北京大学学报(自然科学版) 第 55 卷 第 1 期 2019 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 55, No. 1 (Jan. 2019)
doi: 10.13209/j.0479-8023.2018.057

中文篇章零元素语料库构建

盛晨 孔芳† 周国栋

苏州大学计算机科学与技术学院自然语言处理实验室, 苏州 215006; † 通信作者, E-mail: kongfang@suda.edu.cn

摘要 针对中文零指代问题, 从篇章视角进行理论分析, 并完成中文篇章零元素语料库(Chinese Discourse Zero Corpus, CDZC)的构建工作。首先, 整理和分析已有的理论研究以及语料资源, 探究篇章层面中文零元素语料库标注的必要性。然后, 采用自底向上、前向搜索的标注策略和人机结合的半自动标注方式, 完成 CDZC 语料库的构建。最后, 对该语料库进行一系列详细的统计分析。结果表明, CDZC 能够充分反映出中文零元素省略的语言特点, 为相关研究提供语料资源支持。

关键词 中文零元素; 篇章视角; 语料库构建; 中文篇章零元素语料库

Building Chinese Zero Corpus Form Discourse Perspective

SHENG Chen, KONG Fang†, ZHOU Guodong

Natural Language Processing Lab, School of Computer Science and Technology, Soochow University, Suzhou 215006;
† Corresponding author, E-mail: kongfang@suda.edu.cn

Abstract To better deal with Chinese zero elements, this paper makes a theoretical analysis from discourse perspective and completes the construction of the Chinese Discourse Zero Corpus (CDZC). First, the necessity of corpus construction has been explored based on the research of existing theoretical and data sources. Then, the top-down and forword search annotation strategy and the combination of the human machine are used to complete corpus annotation. Finally, the detailed statistics analysis shows that CDZC can fully reflect the characters of Chinese linguistic and provide corpus resources for related research.

Key words Chinese zero elements; discourse perspective; corpus construction; Chinese discourse zero corpus

随着人工智能如火如荼地发展, 为实现无障碍人机智能交互的终极目标, 自然语言处理作为其至关重要的分支, 承担起语言理解的重任[1]。省略作为一种常见现象广泛存在于汉语表述, 其省略成分称为零元素(zero pronoun)。准确识别出该零元素并理解作者的真实意图, 是自然语言处理面临的重大挑战任务之一。

到目前为止, 尽管这些有关中文零元素的研究已取得一定的成果, 但其效果仍不尽如人意。首先, 汉语的复杂性决定了该任务的难度, 大量的长句以及复杂的句法带来巨大的挑战。此外, 语料资源的稀缺也是限制中文零元素发展的重要原因之一。

本文针对上述问题, 基于篇章理解层面, 对中文省略现象进行深入的探究, 提出篇章零元素的概念。在此基础上, 完成中文篇章零元素语料库构建, 并进行语料库相关的统计分析。

1 相关工作

近年来, 中文零元素现象备受关注, 针对中文的零元素识别与消解任务取得一系列的研究成果。然而, 这些研究主要侧重于方法, 对语料库的构建则考虑较少。

Zhao 等[2]给出一个完整的基于机器学习的中文零指代识别及消解方案, 并提出一套有效的特征集合。Yang 等[3]基于 CTB 语料对零元素识别进行研究, 采用序列化标注模型来识别句中存在的零元

国家自然科学基金(61472264, 61751206)和国家重点研发计划(2017YFB1002101)资助
收稿日期: 2018–04–19; 修回日期: 2018–08–07; 网络出版日期: 2018–08–22

素。Kong 等[4]给出一个基于树核函数的中文零元素消解的完整框架, 将中文零指代消解任务清晰地划分成 3 个子任务: 零元素识别、待消解项识别和零元素消解, 分别给出每一个子任务适用的结构化特征集。Chen 等[5−7]首次给出完整的端到端的全自动状况下的中文零指代消解平台, 并提出一组更有效的句法和上下文特征; 而后, 为了避免有监督学习下对语料的依赖性, 在之前的工作基础上, 又给出一个无监督方法的生成式模型, 取得较好的性能。

