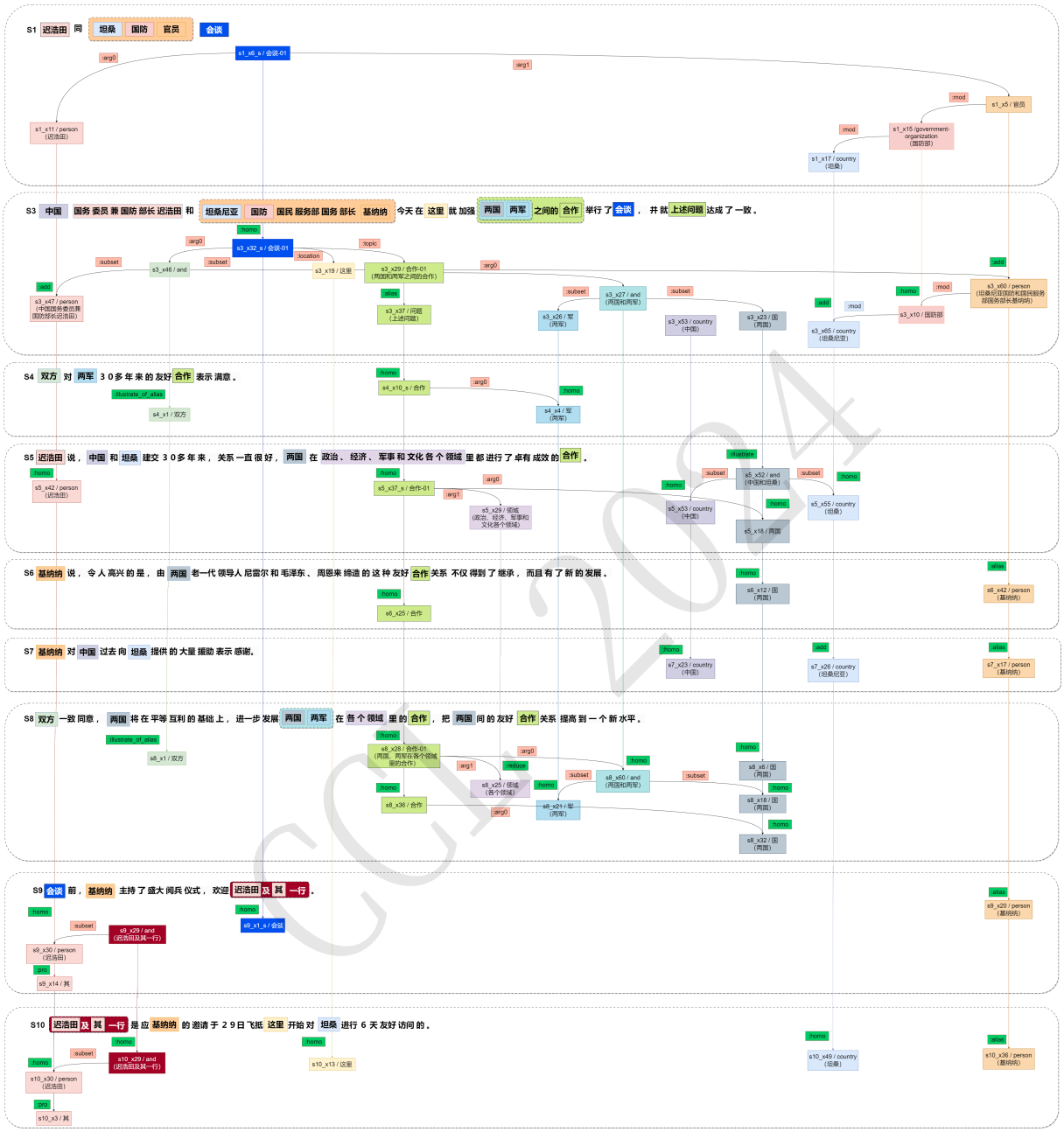

图 1: 篇章共指语义图示例

参考文献

Mira Ariel. 1990. Accessing noun-phrase antecedents. vol. 96. londres: Routlege. *Linguistics*, pages 113–118.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839.

Hong Chen, Zhenhua Fan, Hao Lu, Alan L Yuille, and Shu Rong. 2018. Preco: A large-scale dataset in preschool vocabulary for coreference resolution. *arXiv preprint arXiv:1810.09807*.

Loic De Langhe, Orphée De Clercq, and Veronique Hoste. 2023. Constructing a cross-document event coreference corpus for dutch. *Language Resources and Evaluation*, 57(2):819–848.

Vladimir Dobrovolskii, Mariia Michurina, and Alexandra Ivoylova. 2022. Rucoco: a new russian corpus with coreference annotation. *arXiv preprint arXiv:2206.04925*.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Feride Erkü and Jeanette K Gundel. 1987. The pragmatics of indirect anaphors. In *The pragmatic perspective*, pages 533–545. John Benjamins BV.

