软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(7):2052−2072 [doi: 10.13328/j.cnki.jos.005834]

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

汉语篇章理解研究综述*

孔 芳[1,2], 王红玲[1,2], 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: 人们理解自然语言通常是在篇章级进行的,随着词汇级及句子级研究的日益成熟,自然语言处理研究的焦点已转向篇章级.篇章分析的主要任务就是从整体上分析出篇章结构及其构成单元之间的语义关系,并利用上下文理解篇章.根据不同的篇章分析目的,篇章单元及其关系可以表示为不同的篇章基本结构,不同篇章基本结构及其关系的研究可提供不同层面的篇章理解.目前对汉语篇章内在规律的研究较少,缺乏对篇章进行有效分析和深入理解的理论方法体系,这严重制约了篇章级的相关研究及应用.重点关注篇章的两个最基本特征,即衔接性和连贯性,从篇章结构分析的理论研究、资源建设和计算模型这3个方面,分别探讨篇章修辞结构(体现篇章连贯性)和话题结构(体现篇章衔接性),对篇章理解的国内外研究现状进行了归纳和整理,并给出了目前存在的主要问题和研究趋势.

关键词: 自然语言理解;篇章分析;篇章修辞结构;篇章话题结构

中图法分类号: TP391

中文引用格式: 孔芳,王红玲,周国栋.汉语篇章理解研究综述.软件学报,2019,30(7):2052−2072. http://www.jos.org.cn/1000-9825/5834.htm

英文引用格式: Kong F, Wang HL, Zhou GD. Suvery on Chinese discourse understanding. Ruan Jian Xue Bao/Journal of Software, 2019,30(7):2052−2072 (in Chinese). http://www.jos.org.cn/1000-9825/5834.htm

Suvery on Chinese Discourse Understanding

KONG Fang[1,2], WANG Hong-Ling[1,2], 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: Natural language is usually understood from discourse perspective. With the success of research at the lexical and sentence levels, the focus of natural language processing research has shifted to the discourse level. The object of discourse analysis is to analyze the text structure and the semantic relationship between discourse units, and thus understand the text. According to different purposes, discourse units and their relationships can be expressed as different textual structures, and the study of them can provide different levels of text comprehension. Currently, there are few studies on the inherent laws of Chinese texts and the lack of theory and method system for effective analysis and in-depth understanding of Chinese discourse has seriously restricts the relevant research and application. This study focuses on two basic features of a text, namely cohesion and coherence. From three aspects of theoretical research, resource construction, and computational model of discourse analysis, it explores the rhetorical structure (reflecting text coherence) and topic structure (reflecting text cohesion) respectively. It summarizes the current research, and presents the main problems and research trends.

Key words: natural language understanding; discourse analysis; discourse rhetorical structure; discourse topic structure

* 基金项目: 国家自然科学基金(61751206, 61876118, 61673290)

Foundation item: National Natural Science Foundation of China (61751206, 61876118, 61673290)

收稿时间: 2018-11-16; 修改时间: 2019-01-12; 采用时间: 2019-03-12; jos 在线出版时间: 2019-04-10

CNKI 网络优先出版: 2019-04-09 17:32:27, http://kns.cnki.net/kcms/detail/11.2560.TP.20190409.1732.006.html

1 引 言

人们理解自然语言通常是在篇章级进行的.作为自然语言处理的一个核心任务,篇章分析(discourse analysis)的主要任务就是从整体上分析出篇章结构及其构成单元之间的语义关系,并利用上下文理解篇章.根据不同的篇章分析目的,篇章单元及其关系可以表示为不同的篇章基本结构.篇章结构可以是篇章内部关系的不同结构化表达形式,主要包括修辞结构、话题结构、指代结构、功能结构、事件结构等范畴[1].从语言学角度讲,这些不同的结构表达形式从不同的角度对篇章进行描述;从计算的角度来看,它们可用线性序列、树和图等数据结构进行抽象表示.随着词法、句法分析技术的不断成熟,篇章分析已成为制约自然语言处理发展的一个瓶颈.

作为篇章分析的基本概念,篇章(discourse)又称为语篇或文本,是由一系列连续的词、短语、子句、句子或段落构成的语言整体单位[1].这里,词被认为是自然语言中有意义的最小单位,相继可以构成短语、子句和句子,句子又可以构成段落,并最终构成篇章.需要强调的是,篇章不是其构成单元的无序堆砌,只有当构建的整体单位上下连贯相互关联,所含信息整体一致,表达完整的思想和意图,才能具有明确的意义,从而称为篇章.以图 1 给出的两个例子进行对比说明.在例 1 中,尽管每个独立子句语义正确,句法完整,但是顺次连接在一起并不能够构成一个篇章.原因在于,这些子句所表达的意义彼此没有关联,难以形成一个整体,也无法表达明确的主题.与此相比,例 2 中,尽管有些子句的句法成分缺失(例 2 所示的段落由 6 个基本篇章单元构成,基本篇章单元分别用(a)~(f)表示;〈〉扩起的内容表示篇章关系中缺省的连接词;[]表示对应子句在该位置缺少相关的句法成分),然而借助于句子之间的意义关联,可以构建形成一个以“李四”作为中心话题的语言整体,因而构成了一个篇章.

例 1:尽管比尔来自美国,但今天交通非常拥挤,并且长江贯穿中国的多个省市,因此,自然语言处理是计算机科学与语言学的融合.

例 2:李四比较年轻(A),〈而且〉[]既没有丰富的工作经验(B),[]又没有高学历(C),但是[]不论做啥事情(D),他都认真负责(E),所以,领导非常器重他(F).

Fig.1 Chinese discourse examples

图 1 篇章示例

篇章一般围绕某个话题展开.篇章信息的一致性(篇章信息性)和篇章意图的整体性(篇章意图性)通常表现为一个话题,该话题的完整性从形式和内容两方面分别体现为篇章的两大基本特性,即篇章连贯性(coherence)和篇章衔接性(cohesion).篇章衔接性和篇章连贯性分别从内容和形式两个方面保证了篇章所要表达的意图性,即作者所要表达话题的正确性和可理解性,二者相互依赖,相互补充.

具体而言:一方面从篇章连贯性角度,话题在形式上的完整性往往体现为某种篇章基本构成单元通过递归组合,基于不同层面的逻辑关系联接,形成一种修辞上的层次化结构,即篇章修辞结构.如图 2 所示,B 和 C 之间构成并列关系,B 和 C 都是中心,BC 的组合和 A 构成递进关系,ABC 的组合和 DEF 的组合之间构成转折关系,DEF 的组合为中心.各基本篇章单元组合后形成高一级篇章单位,进而通过再组合形成更高一级篇章单位,如此层层组合,最终可以表示成一棵篇章修辞结构树.各层篇章单位赖以组合的原因在于其间存在一些为数不多的、反复出现的修辞结构关系(如并列、递进等),这些修辞结构关系有时以连接成分作为形式标记(如例 2 中的“既…又…”),有时则完全隐含(如例 2 中的缺省连接词,“〈而且〉”).

上述篇章修辞结构的分析结果对篇章话题理解非常重要.例如,在自动问答系统中,通过例 2 中的因果关系,可以较容易地自动抽取出相关问题的答案:“领导非常器重他”的原因是“不论做啥事情,他都认真负责”.又譬如,对于自动文摘而言,根据图 2 中最高层的“转折”关系,可以得出“基本篇章单元 DEF 的组合”比“基本篇章单元 ABC 的组合”更重要;而对于次一级“因果”关系而言,“基本篇章单元 F”可能比“基本篇章单元 DE 的组合”更重要;如此层层推进,最终可以得到该段篇章的核心话题,即为“基本篇章单元 F”.当然,上述推进过程的实现,主要依赖于篇章关系传递性及中心指向原则.

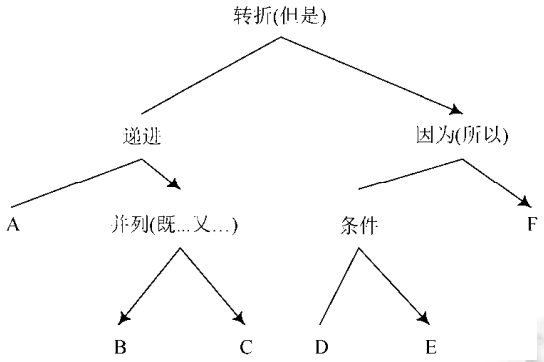

Fig.2 Discourse rhetorical structure of the second example shown in Fig.1
(the EDUs indicated by arrows are nuclearities)
图 2 图 1 中例 2 对应的篇章修辞结构(箭头所指为主要篇章单元)

另一方面,从篇章衔接性角度来看,话题在内容上的完整性往往体现为思维的放射性与表达的线性之间的有机联系.这里所谓“思维的放射性”是指一个话题(或称主题)由若干子话题(或称小主题)构成,而“表达的线性”则是指各分话题的排序应符合思维的逻辑性和次序性,两者一起构成篇章话题结构.

譬如仍然以例 2 作为分析对象,对于自动问答系统而言,我们能够利用图 2 所示的篇章修辞结构为问答系统提供为什么“领导非常器重他”的答案(即回答“Why”问题),但是,如果需要提供“‘他’是谁?”这样的问题答案(即回答“Who”问题)时,图 2 所示的篇章基本结构就显得力不从心了.这时,需要我们构建如图 3 所示的篇章话题指称结构来解决该问题.通过其所含的指称链接关系,我们就能够回答问题“‘他’是谁?”中的“他”即指“李四”.不过,与上述篇章修辞结构类似,图 3 中的单一篇章指称结构也只能够解决“Who”这一类问题,对“Why”问题无能为力.

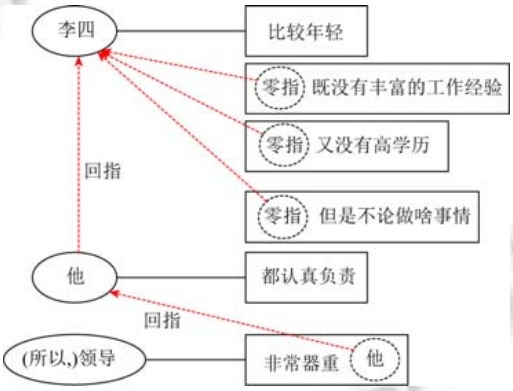

Fig.3 Discourse anaphor structure of the second example shown in Fig.1
(the mentions indicated by arrows are the antecedents)
图 3 图 1 中例 2 对应的篇章指代结构(箭头所指为先行词)

不同篇章基本结构及其关系的研究可以提供不同层面的篇章理解.显然,篇章修辞结构和篇章话题结构这两者相互依赖,相互补充.对于需要解决包含 5W1H 问题(Who,Why,Where,When,What,How)的篇章理解而言,迫切需要联合不同类型的篇章结构共同解决不同类型的篇章理解问题.

2 国内外相关研究

篇章理解是自然语言理解的最终目标.认知科学家和语言学家对这个问题的研究,始于 20 世纪 70 年代.其中,概念依存(concept dependency)理论[2]开启了篇章理解研究的先河,脚本(script)方法紧随其后,用于分析理解

某种具体的场景“故事”.通过对内容的简化处理,类似脚本方法的技术思想已经在信息抽取(information extraction)领域得到成功应用.然而,脚本方法的缺陷在于对领域所在场景存在过度依赖,导致脚本的构建需要随时同步场景变化.这对于有些无法表示为场景的篇章而言,很难采用该类方法加以分析理解,因而进一步需要发现更为通用及开放的结构来表示篇章.为达到此目的,通过探寻篇章的基本特征来寻求解决之道不失为可行方法.