在语料资源方面, 得到大众认可的中文零元素语料是 OntoNotes 语料[8]。该语料是由美国众多科研机构联合创立的权威语料库, 存在中、英、阿拉伯 3 种标注语言。该语料的中文部分标注了汉语中主语位置的零元素省略及其指代链, 为目前已有的中文零元素研究工作提供资源支持。

2 标注动机

首先, 汉语的语言特点决定了篇章视角研究的必要性。从形式上看, 零元素被视为句中省略的词。然而, 从语义理解的角度来看, 省略的语义成分却是依赖于篇章的上下文表述。也就是说, 零元素并非句子内部词汇成分, 而是连贯上下文中特殊语义表述的载体。零元素体现的不是句子内部的语言特点, 而是以篇章为单位的语义表达方式。在省略表述过程中, 只有先在前文中被提及, 后文中才可以省略, 并且前后文间必然存在相应的语义逻辑关系。由此可见, 篇章视角下的中文零元素研究工作有其必要性。

其次, 语料库资源的唯一性限制了研究的进展。中文省略表述属于篇章的范畴, 然而 OntoNotes 语料标注却倾向于句法层面, 以致目前大多数相关研究均是基于句法层面进行的: 研究对象是句子, 所选特征也约束在词法和句法特征之内。众所周知, 语言是文化的载体, 语言的不同反映文化的差异。西方文化特点决定其语言(英语)的表述更倾向于直来直去的方式, 大多时候一句话就可以清晰地表述说话者的意图。然而, 中国的文化特点在于含蓄, 其语言表述方式也与英语大不相同。中文表述过程中, 说话者的意图往往经过多层铺垫和转折加以修饰, 委婉地表达出来。由此可见, 以句子为单位的零元素标注方式在西方语言的语料上取得令人满意的成果, 但对于中文语料的研究, 这种标注方式不尽合理。

此外, 从篇章视角来看, OntoNotes 语料标注存在不足之处。Li 等[9]参考修辞结构理论(rhetorical structure theory, RST)[10]以及宾州篇章树库(Penn Discourse Tree Bank, PDTB)[11]体系, 提出基于连接依存树的汉语篇章结构表示体系, 并标注了中文篇章树库(Chinese Discourse Treebank, CDTB)。以基本篇章单元(elementary discourse unit, EDU)作为叶子节点, 修辞关系作为非叶子节点, 自底向上构成一棵树结构, 用来表示汉语篇章结构。通过对 CDTB 与 OntoNotes 重叠语料部分的统计, 我们发现以下问题。

1) 部分零元素标注不存在对应的指代链标注(chain), 占比约 12.9%。通过对这部分语料的逐一人工分析, 发现该部分零元素标注大多仅是为了句法结构的严谨性, 对于篇章语义的理解影响无关紧要。

2) 尽管已给出零元素对应的指代链标注, 然而其指代链上的指代项均为零元素, 这部分占比为 5.2%。通过分析, 此处省略的成分较特殊, 一般为大众熟知的常识内容, 如“中国”此类概念性实体。

3) 统计结果显示, 大约有 16.8%的零元素标注虽然存在有效指代关系, 但该关系并不在篇章内部。也就是说零元素与其先行词不在同一个篇章关系之中。此类指代属于跨篇章的指代关系, 即便是汉语语言学家进行判断, 也存在较大的歧义性, 不属于本文研究的范围。

4) 该语料存在一定的漏标现象, 如例 1 所示。

例 1 [专家们认为, 在中国的五个经济特区中, **海南**的地理位置、资源条件、经济发展状况较为特殊,]e1 | [**ϕ**应进一步扩大对外开放,]e2 | [率先实现与亚太区域经济一体化和国际贸易自由化的对接。]e3

例 1 选自 chtb_0018 文档, 零元素用 ϕ 标注, 并与其先行词用特殊字体标注(加粗、下划线)。分隔符“|”切分段落为对应基本篇章单元序列, 构成如图 1 所示的篇章修辞结构关系: e2 与 e3 构成条件关系, 进一步与 e1 构成因果关系。例 1 中段落表述的完整语义是: “专家们认为由于**海南**……, 所以应该让[**海南**]进一步扩大对外开放, 才能让[**海南**]率先实现……”。不仅在 e2 中存在语义省略, e3 内部也存在语义省略。OntoNotes 语料仅给出前一处的标注而忽略了后一处。

