文章编号：1003-0077(2016)04-0001-11

篇章关系分析研究综述

严为绒，徐 扬，朱珊珊，洪 宇，姚建民，朱巧明

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：篇章关系研究，旨在推断同一篇章内相邻或跨度在一定范围内的文本片段之间的语义连接关系。语义连接关系对篇章内容理解和结构分析都具有重要作用，成为目前篇章分析领域的重点研究内容。该文针对三个中英文篇章关系研究领域的语料库：基于修辞结构理论的篇章树库（Rhetorical Structure Theory Discourse Treebank，RSTDT）、宾州篇章树库（Penn Discourse Treebank，PDTB）和哈尔滨工业大学中文篇章关系语料库（HIT Chinese Discourse Treebank，HIT-CDTB），主要介绍篇章关系分析理论的语料资源与研究背景、标注与评测体系以及国内外研究现状。此外，总结相关工作，指出目前篇章关系，尤其是隐式篇章关系研究的主要难题。

关键词：篇章关系；篇章修辞结构；RSTDT；PDTB；CDTB

中图分类号：TP391 **文献标识码**：A

A Survey to Discourse Relation Analyzing

YAN Weirong, XU Yang, ZHU Shanshan, HONG Yu, YAO Jianmin, ZHU Qiaoming

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: The research on discourse relation is aimed at inferring the inter-sentential semantic relationship which occurs in the same discourse. This relation plays an important role in discourse content understanding and structure analyzing, becoming research focus in the field of discourse analysis. In this paper, we introduce the corpus and background, annotation and evaluation system as well as in this field based three corpora: Rhetorical Structure Theory Discourse Treebank (RSTDT), Penn Discourse Treebank (PDTB) and HIT Chinese Discourse Treebank (HIT-CDTB). Finally, through analyzing current work, we summarize the main difficulty and challenge in recognizing discourse relation especially implicit discourse relation.

Key words: discourse relation; discourse rhetoric structure; RSTDT; PDTB; CDTB

1 引言

自然语言处理（Natural Language Processing，NLP）的研究从表层词汇理解延伸到更深层次的句法语义，研究粒度从单个词的语义发展到短语、句子，直至篇章。其中，篇章分析的研究主要基于篇章间潜在树形结构，重点研究树的内部节点及其对应的属性，后期倾向于更复杂的图结构。

目前，篇章关系分析的研究尚未成熟，主要包括：篇章语义关系识别（Discourse Relation Recognition，DRR）和基于修辞结构理论（Rhetorical Structure Theory，RST）的篇章结构及修辞关系分析等。本文针对篇章级语义分析，尤其是句间语义关系研究进行综述，对比分析研究现状与意义，并提出这一研究的关键难点。篇章关系的研究意义主要体现在以下两个方面。

（1）有利于篇章文本结构化

篇章不是由句子堆积成的简单序列，而是由一系列结构衔接[1]、语义连贯[2]的短语、子句、句子或段落构成的具有独立语义的自然语言文体。篇章内部的层次关系可实现篇章文本结构化。结构化关系

收稿日期：2014-09-25 定稿日期：2015-01-05

基金项目：国家自然科学基金（61373097，61272259，61272260，90920004）；教育部博士学科点专项基金（2009321110006，20103201110021）；江苏省自然科学基金（BK2011282）；江苏省高校自然科学基金（11KJA520003）；苏州市自然科学基金（SH201212）

树不仅有利于理解篇章的语义关系，而且可用于深层次的篇章分析，如计算篇章间语义相似度等。

（2）具有广泛的应用价值

篇章的因果关系用于自动问答和事件关系抽取[3-4]；对比关系用于研究情感分析[5]；扩展关系用于自动文摘和篇章关键词抽取[6]。另外，在机器翻译中也得到广泛应用[7]。

本文组织结构如下，第二节介绍与篇章分析研究相关的三种权威语言学资源RSTDT、PDTB和HIT-CDTB；第三节概括篇章关系分析任务及评测方法；第四节回顾国内外篇章关系的研究现状；第五节分析目前篇章关系研究所要解决的关键问题和研究难点；第六节总结。

2 语料资源

本节首先介绍与篇章分析尤其是篇章关系分析研究紧密相关的三种语言学资源RSTDT、PDTB和HIT-CDTB之间的差异，其次介绍三种语料的标注过程、基本组成及相应的实例分析。

2.1 三种语言学资源的区别

显而易见，三种语言资源间存在着下列不同之处。

（1）语言种类不同

RSTDT和PDTB是由美国南加州大学和美国宾夕法尼亚大学在语言数据联盟（Linguistic Data Consortium，LDC）①上分别于2002年和2008年发布的两种英文篇章关系分析语言学资源。HIT-CDTB②则是由哈尔滨工业大学于2013年发布的中文篇章关系分析语言学资源。