篇章的 7 个基本特征[1]已被自然语言处理领域的研究者广为接受,其中,前 4 个基本特征,即连贯性(coherence)、衔接性(cohesion)、信息性(informativity)及意图性(intentionality)更是有力地促进了自然语言处理研究的发展[3−9].通过分析篇章的衔接性和连贯性,可以发现篇章表层的形式表示;而通过分析篇章的信息性和意图性,则可以挖掘篇章的语义特征.同时,后两者的分析过程需要以前两者为基础关联起来综合考虑.例如,从内容表示角度,篇章的信息性注重新旧信息的变化推进,强调在符合衔接和连贯的特点下,如何合理、恰当地向读者传递新信息.相比于传递新信息的篇章信息性,篇章意图性更关注作者通过传递新信息后所产生的某种期望影响,这也反映了读者对篇章的理解程度.因此,篇章的信息性和意图性与篇章理解存在着密切的深层关系.

无论西方语言或者汉语,篇章的衔接性和连贯性都是最需要关注的两个问题,是篇章的两个最基本特征[1].连贯体现篇章的整体性,是篇章中句子级的关联,采用句子间的语义连接来表示篇章的关联.而衔接是一种词汇级的关联,采用词汇(或短语)之间的语义关联来表示篇章中各语言单元之间的关联.从表达和内容两个角度,通过篇章的连贯性和衔接性的共同作用,篇章的信息性和意图性得以体现,即作者所要表达话题的正确性和可理解性得到保证.

可以看到,篇章的信息性和意图性的研究是以篇章的衔接性和连贯性研究为基础的,目前,篇章分析的研究主要集中在衔接性和连贯性的研究方面,下面分别从篇章结构分析的理论研究、资源建设、计算模型这 3 个方面,重点探讨篇章修辞结构(体现篇章连贯性)和话题结构(体现篇章衔接性)这两种结构,从而充分展现国内外研究现状.

2.1 理论研究

篇章结构理论主要有浅层衔接理论[10]、Hobbs 模型[4,5]、修辞结构理论(rhetorical structure theory)[6,7]、宾州篇章树库理论(Penn discoursetreebank)[11,12]、意图结构理论(intentional structure theory)[8]、主述位结构理论[13]、主位推进理论(thematic progression theory)[14,15]、句群理论[16]、复句理论[17,18]、基于连接依存树的汉语篇章结构(connective-drivendependency tree)理论[19,20]、广义话题结构理论[21−23]等.

2.1.1 篇章修辞结构理论体系

涉及篇章修辞结构理论体系的理论主要包括 Hobbs 模型、修辞结构理论、宾州篇章树库理论、汉语句群理论、汉语复句理论、基于连接依存树的汉语篇章结构理论等.

(1) Hobbs 模型

Hobbs 模型[4,5]提出篇章单元和篇章单元间的连接关系是组成篇章结构的基本部分.其中,篇章单元可以是子句、句子、句群,甚至是篇章本身,而连接关系是指篇章单元间的语义关联性.Hobbs 定义了 12 类关系,包括:详述、并列、结果、背景和时机等.

(2) 修辞结构理论

修辞结构理论(RST)[6,7]是一种基于树状模型的修辞结构理论,早期应用于计算机文本自动生成,目前主要作为篇章结构和功能描述研究的理论基础.RST 与 Hobbs 模型具有很大的相似性,共定义了 4 大类、25 小类修辞关系,每个关系可连接两个或多个篇章单元.如果修辞关系连接的篇章单元间存在主次,那么中心信息单元称作“核(nucleus)”,传达支撑信息的其他单元称作“卫星(satellite)”.当修辞关系连接的单元无主次之分时,则称其为“多核”关系.与 Hobbs 模型相比,RST 更注重句子内部的结构,篇章单元可以小到短语或语块.RST 认为功能语块是最基本的篇章单元(elemental discourse unit,简称 EDU),EDU 间的语义关系具有开放性和可扩充性.在 RST 构造出来的树形结构中,叶节点、非叶节点、弧线和垂直线分别表示 EDU 单元、连续文本块、修辞关系和核心语块.这里的“核心”与 RST 中的 3 个基本概念之一,核心性有关.核心性是指篇章由辅助单元和核心单元构成,

具有不对称性.RST 的另外两个概念分别是“制约因素”和“效果”,前者表示辅助篇章单元及核心篇章单元至少有一个具有制约特性,从而表明命题存在的必要性;后者表示篇章关系的解释机制,即可以用关系达到的效果反向解释关系本身.

(3) 宾州篇章树库理论

宾州篇章树库(PDTB)[11,12]理论将源自修辞结构理论的篇章修辞关系作了改进,将其划分成 3 层,其中,第 1 层共 4 大类,第 2 层 16 类,第 3 层 23 类.相比 RST,PDTB 体系凸显了篇章修辞关系中连接词的作用,它以连接词为核心,根据有无显式的连接词将篇章关系区分为显式和隐式关系,并对隐式关系人工添加了可表示当前语义关系的连接词,在此基础上再标注相关的篇章单元.另外,PDTB 体系中的篇章单元不再考虑短语级,将从句作为最小篇章单位,从而大幅度增加了实用性.

(4) 汉语复句理论

汉语复句理论起始于 19 世纪末,普遍认为是以 1898 年马建忠的《马氏文通》出版为标志[24],创建了汉语复句理论.《马氏文通》是最早讨论到复句问题、首次把复句问题引入汉语语法理论领域的语法著作.然而,另外也有人认为《马氏文通》在分析句子成分时使用的是自己的一套“句读论”,固然已经分析出了许多基本复句类型,但并未明确提出“复句”的概念,是“有实无名”.真正最先提出汉语复句系统之“名”的是严复的《英文汉诂》.

复句由两个或两个以上意义相关、结构上互不作为句子成分的分句组成.分句是结构上类似单句而没有完整句调的语法单位.复句中的各个分句之间一般有停顿,书面上用逗号、分号或冒号表示;复句前后有隔离性语音停顿,书面上用句号或问号、叹号表示.语法上是指能分成两个或两个以上相当于单句的分段的句子.同一复句里的分句,说的是有关系的事.一个复句只能有一个句终语调,不同于连续几个单句[17,18].

(5) 汉语句群理论

句群也叫句组,由前后连贯共同表示一个中心意思的几个句子组成.如同分句组成复句,句子组合成为句群一样的道理[16].语法学对句群的研究最早始于黎锦熙等人[25],在我国汉语语法研究史上首次详尽地论述句群,并提出了“句群是介乎复式句和段落之间的一种语言单位”的定义.

从构成成分来看,句群是句子的组合,至少需要有两个句子组合而成的语言单位才能叫作句群.从语义联系上看,组成句群的句子之间要有紧密的逻辑关系,它们必须共同拥有一个中心思想.从组合方式来看,几个句子运用一定的方式组合在一起成为一个句群,组合方式有两种:语义组合和关联组合.

句群的分类角度有很多,例如:根据句群中句子的结构关系分类,可以将其分为“并列关系”“连贯关系”“递进关系”等 12 种类别.从句群的功能角度来看,则可将其分为主题句群、过度句群和插入句群三大类.句群分类大都是借鉴句子和复句的分类方法,分类方法众多,还未形成统一的标准.

(6) 基于连接依存树的汉语篇章结构理论

苏州大学自然语言处理实验室结合 PDTB 体系中连接词驱动策略和 RST 体系中篇章树形表示结构的优势,同时结合汉语复句和句群理论,提出了一种基于连接依存树(connective-driven dependency tree,简称 CDT)的汉语篇章结构表示体系[19,20,26].该理论对完整的篇章结构(包括篇章单位、连接词、篇章结构、篇章关系、篇章主次)进行了系统的定义和描述.在该基于连接依存树的篇章结构中,叶子节点表示基本篇章单位(elementary discourse units,简称 EDUs),内部节点为连接词(connective),由连接词连接的基本篇章单位组合称为篇章单位(discourse units,简称 DUs).各子句之间通过连接词形成更高一级的篇章单位,层次组合直至形成一棵完整的篇章结构树.连接词既可以表示篇章单位层次,也可以表示篇章单位之间的逻辑语义关系,一个连接词可以连接两个或多个篇章单位,篇章单位根据在篇章中的重要程度可分为主要篇章单位和次要篇章单位.

2.1.2 篇章话题结构理论体系

涉及篇章话题结构理论体系的主要包括浅层衔接理论[10]、主述位结构及推进模式理论[13−15]、意图结构理论[8]、话题链理论[27−32]、广义话题结构[21−23]、微观话题结构理论[33,34]等.

(1) 浅层衔接理论

浅层衔接理论是最早研究篇章衔接关系的理论体系.浅层衔接理论[10]指出,“当篇章中的某个成分的解释

依赖于篇章中另一个成分的解释时,这两个成分之间就产生了衔接关系";衔接方式通常分为语法衔接和词汇衔接两大类,其中语法衔接手段包括指称、省略、替代和(逻辑)连接,连接又划分为增补型(additive)、转折型(contrastive)、原因型(causal)、时间型(temperal)4 类,词汇衔接手段包括词汇的重复和搭配.

Grimes 在深化 Halliday 的浅层衔接理论时考虑了非词汇化的命题关系,给出了更详细的衔接关系类别.此外,Grimes 首次提出了衔接关系的论元有主次之分,并明确指出,并列(paratactic)关系的论元同等重要,而主从(hypotactic)关系的论元有主次之分.

(2) 主述位理论

主述位理论中的主位、述位两个概念,最早来自于布拉格学派提出的功能语句观理论框架[13−15].Mathesius 从功能语句观的角度提出主位、述位信息理论,用于描述句子所传递的信息结构.主位是指在既定语境中已知或至少是明显的信息,是说话人信息的出发点;述位是话语的核心,是说话人对主位的阐发.

Mathesius 对主位的界定涉及 3 个方面的内容:句首性(sentence-initialness)、相关性(aboutness)、信息的新旧性(informational status).随后,Firbas 又从"交际动力"的角度对主位作了进一步阐释:他提出主位是已知信息,所承载的交际动力低;述位是新信息,所承载的交际动力高;主位-述位的推进更替推动了篇章交际动力的动态传递.

此后,以 Halliday(1994 年)为代表的系统功能语言学派认为布拉格学派对主位的界定有些含混,故区分了主位研究的两个层次:句法层次上的主位-述位结构和语意层次上的信息结构.主位-述位结构是从篇章产生的角度来界定的,突出小句或话语的起点,而信息结构(已知/未知信息)是从篇章接受的角度来界定的,侧重篇章解读者对信息的处理.

从篇章功能的角度来看,每个小句和小句复合体的第 1 个句法成分是主位,其余成分是述位.从系统功能语法学角度来看,主位和述位一起构成一则信息,主位是信息的起点,是小句组合的基础;述位是对主位的阐释和发展.

(3) 意图结构理论

意图结构理论由 Grosz 和 Sidner 最早提出[8],他们认为篇章是包含意图的,原因在于篇章的作者就是怀有表达自身意图的目的开始写作的.所以,篇章意图的解释应该和篇章内容一样纳入篇章结构理论的研究范畴,因而意图结构完全可以成为篇章结构理论的基础.在他们提出的篇章结构中,包括 3 个方面,分别是语言结构(linguistic structure)、意图结构(intentional structure)、焦点状态(attentional state).

根据 Grosz 和 Sidner 对篇章结构的定义,篇章意图(discourse purpose,简称 DP)由篇章段意图(discourse segment purpose,简称 DSP)分解和表达,显示出篇章意图的层次性特点.同一个意图层,如果 DSP1 有助于表达 DSP2,则 DSP2 占主导地位,称为支配(dominance)关系,支配关系与修辞结构理论中的"核心-卫星"结构相似,因此可以看作是主次关系在篇章意图层上的定义.

