软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(12):3782−3801 [doi: 10.13328/j.cnki.jos.006119]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

篇章视角的汉语零指代语料库构建*

孔 芳[1,2], 葛海柱[1], 周国栋[1,2]

[1](苏州大学 计算机科学与技术学院 自然语言处理实验室,江苏 苏州 215006)

[2](江苏省计算机信息处理技术重点实验室,江苏 苏州 215006)

通讯作者: 周国栋, E-mail: gdzhou@suda.edu.cn

摘 要: 零指代是汉语中普遍存在的一个现象,在汉英机器翻译、文本摘要以及阅读理解等众多自然语言处理任务中都起着重要作用,目前已成为自然语言处理领域的一个研究热点.提出了篇章视角的汉语零指代表示体系,从服务于篇章分析的角度出发,首先以基本篇章单元为考察对象,判别其是否包含零元素;再根据零元素在基本篇章单元中承担的角色将零元素划分成主干类和修饰类两类;接着以段落对应的篇章修辞结构树为考察指代关系的基本单元,依据先行词与零元素间的位置关系将指代关系分成基本篇章单元内和基本篇章单元间两种,并针对基本篇章单元间的指代关系,根据零元素对应的先行词的状况将指代关系分成实体类、事件类、组合类和其他等 4 类;最后,基于篇章视角的汉语零指代表示体系,选取汉语树库 CTB、连接词驱动的汉语篇章树库 CDTB 和 OntoNotes 语料中重叠的 325 篇文本进行了汉语零指代的标注,构建了服务于篇章分析的汉语零指代语料库.一方面,借助系统检测来说明所提出的表示体系合理有效,构造的语料库质量上乘;另一方面构建了完整的汉语零指代消解基准平台,从可计算的角度验证了所构建的汉语零指代语料库能够为篇章视角的汉语零指代研究提供必要的支撑.

关键词: 零指代;语料库构建;篇章分析;基本篇章单元;零元素

中图法分类号: TP18

中文引用格式: 孔芳,葛海柱,周国栋.篇章视角的汉语零指代语料库构建.软件学报,2021,32(12):3782−3801. http://www.jos.org.cn/1000-9825/6119.htm

英文引用格式: Kong F, Ge HZ, Zhou GD. Corpus construction for Chinese zero anaphora from discourse perspective. Ruan Jian Xue Bao/Journal of Software, 2021,32(12):3782−3801 (in Chinese). http://www.jos.org.cn/1000-9825/6119.htm

Corpus Construction for Chinese Zero Anaphora from Discourse Perspective

KONG Fang[1,2], GE Hai-Zhu[1], ZHOU Guo-Dong[1,2]

[1](Laboratory for Natural Language Processing, School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

[2](Jiangsu Key Laboratory of Computer Information Processing Technology, Suzhou 215006, China)

Abstract: As a common phenomenon in Chinese, zero anaphora plays an important role in many natural language processing tasks, such as machine translation, text summarization and machine reading comprehension. Currently, it has become a research hotspot in the field of natural language processing. Towards better discourse analysis, this study proposes a representation architecture for Chinese zero anaphora from the discourse perspective. Firstly, the elementary discourse unit is taken as the investigation object to determine whether it contains zero elements. Secondly, according to the roles of zero elements in the elementary discourse unit, the zero elements are divided into two categories: the core type and the modifier type. Thirdly, the discourse rhetorical tree of the paragraph is used as the basic unit to evaluate the Chinese zero coreferential relationship. According to the positional relationship between the antecedent and the zero element, the coreferential relationship is classified into two types, i.e., Intra-EDU and Inter-EDU. After that, for Inter-EDU type, the coreferential

* 基金项目: 国家自然科学基金(61876118, 61751206); 江苏高校优势学科建设工程

Foundation item: National Natural Science Foundation of China (61876118, 61751206); A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD)

收稿时间: 2020-05-15; 修改时间: 2020-06-22; 采用时间: 2020-07-17

relationship is furtherly divided into four categories according to the status of the antecedent, i.e., entity, event, union, and others. Finally, this study selects the overlapped 325 texts of the Chinese treebank (CTB), the connective-driven Chinese discourse treebank (CDTB), and the OntoNotes corpus to annotate the Chinese zero anaphora. System evaluation shows the high quality of the constructed corpus for Chinese zero anaphora. Moreover, a complete zero anaphor resolution baseline system is constructed to show the appropriateness and the effectiveness of the proposed representation architecture for Chinese zero anaphora from computability perspective.

Key words: zero anaphora; corpus construction; discourse analysis; elementary discourse unit; zero pronouns

篇章中提及了某个事物后,当再次论及这个事物时,会采用各种方式来进行上下文的照应,这一现象称为回指(anaphor).当回指在篇章上没有任何的形式层体现时,就称这种回指为零指代(zero anaphor),它是一种特殊的指代现象.相比英文,汉语中零指代出现的频度很高,正如 Kim[1]所统计:汉语中在主语位置出现零指代的情况约占 36%,而英文主语位置的零指代不足 4%.正因如此,汉语零指代的研究对汉英文机器翻译、文本摘要以及阅读理解等众多自然语言处理任务意义重大,已成为自然语言处理领域针对汉语研究的一个热点.

例 1 给出了一个汉语零指代的具体实例,该实例摘自 OntoNotes 中文语料的 chtb_0009 文件,其中:零元素用“#”表示,位于相同指代链(即具有指代关系)的实体表述、零元素用相同颜色表示.

例 1:针对[甘肃]旅游业的发展需求,[人保公司]积极推出海外游客保险,$[\#]_1$ 还在国内首家推出海外散客保险办法,$[\#]_2$ 使“八五”期间到[甘肃]观光游览的海外游客全部得到保险保障.[甘肃省]还积极探索高风险业务,“八五”期间,$[\#]_3$ 参与卫星发射的共保,$[\#]_4$ 分担的风险金额达一千万元,$[\#]_5$ 支付赔款五百万元,$[\#]_6$ 成为西北首家参与航天业务的公司.

例 1 共包含 6 个零元素,其中:第 1 号、第 4 号~第 6 号零元素指代相同的实体“人保公司”;第 3 号零元素指代“甘肃省”;第 2 号零元素并不指代某个具体的实体,而是指代前文提到的“人保公司积极推出海外游客保险,还在国内首家推出海外散客保险办法”这两个举措.由此可以看到,完整的零指代消解任务由 3 个子任务构成:(1) 零元素识别,即识别出篇章语义上存在、但形式上未出现的那些“成分”;(2) 待消解的零元素识别,即根据零元素出现的篇章上下文来确定其是否回指某个具体的实体;(3) 零指代消解,确定待消解的零元素回指的具体实体,即先行词的确定.

众所周知,指代结构属于篇章的范畴.随着句子级词法、句法研究的日益成熟,特别是 MUC[2],ACE[3]以及 CoNLL-shared Task2011 和 2012 系列会议和比赛[2,3]相继开展,多语言的实体指代语料库日益丰富,与篇章密切相关的指代结构成为自然语言处理领域的研究热点之一.相比普通的实体指代消解,零指代消解任务更具挑战:首先,零元素在形式上不存在,没有任何显式的提示信息,需要从篇章语义的上下文中识别出这些隐式的“成分”;其次,汉语在形式的组织上相对松散,相同的语义成分可以以多种不同形式存在,这就造成了某些零元素可以出现的位置并不唯一;最后,因为形式上不存在,很多在实体指代消解中非常有用的词、数、性等特征都无法提取.因此,如何高效表征上下文特征成为研究的核心.要应对这些挑战,首先亟需解决的就是零指代结构体系及对应语料资源的构建问题.本文从服务于篇章分析的角度出发,对汉语零指代进行了表示体系的研究,并基于这一体系构建了中等规模的服务于篇章分析的汉语零指代语料库,为后续在篇章视角下开展汉语零指代消解的研究奠定了扎实的基础.

本文第 1 节介绍汉语零指代资源建设的相关研究.第 2 节对篇章视角汉语零指代的表示体系进行详细说明,并给出基于这一体系进行语料构建的标注规范和标注流程.第 3 节介绍了据此构建完成的语料资源.第 4 节以该语料为基础给出了一个完整的汉语零指代消解基准平台.第 5 节对本文的工作进行总结和展望.

1 相关工作

虽然在语言学领域,对汉语零指代现象已经进行了一些比较系统全面的理论研究[6,7],许多研究者从作者和阅读者的角度出发,以话题链为描述手段,归纳总结了汉语话题凸显的语言描述特点,并强调汉语中零指代现象广泛存在且没有任何约束,只能借助语义和语用知识,根据篇章中出现的信息进行零指代的解释.正是由于零指代的灵活多样且没有约束,相关语料资源的标注非常困难.因此,受限于汉语零指代语料资源,在计算语言学领

域的相关研究较少,主要包括以下 3 类代表性工作.

(1) 针对某一类或多类零指代现象自行构建小规模语料并进行可计算模型的探索.

典型的工作包括:Converse[8]在其博士论文研究中选取 CTB3.0 中的 205 篇新闻文本进行了第三人称代词和零指代的标注.Converse 的标注直接在句法树上进行,句法是进行各类现象判断的标准.此外,因为 Converse 的研究主要关注第三人称代词和零指代,对于不包含第三人称代词和零指代的其他指代关系并未进行标注,但各类指代现象间存在明显的互补性,孤立地进行两种类别指代的标注可能会隐藏部分重要信息.对标注结果的分析也发现,Converse 标注的零指代只涵盖了部分句法树上处于主语位置的零指代现象.使用 Converse 的语料,Zhao 和 Ng[9]首次提出了一个基于机器学习的汉语零指代方法,并探讨了这一任务的困难之处.他们将零元素消解分为零元素的识别和消解,通过与标准句法树进行对比,构建正例和负例作为训练实例,借助决策树来进行分类.但是由于正例和负例的分布严重不平衡,因此实验结果并不理想.Kong 等人[9]在研究了与汉语零指代相关的几种句法结构的基础上,选取 CTB6.0 中的 100 个文档进行了零指代的标注.Kong 的研究主要关注结构化句法信息对汉语零指代的影响,因此其标注也是以句法树为参考依据.与 Converse 不同的是:他们不仅标注了主语位的零指代现象,也考虑了宾语等其他位置.但 100 个文档的规模较小,也仅仅标注了零元素的先行词,零指代与普通实体指代间的关系并未进行标注.他们基于这一小规模的语料进一步提出了一个统一的框架进行零指代消解,将这一任务分为零元素识别、待消解项确定和零元素消解.不同于 Zhao 和 Ng[9]提出的基于特征的方法,他们使用基于句法树的方法,在零元素识别和消解上相较于 Zhao 和 Ng[9]都有了明显的提升.