（2）语料来源和规模不同

RSTDT和PDTB两种语料均选自PTB③（Penn Treebank）[8]语料，根据各自定义的规则及目标进行标注。PTB语料内容来自美国《华尔街日报》（Wall Street Journal，WSJ）的新闻报道，包含多种不同的新闻主题，例如商业经济、文化报道及理财投资等。PDTB相较于RSTDT来说，语料规模更大。HIT-CDTB语料针对OntoNotes④语料的中文文档进行标注，语料内容来源于广播新闻、杂志和网络等。

（3）关系体系不同

RSTDT主要针对篇章中的修辞结构关系进行标注，共定义18种修辞结构关系，有的修辞关系在篇章中只能出现一次且横跨整个篇章，如“摘要关系”；有的修辞关系在篇章中可能出现多次且跨度在一定范围内，如“对比关系”。

PDTB对篇章内毗邻或跨度在一定范围内的各片段之间，以连接词为核心，构成整体篇章关系的层次结构。PDTB共定义五种关系体系：显式关系Explicit、隐式关系Implicit、可被推导而加入连接词则表达冗余的篇章关系AltLex、不可推导篇章关系且后一论元扩充前一论元实体信息EntRel、既不存在篇章关系且无论元间的实体一致性NoRel。

与PDTB类似，HIT-CDTB也凸显了连接词的重要性，但是HIT-CDTB中只定义显式和隐式两种关系体系，并根据不同的颗粒度，将篇章关系的语义结构类型分为六大类。

2.2 RST篇章树库（RSTDT）⑤概述

基于Mann和Thompson等[9]1988年提出的修辞结构理论（RST），标注篇章修辞结构关系的语料资源RSTDT[10]于2002年由LDC发布。该语料库基于RST框架，标注文本的修辞结构，用于表示文本一致性类别、连贯性及文本各片段的独立作用。其中，修辞结构是指篇章内各片段间，依靠语义修辞关系进行相互连接，构成整体篇章关系层次结构。RSTDT是标注多层语言学信息的大规模、高质量语料库，为研究子句间的结合形式及各自功能、隐式衔接间的篇章结构与修辞关系提供基础资源。

- 构建过程

RSTDT主要对WSJ中的385篇文章进行标注，其中53篇（13.8%）被重复标记，目的是检测不同标注者标记结果的一致性。语料库构建过程主要包括两个基本子任务。

1）对篇章文本进行切分，目的是形成若干句型独立且能表达一定语义的片段，称为基本篇章单元（Elementary Discourse Units，EDU）。

2）构建修辞结构树，确定同一篇章内相邻单元间修辞关系，层层叠加最终形成树形结构，树中叶节点是上一步切分的EDU，内部节点表示具有具体修辞关系的一段连续文本跨度。

① http://www.ldc.upenn.edu/

② http://ir.hit.edu.cn/hit-cdtb/index.html

③ 宾州树库（PTB）是对WSJ语料进行句法结构标注的公认语料资源 http://www.cis.upenn.edu/~treebank/

④ https://catalog.ldc.upenn.edu/LDC2011T03

⑤ http://www.isi.edu/~marcu/discourse

• 基本组成

RSTDT 将篇章中的修辞结构关系分为两种：单核(Mononuclear)和多核(Multinuclear)。

单核是指包含修辞关系的两个 EDU 间存在主次之分，体现出一种“中心-卫星”理论(Nucleus-Satellite Theory)。其中，“中心”指修辞中心，称为“核”；而“卫星”则是修饰衬“中心”，从属于“核”。每个“中心-卫星”结构包含两种状态 NS(左单元为 Nucleus，右单元为 Satellite)和 SN 则反之。

多核是指包含修辞关系的两个或多个 EDU 之间，彼此权重相等，例如，修辞关系“列表(List)”中，各 EDU 属于并列关系，无主次之分。该结构只有一种状态 NN(左右单元都为 Nucleus)。

RSTDT 语料共 78 种(53 种单核与 25 种多核)篇章修辞关系。根据修辞关系之间的相近程度，将其划分为 18 种类别，并附加核状态信息(NS、SN、NN)共得到 41 种不同的关系，如表 1 所示。

表 1　18 个类别及其核状态形成的 41 种关系

Relation	核状态	Relation	核状态
Cause Comparison Condition Contrast Evaluation Explanation Temporal Topic-Comment	NN NS SN	Attribution Background Elaborate Enablement Manner-Means Summary Topic-Change	NS SN
		Joint Same-Unit Textual-Organization	NN

• 实例分析

针对 RSTDT 中定义的修辞结构关系，列举实例如图 1 所示。根据上述标注方法，首先根据文本语义将原句切分为三个 EDU；然后识别相邻 EDU 间的修辞关系，并将原句表示成层次化的树形结构。如图 1 所示，EDU2 与 EDU3 存在“对比(*Contrast*)”关系，而 EDU2 和 EDU3 整体与 EDU1 存在“时序-之后(*Temporal-After*)”关系，形成修辞关系结构树。根据“卫星-中心”理论，箭头由表示修饰的辅助成分(Satellite)指向语义关系的中心(Nucleus)。