Moser 和 Moore 的研究表明,意图结构理论和修辞结构理论之间存在共性,如意图结构中的支配和修辞结构理论中的核相对应.

(4) 话题链理论

曹逢甫[27]最早提出了汉语话题链(topic chain)的概念,细致地分析了话题在控制小句连接方面的作用.话题链的形成主要依赖各种指代回指(anaphor)形式,即零形回指(zero anaphor,简称 ZA)、代词回指(pronoun anaphora,简称 PA)和名词回指(nominal anaphor,简称 NA)的选择方法.曲承熹[28]总结了前人的研究成果,提出了操作性较强的话题链定义"一组以零回指 ZA 形式的话题连接起来的小句".

刘礼进[29]使用人工标注的小规模汉英篇章对比语料库,深入分析了话题链在汉英篇章的宏观语义结构描述功能上的差异情况;孙坤[30]对英汉篇章组织模式进行了对比研究;王建国[31]把话题链的描述作用从句子拓展到句群和篇章,重新定义话题链为"由同一话题引导的系列语句",并深入分析了话题链在汉英篇章中的不同描述特点;周强[32]引入话题链描述形式,设计不同类型的话题评述关系集,构建了以话题链为主,融合关联词语和其他连贯形式的描述机制.

话题链是指由各个话题连接而成的链条.根据话题相同与否以及是否包含不同话题,话题链可分为“同题链”“异题链”和“包题链”3 种基本类型.同题链是相同的话题形成的话题链;异题链是由不同的话题形成的话题链;包题链是由有包容关系的话题形成的话题链.在实际的篇章中,同题链、异题链、包题链层层相套,互相交错,交织形成话题网,共同推进篇章的发展(生成).

(5) 广义话题结构理论

宋柔等人针对汉语篇章话题结构进行了比较深入的研究,根据汉语篇章的特点,以标点句为基础,给出了广义话题结构的概念和相应的表示方法,提出了“话题的不可穿越性”和“话题句的成句性”两个广义话题结构性质;描述了汉语的话题结构和话题句特征,给出了话题句动态堆栈模型[21−23].这一研究成果是汉语篇章分析领域的一项开创性工作.但同时,广义话题理论的动态堆栈模型,强调子句语法成分的完整性,在分析层面描述粒度过细,在操作层面也面临可计算问题.

(6) 微观话题结构理论

苏州大学自然语言处理实验室在分析话题结构相关理论的基础上提出了基于主述位理论的篇章微观话题结构表示体系[33,34].该体系从篇章视角确立基本微观话题单元,将该单元表示成包含主位和述位的实体形式化表示模式,并基于主位推进理论搭建基本微观话题的上下文关联模式,再融合实体和上下文关联形成完整的汉语篇章话题结构表示体系.

2.2 资源建设

目前篇章结构的资源建设主要与上述篇章修辞结构(篇章连贯性)和篇章话题结构(篇章衔接性)理论体系相关,代表性资源包括修辞结构篇章树库(rhetorical structure theory discourse treebank,简称 RST-DT)[35]、宾州篇章树库(Penn discourse treebank,简称 PDTB)[36]、ACE(automatic content extraction)评测语料[37]、ARRAU[38]、OntoNotes[39]和篇章图库(GraphBank)[40]等.

2.2.1 篇章修辞结构资源建设

目前与篇章修辞结构有关的英文资源主要包括宾州篇章树库 PDTB[36]和修辞结构篇章树库 RST-DT[35].

(1) PDTB:由美国宾夕法尼亚大学、意大利托里诺大学和英国爱丁堡大学联合标注,并由 LDC(linguistic data consortium)于 2006 年正式发布.2008 年 PDTB 2.0 发布,它是目前规模最大的英文篇章语料库,共标注了 40 600 个关系,其中,包括 18 439 个显式篇章关系,16 224 个隐式篇章关系,624 个由非连接词表示的篇章关系,5 210 个通过实体重复或共指表示的关系,还有 254 个相邻句子不存在所定义的关系.

(2) RST-DT:由美国南加州加利福尼亚大学标注,并由 LDC 于 2002 年正式发布.RST-DT 选用宾州树库的文章构建二叉修辞结构树.RST-DT 对 EDU 进行了严格的定义,规定主语或宾语从句不属于 EDU,充当主要动词的补语的从句也不属于 EDU.此外,所有词汇或句法标记的起状语作用的从句属于 EDU,定语从句、后置的名词修辞短语或将其他 EDU 分割开的从句或非谓语动词短语为内置语篇单位.RST-DT 完成了 85 篇文章的标注,共标注了 53 种单核心关系和 25 种多核心关系,这 78 种关系又分成 16 个组别,每组都具有相同的修辞功能.标注的文章内容涉及到财政报道、商业新闻、文化点评、读者来信等多种话题.

相比英语,汉语篇章修辞结构的资源构建主要采用 4 种方法.

(1) 基于 RST 的标注

乐明[41]以 RST 为指导,参考汉语复句和句群理论,进行了篇章结构标注的尝试.他定义了 12 类 47 种汉语修辞关系,以句号、问号、叹号、分号、冒号、破折号、省略号及段落结束符等为标记定义汉语基本篇章单位,完成 97 篇财经评论文章的修辞结构标注,探索了中文篇章分析中采用 RST 的可行性.陈莉萍[42]试图采用 RST 标注汉语篇章,其基本篇章单位以标点分割,如“目前,…”中的“目前”也会作为基本篇章单位.他们的研究都表明 RST 的很多篇章关系无法在汉语中找到与之对应的关系.

(2) 基于 PDTB 体系的标注

Zhou 和 Xue[43]尝试使用 PDTB 体系标注汉语,PDTB 体系以连接词为谓词标注其论元结构,结合汉语自身的特点对 PDTB 体系进行了改进,并以此为参考从中文树库(Chinese Treebank,简称 CTB)中选取了 98 篇新闻语

料进行了标注.2015 年,Zhou 和 Xue[44]进一步将该语料扩大到 164 篇,并最终提交 LDC 对外进行发布.但汉语中连接词大量缺省,PDTB 体系表现出很大的不适应;又由于连接词并不能覆盖每一个篇章单位,PDTB 体系通常不能构建一个完整的篇章结构,这对篇章结构分析而言显然缺少了很重要的内容.张牧宇等人[45]在英文篇章关系研究的基础上分析了中英文的差异,总结了中文篇章语义分析的特点,提出一套面向中文的层次化篇章关系体系,并进行了标注实践,目前发布了哈尔滨工业大学中文篇章关系语料(HIT-CDTB),该语料选取 LDC 发布的 OntoNotes 4.0 中的 525 篇汉语文本按照 PDTB 体系进行了分句、复句和句群 3 个层次的篇章关系的标注.标注内容包括显式篇章关系的关系连接词、关系元素和关系类别信息;以及隐式关系的可插入的连接词和篇章关系类别信息.他们将篇章关系分为时序、因果、条件、比较、扩展和并列这 6 类,标注的关系连接词共 1 472 类.

(3) 采用汉语本土复句和句群理论标注

参考邢福义的汉语复句研究成果[17],华中师范大学标注了汉语复句语料库[46],目前已收有标复句 658 447 句,约 44 395 000 字,语料来源以《人民日报》和《长江日报》为主.但汉语有标复句只占汉语复句的 30%左右,这就使得该语料库的应用受到很大限制.而且该语料库仅关注复句内部关系,没有涉及句子及其以上篇章单位的结构问题,这显然不能满足篇章结构分析的需求.清华汉语树库(Tsinghua Chinese Treebank,简称 TCT)[47]是从大规模的经过基本信息标注的汉语平衡语料库中提取出 100 万汉字规模的语料文本,经过自动断句、自动句法分析和人工校对,形成的高质量汉语句法树库语料.TCT 中标出了复句内各分句之间的关系信息,复句分类采用比较常用的并列关系、连贯关系、递进关系、选择关系、因果关系、目的关系、假设关系、条件关系、转折关系分类方法.但清华汉语树库中没有标注特定复句关系所对应的复句关系词,也没有标注句子之间的关系.

(4) 基于连接依存树的篇章结构资源建设

苏州大学自然语言处理实验室结合 PDTB 和 RST 体系的优势,提出了使用连接依存树(CDT)表示汉语篇章修辞结构的方案,并基于该方案,选取宾州汉语树库 6.0 版(Penn Chinese TreeBank,CTB 6.0)上的 500 篇文章进行了篇章修辞结构的标注,构建了汉语连接词驱动的篇章语料库(CDTB)[19,20],每个段落标注为一棵连接依存树,共有效标注 2 342 个篇章(段落),标注信息包括基本篇章单位、连接词、篇章结构、篇章关系和主次篇章单位.

表 1 给出了篇章修辞结构的 4 种核心体系的对比情况,从中可以看出,CDT 借鉴了 RST、PDTB 和汉语的复句、句群理论,一方面明确了 EDU 和篇章树结构,考虑汉语中的复句,以标点句作为 EDU 判别的基本依据;另一方面兼顾了连接词在篇章关系中的地位,以连接词为关系类别判断的基点,可实现关系不同分类体系的迁移.

Table 1 Comparison of several important architectures of discourse rhetorical structure

表 1 篇章修辞结构的核心体系的对比

体系类别	RST	PDTB	复句、句群理论	CDT
基本篇章单位	EDU,短语、子句都可以作为 EDU;一个关系有一个或多个 EDU	论元,采用谓词-论元模式,一个篇章关系有两个论元	复句内的分句	EDU,以子句为单位;一个关系有多个 EDU
连接词	无	标注显式连接词和可添加的隐式连接词	无	显式连接词;显式连接词是否可删;可添加的隐式连接词
篇章关系	给定语义类别并标注了 78 类	给定 3 层语义类别,标注语义类别和连接词	复句内关系	用连接词代表关系;将连接词映射到一个 3 层的关系体系上
结构	一篇文章对应一棵完整的篇章结构树	浅层篇章结构,以谓词-论元模式体现	复句内浅层关系,无结构	一个段落对应一棵完整的篇章结构树
中心	中心由具体关系类别决定	无(两个论元不分主次)	无	中心由全局意图决定,与关系无直接关系

表 2 给出了 3 个具有一定影响力的汉语篇章修辞结构语料库的对比情况,其中,HIT-CDTB 和 LDC-CDTB 都遵循了 PDTB 体系,进行了篇章关系的浅层标注,SUDA-CDTB 则遵循了 CDT 体系,进行了篇章树结构的标注.

Table 2 Comparison of Chinese corpora for discourse rhetorical structure

表 2 汉语篇章修辞结构语料库对比

数量	HIT-CDTB	LDC-CDTB	SUDA-CDTB
文档	525	164	500
显式关系	11 519	1 223	1 812

Table 2 Comparison of Chinese corpora for discourse rhetorical structure (Continued)
表 2 汉语篇章修辞结构语料库对比(续)

数量	HIT-CDTB	LDC-CDTB	SUDA-CDTB
隐式关系	9 848	4 193	7 310
结构	无	无	2 342
连接词	1 472	256	275
状态	已发布,可免费获取	已发布,可从 LDC 购买	已发布,可免费获取

2.2.2 篇章话题结构资源建设

篇章话题结构方面的语料库相对较少,主要包括面向话题指称结构、面向篇章意图性、汉语篇章广义话题结构和基于主述位理论的汉语微观话题语料库资源建设等.