(2) 将零元素看作句法分析中产生的空语类的一种,借助句法树中标注的空语类信息进行研究.

早期关于空语类的研究大都采用基于规则的方法.CAMPBELL[10]提出一种基于宾州树库的算法来恢复空语类.Chung 等人[11]在研究机器翻译时发现:在句子中添加空语类,可以有效提升翻译准确率.仿照 Johnson[12]和 Gabbard[13]的工作,他们使用基于模式的方法,通过统计语料发现:只有充当代词成分的零元素能够提高语料句法结构的完整性,并且可以帮助提升下游机器翻译任务的准确率.

随着机器学习技术的发展,很多学者开始尝试借助机器学习模型进行空语类的回复.Yang 和 Xue[14]提出组合词汇和句法信息进行空语类恢复,他们将空语类的恢复看作是序列标注问题,通过给空语类建立句法特征和词法特征,使用最大熵模型建立分类器,在每个词的后面判断是否有空语类.在标准句法树上性能较好,*F* 值达到 89%,但在自动句法树上,性能下降至 63.2%,以此说明空语类严重依赖句法信息.受 Yang 和 Xue[14]工作的启发,Cai 等人[15]将空语类的恢复集成到中文句法分析中,通过修改句法解析器,使得它可以用 WordLattice(字格)作为输入,并能够减少人工干预,自动恢复空语类.这使得在自动句法树上的性能较 Yang 和 Xue 有了一定的提升,F 值达到 67.0%.Kong 和 Zhou[16]提出了基于小句的空语类识别方案,认为局部句法信息的准确性相较较高.使用语义角色标注方法获得短句,针对终端短句,采用线性标注的方法;针对非终端短句,使用结构化分析的方法.此外,考虑到中文逗号意义丰富,为了提高短句识别的准确性,还加入了逗号消岐.中文空语类识别在自动句法树上的性能 *F* 值提升至 74.6%.Xiang 等人[17]将恢复空语类的问题转化为分类问题.考虑到空语类对句法结构有很强的依赖性,他们将空类型标签删掉,并将空语类的位置信息和类别信息转移到句法树上层节点,引入句法树特征、词法特征以及空语类特征,借助最大熵模型对预测为包含空语类信息的节点进行恢复.Xue 等人[18]首次引入依赖关系,使用空语类的头信息和后一个词组成训练实例,成功解决 Yang 和 Xue[14]给出的序列标注无法识别连续多个空语类的问题.Zhou 等人[19]通过实验发现:在句法分析中加入空语类标签,能够有效提升准确率.为了更好地描述空语类,他们将空语类标签重新定义,并提出了基于规则、句法分析以及依赖关系的 3 种方法.实验结果表明:使用新的空语类标签后,句法分析准确率明显提高.但空语类表达是成分间的句法依赖关系,与篇章层的零指代存在一定的差异.

(3) 在 OntoNotes 语料上进行零指代研究.

语料资源方面,得到大众认可的汉语零指代语料是 OntoNotes 语料[20].该语料的中文部分标注了主语位置的零元素及其所属的指代链情况,为目前的汉语零元素研究工作提供了资源支持.与前面小规模的语料标注相比,该语料的规模扩大很多,但仍然是基于句法信息的零指代资源.使用 OntoNotes 语料,一些研究者展开了零指

代可计算性的相关研究.典型工作包括:包含零指代识别和消解两个子任务,Chen 等人[21]第 1 个给出了完整的端到端的汉语零指代消解平台,并给出一组有效的句法和上下文特征,借助这些特征实现了全自动的零指代分析.深度网络技术的推进,各类向量嵌入工作的开展,Chen 等人[22]基于深度神经网络模型,将字法、词法、句法等许多已经验证有效的特征以向量嵌入的方式融入零指代消解,以此构建了一个神经网络框架,一定程度上提升了零指代消解的性能.但他们的工作也验证了,零指代消解的性能受到句法分析性能的严重影响.相比标准句法树,在自动句法树下的端到端的汉语零指代消解的 *F* 值下降了近 42%.如何提升自动句法树下零指代的性能,成为了关注焦点.Yin 等人[23]在 Chen 等人[22]的基础上对神经网络模型做了拓展,给出了一个深度记忆网络,利用两个编码器对先行词进行局部编码和全局编码,获取先行词的局部特征和全局特征,再对零代词用上下文向量表示来获取其上下文特征.为了更好地描述零代词,除了零代词的上下文信息外,还引入了候选先行词特征,通过词嵌入获取向量之间的语义特征,并为每一层网络加上注意力信息,实验结果证明了该方法的有效性.Zhang 等人[24]也尝试通过将特征向量化的方式来更好地表征先行词候选以及零元素和先行词候选的上下文语义信息,再借助神经网络模型进行零指代消解.Yin 等人[25]在高效表征各类信息的基础上,还向神经网络平台引入了强化学习策略,通过进一步提升神经网络的学习能力来提升汉语零指代消解的性能.Kong 和 Zhou[26]提出零指代不应该被孤立对待,而是应该与普通名词短语的消解形成完整的整体.基于此观点,他们提出了一种全新的链到链的汉语零指代消解方案:首先将零元素聚类为共指链,每条共指链都作为一个独立的指代词,这样,那些距离较远的零元素和先行词可以通过共指链的传递性进行链接;其次,名词短语也被聚类成不同组,每一个组作为一个先行词独立存在.通过将普通名词短语的指代消解结果看作是对先行词候选进行过滤的一种手段,以指代链为单位进行汉语零指代消解,这样大大减少了搜索空间,使得零指代消解的性能明显提升.

从上述相关研究可以看到,语料资源是开展汉语零指代可计算研究不可或缺的条件.为了降低对标注语料的依赖,一些研究者也开展了各种相关研究.为了解决对标注语料的依赖,Chen[27]提出了一种无监督的方法,借助最大熵构建一个候选先行词排序模型,在包含显性代词的语料上训练得到模型参数后,将其应用到零代词消解上.实验结果表明,该方法取得了比监督模型更好的消解效果.但是该方法的局限性在于:他们并没有研究零代词的识别,提出的模型只能在零代词已经正确识别的基础上进行消解.在此基础上,为了更好地描述先行词特征,Chen[28]又提出一种非监督概率模型,为先行词加入了 4 个语法特征:Number(数量特征),Gender(性别特征),Person(人称特征)和 Animacy(有生性特征),并使用 EM 算法[29]来推测最可能的先行词.为了考虑篇章特征,Chen[30]使用 SalienceModel(凸显模型)为每一个有效实体计算得分,并采用联合的方式识别和消解零代词.Liu 等人[31]为了解决零指代标注语料不足这一问题,将对零指代消解的方法由分类模型转化为阅读理解模型.利用大量的伪语料训练阅读理解模型,并将此模型应用在零指代消解上.但他们的工作也针对零元素的消解进行了研究,提出的基于注意力机制的神经网络模型也只适用于零元素已知的情况.

随着一定规模的 OntoNotes 语料库的发布,汉语零指代消解研究日趋活跃.不过,指代属于篇章级的语言现象,从句法视角构建汉语零指代的结构体系存在着一些明显的问题,正如 Yang 和 Xue[14]分析实验结果得出的结论:仅关注句法信息,中文零元素的判别与句法层的共享主语现象间很难区分.另一方面,随着篇章分析相关研究的展开,研究者开始意识到篇章层的信息对于零指代消解意义重大.例如,Sheng 等人[32]在传统零指代消解平台中考虑了篇章修辞结构信息,在零元素识别、零元素消解等多个环节,都通过提取各类篇章级的信息来提升性能.相应地,也有一些研究表明,零指代对于中文篇章分析意义重大.例如,奚雪峰等人[33,34]提出一种基于主述位理论的篇章微观话题结构,其中的隐式主述位本质上就是零元素,它们在话题链的形成中意义重大.因此,本文提出从服务于篇章分析的视角来构建汉语零指代的体系结构.

2 篇章视角的汉语零指代表示体系

研究者普遍认为:各语义成分是由驱动谓词管辖的,语义成分的缺省(零元素)可以通过“谓词驱动”这一方式进行识别.例如:Cai 等人[15]尝试在句法分析的过程中,依据驱动谓词进行空语类的识别;Kong 和 Zhou 等人[16]提出,借鉴简化的语义角色标注(sematic role labeling,简称 SRL)方法识别子句,再以子句为单位进行空语类和零

元素的识别.不过,本质上零元素并不是单纯的“缺失的语义成分”,而是在上下文衔接中缺失的有意义的语义成分,需要根据上下文进行判断.另外,汉语重意合的特点使得汉语表达更加灵活,许多固定句式虽然从谓词驱动的角度似乎存在语义成分的缺失,但从整体表达的语义信息看又不存在缺失.

例如,例 2 所示的句子包含 3 个谓词:“防止”“出现”和“出台”.其中,“出台”的各语义成分都齐全,未出现任何省略;“防止”的施事者“新区管委会”在篇章后面提及了,可以认为在“防止”前存在一个语义省略,后文进行了恢复;“出现”的施事者在文中并未提及,说明这一语义对象并非当前篇章关注的焦点,不存在上下文衔接中有语义成分缺失.

例 2:为防止出现无序现象,新区管委会及时出台了一系列规范建设市场的文件.

汉语重意合的特点,决定了汉语零指代表示体系的确立必须从篇章的视角进行.从形式上看,零元素是句子中省略的某个成分;而从语义理解的角度看,省略的这个成分一定包含明确的语义信息,承担了一定的语用功能,即这个语义成分是依赖于篇章的上下文表述的,是衔接上下文的特殊语义载体.

需要特别说明的是:盛晨等人[35,36]提出从篇章视角分析汉语零指代,他们从篇章视角将零元素分成主干型和修饰型两大类,同时又根据零元素所处篇章基本单元的句法结构将零元素细分成若干小类.但他们的工作存在两方面缺陷:首先,大类的划分是篇章视角的,而小类的划分是句法层面的,句法虽然利于语料标注质量的控制,但从分类体系的角度,两种视角存在一定的冲突;其次,盛晨等人[35,36]仅对零元素的分类体系进行了研究,但篇章中更重要的是衔接上下文的零元素,离开指代关系独立分析零元素对服务篇章的支撑是有限的.受盛晨等人工作的启发,葛海柱等人[37]进一步梳理了篇章视角的零指代结构.基于盛晨和葛海柱等人的工作,我们从服务于篇章分析和文本理解的目标出发,我们构建了完整的篇章视角的汉语零指代结构体系,它由篇章视角的零元素分类体系和篇章视角的零指代结构两部分构成,下面分别加以说明.