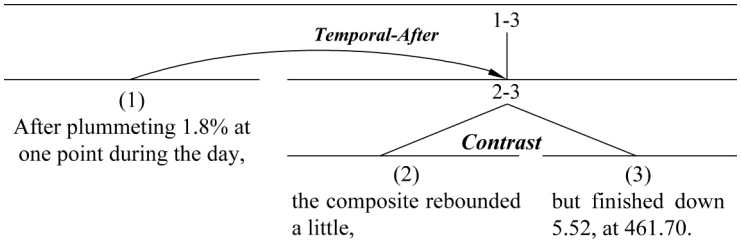

图 1　RSTDT 修辞结构关系实例

2.3　宾州篇章树库(PDTB)[①]概述

PDTB 2.0[11]是 LDC 于 2008 年发布的针对篇章关系标注的语料资源，该语料覆盖 WSJ 中近 2 500 篇文章，共标注 40 600 个篇章关系实例，是目前篇章分析领域规模最大的语言学资源。

• 构建过程

PDTB 主要参照命题库(PropBank)中的“谓词-论元(Predicate-Arguments)”结构，将篇章中带有篇章语义关系的文本片段标记为“连接词-论元(Connective-Arguments)”结构。其中，由连接词衔接的两个片段称为论元，由连接词引导的论元记为 Arg2，另一论元记为 Arg1，由 Arg1 和 Arg2 组成的整体称作“论元对”。

标注过程包括论元边界划分，对于显式实例(即“包含连接词”)，直接判别连接词的关系属性，指定论元篇章关系类型。然而，对于隐式实例(即“不包含连接词”)首先预估论元对的篇章关系，然后指定具有这一关系的连接词。

• 基本组成

显隐式篇章关系，是 PDTB 根据论元间是否包含连接词进行划分的。显式关系表示直接由显式连接词触发的篇章关系；而隐式关系是指相邻句子间不出现显式连接词，但根据上下文语义信息以及相关领域知识可自行推理的篇章关系。

PDTB 针对显式、隐式和 AltLex[12]篇章关系定义具体的语义关系类型体系。根据不同粒度，将篇章关系分为三层：第一层 4 类，第二层 16 类，第三层 23 类，共 43 类。第一层为四种主要的关系类型：Temporal、Comparison、Contingency 和 Expansion；

① http://www.seas.upenn.edu/~pdtb/

第二层和第三层分别在上一层基础上进一步细分，如表2所示。

表2 PDTB三层篇章语义关系体系

第一层	第二层	第三层	第一层	第二层	第三层
Comparison	Pragmatic Contrast	—	Temporal	Synchronous	—
	Pragmatic Concession				
	Contrast	Juxtaposition		Asynchronous	Precedence
		Opposition			
	Concession	Expectation			Succession
		Contra-Expectation			
Contingency	Pragmatic Cause	—	Expansion	Exception	—
	Pragmatic Condition	Relevance		List	
		Implicit Assertion		Conjunction	
	Cause	Reason		Instantiation	
		Result		Restatement	Specification
	Condition	Hypothetical			Equivalence
		General			Generalization
		Unreal Present		Alternative	Conjunctive
		Unreal Past			Disjunctive
		Factual Present			Chosen Alternative
		Factual Past			

• 实例分析

针对目前研究重点关注的显式与隐式关系类型，具体实例分析如下：PDTB中的显式关系，如例1所示，由连接词“*but*（但是）”引导的Arg1和Arg2间的篇章关系属于“对比（Comparison）”关系。相对地，隐式关系如例2所示，论元对之间无连接词，但能根据语义推断论元对之间属于“时序（Temporal）”关系。例2中用方括号注明的“*Implicit = at the time*”是人为添加的，表明论元间的隐式连接词为“*at the time*（当时）”。

例1 bridges need to be repaired or replaced

Arg2：***but*** *there's disagreement over how to do it*

Relation：Comparison. Contrast. Juxtaposition

例2 **Arg1**：By 1982, he was selling thousands of tires

Arg2：[***Implicit = at the time***] Newspapers published articles about him, and he was hailed as "the tire king"

Relation：Temporal. Synchrony

由以上分析可知，PDTB语料明确区分显式与隐式篇章关系，并对各种关系类型给出严格的层次定义，为篇章语义关系研究提供重要的基础资源。

2.4 中文篇章树库（HIT-CDTB）[①]概述

由于英文体系对中文语义覆盖不完整、英文体系对某些关系分类不清和英文时态关系平移困难等。哈尔滨工业大学对此作出分析，于2013年发布中文篇章关系语料资源HIT-CDTB，是目前国内首次公布的大规模篇章分析领域语言学资源。

• 构建过程

HIT-CDTB参照PDTB的标注准则，将篇章关系分为显式和隐式，并按照文本片段的粒度将篇章关系分为分句、复句和句群三种。分句篇章关系是指由篇章关系衔接的两个文本片段位于同一句子内；复句篇章关系是指由篇章关系衔接的两个文本片段是两个独立的句子；句群篇章关系是指由篇章关系衔接的两个文本片段都是句子集合。