综上所述, 一方面中文零元素语料库资源紧缺,

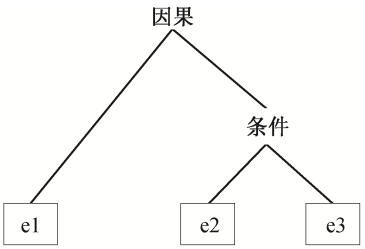

图 1　例 1 对应篇章层次化结构
Fig. 1　Discourse structure of example 1

另一方面, 唯一被认可的语料也存在一系列的欠缺。因此, 构建基于篇章视角的零元素语料库成为研究过程中不可或缺的一步。

3 语料库构建

3.1 篇章零元素

依照零元素是否承担所在EDU主干语义成分, 将其分为两大类。汉语篇章结构表示体系对EDU定义如下: 至少包含一个谓语部分, 至少表达一个命题[9]。我们认为EDU内部的主、谓(宾)结构承担其主干语义。例如, 若零元素作为EDU主干语义成分(例如主语成分), 则定义该零元素为**篇章主干性零元素**; 否则, 认为该零元素作为EDU主干的修饰性成分(例如主语的修饰成分), 定义该零元素为**篇章修饰性零元素**。

例 2　[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

如例2所示, 斜体、双下划线字体标注EDU的驱动谓词, 加粗、下划线字体标注零元素ϕ及其指代先行词。ϕ所在EDU对应主干语义: "**良好的态势**使其面临更多的发展机遇"。该零元素承担EDU内部谓词的主语成分, 符合篇章主干性零元素的定义。

例 3　[**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

例3中零元素ϕ所在EDU表述主干语义为"大量出现的是……的新情况、新问题", ϕ作为宾语"新情况、新问题"的修饰成分, 属于篇章修饰性零元素。与篇章主干性零元素相比, 此处省略成分对篇章关系构建的影响较小, 对EDU内部语义关系抽取以及局部句法分析影响较大。准确地识别该零元素有助于明确局部语义成分, 减少复杂的修饰性成分对篇章理解带来的噪声。

3.2 语料资源

我们从宾州树库语料(CTB 6.0)中抽取325篇文档(chtb0001-chtb0325)作为语料标注资源。采用该语料的原因主要有以下几方面。

1) OntoNotes中存在该部分对应语料。自其发布以来, OntoNotes语料在多类自然语言处理任务中得到广泛应用, 具有较高的认可度。在该语料上完成篇章零元素标注, 有利于与已有的研究工作进行对比。

2) 这部分语料对应的篇章修辞关系语料已经构建。本课题组结合PDTB与RST体系的优势, 将汉语篇章结构表示成一棵树结构, 并基于上述CTB语料发布了对应CDTB标注语料。该语料可提供本研究所需的篇章体系结构以及对应的篇章标注。

3) 该部分语料的来源对应为新华社的新闻语料。与其他领域(例如微博、推特、医学领域等)相比, 作为大众化书面语言, 新闻语料显得更整齐, 其表述更符合中文语法规范, 适合初步研究。

4) CTB语料包含丰富的词法、句法等标注资源, 可以为后期的研究提供不同层次的特征。

3.3 标注策略与流程

3.3.1 中文篇章零元素标注策略

基于汉语篇章结构理论体系[9], 作为篇章的基本单位, EDU上层对应的篇章关系反映全局的语义结构信息, 下层对应的句法信息则可有效地辅助理解局部语义。此外, 零元素对应的语言成分也大多存在于省略之前。上述特点决定了如下的标注策略: 输入与一个段落对应的篇章关系, 以EDU为标注单位, 向上考虑篇章修辞关系, 向下结合句法结构, 判断其内部是否包含零元素。如果存在零元素, 在EDU内部定位该零元素, 并向前搜索其对应的语言成分, 进而完成篇章零元素标注。