2.1 篇章视角的零元素分类

在汉语篇章微观修辞结构表示体系[38,39],将基本篇章单元(elementary discourse unit,简称 EDU)定义成至少包含一个谓语部分,即至少表达一个命题,认为 EDU 是篇章构成的基本单位.从服务于篇章分析的目标出发,我们将 EDU 看作考察是否包含零元素的基本单元.与盛晨等人[35,36]的工作类似,依据 EDU 内是否存在缺失的语义成分,以及缺失的语义成分在 EDU 中是否承担主干成分,我们将零元素划分成两大类,即主干型零元素和修饰型零元素,但不再进行小类的区分.

以例 3 所示的句子为例,从篇章分析的视角看,它由 3 个基本篇章单元构成,图中用“[·]”进行分割,分别记作 e1,e2 和 e3,这 3 个基本篇章单元构建形成的修辞结构树如图 1 所示.

例 3:[国家统计局预测,一九九六年全球经济将继续保持增长,]$_{e1}$ [*这种良好的态势*对中国的发展十分有利,]$_{e2}$ | [φ使其面临很多发展机遇.]$_{e3}$

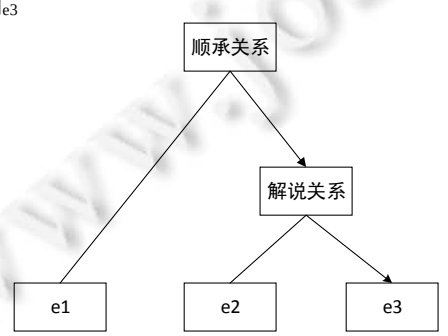

Fig.1 Discourse rethorical structure tree of example 3

图 1 例 3 中各基本篇章单元形成的修辞结构树

可以看到:基本篇章单元 e1 和 e2 在语义成分上是完整的,不存在零元素;但对于 e3 而言,独立观测这一单元,它表达了两层含义:一是“其(中国的发展)面临很多发展机遇”,二是“这种良好的态势造成了其面临很多发展机

遇”.其中,第 1 层含义各语义构成成分完整,不存在零元素;第 2 层含义中的施事者“这种良好的态势”缺失了,因此存在一个零元素,即例 3 的 e3 中所示的“φ”,它指代前一个 EDU 中提及的“这种良好的态势”,形成了一个零指代关系.在这两层含义中,主干语义是第 2 层含义,即“这种良好的态势使得其面临很多发展机遇”,零元素承担了 EDU 内主干语义成分的角色,属于篇章主干型零元素.

例 4 给出了一个包含两个 EDU 的句子示例,这两个 EDU 形成了因果关系.其中,第 2 个基本篇章单元 e2 表达的主干语义信息是“大量出现的是新情况、新问题”,而“以前不曾遇到过的”是“新情况、新问题”的修饰成分,但在这修饰成分中,谓词“遇到”的施事者被省略了,它指代的是前一个 EDU 中出现的“浦东”.因为此处的语义缺省出现在修饰成分中,我们将这一零元素归为修饰型零元素.

例 4:[*浦东*开发开放是一项振兴上海,建设现代化经济、贸易、金融中心的跨世纪工程,]$_{e1}$ [因此大量出现的是φ以前不曾遇到过的新情况、新问题.]$_{e2}$

对比例 3 和例 4 我们可以看到:相比篇章主干型零元素,篇章修饰型零元素对 EDU 内部语义成分间的关系抽取以及局部句法分析的影响较大,它们的存在与 EDU 内部的句法结构,甚至是某一短语内的句法结构关系密切,对 EDU 之上的粒度更大的篇章分析的影响相对较小.但很明确,准确识别修饰型零元素将有助于明确局部语义成分,帮助更好地表征 EDU,从而减少复杂的修饰成分对篇章理解带来的噪声.从可计算的角度考虑,篇章主干型零元素与篇章的衔接性和连贯性关联更大,在这类零元素的自动识别方面,应更多地考虑篇章层面的信息;修饰型零元素更多与 EDU 内部的局部句法信息关系密切,这类零元素的自动识别应更多地考虑句法信息的支撑.

对于零元素的标注还存在定位问题.所谓零元素,是形式上不存在,而语义上存在的某个成分.另外,人为对其进行形式上的添加存在位置的不唯一性.就例 4 给出的例子看,人为将零元素插入在“以前”这一修饰语的前面或后面都可以,具体参见例 5 给出的两种插入结果.零元素的先行词是“浦东”,对于 e2 这个 EDU 而言,语义补全后,“大量出现的是浦东以前不曾遇到过的…”和“大量出现的是以前浦东不曾遇到过的…”,从句法和语义层都是合理的.

例 5:

(1) [*浦东*开发开放是一项振兴上海,建设现代化经济、贸易、金融中心的跨世纪工程,]$_{e1}$ [因此大量出现的是φ以前不曾遇到过的新情况、新问题.]$_{e2}$

(2) [*浦东*开发开放是一项振兴上海,建设现代化经济、贸易、金融中心的跨世纪工程,]$_{e1}$ [因此大量出现的是以前φ不曾遇到过的新情况、新问题.]$_{e2}$

对于上述情况,为了保证语料标注的一致性,对零元素的位置出现多个可选时,要求统一定位在可选的首号位置.当然,在进行可计算研究时,在评测中可考虑在忽略零元素前后的连词和修饰成分的基础上进行位置是否正确的判定.

2.2 篇章视角的零指代结构

零指代结构关注零元素与其先行词之间的关联关系.篇章视角的零指代结构需要从篇章层确定指代结构的几个核心要素,具体包括:

(1) 指代关系的考察范围

指代描述的是篇章层的语言现象,实体指代关系遍布整篇文章.不过,已有的研究(特别是对代词作为待消解项的研究[40])发现,其先行词通常在当前句或前两句.零元素是形式上省略、而读者可以根据上下文进行语义恢复的对象,其聚焦性强于代词.因此,与零元素关联的先行词通常不会与零元素跨越很远.基于这一原则,我们将零指代结构的考察范围限定在相同段落内.由于汉语微观篇章修辞结构[38,39]将每个段落映射成一棵独立的篇章修辞结构树,因此我们将零指代结构的考察范围限定在零元素所在的篇章修辞结构树中.

在上下文中承担了衔接作用的零元素,这类零元素的先行词一定显式地在上下文中出现过.为了从语义层更好地确定零元素指代的先行词,我们首先将实体指代链作为考察对象,确认当前零元素指代的是哪一个实体指代链.众所周知:指代结构并不是两个表述之间的关系,而是若干个表述之间的关系.将零元素关联到具体的

实体指代链,一方面可以充分利用已有的实体指代的标注信息和端到端的自动实体指代消解工具;另一方面,也可以较为容易地对零元素是否在上下文中承担了衔接作用进行准确地判断.当然,在标注过程中可以根据语义选择同一指代链上的任意一个表述进行指称关系的标注,最终的先行词是由该表述对应的指代链来表示的.若不存在某个实体指代链与当前零元素间有指代关系,我们再进行短语级别的其他指代对象的考察.

(2) 指代关系的分类

从服务于篇章的视角,我们从两个维度对零指代关系进行了分类.

一是根据指代关系是否跨越 EDU 将零指代关系分成 EDU 内(intra-EDU)和 EDU 间(inter-EDU)两种,其中:inter-EDU 类型的指代关系发生在两个不同的 EDU 间,衔接的上下文更多的是篇章层对象;而 intra-EDU 类型的指代关系发生在 EDU 内部,受到局部句法信息的影响更大.例 6 给出了一个 EDU,其中包含一个修饰型零元素,其指代的对象是该 EDU 的主干主语"浦东",这一指代关系在 EDU 内部完成,属于 intra-EDU 类型.可以看到,intra-EDU 类型的指代关系中涉及的零元素一定是修饰型零元素.例 7 给出了一个 inter-EDU 类型的零指代关系示例,该例子涉及相邻的两个 EDU,这两个 EDU 之间是并列关系,其中,第 2 个 EDU 的主干主语缺省,指向第一个 EDU 的主干主语.

例 6:[*浦东*不是简单的采取"干一段时间,等φ积累了经验以后再制定法规条例"的做法,]$_{e1}$

例 7:[*这个开发区*位于中国著名风景旅游城——杭州市区内,]$_{e1}$ [φ是一九九一年国务院批准建设的国家级高新技术产业开发区.]$_{e2}$

二是将 inter-EDU 类型的指代关系,根据指代关系关联的对象是实体、事件还是其他抽象概念,分成了以下 4 种.

- EntityType:零元素指代前面提到的实体.例如:在例 7 中,第 2 个 EDU 中的零元素指向前一个 EDU 中提及的实体"这个开发区";
- EventType:零元素指代前面提到的事件,而不是某一个实体.例如:例 8 中包含两个 EDU,后一个 EDU 中包含一个零元素,而它指代的正是前一个 EDU 提及的事件;
- UnionType:零元素指代前面提到多个事件或实体.如例 9 给出的例子,该句子包含 3 个 EDU:前两个 EDU 间构成了并列关系,再与第 3 个 EDU 构成了递进关系.在第 3 个 EDU 中存在一个主干成分的缺失,而这一零元素从语义上指代前面的"从业人员"和"私营企业注册资金"两个实体;
- RETType:零元素指代的单元位于此零元素后面或者未显式出现的某个抽象概念.例 10 给出了一个先行词在待消解项后面出现的示例.

例 8:[但*全民公决不接受这一方案*,]$_{e1}$ [φ也就终止了整个进程.]$_{e2}$

例 9:[*从业人员*有九万七千九百六十三人,]$_{e1}$ [*私营企业注册资金*达到了三十亿零八千多万元,]$_{e2}$ [φ分别比去年同期增长一成至两成.]$_{e3}$

例 10:[φ为了造福社会,]$_{e1}$ [*王码电脑公司*毅然放弃本来可以赚大钱的机会.]$_{e2}$

(3) 指代关系的标注位置

实体与事件之间是可以相互指代的,从服务于篇章理解,进行实体和事件的统一指代消解为目标,在进行零元素指代关系构建时,我们参考 Proposition Bank 中语义角色标注(semantic role labeling,简称 SRL)的标注策略,将指代的先行词定位成篇章修辞句法组合树中对应的某个结点.