① http://ir.hit.edu.cn/hit-cdtb/

与 PDTB 类似，HIT-CDTB 以关联词（连接词）为核心，对篇章语义关系进行标注。该语料将常见的显式关联词分为以下三个类别：可以单独使用来标识篇章关系的普通关联词，如“不论”；普通关联词与副词可搭配使用的带修饰的关联词，如“或许，因为”等；以及由两个或以上部分组成的平行关联词，如“虽然…但是…”等。其中，普通关联词和带修饰关联词共 870 种，平行关联词共 517 种。同时，显式和隐式关联词分别有 1 472 种和 533 种。

• 基本组成

HIT-CDTB 定义了 Explicit 和 Implicit 两种关系，并根据不同粒度，对篇章关系的语义结构进行多层分类。在 PDTB 的基础上，将第一层关系类型从四种调整为六种：时序、因果、条件、比较、扩展和并列，如表 3 所示。

表 3　HIT-CDTB 中文篇章语义关系体系

第一层	第二层	第三层	第四层	第一层	第二层	第三层	第四层
时序	同步	—	—	并列	平行	—	—
	异步	先序	—		选择	相容选择	—
		后序	—			互斥选择	—
因果	直接因果（说明因果）	原因	—	比较	直接对比（事实对比）	同向对比	—
		结果	—			反向对比	—
	间接因果（推论因果）	推论	—		间接对比（转折）	—	—
		证据	—				
	目的	目的在前	—		让步	让步在前	—
		目的在后	—			让步在后	—
条件	直接条件（事实条件）	必要条件	必要条件在前	扩展	细化	例外	例外在前
			必要条件在后				例外在后
		充分条件	充分条件在前			实例	实例在前
			充分条件在后				实例在后
		任意条件	任意条件在前			解释说明	—
			任意条件在后				
	形式条件（说明条件）	形式条件在前	—		泛化	—	
		形式条件在后	—		递进	—	—

• 实例分析

针对 HIT-CDTB 中的显式与隐式关系类型，具体实例（已分词）分析如下：显式关系实例，如例 3 所示，由平行关联词“不仅…同时…”引导的 Arg1 和 Arg2 间的篇章关系属于“扩展. 递进”关系。相对地，隐式关系如例 4 所示，其两个关系元素之间无关联词，但能根据语义推断 Arg1 和 Arg2 之间属于“时序. 异步. 后序”关系。

例 3　Arg1：他/不仅/是/一/名/小说家/和/剧作家

Arg2：同时/也/是/一/名/画家

显式篇章关系：扩展. 递进

例 4　Arg1：现年/60/岁/的/高行健

Arg2：[之前]在/1987 年/逃离/了/中国/，/流亡/到/法国

隐式篇章关系：时序. 异步. 后序

3　篇章关系分析任务及评测方法

根据三种篇章级语料库的侧重点，将篇章关系研究分为两个方面：基于 RSTDT 的篇章修辞结构关系和基于 PDTB 和 HIT-CDTB 的篇章语义关系。

3.1　修辞结构关系分析

目前篇章修辞结构关系分析，着重把握篇章的整体脉络，理解篇章层次结构。

• 任务定义

基于RSTDT的篇章结构关系分析过程与标注过程一致，主要包括文本结构划分和篇章结构生成两个部分。其中，由整体到局部是指将整个文本根据语义结构切分成若干基本篇章单元；而由局部到整体则是指借助篇章单元之间的修辞关系类型，实现由局部单元自底向上构建整体篇章的树形结构。修辞结构关系分析即是通过修辞结构树表示篇章文本间的语义结构信息，将篇章转换为基本篇章单元间的结构化组合。

• 评测方法

由于以上两个基本步骤需顺序执行，即后一步工作的输入依赖于前一步工作的输出。目前的篇章修辞结构关系分析任务，为独立评估每项工作的性能，对各步骤的输出结果分别进行评测，包括基本篇章单元的切分准确性以及各单元之间修辞关系类型判别的准确性。具体评测方法一般采用准确率P、召回率R、F值及精确率Accuracy四项常用指标[13]。

3.2 篇章语义关系分析

目前篇章语义分析，主要针对篇章片段中的语义连接关系进行识别并分类。

• 任务定义

PDTB和HIT-CDTB语料都是针对篇章语义分析研究展开标注的。其中，关于显隐式篇章关系的研究较多，下面以PDTB为例进行介绍。

例5 [No wonder he does well in his all subjects,]*s*1 [he is studying so hard.]*s*2 [He wants to get a scholarship,]*s*3 ***because*** [he had no money to pay his tuition fee.]*s*4

1）识别篇章中所有显式连接词*C*，如例5中的“*because*”，然后对显式连接词进行消歧[14]；

2）针对每个显式连接词*C*，定位其Arg1与Arg2的位置以及范围边界，如例5中文本片段*s*3和*s*4即是由连接词“*because*”所引导的论元组范围[15]；