3.3.2 人机结合的语料标注流程

标注工作由一名导师与两名研究生合作完成, 整个标注过程分为3个阶段。第1阶段, 为保证语料标注的质量以及通用性, 我们制定初步的标注规范, 开发相应的标注工具。第2阶段, 依照初步的标注规范, 所有标注者分别标注相同的20篇文档(111个段落, 237个句子), 然后针对上述标注进行讨论, 讨论涉及零元素的定义、先行词类型、标注

方式以及标注属性等内容。通过小组内的讨论，得到最终的标注规范，并且完成所有的语料标注。第 3 阶段，对最终的标注文档逐一校对，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。

为了简化工作量，提高标注效率以及标注一致性，我们设计开发了零元素标注平台，工作流程如图 2 所示。首先导入生语料，利用计算机辅助工具生成可视化的篇章结构以及对应的句法结构；然后通过人工分析，识别 EDU 内部零元素，并进行相关属性标注，用 XML 文件格式保存标注结果；最后对 XML 文件进行统计分析，得出统计结果。

3.4 标注规范

3.4.1 标注总则

首先通过一个例子来介绍篇章零元素标注的具体内容。

例 4 [崇明是中国第三大岛,]e1 | [具有优越的地理条件和悠久的历史,]e2 | [改革开放以来, 崇明县的经济建设和对外开放发展迅猛,]e3 | [外商投资企业不断增多,]e4 | [进出口货物大量增加,]e5 | [是中国综合实力百强县之一。]e6

如例 4 所示，分隔符“|”将段落切分为 6 个 EDU 并构成图 3 所示的篇章结构。对 e1 进行人工语义判断，其主、谓、宾结构清晰，不存在省略成分；继续判断 e2，存在主语省略，其表达的完整语义是“[**崇明**]具有优越的地理条件和悠久的历史”在此标注相应零元素及其指代先行词。重复上述过程，依次对段落中其他的 EDU 依次进行判断、标注，形成最终对应的 XML 标注文档。

3.4.2 篇章零元素标注

本节结合具体实例详细介绍中文篇章零元素标注(零元素用**ϕ**标出，并将先行词设定为加粗、下划线特殊字体)。标注方案如下:

```
<Zero>     //零元素标签
  ZID=[1…N]   //零元素 ID
  ZOffset=[0…N]     //所在段落中的位置
  Classify =[***]     //划分零元素类别
    <CorefEDU Position=[a…b] Text=[***] >   // 指代先行词对应 EDU 标签
    <ZeroEDU Position=[a…b] Text=[***] >   // 零元素所在 EDU 标签
</Zero>
```

例 5 <Zero ZID=“1” ZOffset=“66” Classify=“VPType”><CorefEDU Position =“22…66”><Text>上海浦东[**近年来颁布实行了涉及经济、贸易、建设、规划、科技、文教等领域的七十一件法规性文件，**]</Text></CorefEDU><ZeroEDU Position=“67…79”><Text>[**zero**]确保了浦东开发的有序进行。</Text></ZeroEDU></Zero>

例 5 所示为语料标注文档实例，相关说明如下。

Zero 中的 ZID 表示零元素在标注文档对应的唯一标识号，起始为 1，递增标注，增幅为 1。

Zero 中的 ZOffset 表示零元素所在段落内部的位置，与 CDTB 语料库位置标注保持一致。

Zero 中的 ZeroEDU 表示零元素所在 EDU 的信息，Position 记录该 EDU 在段落内部的起始位置和终止位置，Text 记录带有零元素标记的文本(论文中用 ϕ 来指代零元素，语料中是用[zero]标出的)。CorefEDU 标注参考 ZeroEDU 的格式记录，零元素指代先行词对应 EDU 的信息。