篇章修辞句法组合树是指以段落为单位,将每个段落映射成一棵独立的树.该树由两部分组合而成:以 EDU 为基本单位,向上通过篇章修辞关系构建形成修辞结构树;再针对每个 EDU,抽取其对应的句法树或句法子树.例如例 3 所示的一个篇章片段,图 1 给出了其对应的篇章修辞树,将其叶子结点对应 EDU 细化成句法子树就形成了图 2 所示的篇章修辞句法组合树.

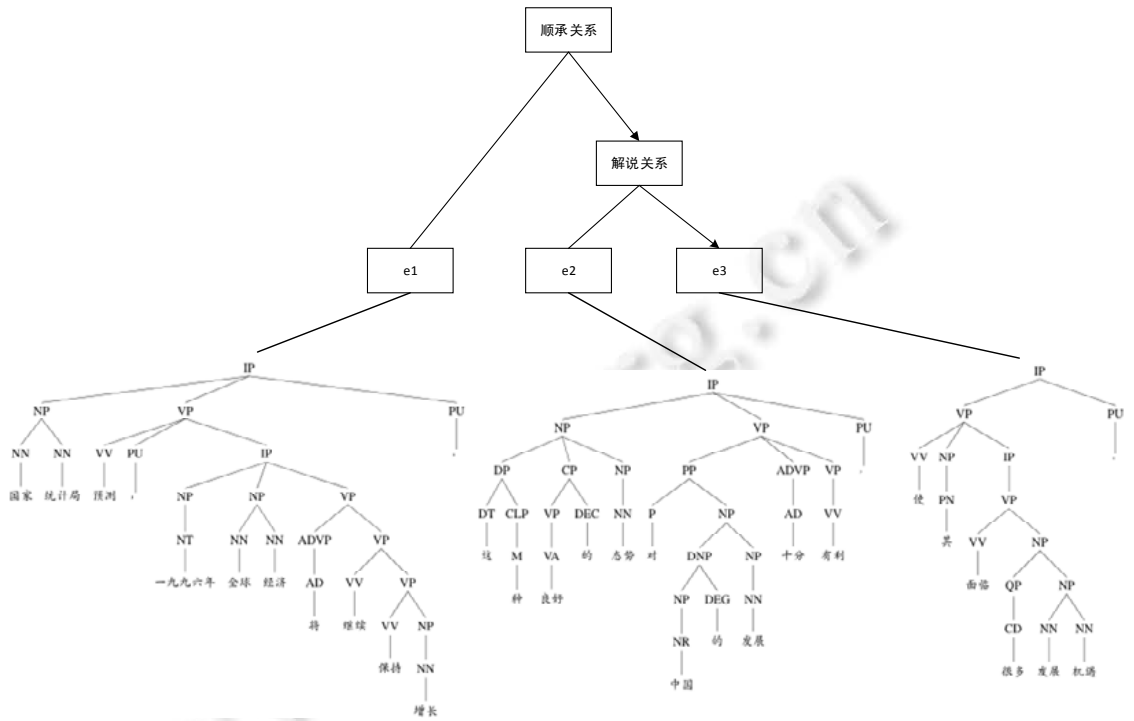

Fig.2 Discourse rethorical and syntatic combination tree of example 3

图 2 例 3 构建形成的篇章修辞句法组合树

若零元素指代的是某个实体,其距离最近的表述形式为一个名词短语,该短语将被映射到修辞句法组合树中的一个对应的结点.实际上,名词短语不会跨越 EDU,因此它是句法子树中的某个结点.例如:图 3 给出了例 7 中第 1 个 EDU 对应的句法子树部分,而先行词“这个开发区”与子树中方框扩起的“NP”结点对应,该结点可以通过起始叶结点的序号与从该结点向上的层次数的形式进行组合定位,其中,叶结点的序号是从整个篇章的角度进行编号(例 7 中的“这个开发区”得到的标注信息是:0+3).

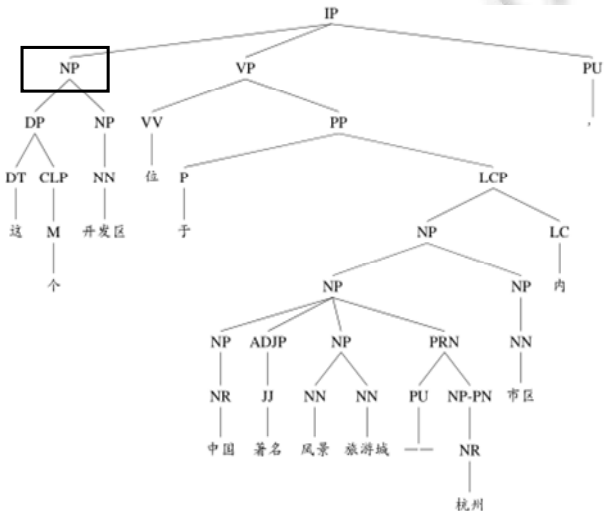

Fig.3 Syntatic subtree of the first EDU in example 7

图 3 例 7 中第 1 个 EDU 对应的句法子树

若零元素指代的是某个事件,其距离最近的事件表述将被标注成先行词.在各种事件抽取任务中,事件表述

被定义为事件触发词与事件论元的组合.我们选取修辞句法组合树中涵盖事件触发词及论元的层次最低的结点作为该事件表述对应的结点.例如:图 4 给出了例 8 中第 1 个 EDU 对应的修辞句法组合树的句法子树部分,而先行词是“全民公决不接受这一方案”这一事件,触发词是“接受”,涉及的论元有“全民公决”“这一方案”,根据这些信息可再定位到图 4 中方框扩起的“IP”结点是该事件表述对应的结点,同样采用起始叶结点在篇章中的序号与向上的层次数的形式来唯一定位该结点.

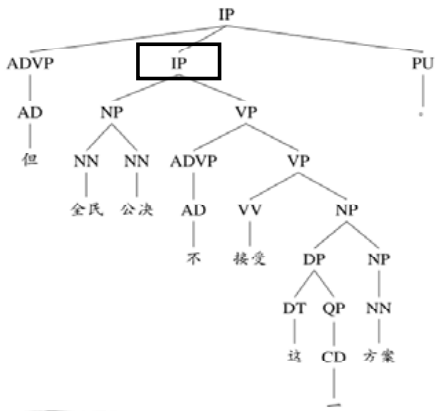

Fig.4 Syntatic subtree of the first EDU in example 8

图 4 例 8 中第 1 个 EDU 对应的句法子树

当零元素指代的是多个事件或实体的组合时,分别找到各个实体和事件对应的篇章修辞句法组合树中的结点,再向上找寻它们共同的最低父结点,将该结点作为映射得到的结点.例如:例 9 对应的篇章修辞句法组合树如图 5 所示,先行词涉及两个实体,它们分别对应句法子树部分方框扩起的两个 NP 结点,再向上找到最低的父结点是圆形扩起的“并列”结点.

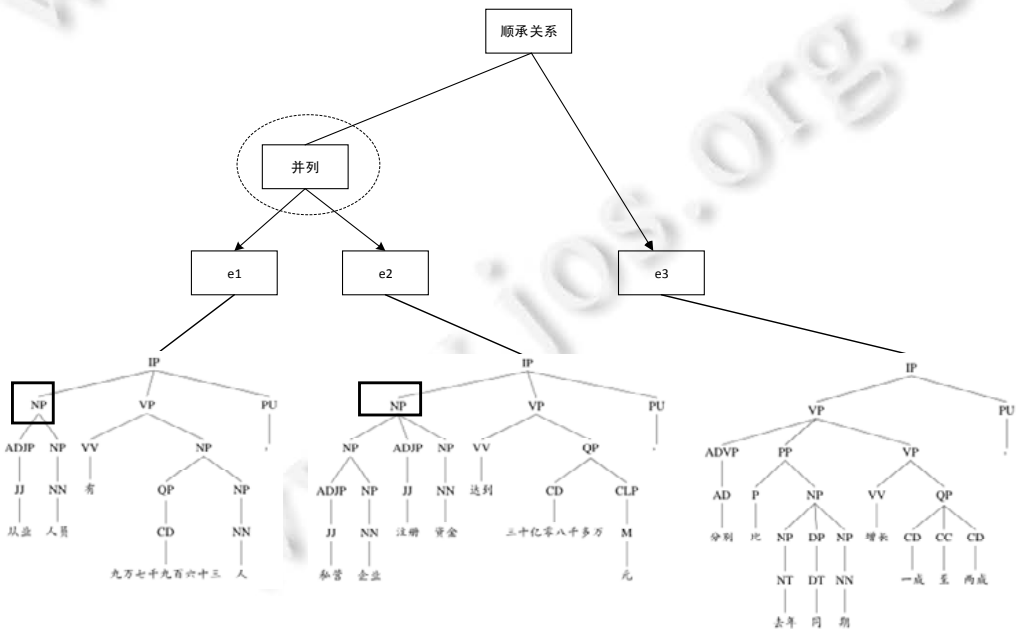

Fig.5 Discourse rethorical and syntatic combination tree of example 9

图 5 例 9 对应的篇章修辞句法组合树

可以看到:通过上述方式,我们可以统一的进行多种先行词的标注.

3 篇章视角的汉语零指代标注规范的制定和语料构建

3.1 文本数据的准备

我们选取宾州汉语树库(Chinese treebank,简称 CTB)[41]中的前 325 篇(chtb0001~chtb0325)文本进行零指代结构的标注,标注的同时进行了成分句法结构、实体指代结构和篇章修辞结构的融合.

CTB 语料由 LDC 正式发布,在 NLP 领域的很多任务中都有广泛应用,经过多年的积累,已经包含句法、浅层语义、可比较语料、实体指代消解等多方面的标注信息.首先,CTB 语料提供了标准的成分句法分析结果,为构建篇章修辞句法组合树奠定了句法部分的基础;其次,OntoNotes 语料给出了实体指代结构、语义角色标注等多方面的信息,其 NW 部分涵盖了 CTB 的这 325 个文档,为零指代结构与普通的实体指代结构的融合奠定了基础;最后,苏州大学自然语言团队发布的基于连接词驱动的篇章树(connective-driven discourse tree,简称 CDT)结构的汉语篇章树库(Chinese discourse treebank,简称 CDTB)[38,39]中也涵盖了这 325 个文档,为从篇章视角进行零指代结构的标注提供了篇章体系结构的支撑.根据其标注的标准段落、句子信息以及 CDTB 中标注的标准 EDU 信息进行统计,该语料总共包含 1 367 个段落(即 1 367 棵篇章修辞结构树),4 098 个句子,6 628 个 EDU.