3）判断每组显式实例中论元间的显式关系类型；

4）识别同一篇章任意相邻片段间是否包含隐式关系，如例3中文本片段*s*1和*s*2之间存在某种隐式关系，并将其分别标记为Arg1和Arg2；

5）检测每组论元对之间的隐式篇章关系类型。

• 评测方法

相关研究分别针对各个篇章语义关系的分类性能及篇章语义关系分类的整体性能进行评测。其中，通过构建多个分类器预测各个篇章语义关系的分类结果。如评估因果关系的分类性能，可将该类别的实例作为正例，其它关系类别的实例作为负例，由此构建二元分类器。通过准确率P、召回率R、F值等评测指标，分析该篇章语义关系分类器性能[16]。另外，在评估篇章语义关系整体分类性能时采用多元分类器，通过精确率衡量分类性能[17]。

4 研究现状

早期篇章关系研究，缺少权威语言学资源，使篇章关系类型定义以及关系类型判别方法的评测欠缺统一标准。自RSTDT语料发布后，篇章结构类型得到明确定义，判别方法统一。此后，PDTB语料的发布，更深层次的篇章关系类型得到明确定义，篇章分析的研究任务和评价策略也随之细化，从而推动篇章分析的进一步发展。HIT-CDTB语料的发布，使中文篇章分析研究迎来更大的突破和挑战。本节首先回顾基于修辞理论的篇章结构关系研究，然后重点分析篇章语义关系研究现状。

4.1 修辞结构关系研究

1988年Mann[18]提出的修辞结构理论(Rhetorical Structure Theory，RST)认为篇章中各句子并非孤立存在，而是通过相互间的修辞关系进行组合，构成篇章内容的连贯性。Marcu等[19]基于RST提出篇章修辞结构分析概念，并针对如何自动地将篇章文本映射到树形结构展开论述。

Soricut等[20]着重对句子内部的修辞结构进行识别和分类，将词汇与句法特征结合，对句子进行片段切分以及结构关系构建。LeThanh等[21]结合句法结构及线索短语，在句子结构划分以及句子内部修辞关系分类方面，均获得相对较优的性能。DuVerle等[13]和Hernault等[22]基于LeThanh的高质量片段切分方法，重点研究片段间修辞关系类型的判定。最终在EDU划分上获得F值为93.8%，在修辞结构和修辞关系分类上精确率达到85%和66.8%。

Feng等[23]在Hernault方法上，通过增加更多有效的语言学特征，如上下文的修辞关系、篇章产生式规则、片段之间语义相似度和线索短语特征，最终提高篇章分析的性能。Joty等[24]结合句内和句间

的修辞结构对文本层次的篇章进行分析，句内采用动态 CRF 模型对修辞关系和修辞结构进行联合训练，句间采用前向后向算法进行修辞结构构建，最终性能获得提升。由于 Joty 方法时间复杂度较高，Feng 等[25]将 Joty 的句内修辞关系和修辞结构联合模型拆分为两个线性链 CRF 模型，并对句内和句间模型进行编辑，最终篇章分析性能达到 58.2%。

此外，针对中文篇章修辞关系的研究，Zhang 等[26]基于启发式规则和向量空间模型提出一种混合型的汉语篇章结构自动分析方法，该方法利用连接成分作为求解篇章结构的形式特征，最终提高处理精度。Tu 等[27]采用序列标注的方法对汉语篇章单元进行切分，通过最大熵模型自动学习篇章结构并判定篇章修辞关系，最终篇章语义单元切分的 F 值达到 89.1%。

4.2　篇章语义关系研究

随着 PDTB 语料库的发布，篇章语义关系研究衍生出三个子任务：显式连接词消歧、论元边界检测以及篇章关系识别。

4.2.1　连接词分类与消歧

由于部分显式连接词存在一词多义现象。Miltsakaki 等[14]对显式连接词歧义进行局部研究，分析“*since*”、“*while*”和“*when*”三种连接词的歧义性，利用句中的助动词，情态动词和动词时态等特征，基于最大熵模型进行简单消歧。随后，Pitler 等[28]指出，篇章连接词的歧义性主要体现在以下两个方面：1）该词在篇章中是否起连接作用；2）该词在表示具体连接关系类型时是否存在歧义。通过提取有效的句法特征进行连接词消歧。Lin 等[15]在句法特征的基础上，增加显式连接词的上下文特征，包括单词以及词性序列特征。与 Pitler 的方法相比，最终性能提高近 2%，准确率达到 96.02%。

4.2.2　论元定位与范围检测

识别篇章关系的前提条件是已知连接词的论元范围。Prasad 等[12]在标记 PDTB 时发现，Arg1 与 Arg2 的位置及其各自跨越的文本范围具有很强的灵活性，主要体现在以下三个方面。

1）Arg1 与 Arg2 相对位置不固定，主要有三种情况：Arg1 与 Arg2 出现在同一句子中；Arg1 出现在 Arg2 前面句子中；Arg1 出现在 Arg2 后面句子中。

2）Arg1 与 Arg2 间距不固定：相邻、嵌套甚至间隔一定距离。

3）Arg1 与 Arg2 本身跨越范围不固定：可能是子句、句子甚至多个句子。

由于 Arg2 在句式结构上与连接词紧密相连，其位置和范围较容易判定，而 Arg1 的出现位置相对随机。论元范围检测的主要任务是自动定位 Arg1 并精确检测其边界范围。