Zero 中的 Classify 表示当前零元素的子类别，存在 4 类取值，分别为 IPType，VPType，MODIFY-Type 和 EDUType。

IPType 类型零元素满足条件：当前零元素为篇章主干性零元素、其所在的 EDU 对应句法节点为

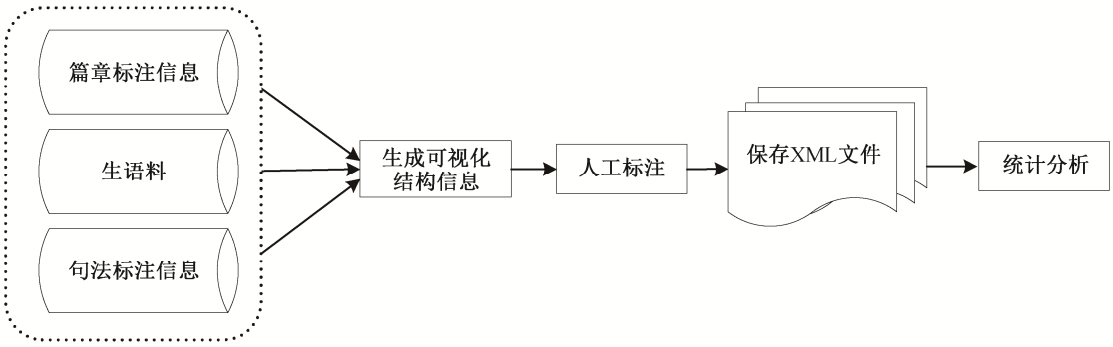

图 2 中文篇章零元素标注平台处理流程

Fig. 2 Processing flow of annotation platform for Chinese discourse zero

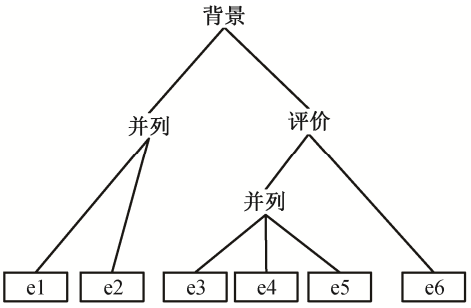

图 3　例 4 对应篇章层次化结构
Fig. 3　Discourse structure of example 4

IP 类型节点、零元素作为 IP 节点的主语成分。

例 6　[建筑是开发浦东的一项主要经济活动,]e1 | [ϕ这些年有数百家建筑公司、四千余个建筑工地遍布在这片热土上。]e2

例 6 所示为 IPType 类型零元素，其所在 EDU 表述的主干语义为: “[浦东]有……”。图 4 为 ϕ 所在 EDU 的句法结构，该句法节点为 IP 类型节点，ϕ 作为主语成分，符合篇章主干性零元素的定义。

VPType 类型零元素满足条件: 当前零元素为篇章主干性零元素，其所在的基本篇章单元对应句法节点为 VP 类型节点，该零元素作为 EDU 驱动谓词的主语成分。

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

例 7 所示为 VPType 类型零元素。该零元素符合篇章主干性零元素的定义，如图 5 所示，ϕ 所在 EDU 对应的句法结构为 VP 类型节点，并且零元素作为驱动谓词的主语成分。

进一步分析 VPType 类型零元素，发现该类型零元素在句法结构中大多呈现为并列 VP 结构，且

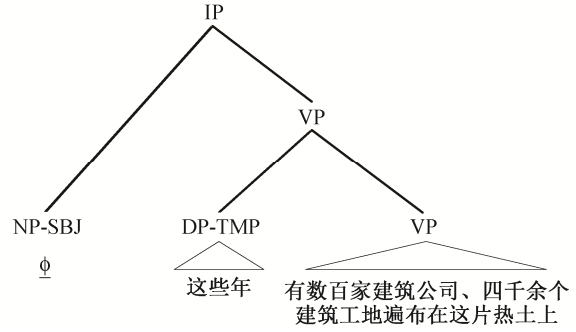

图 4　例 6 零元素所在 EDU 句法结构
Fig. 4　Syntactic structure of EDU including zero in example 6

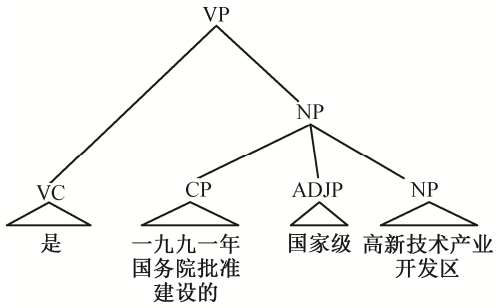

图 5　例 7 零元素所在 EDU 句法结构
Fig. 5　Syntactic structure of EDU including zero in example 7