(1) 面向话题指称结构的语料库资源建设

指称结构是一种存在于篇章中前后两个语言单位之间的特殊语义衔接关系,而确定两者的过程即称为指称消解.目前主要的语料资源有 ACE 评测语料[37]、ARRAU 语料库[38]、OntoNotes 语料库[39].

- ACE 评测语料

ACE 是美国政府支持的自然语言处理重要会议,ACE 语料评测起始于 2000 年,自 2004 年开始引入中文语料.ACE 评测语料基于之前的 MUC 评测语料,其中的指代信息采用指代链的形式标注而成,每个指代链独立编号并被记录在文件中,而相同指代关系的实体都位于同一个指代链上.MUC 和 ACE 评测语料为面向衔接关系的自然语言处理研究提供了重要的语料资源,但在它们通过指代形成的语料衔接关系资源中,仅仅标注了显式实体指代,而忽略了对隐式实体(或称为省略)的指代标注.

- ARRAU 语料库

由 University of Trento(意大利)和 University of Essex(英国)针对较难处理的指代问题,联合建立的指代标注语料库.该语料包括对话、说明文和新闻报道,不仅标注了实体指代,也标注了抽象指代(如事件、行为指代),但并不包含汉语部分.

- OntoNotes 语料库

由 BBN Technologies、University of Colorado(美国)、University of Pennsylvania(美国)和 University of Southern California's Information Sciences Institute(美国)相互合作创立.OntoNotes 集成了多层面的标注,包括词汇层面、句子层面和篇章层面的标注,并不为特定评测服务.OntoNotes 在篇章层面主要包含实体间以及事件的共指关系.OntoNotes 中既包含英语,也包含汉语,汉语部分还标注了主语位置的零指代信息.

虽然面向话题指称结构的语料库资源相对丰富,但是对于汉语中非常突出的零指代问题,资源却非常匮乏.OntoNotes 语料虽然包含了少量的主语位置的零指代信息,但该语料更多关注的是句法成分的缺失,面向篇章分析的零指代标注资源极其匮乏.

(2) 篇章意图性资源建设

为克服子句间的多种篇章关系不能被树模型的篇章结构有效表达这一缺陷,Wolf 和 Gibson 提出了通过图结构表示篇章的方法[40],并研究了篇章图库(discourse graph bank,简称 DGB)的构建问题.同时,以该结构标注了 135 篇文章.该方法主要分为 3 步:首先,根据标点符号将篇章分为基本单元(句子/子句),称为篇章段(discourse segments);然后,再根据标点符号和话题,将上述基本单元归并成组(group),每一个组都集中表达了某个话题;最后,确定基本单元、组之间的连贯关系(coherence).

(3) 汉语篇章广义话题结构资源建设

在针对广义话题结构理论的语料资源方面,宋柔课题组基于他们提出的广义话题结构的概念,以标点句为基本篇章单位,开展了汉语篇章的话题结构标注工作[21−23].目前,已标注了《围城》、《鹿鼎记》和其他语料(涉及章回小说、现代小说、百科全书、法律法规、散文、操作说明书等语体),共约 40 万字.其中,《鹿鼎记》第 1 回的广义话题结构标注及其说明已在网上公开发布(http://clip.blcu.edu.cn/).

(4) 基于主述位理论的汉语微观话题语料库资源建设

苏州大学自然语言处理实验室提出了基于主述位理论的篇章微观话题结构表示体系[33,34],并据此标注形

成了 500 篇文本的微观话题结构语料库 CDTC(Chinese discourse topic corpus)[48,49].该语料从 CTB 6.0 中选取 500 篇文档标注了基本篇章单元、基本篇章话题的主位(theme)和述位(rheme)、篇章微观话题结构(micro-topic scheme)、微观话题联接、微观话题链等信息,为微观话题结构的自动分析奠定了基础.

2.3 计算模型

基于不同的理论体系和相应的语料库,近年来很多有关计算模型的研究工作陆续展开,下面我们就按研究的不同角度分别展开介绍.

2.3.1 篇章修辞结构计算模型

(1) 基于 RST-DT 的研究

基于 RST-DT 的篇章结构分析主要包含两个子任务:EDU 的识别和篇章连接关系的生成.其中,EDU 的识别负责对文本进行切分,提取出 EDU,即构造生成的修辞结构树的树叶;连接关系的生成则采用自底向上的方法生成修辞结构树中的功能节点,并为每一节点确定一个最可能的修辞关系.

关于 EDU 的自动识别研究较多,结果也比较理想.其中比较有代表性的研究包括:Soricut 等人[50]采用基于统计的方法进行识别,EDU 识别在自动句法树上获得 $F1$ 值为 83.1%,在标准句法树上 $F1$ 值为 84.7%.Hernault 等人[51]给出了一个基于序列数据标注的篇章分割模型,使用词汇和句法特征,采用 CRF 进行学习,实验结果表明,作者的序列篇章分割模型 $F1$ 值达到 94%,接近于人工篇章分割的 $F1$ 值 98%.综上可知,目前 RST-DT 上 EDU 识别准确率较高,但进一步提升的空间不大.

在篇章连接关系的生成方面,结果则不理想.Soricut 等人[50]利用语法和词法信息进行句子级的篇章结构分析,他们的算法称为 SPADE,在篇章关系识别时采用概率模型计算各种篇章关系的概率.篇章结构分析模型采用全自动的方法,识别无标注的篇章关系 $F1$ 值为 70.5%,采用正确的基本篇章单位和正确句法树的结果是 96.2%.但是,SPADE 并不对整篇文本进行篇章关系识别.Huong 等人[52]给出了一个文本自动篇章结构生成系统,该系统分为两个层次:句子级的篇章结构分析和文本级的篇章结构分析.句子级的篇章结构分析使用句法和线索词来进行基本篇章单位的识别和篇章结构的生成.对于篇章级别,为缩小篇章结构分析的搜索空间,加入了文本相邻和文本组织限制.最终在缩小搜索空间后,系统的 $F1$ 值达到了 70.1%,其缺点就是计算量较大.Hernault 等人[53]在 RST 上实现了基于 SVM 的篇章结构分析器 HILDA.对篇章切分和关系识别使用 SVM 训练了分类器,采用贪婪的自底向上的方法构建篇章结构树,篇章结构树构建的时间复杂度取决于输入文本的长度.HILDA 在树构建和篇章关系分析上的效果较好,结构识别 $F1$ 值为 72.3%,完整句法树识别 $F1$ 值为 47.3%.Feng[54]在 HILDA 的基础上进行了篇章结构树的构建和关系识别,抽取了更丰富的特征,性能比 HILDA 有所提升.Joty 等人[55]给出一种使用动态条件随机场进行句子级篇章分析的方法,使用人工 EDU 切分结果识别 18 类关系 $F1$ 值为 77.1%.Surdeanu 等人[56]利用感知器模型结合逻辑回归算法进行结构创建和关系预测,同时,该分析器还借助预训练的句法依存树获取句法特征.近几年来,研究人员开始注重用若干篇章中文本的分布特征来表示篇章的内部单元.Braud 等人[57]使用层次神经网络模型(hierarchical bi-LSTM)构建了一个端到端的篇章分析器.Li 等人[58]用基于注意力的层次型双向 LSTM 模型结合 CKY 算法构建了图篇章解析器.Braud 等人[59]使用一种前馈神经网络模型构建了两种过渡型篇章分析器.Ji 和 Eisenstein[60]使用支持向量机结合 shift-reduce 转移系统构建了 DPLP 篇章分析器.导致篇章分析结果较低的主要原因是 RST-DT 中标注的篇章结构树的数量有限,模型没有能力获取深层次的语义信息.

(2) 基于 PDTB 的研究

宾州篇章语料库(PDTB)的构建,以及 CoNLL 2015 和 2016 年 Shared Task 的举办,显著推动了篇章结构分析的研究,在篇章计算方面受到了极大的关注.

基于 PDTB 的篇章分析包含论元的抽取、篇章关系的识别和端到端系统的构建这 3 个方面,下面分别加以介绍.

- 论元的抽取

代表性的工作包括:Dines 等人[61]针对 Subordinate 类型的连接词提出了一种 tree subtraction 算法来自动完

成句内论元的抽取,但该方法使用了一套具有很强针对性的规则,对其他类别的连接词并不完全适用.Lin 等人[62]借鉴 Dinesh 的 tree subtraction 算法,借助机器学习方法首先识别覆盖论元的最小子树,再利用 tree subtraction 算法在子树中抽取论元.但覆盖论元的最小子树也会包含非论元的部分,造成后续的抽取不能完全正确.他们的实验结果也证实了这一点:完全精确匹配的标准下,Arg1 和 Arg2 同时正确的性能仅为 40%,而在部分匹配的标准下,这一性能可达到 80%以上.Wellner 等人[63]提出一种机器学习的方法来确定连接词对应论元 Arg1 和 Arg2 的 head,但是 PDTB 语料中并没有标注论元的 head 信息,因而评测上缺乏一致的标准.Ghosh 等人[64,65]基于条件随机场模型将论元抽取看成序列标注问题,给出了一个论元识别方案,但他们使用了一些来自 PDTB 的标准信息,例如语义类别、Arg2 信息等,给出的结果也只考虑了标准句法树,未对自动句法分析结果进行评测.Kong 等人[66]借鉴 SRL 中的句法树裁剪策略给出了一个论元构成子树的提取方案,并借助 ILP 进行全局最优,大大提升了完全精确匹配下论元识别的性能.

- 篇章关系识别

Pitler 等人[67]指出,在 PDTB 篇章语料库中隐式篇章关系与显式篇章关系大约各占一半.由于显式篇章关系中连接词(connective)的存在且歧义较少(大约只有 2%),因此比较容易识别.这使得隐式篇章关系研究成为篇章结构关系分析成败的关键.识别隐式篇章关系的研究可以归纳为 3 类:基于伪隐式篇章关系语料的研究,基于纯隐式篇章关系语料的研究和基于伪隐式和纯隐式的篇章关系混合语料研究.基于伪隐式关系的研究的代表性工作包括:Marcu 和 Echihabi[68]首次提出使用无监督的方法识别隐式篇章关系.他们使用一系列文本模式从网络上自动获取语料资源,同时去除篇章连接词构成一个伪隐式篇章关系语料.他们的实验结果表明,使用词对(word-pairs)特征为识别隐式篇章关系提供了帮助.Saito 等人[69]扩展了他们的工作,从文本域中提取短语模式特征,实验结果表明,同样有助于提高隐式篇章分析的性能.尽管如此,我们认为伪隐式篇章关系并不能从真正意义上代表纯隐式篇章关系,因为它们在表示关系上存在着很多的不同,比如隐式关系的存在表明上下文的联系足够强而不需要使用篇章连接词来衔接.

随着 PDTB 2.0 的发布,该语料显式地区分了隐式篇章关系和显式篇章关系,并且仅针对段落内相邻句子间的隐式篇章关系进行标注.至此,很多工作开始侧重研究纯隐式篇章关系识别.这方面具有代表性的工作包括:Pitler 等人[67]首次提出使用不同的语言学特征,比如动词、极性和上下文环境等,识别隐式篇章关系.Lin 等人[70]受 Pitler 等人的启发,首次提出使用两类句法特征,即成分句法推导规则和依存句法推导规则,来识别 PDTB 中第 2 层隐式篇章关系.Park 和 Cardie[71]使用了贪婪的特征选择算法确定了识别隐式篇章关系的最优特征子集.他们的实验在第 1 层 4 大类关系上取得了最好的 $F1$ 值.近年来,一些研究表明,样本不平衡问题成为了提高隐式篇章分析性能的重大阻碍.有人提出使用伪隐式和纯隐式关系混合的篇章关系来进行分析.相关工作包括:Zhou 等人[72]使用语言模型来计算困惑度以判断相邻句子间插入连接词的合理性.Biran 和 McKeown[73]使用聚集词对尝试解决特征稀疏问题,但他们的实验结果表明性能提升很小.为了解决隐式关系标注样本缺少的问题,Lan 等人[74]提出使用多任务学习的方法引入伪隐式篇章关系来辅助隐式篇章关系的识别.Zhou 等人[75]提出一种基于信息检索的无监督方法识别隐式篇章关系,他们利用 Web 上的资源提取大量的伪隐式关系辅助识别隐式篇章关系.