Wellner 等[29]和 Elwell 等[30]等提出采用机器学习的方法识别 Arg1 和 Arg2 的中心词，尽管该方法能够准确地定位论元，尤其是 Arg1 的位置，却不能精确标识 Arg1 与 Arg2 的具体边界。其中，Wellner 等对论元中心词的识别精确率达到 69.8%。Prasad[12]单独针对 Arg1 与 Arg2 出现在不同句中的情况，通过识别包含 Arg1 的句子，间接检测其精确范围，但也仅定位到论元所在句子，最终 Arg1 的范围检测准确率为 86.3%。然而 Lin 等[15]不仅对 Arg1 进行定位和范围检测，而且提取识别的论元文本跨度。主要方法是增加显式连接词上下文的单词和词性特征，对 PDTB 中所有连接词的论元予以定位，并利用句法树中内部节点的最高概率特征确定论元范围，论元定位与范围检测结果见表 4。

表 4　Lin 等论元定位与范围检测结果

F 值	论元匹配方式	正确句法树	自动句法树
论元定位	—	97.95%	91.44%
论元范围检测	部分匹配	86.24%	80.96%
	精确匹配	53.85%	40.37%

4.2.3　篇章关系识别

PDTB 语料库的发布为篇章关系的研究提供了有利条件，下面分别介绍显式和隐式关系类型的具体判定方法。

- 英文显式关系类型判定

Pitler 等[31]通过研究证明，绝大多数连接词不存在歧义，可直接根据连接词推断其显式关系类型，在显式篇章关系的识别准确率达到 93.09%。然而，对应的隐式篇章关系由于缺少连接词特征等直接线索，只能从句法、语义、上下文中抽取相关特征进行分析判断。由于上下文信息的不确定性、句子结构的复杂性和语义关系的歧义性，从而影响隐式篇章关系的推理过程。

相对于显式篇章关系而言，目前隐式关系推理系统的性能仍然不高。因此，隐式篇章关系判别成为目前篇章关系分析领域的重点研究内容。

• 英文隐式关系类型判定

1）基于语言学特征的监督学习

现有的基于PDTB的隐式篇章关系类型判别研究，均是采用监督学习的方法。如Pitler等[32]首次单独针对PDTB中隐式篇章关系进行分类，提取句子中的情感词极性、动词短语、句子首尾单词等特征，最终的分类结果优于随机分类的性能。Lin等[33]细化上下文、句法树和依存树，对隐式关系的第二层关系进行分类，最终精确率达到40.2%。Wang等[17]基于树核函数的方法以扩充句法结构特征，尽管分类性能有所提升，但整体性能仍然偏低。

Park等[34]通过特征集优化算法进行特征选择，分类性能有所提高，但句法树和上下文特征在非新闻领域以及问答系统中很难被利用。Biran等[35]提出一种聚合单词对特征，并融入其它语言学特征，最终性能接近Park等的性能。Lan等[36]在交互结构优化多任务学习框架下，基于论元的动词、极性等基本语言学特征，分别使用现实语境的隐式论元对数据和人造伪隐式论元对，训练主分类器和辅分类器，提升隐式关系推理系统性能至42.30%。

2）基于统计学知识的概率推理

Marcu等[37]首次将概率统计方法应用于篇章关系分析，利用共现词特征与篇章关系的隐射概率，估计论元对间的显式连接词。Saito等[38]继承了Marcu的推理机制，验证短语的共现概率特征能提高日文隐式篇章关系推理的性能。Zhou等[39]将统计语言模型应用于隐式论元对间的连接词预测，使用三元语法模型预测共现概率最高的显式连接词，利用多种语言学特征并辅加预测的连接词特征推理隐式关系。与Saito等的方法相比，Zhou等最终分类精确率在Contingency和Temporal上有所提升，分别达到70.79%和70.51%。

• 中文篇章关系研究

目前，国内的篇章关系研究仍处于初级阶段。Xue等[40]提出构建中文篇章树库（Chinese Discourse Treebank，CDPB）的任务，并分析了中文连接词的分布特征以及存在的歧义性问题，指出论元范围鉴定是标注CDPB过程中最主要的困难。

Zhou等[41]结合PDTB语料的标注特征以及中文特点，提出具体的中文篇章关系标注准则，在跨语言篇章关系标注任务中发挥了良好的作用。

Huang等[42]根据中文文本特点，参照PDTB中定义的篇章关系类型，初步构建面向中文的篇章关系分析数据，并采用句子长度特征、标点符号特征、词语特征以及词性特征对中文的四种主要关系类型进行分类，最终判别精确率和F值分别达到88.28%和63.69%。

由于英文语料与中文语料存在显著的区别，比如英文语料中经常出现显式连接词，而中文语料反之；Li等[43]利用中英文语料之间的区别，通过英文关系辅助中文篇章关系推理。