Abbas Ghaddar and Philippe Langlais. 2016. Wikicoref: An english coreference-annotated corpus of wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142.

Talmy Givón. 1983. Topic continuity in discourse. *Topic continuity in discourse*, pages 1–498.

Lynette Hirschman and Nancy Chinchor. 1998. Appendix f: Muc-7 coreference task definition (version 3.0). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

C-T James Huang. 1984. On the distribution and reference of empty pronouns. *Linguistic inquiry*, pages 531–574.

Jun Lang, Bing Qin, Ting Liu, and Sheng Li. 2007. 篇章共指消解研究综述. *Journal of Chinese Language and Computing*, 17(4):227–253.

Ekaterina Lapshinova-Koltunski and Pedro Augusto Ferreira. 2022. *ParCorFull2. 0: A parallel corpus annotated with full coreference*. Saarländische Universitäts-und Landesbibliothek.

Bin Li, Yuan Wen, Li Song, Weiguang Qu, and Nianwen Xue. 2019. Building a chinese amr bank with concept and relation alignments. *Linguistic issues in language technology*, 18.

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, 56(12):14439–14481.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th annual meeting of the association for computational linguistics*. Association for Computational Linguistics (ACL).

Rui Mao, Xiao Li, Mengshi Ge, and Erik Cambria. 2022. Metapro: A computational metaphor processing model for text pre-processing. *Information Fusion*, 86:30–43.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, et al. 2006. Annotation on the tectogrammatical level in the prague dependency treebank. *annotation manual. Technical Report*, 30:5–11.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtskỳ, Amir Zeldes, and Daniel Zeman. 2022. Corefud 1.0: Coreference meets universal dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872.

Tim O' Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. 2018. Amr beyond the sentence: the multi-sentence amr corpus. In *Proceedings of the 27th international conference on computational linguistics*, pages 3693–3702.

Massimo Poesio, Ron Artstein, et al. 2008. Anaphoric annotation in the arrau corpus. In *LREC*.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789.

Ina Rösiger. 2016. Scicorp: A corpus of english scientific articles annotated for information status analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1743–1749.

Ina Rösiger. 2018. Bashi: A corpus of wall street journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Kepa Rodriguez, and Massimo Poesio. 2016. Arrau: Linguistically-motivated annotation of anaphoric descriptions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2058–2062.

Renata Vieira and Massimo Poesio. 2000. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23:170.

冉晨. 2024. 现代汉语中数量名回指语的指称性质与回指确认方式. 语言教学与研究, pages 59–68.

刘东虹. 2014. 书面语篇中的抽象实体回指研究. 华中师范大学出版社.

吴丹苹. 2021. 基于语料库的现代汉语概述回指研究. Ph.D. thesis, 博士学位论文]. 杭州: 浙江大学.

周炫余, 刘娟, and 卢笑. 2014. 篇章中指代消解研究综述. 武汉大学学报: 理学版, (1):24–36.

孔芳, 葛海柱, and 周国栋. 2021. 篇章视角的汉语零指代语料库构建. 软件学报, 32(12):3782–3801.

宋宏. 2010. 人称代词语篇回指研究.

宋洋 and 王厚峰. 2015. 共指消解研究方法综述. 中文信息学报, 29:1–12.

寇鑫 and 徐坤宇. 2023. 抽象回指的指称内容与可及性研究——以“这”和“这件事”为例. 语言教学与研究, pages 88–100.

徐赳赳. 2003. 现代汉语篇章回指研究. 中国社会科学出版社.

曹娜 and 曹贤文. 2024. 汉语二语学习者主语回指语显隐的多因素分析. 语言教学与研究, pages 24–34.

李榕 and 王元鑫. 2021. 中高级阶段韩国留学生汉语篇章第三人称回指的习得研究. 世界汉语教学, 35:276–288.

杨永生 and 肖奚强. 2020. 韩国学生汉语“这/那”句习得考察. 华文教学与研究, (1):69–75.

王军. 2004. 英语叙事篇章中间接回指释义的认知研究. 苏州大学出版社.

王军. 2013. 英汉语篇间接回指. 商务印书馆.

王秀丽. 2012. 篇章分析中的概述回指. 当代语言学, 14(3):301–306.

石艳华. 2014. 认知激活框架下的汉语篇章回指研究.

胡霄钦 and 王秀丽. 2021. 法汉指称链条平行语料库的建设与应用. 语料库语言学, 8:112–121.

蒋平. 2021. 存现宾语的零形回指及其认知语言学解释. 山东外语教学, 42:22–30.

许余龙. 2004. 篇章回指的功能语用探索*:* 一项基于汉语民间故事和报刊语料的研究. 上海外语教育出版社.

陈远哲, 匡俊, 刘婷婷, 高明, 周傲英, et al. 2019. 共指消解技术综述. 华东师范大学学报 *(*自然科学版*)*, 5:16–35.