近几年,越来越多的研究人员开始寻求用神经网络的方法来完成隐式篇章关系识别的任务.同时,为了缓解有标数据缺少带来的问题,很多传统算法和神经网络算法都借助没有标注的数据,辅助完成隐式篇章关系识别.Lan 等人[76]提出了一种基于多任务注意力机制的神经网络来解决隐式篇章关系的表示和识别问题,并取得了当前最好的性能.

- 端到端的篇章结构分析

Lin[77]研究如何在 PDTB 上进行篇章结构分析,对于难度较大的隐式篇章关系识别,采用上下文、词对、句法特征、依存树特征进行识别.整个系统包括连接词识别、论元识别、显式关系分类、隐式关系分类、属性标注,这是第一个端到端的 PDTB 分析工作.此后,随着 CoNLL 2015 和 2016 年 Shared Task 以端到端的篇章逻辑语义分析为任务,大量工作随之展开,主要可以分成 3 类:一是跟随 Lin 等人的工作,进一步完善各个模块;二是借

助 ILP、Structured Perceptron 等全局优化策略对系统进行全局优化;三是引入神经网络、深度学习框架对平台中影响性能的论元识别和隐式关系识别进行改进.

(3) 汉语篇章修辞结构分析

由于语料缺乏,这部分研究受到了制约.代表性的工作包括:张牧宇等人[78]在哈尔滨工业大学中文篇章关系语料(HIT-CDTB)上进行显式篇章句间关系和隐式篇章句间关系识别,并给出初步的实验结果,但其所标语料参考英语 PDTB 体系,不能进行完全的篇章结构分析,只能进行部分篇章分析.CoNLL 2016 的 Shared Task 中以 Zhou 和 Xue[44]标注的、LDC 发布的 CDTB V0.5 为语料,引入了汉语浅层篇章修辞结构分析的任务,使得汉语浅层篇章修辞结构分析得到了一定的关注,但大部分工作都采取用英文一致的体系进行.涂眉等人[79]在 TCT 上进行了基于最大熵的汉语篇章结构自动分析方法,实验结果表明,篇章语义单元自动切分的 *F*1 值能达到 89.1%,当篇章语义结构树高度不超过 6 层时,篇章语义关系标注的 *F*1 值为 63%.Kong 等人[80]基于苏州大学的 CDTB 语料采用流水线的方式构建的端到端的中文篇章解析器,该平台包括子句识别、连接词识别与分类、隐式篇章关系识别、篇章单位主次识别等部件,最终输出构建完成的篇章结构树.在 CDTB 上的结构性能的 *F*1 值达到了 46.7%,但若再综合进篇章树中的每个关系的具体属性,整个分析器的 *F*1 性能只有 20.0%.Jia 等人[81]利用转移系统和深度学习的方法,给出了一个完整的从平文本到树形结构的篇章结构自动解析框架,在英文 RST 和苏州大学的 CDTB 语料上都取得了较好的性能.孙成等人[82]给出了一个完整的基于转移系统的篇章结构树的生成框架,并参考 RST 上相关评价体系给出了完整的汉语篇章结构树的评价体系.

2.3.2 篇章话题结构计算模型

受限于理论体系的可计算性和相应语料资源的匮乏,目前有关篇章话题结构的计算模型研究主要集中在指代结构的研究,而指代结构的研究又分别从实体指代、事件指代和零指代 3 方面展开.

(1) 实体指代消解研究

作为信息抽取的核心组成部分之一,指代消解一直都是自然语言处理领域的一个研究热点.早期指代消解方法均采用启发式规则方法,从 20 世纪 90 年代开始,随着各类指代消解标注语料的不断发布以及一些有影响力的自然语言处理会议和公开评测的召开,例如 MUC(Message Understanding Conf.)[83,84]、ACE(automatic content extraction)[37]、CoNLL shared task[85,86]等,指代消解的研究重点也转向了数据驱动的指代消解方法研究.目前主流的方法有:

- 基于规则的方法:2010 年,Raghunathan 等人[87]提出了一个基于多重过滤框架的共指消解模型.这个框架是由 7 个消解模块组成,这些模块按照精度从高到低进行排列,每一层的输入以上一层输出的实体聚类体为基础.该框架通过共享属性传递全局信息保证了强属性信息的功能要优于弱属性,也使得过滤模型做出共指判断时能使用所有的属性信息.2011 年,Lee 等人[88]基于 Raghunathan 的思想进行了扩展,通过添加过滤器,增加候选先行语的抽取和确定以及全局优化,使得系统在 CoNLL-2011 Shared Task 测评中获得最高的准确率.
- 基于统计的方法:1999 年,Cardie 等人[89]提出通过聚类方法进行名词短语的同指消解,其基本思想是收集篇章中的基本名词短语,根据短语的特征对名词短语聚类,判断两个名词是否属于同一个类.
- 基于分类的方法:1995 年,McCarthy[90]把判断先行语的问题转换成分类问题,通过分类器判断指代语与每个先行词候选之间是否存在指代关系.这一思想为日后指代消解的研究开辟了一条全新的道路.Soon 等人[91]则给出了详尽且完整的实现步骤,并开发出实用的系统.在此基础上,许多研究者进行了不同程度的扩充和改进,主要包含3类:(1) 抽取强而有力的平面特征以及篇章中结构化信息支持学习模型.例如,2012年,孔芳等人[92]提出基于树核函数的中英文消解方法;(2) 单一模型向多重模型融合逐渐演变,并以此增强分类器效果.例如,2012 年,Xu 等人[93]提出融合基于规则与基于分类的方法用于指代消解;(3) 优化共指链的形成.2012 年,Belder 等人[94]提出一种新的方法优化二元分类后共指链链接问题,把共指链接问题看成是一个线性规划问题,并提出用列生成的方法获取最优解以此达到准确消解的目的.
- 深度学习方法:深度学习是通过模拟人脑神经元和突触处理感知信号的过程,构建含多个隐层的机器学习模型.其主要优势在于能自动地学习数据中比浅层特征更加抽象的高层特征表示.Wiseman[95]提出利用循环

神经网络来学习潜在的、全局的实体聚类的特征表示,利用贪婪搜索算法实现实体-实体表达模型.Clark[96]使用增强学习方法结合神经网络对实体表达排序模型进行直接优化,并提出了两种优化算法:增强策略梯度算法和奖励重调最大化算法,后者实现了更好的性能.Lee[97]利用循环神经网络对实体表达的上下文信息进行编码,结合单词的分布式表达,利用注意力机制形成 mention 的有效表示,然后最大化得分函数来训练神经网络,在 CoNLL 2012 任务上取得了最好的结果.

上述研究主要针对英文.相比英文指代消解,目前汉语指代消解的研究要少很多,主要属于跟进型研究.代表工作包括:王厚峰等人[98−100]分别从领域和语义等知识出发,提取规则进行了指代消解的研究;李国臣等人[101]将英文平台的类似做法移植到中文指代消解中,采用决策树方法对中文人称代词的消解进行了研究.周俊生等人[102]提出了一种基于图划分的无监督的汉语指代消解算法,其性能与监督的汉语指代消解性能相当;杨勇等人[103]给出了一个基于机器学习的指代消解平台,并对指代消解中各类距离特征对指代消解性能的影响进行了深入的探索;王海东等人[104]探索了语义角色对指代消解性能的影响,他们的研究表明,语义角色信息的引入能够显著提高指代消解的性能;李渝勤等人[105]针对基于机器学习的中文共指消解中不同类别名词短语特征向量的使用差异,提出一种基于特征分选策略的方法,提高了共指消解的性能.张牧宇等人[106]提出一种利用中心语信息的新方法.该方法首先引进一种基于简单平面特征的实例匹配算法用于共指消解.在此基础上,又引入了先行语与照应语的中心语字符串作为新特征,并提出一种竞争模式,将中心语约束融合进实例匹配算法,提升了消解效果.Song 等人[107]提出一种基于马尔可夫逻辑网的共指消解模型.

(2) 零指代研究

除上述名词短语的指代消解外,零指代现象在中文中频繁出现,近年来,中文零指代成为研究热点.代表性的工作有:Zhao 等人[108]给出一个完整的基于机器学习的中文零指代消解方案,并提出一套有效的适用于中文零指代任务的特征集合.但是他们的工作主要关注零指代的消解子任务,对零指代项的识别仅给出一个保证高召回率的规则方法.他们的实验结果也表明,过低的零指代项识别准确率会严重影响后续消解的性能.Kong 等人[109]给出一个中文零指代消解的完整框架,将中文零指代消解清晰地划分成零元素识别、零待消解项识别和零元素消解 3 个子任务,并采用基于树核函数的方法分别给出每一个子任务适用的结构化特征集.但是,他们仅关注平台的统一性,只给出了标准句法树上平台的性能,未给出完全自动状况下方法有效性的验证.Chen 等人[110]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.Chen 等人[111]给出一个无监督方法的生成式模型,并借助它进行中文零指代消解.基于这一工作,Chen 等人[112]进一步在生成式模型中基于概率将零待消解项识别和消解任务进行联合学习,取得了一定性能的提升.Chen 等人[113]又进一步在该平台中引入深度学习方法,取得了更好的性能.Sheng 等人[114]在传统零指代消解平台中考虑了篇章修辞结构信息,从篇章修辞树结构中提取各类篇章级的信息来帮助中文零指代,并通过一系列实验验证了修辞结构信息的引入能够提升中文零指代的性能.Kong 和 Zhou[115]参考普通名词短语消解平台的研究进展,提出了一种全新的链到链的中文零指代消解方案,其基本思想是将普通名词短语的指代消解结果看作对中文零元素的先行词候选的一种过滤,并以指代链为单位进行中文零指代消解,实验取得了目前最好的性能.Yin 等人[116]提出了一个借助深度记忆网络将零元素的上下文信息向量化,从而自动学习相关的语义信息来帮助零指代.Zhang 等人[117]给出了一种深度神经网络方法,通过对零元素的上下文和可能的先行词候选及其上下文进行高效的向量化表征来提升零指代的性能.Liu 等人[118]为了解决零指代标注语料不足这一问题提出了一种自动生成大规模伪训练语料的方法,使用这些伪语料,借助神经网络方法提升汉语零指代消解的性能.进一步地,Yin 等人[119]在神经网络平台中引入强化学习策略,进一步提升了汉语零指代消解的性能.