• 其他小语种的篇章关系研究

随着篇章关系研究在英文中逐渐盛行，部分研究者尝试将篇章分析研究扩展到其它语言。如土耳其语[44]、北印度语[45]等。但由于其它语言缺少公认语料，主要采用基于规则的方法，从生语料或维基百科等大规模网络语料中抽取篇章关系论元对进行篇章分析。Alsaif等[46]提出自动识别阿拉伯文本中的显式连接词，采用最优句法特征时，识别性能近似人工标注。

5 关键问题及研究难点

综上所述，虽然对篇章结构化划分和显式语义关系识别等任务已经取得显著效果，但对于诸如篇章修辞关系分类，以及隐式语义关系推理仍难以得到有效提高。本节针对现有篇章关系分析尚存的关键问题和难点予以介绍。

• 关键问题

1）篇章修辞关系分类

目前，文本结构化划分已取得较优性能，但篇章修辞关系的分类性能仍然偏低，主要是由于篇章内容和结构的复杂性、不确定性，修辞关系类型的多样性、交叉性造成的。因此，篇章修辞关系识别，以及如何由各层关系构建篇章整体结构，仍然是篇章分析领域待解决的关键问题。

2）隐式语义关系推理

由于各关系类别的数据不平衡现象，容易导致样本稀疏的类别难以被正确分类，导致整体分类性能大幅度下降。虽然目前已有一些针对数据不平衡问题的解决办法，例如重采样、半监督学习等，但整体效果依然不理想。借助监督的机器学习算法，对训练语料具有较强的依赖性，很难保证方法的可扩展性和健壮性，而无监督的关系推理研究在国内外尚属空白。

• 研究难点

1）歧义性与主观性

观察发现，在语料构建过程中，由于标注者自身存在主观性，造成不同标注者的标记结果之间存在

歧义。因此，如何依据上下文，以及选取多少上下文进行篇章关系消歧已经成为隐式篇章关系分析过程中的主要难点。

2）上下文特征抽取

针对具有歧义性的文本片段，若不考虑上下文，则很难正确识别其篇章关系类型，尤其在隐式关系推理过程中，上下文特征发挥着重要作用。但若要挖掘更多相关上下文必将包含更复杂的特征分析，甚至会引入部分噪声。所以，如何抽取有效的上下文特征辅助推理篇章关系也是至今隐式篇章关系分析过程中的重要难点。

6　总结

本文主要介绍篇章关系识别的研究背景、研究意义以及任务描述，并基于已标注的国际公认语料 RSTDT、PDTB 和 HIT-CDTB，详细论述国内外在该领域的现有研究方法。目前篇章关系研究在国内外仍然处于发展阶段，如隐式篇章关系分析仍然无法满足实际应用的需要。但随着语义研究的不断深入和语用研究的不断多元化，篇章关系分析将成为自然语言处理领域中的重要研究方向，具有极高的研究价值和广泛的应用前景。

参考文献

[1] E Pitler, A Nenkova. Revisiting readability: A unified framework for predicting text quality[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2008:186-195.

[2] Z Lin, H T Ng, M Y Kan. Automatically Evaluating Text Coherence Using Discourse Relations[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), 2011: 997-1006.

[3] M Riaz, R Girju. Another look at causality: Discovering scenario-specific contingency relationships with no supervision[C]//Proceedings of the 4th International Conference on Semantic Computing (ICSC), 2010: 361-368.

[4] Q X Do, Y S Chan, D Roth. Minimally supervised event causality identification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011:294-303.

[5] L Zhou, B Li, W Gao, Z Wei, et al. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011:162-171.

[6] 王继成，武港山. 一种篇章结构指导的中文 Web 文档自动摘要方法[J]. 计算机研究与发展, 2003, 40(3): 398-405.

[7] D Y Xiong, D Yang, M Zhang, et al. Lexical Chain Based Cohesion Models for Document-Level Statistical Machine Translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013:1563-1573.

[8] M P Marcus, M A Marcinkiewicz, B Santorini. Building a large annotated corpus of English: The Penn Treebank[J]. Computational linguistics, 1993, 19(2): 313-330.

[9] W C Mann, S A Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization[J]. Text, 1988, 8(3):243-281.

[10] L Carlson, D Marcu, M E Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory[C]//Proceedings of 2nd SIGdial Workshop on Discourse and Dialogue, 2001:1-10.

[11] R Prasad, N Dinesh, A Lee, et al. The Penn Discourse TreeBank 2.0[C]//Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), 2008:2961-2968.

[12] R Prasad, A Joshi, B Webber. Exploiting scope for shallow discourse parsing[C]//Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC), 2010:2076-2083.

[13] D A DuVerle, H Prendinger. A novel discourse parser based on support vector machine classification [C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009:665-673.

[14] E Miltsakaki, N Dinesh, R Prasad, et al. Experiments on sense annotations and sense disambiguation of discourse connectives[C]//Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT), 2005:1-12.

[15] Z Lin, H T Ng, M Y Kan. A PDTB-Styled End-to-End Discourse Parser[J]. Natural Language Engineering, 2012, 1(1):1-35.

[16] M Lan, Y Xu, Z Y Niu. Leveraging Systhetic Discourse Data via Multi-task Learning for Implicit Discourse Relation Recognition[C]//Proceeding of the 51st of ACL, 2013: 476-485.