共享同一个主语成分。我们称这类现象为句法层面的共享主语现象，其对应的句法结构如图 6 所示，VP1，VP2 和 VP3 节点共享主语节点 SBJ。然而，共享主语现象是句法层面的概念，应与篇章零元素严格区分开来。我们认为，若该 VP 节点与其主语位于同一个 EDU 内部，对上层篇章来说该 EDU 表述是完整的，当前省略表述不作为篇章零元素。

例 8　他说，公署还积极配合中国驻外使领馆，密切与特区政府有关部门联系与合作，

图 7 为例 8 对应的句法结构，表述的主干语义为“他说……”，驱动谓词“说”引导宾语从句，其内部存在共享主语现象，表述的完整语义为“他说，公署还积极配合中国驻外使领馆，[公署]密切与特区政府有关部门联系与合作”。然而，该语义省略仅表现在 EDU 的句法层面，不属于篇章层面的零元素，故忽略此处的语义省略标注。

MODIFYType 与 EDUType 的判断条件: 当前零元素为篇章修饰性零元素，进一步判断指代关系。若先行词与零元素位于不同的 EDU，划分为 MODIFYType，否则为 EDUType。

例 9(a)　[以茂名三十万吨乙烯工程为依托的水东开发区，不断加大ϕ招商引资的力度,]e1

例 9(b)　[浦东开发开放是一项振兴上海，建设

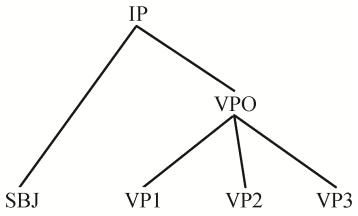

图 6　句法层面的共享主语结构
Fig. 6　Structure of share subject from sentence perspective

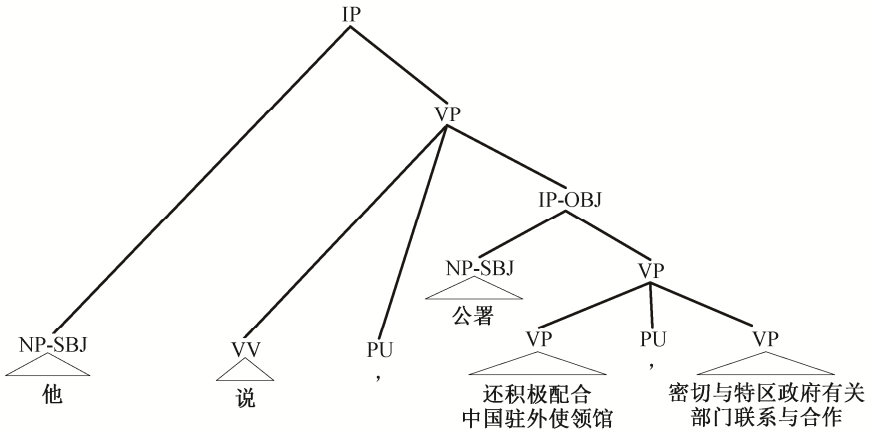

图 7 例 8 对应句法结构

Fig. 7 Syntactic structure of example 8

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

例 9(a)和(b)中零元素均为篇章修饰性零元素(作为宾语的修饰成分)，依据其指代关系的类型，各自标注为 EDUType 和 MODIFYType。

3.5 统计与分析

3.5.1 标注一致性检测

尽管标注人员遵循统一的标注规范，由于个人的主观性差异，导致语料的标注结果依旧存在不一致的现象。语料的一致性检验是用来衡量上述一致性的重要标准，反映语料的标注质量。本研究采取 Kappa 检验进行一致性检验，计算公式为

$$\text{Kappa}=\frac{P_{\text{O}}-P_{\text{C}}}{1-P_{\text{C}}}, \tag{1}$$

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 表示标注具有较好的一致性，Kappa 值小于 0.4 则表示一致性较差。

我们选取两名标注人员 A 和 B，对相同的 30 篇文档(chtb0101~chtb0130)进行独立标注，根据标注结果进行一致性测试。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。通过计算，零元素标注的 Kappa 值为 0.85，表明该语料的标注结果是可信的。