(3) 事件指代消解研究

受限于标注语料及任务的复杂度,相比实体指代消解而言,事件指代消解的相关研究刚刚起步,大多参考实体指代消解的解决思路.主要的代表性工作有:2006 年,Ahn[120]通过构建事件对,计算事件对之间的相似度来判断事件的同指关系.随着机器学习方法的推进,事件指代消解任务的研究转向通过人工构建事件的特征来计算事件之间的“距离”,进而判断同指关系.Chen 等人[121]利用最大熵模型建立事件指代消解系统,并在各项评测指

标下评估了系统的性能.Bejan 和 Harabagiu[122]运用无监督的非参贝叶斯模型将词汇特征和 WordNet 中的语义相似度引入事件指代消解任务中.2015 年,Araki 等人[123]首次提出一种联合学习模型,即将事件抽取任务和事件指代消解任务同时研究.随后 Lu 和 Ng[124]也构建了一个基于一元二元以及三元特征融合的联合学习模型.近年来,神经网络在自然语言处理的各个领域都取得不错的研究成果,Nguyen[125]通过非连续卷积模型在 KBP[126]语料上完成事件指代消解任务的研究.同年,Krause 等人[127]也搭建了卷积神经网络模型,并在 ACE 和 ACE++语料进行了相关任务研究.在中文事件指代消解方面,受限于语料,目前只有少量工作,代表性工作包括:Lu 和 Ng[124]构建的平台不仅汇报了英文事件指代消解的性能,也汇报了 KBP 中文语料上的性能;滕佳月等人[128,129]基于 ACE 中文语料进行了中文事件指代消解的研究,并提出了基于全局优化进行性能改善的策略.

除指代外,针对篇章意图性的计算模型的研究很少,代表性工作是 Pustejovsky 等人[130]在 GraphBank 上的相关工作,他们对 GraphBank 进行了分析,认为篇章连接词和两个句子间的跨度距离是高效识别显式和隐式篇章关系的关键因素.

2.4 存在的问题和研究趋势

从上述国内外研究现状的分析中我们可以看到,相比英语,汉语的篇章研究刚刚起步,汉语篇章阅读理解研究鲜有见诸文献.目前汉语篇章理解还存在如下一些主要问题.

(1) 适用于汉语篇章阅读理解的篇章结构理论体系很不完善.有必要借鉴英语的相关篇章理论,并结合汉语特点和复句、句群、广义话题结构等本土理论,逐步建立汉语篇章结构理论体系.

(2) 适用于汉语篇章阅读理解的篇章结构大规模标注资源非常缺乏.虽然有一些研究者,或基于英语篇章理论体系,或基于汉语的复句、句群和广义话题结构等理论,对汉语篇章结构资源库展开了研究,但相关研究比较分散,大多属于探索性工作,有待进一步深入、系统地进行研究.

(3) 适用于汉语篇章阅读理解的篇章结构分析关键技术十分匮乏.由于适用于汉语篇章结构分析的理论体系尚未有效建立,相关标注资源缺乏,因此很难大规模有效地进行关键技术研究.

(4) 篇章理解需要涉及不同视角、不同层次的篇章结构分析结果,各种结构间也存在明显的互补关系,构建统一体系(包括理论体系和资源)进行多视角、多层次的联合分析研究,有待进一步深入.

2.5 机器阅读理解的相关研究

虽然适用于汉语篇章阅读理解的篇章结构分析研究处于起步阶段,机器阅读理解的相关研究却吸引了众多研究者.目前,机器阅读理解方面已经开展了一些工作,具体包括:Hermann 等人[131]借助爬虫技术从 CNN 和每日邮报新闻网页爬取数据,构建了一个完形填空类型(cloze-style)的阅读理解数据库 CNN and Daily Mail.2016 年,斯坦福大学通过亚马逊众包平台建立了一个新的阅读理解数据集 SQuAD[132],它包含 536 篇维基百科文章,100 000 多个问题,而且每篇文章都是经过人工阅读,提出问题并给出答案片段.微软公司选取了 100 000 多名用户通过 Bing 搜索引擎提出的问题,每一个问题都会对应大约 10 篇相关的从网页抽取的文章,相关人员会根据 10 篇文章给出问题的答案,以此构建了 MS MARCO[133]语料库.随着这些语料的正式发布,各种机器学习方法、深度神经网络方法和 attention 机制都不断被提出并被应用到这一任务中[134−142].此外,Cui 等人[143]发布了第一个中文 cloze-style 阅读理解语料 People Daily News 数据集和 Children's Fairy Tale(CFT)数据集.从 2017 年至今,"讯飞杯"中文机器阅读理解评测已经成功举办两届,从第 1 届以填空型阅读理解问题为主,到第 2 届关注基于篇章片段抽取的阅读理解,评测会议发布了人工标注的中文填空型和篇章片段抽取型阅读理解的数据集[144],很多的相关研究也在这些数据集上有所展开.但本质上,这些工作只是把篇章看作一个词符号序列,缺乏真正意义上的篇章理解.当然,从另一层面而言,这些研究也大大推动了人们对篇章理解的关注和重视.例如,NSFC 最近几年就批准了多个汉语篇章理解方向的重点项目和人工智能应急重点项目,包括哈尔滨工业大学刘挺主持的篇章级中文语义分析理论与方法,中国科学院自动化研究所宗成庆主持的汉语多层次语篇分析理论方法研究与应用,苏州大学张民主持的面向多层次篇章语义的机器翻译理论、方法与实现,北京理工大学黄河燕主持的中文语义深度计算与阅读理解,以及苏州大学周国栋主持的话题驱动的汉语篇章机器阅读理解等.

3 总 结

综上所述,在自然语言处理领域,与词法分析、句法分析等研究相比,篇章结构分析研究相对滞后.特别是适用于汉语篇章阅读理解的篇章结构分析研究还处于起步阶段,尚未形成一套有效的理论体系,相应语料库资源建设薄弱,关键技术研究严重滞后.相应地,机器阅读理解的相关研究也刚刚起步,目前主要是基于检索技术的相关片段抽取,缺乏真正意义上的篇章理解.众所周知,与英语等西方语言相比,汉语无论是篇章结构和信息意图表达方式,还是事件描述方式和话题表述方式等方面都有较大的差异.这就迫切需要进一步完善适用于汉语篇章阅读理解的篇章结构理论体系,建立一定规模的适用于汉语篇章阅读理解的汉语篇章结构资源库,并在此基础上建立汉语篇章结构分析的计算模型,实现高性能的汉语篇章结构分析和篇章深度理解平台,为自然语言理解和篇章级应用提供基础支撑.

References:

[1] De Robert-Alain B, Wolfgang DU. Introduction to Text Linguistics. London, New York: Longman Paperback, 1981.

[2] Schank RC. Conceptual dependency: A theory of natural language understanding. Cognitive Psychology, 1972,3(4):552–631.

[3] Halliday MAK, Ruqaiya H. Cohesion in English. Harlow: Longman PubGroup, 1976.

[4] Hobbs JR. Coherence and coreference. Cognitive Science, 1979,3(1):67–90.

[5] Hobbs JR. Information, intention and structure in discourse. In: Proc. of the NATO Workshop on Burning Issues in Discourse. 1993.

[6] Mann WC, Thompson SA. Rhetorical structure theory: Toward a functional theory of text organization. Text-interdisciplinary Journal for the Study of Discourse, 1988,8(3):243–281.

[7] Mann WC, Thompson SA. Relational propositions in discourse. Discourse Processing, 1986,9(1):57–90.

[8] Grosz BJ, Sidner CL. Attention, intentions, and the structure of discourse. Computational Linguistics, 1986,12(3): 175–204.

[9] Grosz BJ, Weinstein S, Joshi AK. Centering: A framework for modeling the local coherence of discourse. Computational Linguistics, 1995,21(2):203–225.

[10] Halliday MAK. Linguistic function and literary style: An inquiry into the language of William Golding's the inheritors. In: Chatman S, eds. Proc. of the Symp. on Literary Style. New York: Oxford University Press, 1971. 330.

[11] Marcu D. The rhetorical parsing, summarization, and generation of natural language texts [Ph.D. Thesis]. Toronto: Department of Computer Science, University of Toronto, 1997.

[12] Marcu D. The Theory and Practice of Discourse Parsing and Summarization. Cambridge: The MIT Press, 2000.

[13] Halliday MAK. An Introduction to Functional Grammar. 3rd ed., London: Arnold, 2004.

[14] Daneš F. Functional sentence perspective and the organization of text. In: Functional Sentence Perspective. Prague: Academica, 1974. 106–128.

[15] Fries PH. On the status of theme in English: Arguments from discourse. Forum Linguistica, 1981,6(1):1–38.

[16] Wu WZ. Chinese Sentence Group. Beijing: The Commercial Press, 2000 (in Chinese).

[17] Xing FY. Research on Chinese Complex Sentences. Beijing: The Commercial Press, 2001 (in Chinese).

[18] Yao SY. A research on the collocation of the relation markers of Chinese compound sentences and some relevant explanation [Ph.D. Thesis]. Wuhan: Central China Normal University, 2006 (in Chinese with English abstract).

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

[20] Li YC, Feng WH, Sun J, Kong F, Zhou GD. Building Chinese discourse corpus with connective-driven dependency tree structure. In: Proc. of the EMNLP. 2014. 2105–2114.

[21] Jiang YR, Song R. Topic clause identification based on generalized topic theory. Journal of Chinese Information Processing, 2012, 26(5):114–120 (in Chinese with English abstract).

[22] Song R. The flowing water model of Chinese discourse generalized topic structure. Studies of the Chinese Language, 2013,6: 483–494 (in Chinese with English abstract).

[23] Shang Y, Song R, Lu DW. Study on the sentence-formability of topic sufficient sentence under the perspective of generalized topic structure theory. Journal of Chinese Information Processing, 2014,28(6):107–113 (in Chinese with English abstract).

[24] Ma JZ. Ma's Grammar. Beijing: The Commercial Press, 1929 (in Chinese).

[25] Li JX, Liu SR. The Chinese Grammar Textbook. Tianjing: Tianjin Dazhong Press, 1952 (in Chinese).

[26] Chu XM, Zhu QM, Zhou GD. Discourse primary-secondary relationships in natural language processing. Chinese Journal of Computers, 2017,40(4):842−859 (in Chinese with English abstract).

[27] Cao FF. Clause and Sentence Structure in Chinese: A Functional Perspective. Student Book Co., 1990.

[28] Qu CX. Chinese Discourse Grammar. Beijing: Beijing Language and Culture University Press, 1998 (in Chinese).

[29] Liu LJ. A Contrastive Study of Discourse Structures in English and Chinese. Guangzhou: Sun Yat-sen University Press, 1998. 166−178 (in Chinese).

[30] Sun K. A contrastive study of discourse constructional patterns between Chinese and English: spective of topic chains. Journal of PLA University of Foreign Languages, 2013,36(3):12−18 (in Chinese with English abstract).

[31] Wang JJ. On the Continuation of the Topic: Research on Topic-chain-based Chinese-English Discourse. Shanghai: Shanghai Jiaotong University Press, 2013,2013 (in Chinese).

[32] Zhao Q, Zhou XC. Topic-chain-based coherence annotation scheme for Chinese text. Journal of Chinese Information Processing, 2014,28(5):102−111 (in Chinese with English abstract).

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

[34] Xi XF, Sun QY, Zhou GD. Research and prospect of discourse topic structure analysis for discourse intentionality. Chinese Journal of Computers, 2017 (in Chinese with English abstract). http://kns.cnki.net/KCMS/detail/11.1826.TP.20170601.1917.016.html

[35] Carlson L, Marcu D, Okurowski ME. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: van Kuppevelt J, Smith R, eds. Proc. of the Current Directions in Discourse. New York: Kluwer, 2003. 85−112.

[36] Prasad R, Dinesh N, Lee A, Miltsakaki E, Robaldo L, Joshi A, Webber B. The Penn Discourse Treebank 2.0. In: Proc. of the Int'l Conf. on Language Resources and Evaluation. 2008.

[37] Doddington GR, Mitchell A, Przybocki MA, *et al*. The automatic content extraction (ACE) program-tasks, data, and evaluation. In: Proc. of the LREC. 2004,2:1.