3.2 规范制定和标注过程

标注工作分为 3 个阶段.

- 第 1 阶段确定初步的标注规范,并设计开发相应的标注平台.这部分工作的主要参与者是对可计算有一定理解的资深语言学家,在大量生语料分析的基础上,同时考虑语料标注的质量以及通用性,充分讨论的基础上形成初步的标注规范.然后对将要参与标注的人员进行初步培训,确保他们真实理解这一规范;
- 第 2 阶段是预标注阶段,主要希望通过实践来确认参与标注的人员对规范的理解,同时检验规范的可实施性,并在标注过程中对规范进行微调,并得到最终的标注规范;
- 第 3 阶段是正式标注和质量保证阶段.根据最终的标注规范完成所有文档的标注,对最终的标注文档逐一校对,通过一致性分析确定分歧较大的语篇,以讨论的形式进行修正或删除不合理项,形成完整的可发布的中文篇章零元素语料库.

篇章视角的零指代结构的标注是在以段落为单位的篇章修辞句法组合树上进行,以给定的实体指代链为辅助信息.标注过程分 3 步进行:(1) 零元素及其类型的确定;(2) 先行词的确定;(3) 指代关系类型的确定.

为了简化工作量、提高标注效率以及标注一致性,我们首先将标注工作流程化,在恰当的场合提供必要的辅助信息.给定文本后,从 CTB,CDTB 中提取句法和篇章修辞信息,以段落为单位,构建形成篇章修辞句法组合树.当用户确定当前段落后,进入标注的 3 个阶段.

(1) 在零元素及其类型的确定阶段,EDU 是零元素确定的基本单位,篇章间的修辞结构或 EDU 内的局部句法信息是确定零元素类别的参考依据.因此,将段落以切分好的 EDU 为单位进行篇章修辞结构的展现,在标注者指定相应的 EDU 后,再进一步展现 EDU 对应的句法子树,让标注者依据相关信息进行零元素及其类型的确定;

(2) 设定零元素后,进入先行词的确定环节.完整展现零元素前对应篇章修辞句法组合树的内容,同时读取 OntoNotes 中标注的实体指代关系,并将相关的表述映射到修辞句法树中的各结点,在用户进行先行词对应结点选择时,进行实体指代信息的提示;

(3) 选定先行词后,根据 EDU 跨度情况自动确定是 inter-EDU 还是 intra-EDU 类型,同时让用户确定指代对象的类型.根据用户指定的类型信息,结合零元素位置(段落中第几个词的前面,词的划分以修辞句法树中的叶节点为标准)和类型,先行词对应的结点,形成完整的指代结构信息,将这些信息以独立的 XML 文件格式保存.

根据上述标注流程,我们设计并开发实现了篇章视角的零元素标注平台,平台的基本工作流程如图 6 所示.从工作流程可以看到:在标注过程中,标注者对于零元素的位置、类型、先行词的结点以及指代关系的类型等

信息的确定均以“选择”动作为主.此外,通过标注平台将一些不可能的位置屏蔽,设定一些必要的约束,例如零元素不能出现在某个词的内部,一个 EDU 最多只能有一个主干型零元素等,以此来保障标注质量,提升标注结果的一致性.

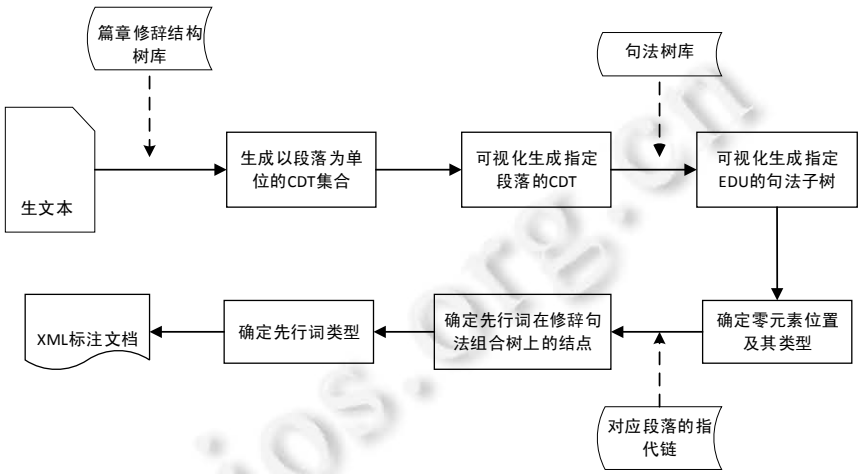

Fig.6 Annotation procedure of chinese zero elments from discourse perspective

图 6 中文篇章零元素标注平台的基本处理流程图

最终形成的 XML 格式的标注信息如图 7 所示.每一个零指代关系对应形成一个 ZLink 标签,而 ZLink 标签中,EDUType 表明的是当前这一指代关系是 EDU 内部的,还是跨 EDU 的;ANTType 表明的是先行词属于哪种类型,具体对应第 2.2 节中给出的类别信息;ID 是以段落为单位顺序递增方式的序号.在每个 ZLink 中包含两个基本元素,即零元素和其指代的距离最近的先行词.零元素给出的是位于哪个词的前面,position 记录的是这个词在段落中的序号,type 用于表明零元素是主干型还是修饰型零元素.先行词则通过起始位置(position)和层次(level)定位了篇章修辞句法组合树中对应结点的状况,若先行词是 OntoNotes 中已标注的实体链上的某个表述,则 EntityID 用于记录这个指代链的序号.

```
〈ZLink ID=“·” EDUType=“inter/intra” ANTType=“Entity/Event/Union/RET”〉
    〈Zero position=“idx” type=“Main/Modify”/〉
    〈Antecend position=“idx” level=“idx” EntityID=“·”/〉
〈/ZLink〉
```

Fig.7 Annotation result in XML format of chinese zero anaphora structure

图 7 汉语零指代结构对应的 XML 标注

3.3 标注语料一致性评价及分析

本文采用语料标注领域大家广泛接受的 Kappa 检验[42]进行一致性检验,以此来评估语料标注的质量.Kappa 计算公式如下:

$$Kappa = \frac{P_O - P_C}{1 - P_C} \tag{1}$$

其中,P_O 表示观察一致率,P_C 表示偶然一致率.通常认为:Kappa 值大于 0.75,则表示标注具有较好的一致性;如果 Kappa 值小于 0.4,则表示一致性较差.

从标注语料中随机抽取 30 篇文档,再选取两名标注人员对它们进行独立标注,再根据标注结果计算标注的一致性.汉语零指代语料的一致性主要包括以下 4 个方面.

(1) 零元素位置的一致性:以 EDU 为单位,当标注零元素的在 EDU 内部的位置相同时,认为零元素标注是一致的;

(2) 零元素类别的一致性:当零元素位置一致,再检测主干型和修饰型类别是否一致;

(3) 先行词的一致性:如果标注的先行词位置相同,认为标注的先行词一致;此外,当标注的零元素先行词具有 EntityID,且 EntityID 相同,即使先行词位置不同(也就是选取了相同实体链上不同的表述作为其先行词),我们仍然认为这个标注是一致的;

(4) 指代关系类型的一致性:当先行词一致,再检测指代的类型 Entity,Event,Union 和 RET 是否一致.

通过计算,本语料的零元素位置标注的一致性的 Kappa 值为 0.88,零元素类别标注的一致性 Kappa 值为 0.85,先行词的标注一致性的 Kappa 为 0.82,指代关系类型的一致性 Kappa 值为 0.81,4 个指标均超过了 0.8,表明该语料的标注质量可靠.

3.4 语料规模的统计说明

篇章视角的汉语零指代语料库共包含 325 篇文档(chtb0001~chtb0325),全部来源于 CTB 语料,我们共标注了零指代链 2 672 个,平均每个段落包含零指代关系 1.95 个.因为标注过程中进行了约束,每个 EDU 最多只有一个主干型零元素,而实际上包含多个零元素的 EDU 极少,只出现 2 个,可以看到,包含零元素的 EDU 约占 EDU 总数的 40.31%.

下面对篇章视角的汉语零指代语料库中零元素的分布情况以及指代链的分布情况进行了统计分析.

(1) 零元素的段落分布

基于段落对零元素分布进行统计,对应结果见表 1.可以看到:在所有的 1 367 个段落中,不包含零元素的段落仅占总数的 31.09%,有 425 个段落.也就是说,汉语篇章表述中,约有 68.91%的段落中存在零元素.这也说明了汉语中省略是普遍存在的,汉语零指代是汉语的重要特效之一.

Table 1 Zero elements distribution over paragraphs

表 1 以段落为单位包含零元素数量的分布统计

零元素个数	数量	比例(%)
m=0	425	31.09
m=1	417	30.50
m=2	250	18.29
m=3	131	9.58
m=4	59	4.32
m=5	35	2.56
m=6	19	1.39
m=7	17	1.24
$m\geqslant 8$	14	1.02
Overall	1 367	100

(2) 零元素的类别分布

针对零元素类别分布进行统计,其分布结果见表 2.可以看到:主干型零元素(Main)占据了绝大部分,其比例高达 80.16%,这部分零元素对篇章语义的理解以及篇章层的分析起到至关重要的作用;剩余的修饰型零元素所占比例约为 19.84%,该部分主要关联的是 EDU 内部的细节语义,能辅助局部句法和语义分析,在后续的研究中依旧存在不可替代的作用.

Table 2 Zero elements distritution over categories

表 2 零元素类别分布统计

零元素类别	数量	比例(%)
Main	2 142	80.16
Modify	530	19.84
Overall	2 672	100

(3) 零指代链的类别分布

表 3 给出了零指代链在 EDU 内和跨越 EDU 这两种情况的数量及比例,可以看到,跨越 EDU 的零指代关系占到了绝大多数.这也进一步说明指代是篇章层面的特性,是篇章衔接性的一种体现.