3.5.2 语料库统计

CDZC 共有 325 篇文档(chtb0001~chtb0325)，全部来源于 CTB 语料，总共包含 1367 个段落，4098 个句子，标注零元素 2088 个，平均每个段落包含零元素 1.53 个。下面从零元素分布以及零元素类别两个方面对 CDZC 进行统计分析。

1) 零元素分布统计。基于段落对零元素分布进行统计，对应结果如表 1 所示。1367 个段落中，有 425 个段落不包含零元素，占总数的 31.09%。也就是说，中文篇章表述中，68.91%的篇章中存在零元素。该数据直接地反映出中文省略的普遍性，肯定了中文零元素的研究价值。

2) 零元素类别统计。对零元素类别 Classify 进行统计，分布结果见表 2 。IPType 与 VPType 占据

表 1 基于段落的零元素分布统计

Table 1 Chinese zero distribution statistics based on paragraph

每个段落包含 m 个零元素	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
$m\geq 8$	14	1.02

表 2 零元素类别分布统计

Table 2 Classify of Chinese zero distribution statistics

零元素类别	数量	比例/%
IPType	456	21.84
VPType	1296	62.07
MODIFYType	177	8.48
EDUType	159	7.61

绝大部分，比例高达83%以上。这部分零元素对应为篇章主干性零元素，对篇章语义理解分析起至关重要的作用。剩余的零元素占比约为17%，体现EDU层面的细节语义，辅助局部句法语义分析，在后续的研究中有不可替代的作用。

4 结束语

本文针对汉语表述的语言特点，结合汉语篇章结构体系，对中文省略现象进行理论分析，提出篇章层面的零元素概念，并基于此构建中文篇章零元素语料库(CDZC)。我们选取较有认可度的CTB语料进行标注。为确保标注一致性，我们制定了一整套标注规范，并采用合理的标注策略以及人机结合的标注方法进行语料标注。最终对该语料进行一致性检测以及详细的统计分析，结果表明该语料较好地体现了零元素省略的语言现象以及其对应的语言特点。

目前CDZC语料主要来源于新闻类的文本，数量相对有限，仅能满足初步阶段的研究需要。下一步的研究重点将放在扩大语料库的规模以及生语料文本的类型上，以便满足进一步的研究需要。

参考文献

[1] Chowdhury G G. Natural language processing. Annual Review of Information Science & Technology, 2003, 37(1): 51–89

[2] Zhao S, Ng H T. Identification and resolution of Chinese zero pronouns: a machine learning approach // EMNLP-CoNLL. Prague: DBLP, 2007: 541–550

[3] Yang Y, Xue N. Chasing the ghost: recovering empty categories in the Chinese Treebank // COLING 2010, International Conference on Computational Linguistics. Beijing: DBLP, 2010: 1382–1390

[4] Kong F, Zhou G. A tree kernel-based unified framework for Chinese zero anaphora resolution // EMNLP. Massachusetts: DBLP, 2010: 882–891

[5] Chen C, Ng V. Chinese zero pronoun resolution: some recent advances // Proceedings of the Conference on Empirical Methods in Natural Language Proceedings. Seattle, 1360–1365

[6] Chen C, Ng V. Chinese zero pronoun resolution: an unsupervised approach combining ranking and integer linear programming // Twenty-Eighth AAAI Conference on Artificial Intelligence. Quebéc: AAAI Press, 2014: 1622–1628

[7] Chen C, Ng V. Chinese zero pronoun resolution: an unsupervised probabilistic model rivaling supervised resolvers // Conference on Empirical Methods in Natural Language Processing. Doha, 2014: 763–774

[8] Hovy E, Marcus M, Palmer M, et al. OntoNotes // The Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. New York, 2006: 57–60

[9] Li Y, Feng W, Sun J, et al. Building Chinese discourse corpus with connective-driven dependency tree structure // Conference on Empirical Methods in Natural Language Processing. Doha, 2014: 2105–2114

[10] Mann W C, Thompson S A. Rhetorical structure theory: toward a functional theory of text organization. Text & Talk, 2009, 8(3): 243–281

[11] Miltsakaki E, Prasad R, Joshi A, et al. The Penn Discourse Treebank // LREC. Lisbon, 2004