[38] Poesio M, Artstein R. Anaphoric annotation in the ARRAU corpus. In: Proc. of the Int'l Conf. on Language Resources and Evaluation. 2008.

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

[40] Wolf F, Gibson E. Representing discourse coherence: A corpus-based study. Computational Linguistics, 2005,32(2):249−287.

[41] Yue M. Rhetorical structure annotation of Chinese news commentaries. Journal of Chinese Information Processing, 2008,22(4): 19−23 (in Chinese with English absrtact).

[42] Chen LP. The theoretical framework grounding the annotation of Chinese text structures. Journal of Nanjing University of Aeronautics & Astronautics (Social Sciences), 2008,10(3):68−71 (in Chinese with English abstract).

[43] Zhou YP, Xue NW. PDTB-style discourse annotation of Chinese text. In: Proc. of the ACL 2012. 2012. 69−77.

[44] Zhou YP, Xue NW. The Chinese discourse TreeBank: A Chinese corpus annotated with discourse relations. Language Resources and Evaluation, 2015,49(2):397−431.

[45] Zhang MY, Qin B, Liu T. Chinese discourse relation semantic taxonomy and annotation. Journal of Chinese Information Processing, 2014,28(2):28−36 (in Chinese with English abstract).

[46] Shu JB. Research of auto-identifying the relation markers of compound sentence for Chinese information processing [Ph.D. Thesis]. Wuhan: Central China Normal University, 2011 (in Chinese with English abstract).

[47] Zhao Q. Annotation scheme for Chinese Treebank. Journal of Chinese Information Processing, 2004,18(4):2−9 (in Chinese with English abstract).

[48] Xi XF, Chu XM, Sun QY, Zhou GD. 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).

[49] Hannah R, *et al*. Discourse coherence: Concurrent explicit and implicit relations. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1. 2018.

[50] Soricut R, Marcu D. Sentence level discourse parsing using syntactic and lexical information. In: Proc. of the NAACL-HLT 2003. 2003. 149−156.

[51] Hernault H, Bollegala D, Ishizuka M. A sequential model for discourse segmentation. In: Proc. of the CICLing 2010. 2010. 315−326.

[52] LeThanh H, Abeysinghe G, Huyck C. Generating discourse structures for written texts. In: Proc. of the COLING 2004. 2004. 329.

[53] Hernault H, Prendinger H, duVerle DA, *et al*. Hilda: A discourse parser using support vector machine classification. Dialogue and Discourse, 2010,1(3):1–33.

[54] Wei FW, Graeme H. Text-level discourse parsing with rich linguistic features. In: Proc. of the ACL 2012. 2012. 60–68.

[55] Joty S, Carenini G, Ng RT. A novel discriminative framework for sentence-level discourse analysis. In: Proc. of the EMNLP-CoNLL 2012. 2012. 904–915.

[56] Surdeanu M, Hicks T, Valenzuela-Escarcega MA. Two practical rhetorical structure theory parsers. In: Proc. of the 2015 Conf. of the North American Chapter of the Association for Computational Linguistics: Demonstrations. 2015. 1–5.

[57] Braud C, Plank B, Søgaard A. Multi-view and multi-task training of rst discourse parsers. In: Proc. of the 26th Int'l Conf. on Computational Linguistics: Technical Papers. 2016. 1903–1913.

[58] Li Q, Li TS, Chang BB. Discourse parsing with attention-based hierarchical neural networks. In: Proc. of the EMNLP. 2016. 362–371.

[59] Braud C, Coavoux M, Søgaard A. Cross-lingual RST discourse parsing. In: Proc. of the 15th Conf. of the European Chapter of the Association for Computational Linguistics, Vol. 1. 2017. 292–304.

[60] Ji YF, Eisenstein J. Representation learning for text-level discourse parsing. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1. 2014. 13–24.

[61] Dines N, Lee A, Miltsakaki E, *et al*. Attribution and the (non-) alignment of syntactic and discourse arguments of connectives. In: Proc. of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky. 2005. 29–36.

[62] Lin Z, Ng HT, Kan MY. A PDTB-styled end-to-end discourse parser. Natural Language Engineering, 2014,20(2):151–184.

[63] Wellner B, Pustejovsky J. Automatically identifying the arguments of discourse connectives. In: Proc. of the EMNLP-CoNLL 2007. 2007. 92–101.

[64] Ghosh S, Johansson R, Tonelli S. Shallow discourse parsing with conditional random fields. In: Proc. of the IJCNLP. 2011. 1071–1079.

[65] Ghosh S. End-to-end discourse parsing using cascaded structured prediction [Ph.D. Thesis]. University of Trento, 2012.

[66] Kong F, Ng HT, Zhou G. A constituent-based approach to argument labeling with joint inference in discourse parsing. In: Proc. of the EMNLP 2014. 2014. 68–77.

[67] Pitler E, Louis A, Nenkova A. Automatic sense prediction for implicit discourse relations in text. In: Proc. of the ACL-AFNLP 2009. 2009. 683–691.

[68] Marcu D, Echihabi A. An unsupervised approach to recognizing discourse relations. In: Proc. of the ACL 2002. 2002. 368–375.

[69] Saito M, Yamamoto K, Sekine S. Using phrasal patterns to identify discourse relations. In: Proc. of the NAACL 2006. 2006. 133–136.

[70] Lin Z, Kan MY, Ng HT. Recognizing implicit discourse relations in the Penn discourse Treebank. In: Proc. of the EMNLP 2009. 2009. 343–351.

[71] Park J, Cardie C. Improving implicit discourse relation recognition through feature set optimization. In: Proc. of the SIGDIAL 2012. 2012. 108–112.

[72] Zhou ZM, Xu Y, Niu ZY, *et al*. Predicting discourse connectives for implicit discourse relation recognition. In: Proc. of the COLING 2010. 2010. 1507–1514.

[73] Biran O, McKeown K. Aggregated word pair features for implicit discourse relation disambiguation. In: Proc. of the ACL 2013. 2013. 69–73.

[74] Lan M, Xu Y, Niu Z. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In: Proc. of the ACL 2013. 2013. 476–485.

[75] Zhou XP, Hong Y, Che TT, *et al*. An unsupervised approach to inferring implicit discourse relation. Journal of Chinese Information Processing, 2013,27(2):17–26 (in Chinese with English abstract).

[76] Lan M, Wang JX, Wu YB, Niu ZY, Wang HF. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. 2017. 1299–1308.

[77] Lin Z, Ng HT, Kan MY. A PDTB-styled end-to-end discourse parser. Natural Language Engineering, 2014,20(2):151–184.

[78] Zhang MY, Song Y, Qin B, *et al*. Chinese discourse relation recognition. Journal of Chinese Information Processing, 2013,27(6): 51–58 (in Chinese with English abstract).

[79] Tu M, Zhou Y, Zong CQ. Automatically parsing Chinese discourse based on maximum entropy. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014,50(1):125–132 (in Chinese with English abstract).

[80] Kong F, Zhou G. A CDT-styled end-to-end Chinese discourse parser. ACM Trans. on Asian and Low-resource Language Information Processing (TALLIP), 2017,16(4):26.

[81] Jia Y, Feng Y, Ye Y, *et al*. Improved discourse parsing with two-step neural transition-based model. ACM Trans. on Asian and Low-resource Language Information Processing (TALLIP), 2018,17(2):11.

[82] Sun C, Kong F. A transition-based framework for Chinese discourse structure parsing. Journal of Chinese Information Processing, 2018,32(12):26–34 (in Chinese with Englllish abstract).

[83] Vilain M, Burger J, Aberdeen J, *et al*. A model-theoretic coreference scoring scheme. In: Proc. of the 6th Conf. on Message Understanding. 1995. 45–52.

[84] Hirschman L, Robinson P, Burger J, *et al*. Automating coreference: The role of annotated training data. In: Proc. of the AAAI Spring Symp. on Applying Machine Learning to Discourse Processing. 1997. 118–121.

[85] 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. 2011. 1–27.

[86] Pradhan S, Moschitti A, Xue N, *et al*. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In: Proc. of Joint Conf. on EMNLP and CoNLL-Shared Task. 2012. 1–40.

[87] Raghunathan K, Lee H, Rangarajan S, *et al*. A multi-pass sieve for coreference resolution. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing, EMNLP 2010. 2010. 492–501.

[88] Lee H, Peirsman Y, Chang A, *et al*. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In: Proc. of the 15th Conf. on Computational Natural Language Learning: Shared Task. 2011. 28–34.

[89] Cardie C, Wagstaft K. Noun phrase coreference as clustering. In: Proc. of the Joint Conf. on Empirical Methods in NLP and Very Large Corpora. 1999. 277–308.

[90] Mccarthy JF, Lehnert WG. Using decision trees for coreference resolution. In: Proc. of the 14th Int'l Joint Conf. on Artificial Intelligence. 1999. 1050–1055.

[91] Soon WM, Ng HT, Lim DCY. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 2001,27(4):521–544.

[92] Kong F, Zhou GD. Pronoun resolution in English and Chinese languages based on tree kernel. Ruan Jian Xue Bao/Journal of Software, 2012,34(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]

[93] Xu R, Xu J, Liu J, *et al*. Incorporating rule-based and statistic-based techniques for coreference resolution. In: Proc. of the Joint Conf. on EMNLP and CoNLL-Shared Task. 2012. 107–112.

[94] Belder JD, Moens MF. Coreference clustering using column generation. In: Proc. of the COLING 2012: Posters. 2012. 245–254.

[95] Wiseman S, Rush AM, Shieber SM. Learning global features for coreference resolution. In: Proc. of the NAACL-HLT. 2016. 994–1004.

[96] Clark K, Manning CD. Deep reinforcement learning for mention-ranking coreference models. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. 2016. 2256–2262.

[97] Lee K, He L, Lewis M, *et al*. End-to-end neural coreference resolution. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. 2017. 188–197.

[98] Wang HF. Survey: Computational models and technologies in anaphora resolution. Journal of Chinese Information Processing, 2002,16(6):10–18 (in Chinese with English abstract).

[99] Wang HF, He TT. Research on Chinese pronominal anaphora resolution. Chinese Journal of Computers, 2001,24(2):136–143 (in Chinese with English abstract).

[100] Wang HF, Mei Z. Robust pronominal resolution within Chinese text. Ruan Jian Xue Bao/Journal of Software, 2005,16(5): 700–707 (in Chinese with English abstract). http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=20050508&journal_id=jos [doi: 10.1360/jos160700]

[101] Li GC, Luo YF. Chinese pronominal anaphora resolution via a preference selection approach. Journal of Chinese Information Processing, 2005,19(4):24–30 (in Chinese with English abstract).

[102] Zhou JS, Huang SJ, Chen JJ, *et al*. A new graph clustering algorithm for Chinese noun phrase coreference resolution. Journal of Chinese Information Processing, 2007,21(2):77–82 (in Chinese with English abstract).

[103] Yang Y, Li YC, Zhou GD, *et al*. Research on distance information for anaphora resolution. Journal of Chinese Information Processing, 2008,22(5):39–44 (in Chinese with English abstract).

[104] Wang HD, Hu NQ, Kong F, *et al*. Research on semantic role information in anaphora resolution. Journal of Chinese Information Processing, 2009,23(1):23 (in Chinese with English abstract).

[105] Li YQ, Gan RS, Yng YH, *et al*. Chinese coreference resolution method based on feature respective selection strategy. Computer Engineering, 2011,37(18):180–182 (in Chinese with English abstract).

[106] Zhang MY, Li YB, Qin B, *et al*. Coreference resolution based on head match. Journal of Chinese Information Processing, 2011, 25(3):3–9 (in Chinese with English abstract).