Table 3 Zero anaphora distribution over distances

表 3 零指代关系的距离类别分布统计

指代关系类别	数量	比例(%)
Intra-EDU	177	6.62
Inter-EDU	2 495	93.38
Overall	2 672	100

我们对 Inter-EDU 类型的零指代关系进行了进一步的类别统计,表 4 给出了按先行词类别进行统计得到的数量分布.从表 4 所示的结果可以看到:先行词是 Entity 类别的情况占到了绝大多数,约为 94.91%.对这类零指代进行进一步统计发现,先行词是 OntoNotes 中已标注的某个实体指代链的零指代链有 2 188 个,约占实体类零指代的 92.41%;还有 180 个零元素的先行词是由未构成实体指代链的独立名词短语承担,约占实体类零指代链的 7.60%.

Table 4 Inter-EDU zero anaphora distribution over types

表 4 Inter-EDU 类型的零指代关系的指代类别分布统计

指代关系类别	数量	比例(%)
Entity	2 368	94.91
Event	55	2.21
Union	27	1.08
RET	45	1.80
Overall	2 495	100

(4) 跨 EDU 的零指代链的距离分布

表 5 给出了 Inter-EDU 类型的零指代关系跨 EDU 数量的分布情况.从统计结果可以看到:零指代关系跨度小于等于 3 个 EDU 的情况占到了总情况的 92.71%,而超过 3 个 EDU 的零指代关系通常为 Entity 类型.

Table 5 Inter-EDU zero anaphora distribution over distances

表 5 Inter-EDU 类型的零指代关系的距离分布统计

跨越 EDU 的数量	指代关系的数量					比例(%)
	Entity	Event	Union	PER	Overall	
1	1 624	51	8	35	1 718	68.86
2	412	3	15	6	436	17.47
3	155	1	1	2	159	6.37
4	71	0	1	1	73	2.93
5	32	0	0	1	33	1.32
≥6	74	0	2	0	76	3.05
Overall	2 368	55	27	45	2 495	100

3.5 与OntoNotes中标注的零指代结构的对比

最后,我们将篇章视角的零指代结构的标注结果与 OntoNotes 中已标注的句法视角的零指代结构进行了对比.OntoNotes 中选取了*pro*部分进行了零指代信息的标注.在我们选取的 325 篇来源 CTB 的文档中,*pro*共有 1 077 个,其中,在实体指代链上的*pro*为 944 个,有 133 个*pro*被认为是非待消解的零元素.而我们的篇章视角的零指代语料库共标注了 2 672 个零元素,其中,有 1 010 个与 OntoNotes 中标注的零元素重叠,与 OntoNotes 中标注的实体指代链上的零元素重叠的有 900 个.这 1 010 个重叠的零元素按照我们给出的零元素分类体系进行分类,具体的分布见表 6.

进一步观察这 1 010 个重叠的零元素,发现有 110 个零元素在 OntoNotes 中被视为非待消解项.与 OntoNotes 语料只关注实体指代不同,在我们的语料中,为了后续进行多种类型指代的联合学习,语料标注涵盖了 Event, Union 和 RET 类型.表 7 给出了 1 010 个重叠零元素形成的指代关系的类别分布情况.

从表 7 给出的类别分布统计结果可以看到:修饰型零元素在 EDU 内就完成了指代的消解的情况占到总数的 17.03%,而占据绝大多数的仍然是跨 EDU 的实体类的指代关系.

Table 6 Distribution over categories of zero elements overlapping with the OntoNotes corpus

表 6 与 OntoNotes 中重叠的零元素类别分布统计

零元素类别	数量	比例(%)
Main	503	49.80
Modify	507	50.20
Overall	1 010	100

Table 7 Distribution over types of zero anaphora overlapping with the ontonotes corpus

表 7 与 OntoNotes 中重叠的零元素对应的指代关系的类别分布统计

指代关系类别		数量	比例(%)
Intra-EDU		172	17.03
Inter-EDU	Entity	811	80.30
	Event	5	0.50
	Union	15	1.49
	RET	7	0.69
	Overall	1 010	100

除上述重叠部分,我们进一步分析了不重叠的情况,可以分为两种情况.

(1) OntoNotes 中未标注零元素,而在我们的语料中将其视为零元素.

例 11 给出了一个典型的例子.从统计数据可以看到:我们的语料中包含了 2 672 个零元素,是 OntoNotes 中标注的零元素 2.48 倍.对比标注结果发现:多出的零元素部分,例 11 给出的情况占据了很大的比例.这也是 Yang 和 Xue[14]基于多种句法信息进行了零元素识别和恢复的可计算研究后,分析他们的实验结果得出的一个结论——很难区分是出现了零元素还是句法层面的共享主语.

例 11:[去年外商投资企业出口商品中,工业制成品占九成以上,]$_{e1}$ [φ达四百三十八点八亿美元,]$_{e2}$ [φ比上年增长了百分之三十六点七,]$_{e3}$ [φ明显高于全国平均水平.]$_{e4}$

在篇章表示体系中,EDU 被认为是篇章构成的基本单位,因此篇章视角的零指代表示体系也以 EDU 为考察零元素存在与否的基本单元.若它有语义成分上的缺失,而且这个缺失可以从上下文中恢复,我们就将这一缺失的语义成分看作零元素.例 11 所示的句子包含 4 个 EDU,后 3 个 EDU 中存在明显的语义缺失,而缺失的对象可以从前面的 EDU 中恢复,因此我们认为后 3 个 EDU 中出现的是零元素,而不是主语共享.而且从指代链类型看,出现在 e2 中的第 1 个零元素和出现在 e3 中的第 2 个零元素指代的是“工业制成品”,属于 Entity 类型;而出现在 e4 中的第 3 个零元素指代的是“增长”这件事,属于 Event 类型.

例 12 给出了一个篇章视角不存在零元素,句法视角是共享主语的示例,图 8 给出了对应的句法分析结果.例 12 仅包含一个 EDU,这个 EDU 表述的内容是完整的.而“会积极配合学校发展中心”和“密切与学校相关部门联系与合作”间共享了主语“公司”.我们认为:若 VP 节点与其主语位于同一个 EDU 内部时,对上层篇章来说,该 EDU 表述是完整的,则当前省略表述不作为篇章零元素,而是句法层的共享主语现象.

例 12:[他说,公司会积极配合学校发展中心,密切与学校相关部门联系与合作.]$_{e1}$

(2) OntoNotes 中标注了零元素,而在我们的语料中未将其视为零元素.

对比语料发现,这一现象共有 67 处.其中,位于 OntoNotes 标注的实体指代链上的零元素有 44 处.例 13~例 19 给出了一些 OntoNotes 中进行了标注(*pro*),而我们的语料未标注的零元素示例.从这些例子可以看到:关联某个具体的驱动谓词,确实存在句法层面的成分缺失.然而从篇章视角看,这些 OntoNotes 中标注的缺失成分都不是衔接上下文的语义成分,它们通常指代的是一些常识性的实体,对篇章的理解几乎没有影响.例如,例 14 中标注的*pro*与驱动词“有”相关联,从句法层看缺失了“有”的施事者,但这个施事者在上下文中是没有衔接角色的,因此对篇章理解没有意义.同样,例 16 中,谓词“出台”的施事者缺失了,但这个施事者在上下文中并未承担衔接作用,对篇章理解是没有影响的.

例 13:据了解,目前,*pro*在外商投资企业获得的人民币贷款中,有近一半是中国银行提供的.

例 14:*pro*有人预言,随着九江的进一步开放开发,王翔将从政府划给他的土地中获得可观的利润.

例 15:董建华在*pro*评论该指数时表示,香港特区已连续四年成为全球最自由的经济体.

例 16:如*pro*省里出台并实施的《四川省鼓励外商投资优惠政策》等,为外商提供了优惠、宽松的政策环境.

例 17:*pro*在*pro*与中国缔结友好城市中,以日本为最多.

例 18:研究人员介绍说,*pro*国外目前普遍使用的各种化学合成降糖药对糖尿病并发症均无多大的防治作用.

例 19:镍储量占*pro*中国国内已探明储量的百分之七十.

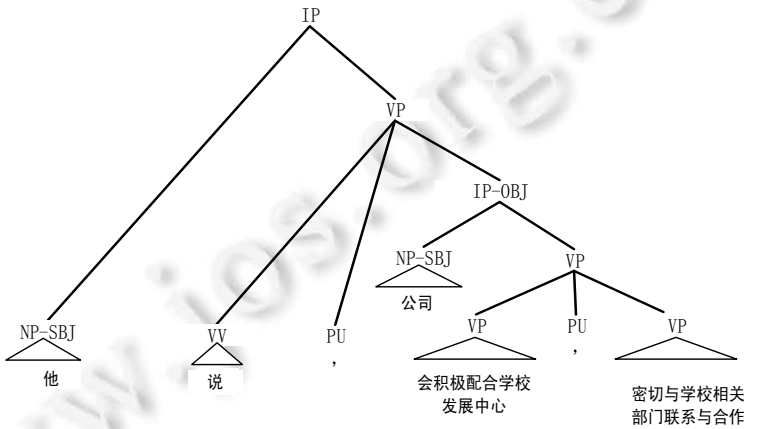

Fig.8 Syntatic parse tree of example 12

图 8 例 12 对应的句法树

4 篇章视角的零指代消解基准平台

完整的零指代消解平台由零元素识别和零元素消解两部分构成,其中:已有的零元素识别相关研究多以句子或子句为单位,依据句法分析的结果从句法成分的缺失这一角度进行,使得零元素识别的性能对句法分析的结果有着严重的依赖;另一方面,零元素消解的相关研究则集中在如何更好地表征零元素所处的上下文信息.从篇章视角构建零指代消解基准平台需要进行以下几个方面的考虑:首先,既然是篇章层的语言现象,高效的零指代消解必然需要多粒度的篇章信息的支持;其次,篇章级的任务更丰富、更复杂,它们之间必然存在密切的联系,而这些联系决定了不能孤立地讨论零指代;最后,零指代归根结底是指代的一种,进行包括实体指代、事件指代在内的多种指代的联合消解势在必行.因此,零指代消解应借鉴较为成熟的实体指代框架.

基于上述考虑,我们选择了 Kong 和 Zhou[26]给出的链到链的汉语零指代消解方案作为基准平台构建的基本方法,在实现上进行了以下改动:(1) 用基于篇章单元(EDU)的零元素识别模块替换了原来的零元素识别模块;(2) 将实体指代消解模块替换成了性能更好的基于神经网络的实体指代消解平台[26];(3) 在零元素链接环节,将原有的人工特征都作为附加特征进行了向量表征,同时增加了基于 Mask 机制的零元素表征;(4) 零指代关系的确定替换成了实体指代消解平台中的前馈神经网络方法.关于链到链的汉语零指代消解方法的细节,请参考 Kong 和 Zhou 的论文[26];实体指代消解平台及前馈神经网络方法,请参考 Kong 和 Fu 的论文[26].本节主要介绍基于 EDU 的零元素识别和基于 Mask 机制的零元素表征.