[17] W T Wang, J Su, C L Tan. Kernel Based Discourse Relation Recognition with Temporal Ordering Information[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics

(ACL), 2010:710-719.

[18] W C Mann, S A Thompson, Rhetorical Structure [J], Theory: Toward a Functional Theory of Text Organization Text, 1988,8:(3): 243-281.

[19] D Marcu. The rhetorical parsing of natural language texts[C]//Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics (EACL), 1997:96-103.

[20] R Soricut, D Marcu. Sentence level discourse parsing using syntactic and lexical information[C]//Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT-NAACL), 2003:149-156.

[21] H LeThanh, G Abeysinghe, C Huyck. Generating discourse structures for written texts[C]//Proceedings of the 20th International Conference on Computational Linguistics, 2004:329-335.

[22] HHernault, H Prendinger, A D Verle. HILDA: A discourse parser using support vector machine classification[J]. Dialogue and Discourse, 2010, 1(3):1-33.

[23] V W Feng, G Hirst. Text-level Discourse Parsing with Rich Linguistic Features[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL). 2012:60-68.

[24] S Joty, G Carenini, R Ng. Combining Intra-and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL). 2013:486-496.

[25] V W Feng, G Hirst. A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL). 2014:511-521.

[26] **张益民，陆汝占，沈李斌. 一种混合型的汉语篇章结构自动分析方法**[J]. **软件学报**, 2000, 11(11): 1527-1533.

[27] **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**[J]. **北京大学学报:自然科学版**, 2014, 50(1):125-132.

[28] E Pitler, A Nenkova. Using syntax to disambiguate explicit discourse connectives in text[C]//Proceedings of the ACL-IJCNLP Conference, 2009:13-16.

[29] B Wellner, J Pustejovsky. Automatically identifying the arguments of discourse connectives[C]//Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007: 92-101.

[30] R Elwell, J Baldridge. Discourse connective argument identification with connective specific rankers[C]//Proceedings of the IEEE International Conference of Semantic Computing, 2008: 198-205.

[31] E Pitler, M Raghupathy, H Mehta, et al. Easily identifiable discourse relations[R]. Technical Reports (CIS), 2008:884.

[32] E Pitler, A Louis, A Nenkova. Automatic Sense Prediction for Implicit Discourse Relations in Text [C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-AFNLP), 2009:683-691.

[33] Z Lin, M Y Kan, H T Ng. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2009:343-351.

[34] Park J, Cardie C. Improving Implicit Discourse Relation Recognition Through Feature Set Optimization [C]//Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), 2012:108-112.

[35] Biran O, McKeown K. Aggregated Word Pair Features for Implicit Discourse Relation Disambiguation [C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), 2013:69-73.

[36] Lan M, Xu Y, Niu Z Y. Leveraging Synthetic Discourse Data via Multi-task Learning for Implicit Discourse Relation Recognition[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013:476-485.

[37] D Marcu, A Echihabi. An Unsupervised Approach to Recognizing Discourse Relations[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), 2002: 368-375.

[38] M Saito, K Yamamoto, S Sekine. Using Phrasal Patterns to Identify Discourse Relations[C]//Proceedings of the Human Language Technology Conference of the NAACL, 2006: 133-136.

[39] Z M Zhou, Y Xu, Z Y Niu. Predicting Discourse Connectives for Implicit Discourse Relation Recognition[C]//Proceedings of the 23rd International Conference on Computational Linguistics (CL): Posters, 2010:1507-1514.

[40] N Xue. Annotating discourse connectives in the Chinese Treebank[C]//Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky, 2005:84-91.

[41] Y Zhou，N Xue. Pdtb-style discourse annotation of Chinese text[C]//Proceedings of the 50th Annual Meeting of the ACL，2012：69-77.

[42] H H Huang，H H Chen. Chinese Discourse Relation Recognition[C]//Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)，2011：1142-1146.

[43] J Li，M Carpuat，A Nenkova. Cross-lingual Discourse Relation Analysis A corpus study and a semi-supervised classification system[C]//Proceedings of the 25th International Conference on Computational Linguistics：Technical Papers (COLING)，2014：577-587.

[44] D Zeyrek，B Webber. A Discourse Resource for Turkish：Annotating Discourse Connectives in the METU Corpus[C]//Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP)，2008.

[45] U Oza，R Prasad，S Kolachina，et al. Experiments with Annotating Discourse Relations in the Hindi Discourse Relation Bank[C]//Proceedings of the 7th International Conference on Natural Language Processing (ICON)，2009.

[46] A Alsaif，K Markert. Modelling discourse relations for Arabic[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)，2011：736-747.

严为绒(1988－)，**硕士研究生，主要研究领域为自然语言处理、篇章分析。**

E-mail：sallyrong8521@gmail.com

朱珊珊(1992—)，**硕士研究生，主要研究领域为自然语言处理、篇章分析。**

E-mail：zhushanshan063@gmail.com

徐扬(1993—)，**硕士研究生，主要研究领域为自然语言处理、事件抽取及关系分析。**

E-mail：andreaxu41@gmail.com