[107] Song Y, Wang HF. Chinese zero anaphora resolution with Markov logic. Journal of Computer Research and Development, 2015, 52(9):2114–2122 (in Chinese with English abstract).

[108] Zhao S, Ng HT. Identification and resolution of Chinese zero pronouns: A machine learning approach. In: Proc. of the EMNLP-CoNLL 2007. 2007. 541–550.

[109] Kong F, Zhou G. A tree kernel-based unified framework for Chinese zero anaphora resolution. In: Proc. of the EMNLP 2010. 2010. 882–891.

[110] Chen C, Ng V. Chinese zero pronoun resolution: some recent advances. In: Proc. of the EMNLP 2013. 2013. 1360–1365.

[111] Chen C, Ng V. Chinese overt pronoun resolution: A bilingual approach. In: Proc. of the AAAI 2014. 2014. 1615–1621.

[112] 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. 2015. 320–326.

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

[114] Sheng C, Kong F, Zhou GD. Toward better Chinese zero pronoun resolution from discourse perspective. In: Proc. of the NLPCC-ICCPOL 2017, Lecture Notes in Computer Science, 2017.

[115] Kong F, Zhou GD. Chinese zero pronoun resolution: A chain to chain approach. In: Proc. of the NLPCC-ICCPOL 2017. Lecture Notes in Computer Science, Springer-Verlag, 2017.

[116] Yin QY, Zhang Y, Zhang WN, Liu T. Chinese zero pronoun resolution with deep memory network. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. 2017. 1309–1318.

[117] Zhang Y, Liu T, Yin QY, Zhang WN. A deep neural network for Chinese zero pronoun resolution. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. 2017. 3322–3328.

[118] Liu T, Cui YM, Yin QY, Zhang WN, Wang SJ, Hu GP. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1. 2017. 102–111.

[119] Yin QY, Zhang Y, Zhang WN, Liu T, Yang WW. Deep reinforcement learning for Chinese zero pronoun resolution. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1. 2018. 569–578.

[120] Ahn D. The stages of event extraction. In: Proc. of the Workshop on Annotating & Reasoning about Time & Events. 2006. 1–8.

[121] Chen Z, Ji H, Haralick R. A pairwise event coreference model, feature impact and evaluation for event coreference resolution. In: Proc. of the Workshop on Events in Emerging Text Types. 2009. 17–22.

[122] Bejan CA, Harabagiu S. Unsupervised event coreference resolution with rich linguistic features. In: Proc. of the Meeting of the Association for Computational Linguistics, Association for Computational Linguistics. 2010. 1412–1422.

[123] Araki J, Mitamura T. Joint event trigger identification and event coreference resolution with structured perceptron. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. 2015. 2074–2080.

[124] Lu J, Ng V. Joint learning for event coreference resolution. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1. 2017. 90–101.

[125] Nguyen TH, Meyers A, Grishman R. New York University 2016 system for KBP event Nugget: A deep learning approach. In: Proc. of the 9th Text Analysis Conf. 2016.

[126] Joe E, Jeremy G, Dana F, *et al*. Overview of linguistic resources for the TAC KBP. 2015. https://tac.nist.gov//2015/KBP/

[127] Krause S, Xu F, Uszkoreit H, *et al*. Event linking with sentential features from convolutional neural networks. In: Proc. of the Signll Conf. on Computational Natural Language Learning. 2016. 239–249.

[128] Teng JY, Li PF, Zhu QM. Global inference for co-reference resolution between Chinese events. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016,52(1):97–103 (in Chinese with Engllish abstract).

[129] Teng JY. Research on Chinese event coreference resolution [MS. Thesis]. Suzhou: Soochow University. 2016 (in Chinese with English abstract).

[130] Wellner B, Pustejovsky J, Havasi C, Rumshisky A, Sauri R. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In: Proc. of the 7th SIGDIAL Workshop on Discourse and Dialogue. 2006. 117–125.

[131] Hirschman L, Light M, Breck E, *et al*. Deep read: A reading comprehension system. In: Proc. of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, 1999. 325–332.

[132] Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ questions for machine comprehension of text. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. 2016. 2383–2392.

[133] Bajaj P, Campos D, Craswell N, *et al*. MS MARCO: A human generated machine reading comprehension dataset. arXiv Preprint arXiv: 1611.09268, 2018.

[134] Ng HT, Teo LH, Kwan JLP. A machine learning approach to answering questions for reading comprehension tests. In: Proc. of the 2000 Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, Vol. 13. Association for Computational Linguistics, 2000. 124–132.

[135] Chen D, Bolton J, Manning CD. A thorough examination of the CNN/daily mail reading comprehension task. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics, Vol. 1. 2016. 2358–2367.

[136] Chu Z, Wang H, Gimpel K, *et al*. Broad context language modeling as reading comprehension. In: Proc. of the 15th Conf. of the European Chapter of the Association for Computational Linguistics, Vol. 2. 2017. 52–57.

[137] Cui Y, Chen Z, Wei S, *et al*. Attention-over-attention neural networks for reading comprehension. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1. 2017. 593–602.

[138] Hewlett D, Jones L, Lacoste A. Accurate supervised and semi-supervised machine reading for long documents. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. 2017. 2011–2020.

[139] Long T, Bengio E, Lowe R, *et al*. World knowledge for reading comprehension: Rare entity prediction with hierarchical LSTMs using external descriptions. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. 2017. 825–834.

[140] Shen YL, Liu XD, Duh K, *et al*. An empirical analysis of multiple-turn reasoning strategies in reading comprehension tasks. In: Proc. of the 8th Int'l Joint Conf. on Natural Language Processing, Vol. 1. 2017. 957–966.

[141] Wang W, Yang N, Wei F, *et al*. Gated self-matching networks for reading comprehension and question answering. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017,1:189–198.

[142] Xie P, Xing E. A constituent-centric neural architecture for reading comprehension. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017,1:1405–1414.

[143] Cui Y, Liu T, Chen Z, *et al*. Consensus attention-based neural networks for Chinese reading comprehension. In: Proc. of the COLING 2016, the 26th Int'l Conf. on Computational Linguistics: Technical Papers. 2016. 1777–1786.

[144] Cui YM, Liu T, Xiao L, Chen ZP, Ma WT, Che WX, Wang SJ, Hu GP. A Span-extraction dataset for Chinese machine reading comprehension. arXiv Preprint arXiv: 1810.07366, 2018.

附中文参考文献:

[16] 吴为章.汉语句群.北京:商务印书馆,2000.

[17] 邢福义.汉语复句研究.北京:商务印书馆,2001.

[18] 姚双云.复句关系标记的搭配研究与相关解释[博士学位论文].武汉:华中师范大学,2006.

[19] 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.

[21] 蒋玉茹,宋柔.基于广义话题理论的话题句识别.中文信息学报,2012,26(5):114–120.

[22] 宋柔.汉语篇章广义话题结构的流水模型.中国语文,2013,6:483–494.

[23] 尚英,宋柔,卢达威.广义话题结构理论视角下话题自足句成句性研究.中文信息学报,2014,28(6):107–113.

[24] 马建忠.马氏文通.北京:商务印书馆,1929.

[25] 黎锦熙,刘世儒.中国语法教程.天津:天津大众出版社,1952.

[26] 褚晓敏,朱巧明,周国栋.自然语言处理中的篇章主次关系研究.计算机学报,2017,40(4):842–859.

[28] 曲承熹.汉语篇章语法.北京:北京语言大学出版社,1998.

[29] 刘礼进.英汉篇章结构模式对比研究.广州:中山大学出版社,2011.166–178.
[30] 孙坤.话题链视角下的汉英篇章组织模式对比研究.解放军外国语学院学报(社会科学版),2013,36(3):12–18.
[31] 王建国.论话题的延续:基于话题链的汉英篇章研究.上海:上海交通大学出版社,2013.
[32] 周强,周骁聪.基于话题链的汉语语篇连贯性描述体系.中文信息学报,2014,28(5):102–111.
[33] 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.
[34] 奚雪峰,孙庆英,周国栋.面向意图性的篇章话题结构分析研究与展望.计算机学报(优先出版),2017. http://kns.cnki.net/KCMS/detail/11.1826.TP.20170601.1917.016.html
[41] 乐明.汉语篇章修辞结构的标注研究.中文信息学报,2008,22(4):19–23.
[42] 陈莉萍.汉语篇章结构标注的理论支撑.南京航空航天大学学报(社科版),2008,10(3):68–71.
[45] 张牧宇,秦兵,刘挺.中文篇章级句间语义关系体系及标注.中文信息学报,2014,28(2):28–36.
[46] 舒江波.面向中文信息处理的复句关系词自动标识研究[博士学位论文].武汉:华中师范大学,2011.
[47] 周强.汉语句法树库标注体系.中文信息学报,2004,18(4):2–9.
[48] 奚雪峰,褚晓敏,孙庆英,周国栋.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.
[75] 周小佩,洪宇,车婷婷,等.一种无指导的隐式篇章关系推理方法研究.中文信息学报,2013,27(2):17–26.
[78] 张牧宇,宋原,秦兵,等.中文篇章级句间语义关系识别.中文信息学报,2013,27(6):51–58.
[79] 涂眉,周玉,宗成庆.基于最大熵的汉语篇章结构自动分析方法.北京大学学报(自然科学版),2014,50(1):125–132.
[82] 孙成,孔芳.基于转移的中文篇章结构解析研究.中文信息学报,2018,32(12):26–34.
[92] 孔芳,周国栋.基于树核函数的中英文代词消解.软件学报,2012,34(5):1085–1099. http://www.jos.org.cn/1000-9825/4044.htm [doi: 10.3724/SP. J.1001.2012.04044]
[98] 王厚峰.指代消解的基本方法和实现技术.中文信息学报,2002,16(6):10–18.
[99] 王厚峰,何婷婷.汉语中人称代词的消解研究.计算机学报,2001,24(2):136–143.
[100] 王厚峰,梅铮.鲁棒性的汉语人称代词消解.软件学报,2005,16(5):700–707. http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=20050508&journal_id=jos [doi: 10.1360/jos160700]
[101] 李国臣,罗云飞.采用优先选择策略的中文人称代词的指代消解.中文信息学报,2005,19(4):24–30.
[102] 周俊生,黄书剑,陈家骏,等.一种基于图划分的无监督汉语指代消解算法.中文信息学报,2007,21(2):77–82.
[103] 杨勇,李艳翠,周国栋,等.指代消解中距离特征的研究.中文信息学报,2008,22(5):39–44.
[104] 王海东,胡乃全,孔芳,等.指代消解中语义角色特征的研究.中文信息学报,2009,23(1):23.
[105] 李渝勤,甘润生,杨永红,等.基于特征分选策略的中文共指消解方法.计算机工程,2011,37(18):180–182.
[106] 张牧宇,黎耀炳,秦兵,等.基于中心语匹配的共指消解.中文信息学报,2011,25(3):3–9.
[107] 宋洋,王厚峰.基于马尔可夫逻辑的中文零指代消解.计算机研究与发展,2015,52(9):2114–2122.
[128] 滕佳月,李培峰,朱巧明.基于全局优化的中文事件同指消解方法.北京大学学报(自然科学版),2016,52(1):97–103.
[129] 滕佳月.中文事件同指消解方法研究[硕士学位论文].苏州:苏州大学,2016.

孔芳(1977－),女,江苏扬州人,博士,教授,博士生导师,CCF 专业会员,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.

王红玲(1975－),女,博士,副教授,CCF 专业会员,主要研究领域为自然语言理解,信息抽取.