4.1 基于EDU的零元素识别

给定一个 EDU,我们认为:构成 EDU 的每个词的前面均有可能存在零元素,唯一不可能存在零元素的位置是最后一个词的后面.因此,我们将零元素识别看作一个边界点识别问题,通过编码-解码框架来进行,图 9 给出了这一框架的具体构成.

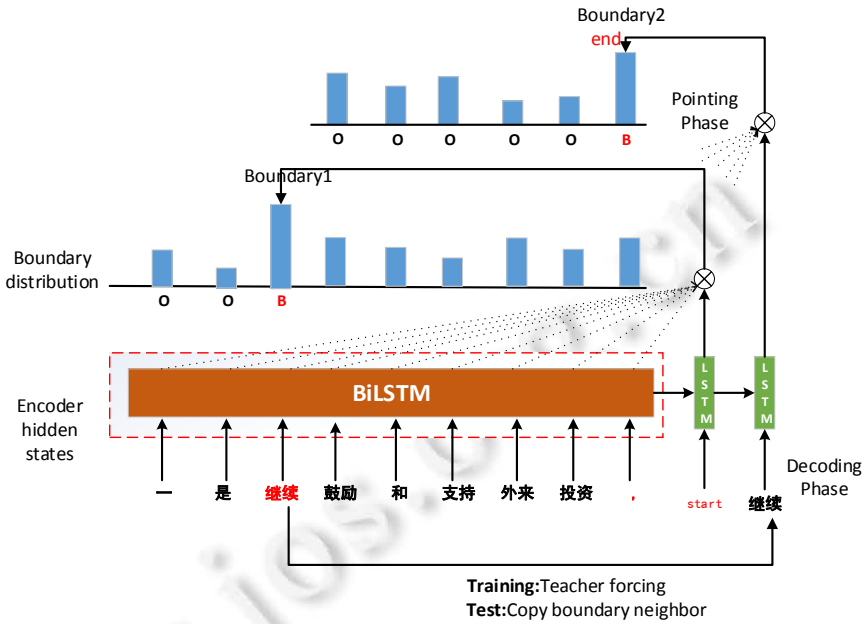

Fig.9 EDU based zero element detection framework

图 9 基于 EDU 的零元素识别框架

在编码阶段,以 EDU 为基本单元作为模型的输入.将含有 n 个词的 EDU 记做 $E=\{w_1,w_2,w_3,\ldots,w_n\}$,其中,$w_i$ 表示 EDU 中的第 i 个词.利用预训练的 Embedding 矩阵将每个词 w_i 映射为低维稠密的词向量,再将该词向量与随机初始化的词性向量拼接作为 BiLSTM 的输入,借助 BiLSTM 动态捕获文本的序列信息,其在两个方向上的最后一个隐状态的拼接 $\tilde{d}$ 将承担解码器初始状态的角色:

$$[h_1,h_2,\ldots,h_n]=BiLSTM(E,\theta) \tag{2}$$

$$\tilde{d}=h_{\vec{1}}\oplus h_{\overleftarrow{n}} \tag{3}$$

解码环节采用指针网络模型实现,它由解码器(decoding phrase)和定位器(pointing phrase)两部分构成.解码器将启动单元 U_m 作为输入,经过一个单向 LSTM 后获得对应输出 d_m,其中首次启动单元为 $\tilde{d}$,之后的启动单元为前一次定位器确定位置的词 w_i 对应的编码 h_i:

$$d_m=LSTM(U_m,\theta) \tag{4}$$

解码时,由于每个输入序列中包含的零元素数量不确定,在得到解码器的输出向量 d_m 后,我们使用指向机制(pointing mechanism)[43]计算输入序列中位于启动单元之后的零元素的位置,具体公式如下:

$$u_j^m=v^T\tanh(W_1h_j+W_2d_m),\ \text{for}\ j\in(i+1,\ldots,n) \tag{5}$$

$$p=softmax(u^m) \tag{6}$$

其中,h 和 d_m 分别为编码层和解码器(decoding phase)的对应输出,j 表示输入序列中词的位置.假设此时的启动单元为原序列中的第 i 个词,v^T,W_1,W_2 均为固定维度的参数,可由训练得到 p,即启动单元为 U_m(原序列中 w_i)时,各位置前包含零元素的概率,最后,取概率最大的位置作为零元素.

训练时采用“teacher forcing”机制[44]来训练模型,即:为解码器提供正确的零元素的位置和启动单元 U_m,测试时则使用模型的当前输出来确定下一步的输入.以图 9 给出的输入序列“一是继续鼓励和支持外来投资,”为例,具体解码过程为:

- 首先,将编码器的输出 $\tilde{d}$ 作为起始启动单元送入解码器端的 LSTM 得到 d_0;然后,通过公式(5)和公式(6)计算输入序列中所有位置的零元素分布概率,得到“继续”一词前面存在零元素的概率最高,因此可以确认第一个分割边界“Boundary1”,第 1 轮解码完成;
- 再将“继续”作为第 2 轮次的启动单元,将其对应的编码端向量送入解码器端的 LSTM 得到 d_8,同前一

步,利用公式(5)和公式(6)计算其右侧各位置的零元素分布概率,发现最后一个词“,”的概率最大,此时,我们认为该 EDU 已没有零元素存在,解码结束.

4.2 基于Mask机制的零元素表征

与传统的实体指代消歧相比,在零指代消解中,如何高效地表征零元素是一个难点.本文采用的基于 Mask 机制的零元素表征方法,其思路来自于 BERT 模型[45].该模型训练时采用 Masked Language Model 的方法,即:随机使用 MASK 标记覆盖每个句子中约 15%的词,用其上下文来预测这些词.很自然地想到:零元素本质上可以看作被 MASK 掉的词,当有足够上下文可以预测这些词时,该 MASK 标记对应的向量可以看作是零元素的表征结果.因此,我们可以借助预训练的 BERT 模型来进行零元素的表征.具体做法是:在预处理阶段,给零元素所在的位置增加一个“[MASK]”标记,来显式地表示零元素(训练时已知正确的零元素位置,测试时借助零元素识别模块自动识别零元素).

图 10 给出了“一是继续鼓励和支持外来投资,”示例中“继续”前的零元素表征的示意图.在获得零元素表征后,与原有的已经向量化的人工特征进行拼接,得到完整的表征后即可进行链接消歧.

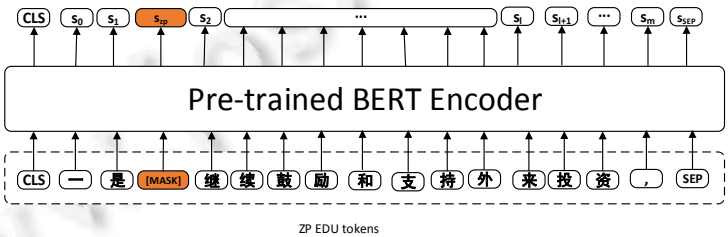

Fig.10 Mask mechanism based zero element representation

图 10 基于 Mask 机制的零元素表征

4.3 基准平台的实验结果

由于语料规模有限,基准平台的实验采用 5 倍交叉验证的方式进行,使用 Precision(P),Recall(R),*F*1-score(F)作为评测标准.验证集是从训练集中随机划分出的数据,占比为 15%,使用早停法(early stopping)来保存验证集上最好的模型,置信度设为 10.模型采用批训练的方法,训练轮次为 100,批次大小为 32,使用 adam 学习器进行参数迭代更新,学习率为 0.001.在 embedding 和 LSTM 层后引入 dropout 机制,dropout 大小为 0.5,LSTM 的层数为 1,使用 BERT 的“BERT-Base-uncased”版本来生成原始的嵌入,词嵌入维度 300,词性嵌入维度 20,隐藏层维度 128.

表 8 给出了基于 EDU 的零元素识别的性能.从结果可以看出,主干零元素的识别性能远远高于修饰型零元素的识别性能.可能的原因有两方面:一是修饰型零元素占比较低,相应的训练实例缺乏;二是直觉上修饰型零元素更多地依赖局部的句法信息,基准平台主要考虑了词与词之间的序列信息,后续可通过句法信息的融入进一步改善修饰型零元素的识别性能.此外,从面向篇章理解的视角来看,主干零元素在篇章的组织、话题的演变等方面起着更为重要的衔接作用,高效地识别出主干型零元素,能够有助于对整个篇章的理解.

Table 8 Performance of EDU based Zero Element Detection

表 8 基于 EDU 的零元素识别的性能

零元素类别	P (%)	R (%)	F (%)
Main	94.79	92.02	93.39
Modify	53.47	50.62	52.01
Overall	92.15	79.04	85.09

表 9 给出了零指代消解的性能.所谓“标准实体链”,我们抽取了 OntoNotes 中标注的实体指代链作为已知信息,仅仅完成将零元素链接到对应实体链上的工作;而“自动实体链”则使用 Kong 和 Fu[26]的系统自动获取实体指代链(使用 OntoNotes 语料重新训练该系统,将本语料的 325 篇文本作为测试集,使用 CoNLL 评测得到的实体

指代消解的性能为 69.66%).从表中列出的实验结果可以看到:不论是标准还是自动实体指代链,零元素的识别性能都对零指代消解的性能产生很大的影响,*F*1 值下降了大约 10%.但相比已有的从句法视角进行的研究(Chen 等人[21,27]以及 Kong 和 Zhou[26]等,自动零元素下消解性能下降了约 20%),下降幅度有所减小,后续将考虑融入更多的篇章级信息来增强系统的鲁棒性.

Table 9 Performance of Zero Anaphor Resolution

表 9 零指代消解的性能

设置		P (%)	R (%)	F (%)
标准实体链	标准 ZP	91.23	85.42	88.23
	自动 ZP	86.23	72.36	78.69
自动实体链	标准 ZP	61.68	60.78	61.23
	自动 ZP	58.07	47.29	52.13

5 总结与展望

从服务于篇章分析和文本理解出发,本文给出了汉语零指代结构的表示体系,并基于这一表示体系选取汉语树库 CTB、连接词驱动的汉语篇章树库 CDTB 和 OntoNotes 语料中重叠的 325 篇文本进行了汉语零指代的标注,构建了一定规模的汉语零指代语料库.系统检测表明:本文提出的表示体系合理有效,构造的语料库质量上乘,能够为篇章视角的汉语零指代研究提供必要的支撑.

本文的主要贡献体现在3个方面:(1) 从篇章视角构建了汉语零指代表示体系,并据此构建了一定规模的汉语零指代语料库,为篇章视角的汉语零指代研究提供了支持;(2) 提出的汉语零指代表示体系使用了汉语篇章微观修辞结构表示体系中定义的基本篇章单元 EDU 和篇章修辞结构树,为探索汉语篇章微观修辞结构与汉语零指代之间的关系,开展两者的联合学习奠定了扎实的基础,同时也为构建多视角的汉语篇章结构的统一表示体系做了初步的探索;(3) 给出了一个基于 EDU 进行汉语零指代的基准平台,为与实体指代的联合以及融入更多的篇章级信息奠定了基础.

接下来我们将进一步修正语料并正式对外发布,同时开展两个核心工作.一是进行篇章视角的汉语零指代消解研究,侧重考虑两方面:(1) 如何借助丰富的篇章信息来更好地表征零元素及其上下文,从而提升零元素识别及消解的性能;(2) 主干型和修饰型零元素对篇章信息和句法信息的依赖度是不一样的,后续将对它们分别建模,再借助多任务学习框架进行结合;二是进行汉语篇章零指代和微观修辞结构的联合学习研究,侧重考虑零指代在篇章逻辑语义关系推进中的作用.

References:

[1] Kim YJ. Subject/Object drop in the acquisition of Korean: A cross-linguistic comparison. Journal of East Asian Linguistics, 2000,9: 325−351.

[2] Beaugrande RAD, Dressler W. Introduction to Text Linguistics. London and New York: Longman Paperback, 1981.

[3] Schank Roger C. Conceptual dependency: A theory of natural language understanding.Cognitive Psychology, 1972,3(4):552−631.

[4] Pradhan S, Ramshaw L, Marcus M, *et al*. CoNLL-2011 shared task: Modeling unrestricted coreference in ontonotes. In: Proc. of the 15th Conf. on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011. 1−27.

[5] Pradhan S, Moschitti A, Xue N, *et al*. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In: Proc. of the Joint Conf. on EMNLP and CoNLL-Shared Task. Association for Computational Linguistics, 2012. 1−40.

[6] Li CN, Thompson SA. Third-person pronouns and zero-anaphora in Chinese discourse. Syntax and Semantics, 1979,12:311−335.

[7] Li WD. Topic chains in Chinese discourse. Discourse Processes, 2004,37:25−45.

[8] Converse S. Pronominal anaphora resolution in Chinese [Ph.D. Thesis]. University of Pennsylvania, 2006.

[9] Zhao SH, Ng HT. Identification and resolution of Chinese zero pronouns: A machine learning approach. In: Proc. of the EMNLP-CoNLL 2007. Association for Computational Linguistics, 2007. 541−550.

[10] Campbell R. Using linguistic principles to recover empty categories. In: Proc. of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004. 645.

[11] Chung T, Gildea D. Effects of empty categories on machine translation. In: Proc. of the 2010 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010. 636−645.

[12] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic modelsfor segmenting and labeling sequence data. In: Proc. of the 18th Int'l Conf. on Machine Learning (ICML 2001). 2001

[13] Gabbard R, Kulick S, Marcus M. Fully parsing the penn treebank. In: Proc. of the Human Language Technology Conf. of the NAACL. New York: Association for Computational Linguistics, 2006. 184−191.

[14] Yang YQ, Xue NW. Chasing the ghost: Recovering empty categories in the Chinese treebank. In: Proc. of the COLIN 2010. Posters, 2010. 1382−1390.

[15] Cai S, Chiang D, Gold-berg Y. Language-independent parsing with empty elements. In: Proc. of the ACL 2011. 2011. 212−216.

[16] Kong F, Zhou GD. A clause-level hybrid approach to Chinese empty element recovery. In: Proc. of the IJCAI 2013. 2013. 2113−2119.

[17] Xiang B, Luo X, Zhou B. Enlisting the ghost: Modeling empty categories for machine translation. In: Proc. of the 51st Annual Meeting of the Association for Computational Linguistics, Vol.1: Long Papers. 2013. 822−831.

[18] Xue N, Yang Y. Dependency-based empty category detection via phrase structure trees. In: Proc. of the 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013. 1051−1060.

[19] Zhou GD, Li PF. Improving syntactic parsing of Chinese with empty element recovery. Journal of Computer Science and Technology, 2013,28(6):1106−1116.

[20] Hovy E, Marcus M, Palmer M, *et al*. OntoNotes: The 90% solution. In: Proc. of the NAACL 2006. 2006. 57−60.

[21] Chen C, Ng V. Chinese zero pronoun resolution: Some recent advances. In: Proc. of the EMNLP 2013. 2013. 1360−1365.

[22] Chen C, Ng V. Chinese zero pronoun resolution: A joint unsupervised discourse-aware model rivaling state-of-the-art resolvers. In: Proc. of the ACL-IJCNLP 2015. 2015. 320−326.

[23] Yin QY, Zhang Y, Zhang WN, *et al*. Chinese zero pronoun resolution with deep memory network. In: Proc. of the EMNLP 2017. 2017. 1309−1318.

[24] Zhang Y, Liu T, Yin QY, *et al*. A deep neural network for Chinese zero pronoun resolution. In: Proc. of the IJCAI 2017. 2017. 3322−3328.

[25] Yin QY, Zhang Y, Zhang WN, *et al*. Deep reinforcement learning for Chinese zero pronoun resolution. In: Proc. of the ACL 2018. 2018. 569−578.

[26] Kong F, Zhang M, Zhou GD. Chinese zero pronoun resolution: A chain to chain approach. ACM Trans. on Asian Low-Resource Language Information Processing, 2019,19(2):21.

[27] Chen C, Ng V. Chinese zero pronoun resolution with deep neural networks. In: Proc. of the ACL 2016. 2016. 778−788.

[28] Chen C, Ng V. Chinese zero pronoun resolution: An unsupervised probabilistic model rivaling supervised resolvers. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). 2014. 763−774.

[29] Dempster AP, *et al*. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological), 1977.

[30] Chen C, Ng V. Chinese zero pronoun resolution: A joint unsupervised discourse-aware model rivaling state-of-the-art resolvers. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing, Vol.2: Short Papers. 2015. 320−326.

[31] Liu T, Cui YM, Yin QY, *et al*. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In: Proc. of the ACL 2017. 2017. 102−111.

[32] Sheng C, Kong F, Zhou GD. Toward better Chinese zero pronoun resolution from discourse perspective. In: Proc. of the NLPCC-ICCPOL 2017. Springer-Verlag, 2017.

[33] Xi XF, Chu XM, Sun QY, *et al*. Corpus construction for chinese discourse topic via micro-topic scheme. Journal of Computer Research and Development, 2017,54(8):1833−1852 (in Chinese with English abstract).

[34] Xi XF. Research on Chinese discourse topic structure: Representation, resource construction and its analysis [Ph.D. Thesis]. Suzhou: Soochow University, 2017 (in Chinese).

[35] Sheng C, Kong F, Zhou GD. Building Chinese zero corpus form discourse perspective. Acta Scientiarum Naturalium Universitatis Pekinensis, 2019,55(1):15−21 (in Chinese with English abstract).

[36] Sheng C. Research of Chinese zero elements detection based on discourse perspective [MS. Thesis]. Suzhou: Soochow University, 2018 (in Chinese with English abstract).

[37] Ge HZ. Research on key issues of Chinese zero anaphora for text understanding [MS. Thesis]. Suzhou: Soochow University, 2020 (in Chinese with English abstract).

[38] Li YC, Feng WH, Kong F, *et al*. Build Chinese discourse corpus with connective-driven dependency tree structure. In: Proc. of the EMNLP 2014. 2014. 2105−2114.

[39] Li YC. Research of Chinese discourse structure representation and resource construction [Ph.D. Thesis]. Suzhou: Soochow University, 2015 (in Chinese with English abstract).

[40] Kong F, Zhou GD. Pronoun resolution in english and chinese languages based on tree kernel. Ruan Jian Xue Bao/Journal of Software, 2012,23(5):1085−1099 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4044.htm [doi: 10.3724/SP.J.1001.2012.04044]

[41] Xue NW, Xia F, Chiou FD, *et al*. The pennchinese treebank: Phrase structure annotation of a large corpus. Natural Language Engineering, 2005,11:207−238.

[42] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics, 1977,33:159−174.

[43] Li J, Sun A, Joty S. SegBot: A generic neural text segmentation model with pointer network. In: Proc. of the IJCAI 2018. 2018. 4166−4172.

[44] Lamb AM, Goyal AGAP, Zhang Y, *et al*. Professor forcing: A new algorithm for training recurrent networks. In: Proc. of the Advances in Neural Information Processing Systems. 2016. 4601−4609.

[45] Declin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805, 2018.

附中文参考文献:

[33] 奚雪峰,褚晓敏,孙庆英,等.汉语篇章微观话题结构建模与语料库构建.计算机研究与发展,2017,54(8):1833−1852.

[34] 奚雪峰.汉语篇章话题结构:表示体系、资源构建及其分析研究[博士学位论文].苏州:苏州大学,2017.

[35] 盛晨,孔芳,周国栋.中文篇章零元素语料库构建.北京大学学报(自然科学版),2019,55(1):15−21.

[36] 盛晨.篇章视角的中文零元素识别研究[硕士学位论文].苏州:苏州大学,2018.

[37] 葛海柱.面向文本理解的汉语零指代关键问题研究[硕士学位论文].苏州:苏州大学,2020.

[39] 李艳翠.汉语篇章结构表示体系及资源构建研究[博士学位论文].苏州:苏州大学,2015.

[40] 孔芳,周国栋.基于树核函数的中英文代词消解.软件学报,2012,23(5):1085−1099. http://www.jos.org.cn/1000-9825/4044.htm [doi: 10.3724/SP.J.1001.2012.04044]

孔芳(1977－),女,博士,教授,博士生导师,主要研究领域为自然语言理解,篇章分析.

葛海柱(1994－),男,硕士,主要研究领域为自然语言理解,篇章分析.

周国栋(1967－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为自然语言理解,机器学习.