文章编号：1003-0077(2013)03-0020-13

篇章分析技术综述

徐 凡，朱巧明，周国栋

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

摘 要：篇章作为词和句子之后的一种文本分析粒度在自然语言理解和自然语言生成中起到至关重要的作用。该文从计算语言学角度出发，对中英文篇章分析技术的研究现状进行了综述。介绍了中英文篇章分析技术在自然语言处理中的应用，并分别从篇章理论、篇章语料库及评测、篇章分析器的自动构建等方面详细阐述了中英文篇章分析技术。最后归纳出篇章分析技术后续研究的几个方向。

关键词：篇章；篇章分析；语料库；评测

中图分类号：TP391　　**文献标识码**：A

Survey of Discourse Analysis Methods

XU Fan, ZHU Qiaoming, ZHOU Guodong

(School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu 215006, China;
Natural Language Processing Lab, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: Discourse, a kind of text analysis granularity beyond word and sentence, plays a crucial role in natural language understanding and generation. This paper surveys the state-of-the-art researches in Chinese and English discourse analysis under the perspective of computational linguistics, including the applications of Chinese and English discourse analysis, the process of constructing a full Chinese and English discourse parser according to different discourse theories, discourse corpus and evaluation, as well as algorithms and detailed implementation. Also, this paper outlines several directions for further researches on discourse analysis.

Key words: discourse; discourse analysis; corpus; evaluation

1 引言

篇章分析(Discourse Analysis)旨在研究自然语言文本的内在结构并理解文本单元(可以是句子、从句或段落)间的语义关系。它是一种续词、句子之后的文本分析粒度，需要对文本单元的上下文进行全局分析。因而，篇章分析更能挖掘出文本内部丰富的结构化信息，对自然语言理解和自然语言生成有着至关重要的作用。篇章分析技术自底向上可以分为三个研究子方向：其一是面向语言学为主的篇章理论研究，主要解决篇章的表示问题，即篇章的建模。在英文方面，代表性的篇章理论主要有基于实体关系的中心理论(Centering)[1]、基于树状模型的修辞结构理论(Rhetorical Structure Theory，简称RST)[2]、篇章词汇化树型连接语法(Discourse Lexicalized Tree Adjoining Grammar，简称 D-LTAG)[3]和基于图的篇章模型[4-5]等。在中文方面，代表性的

收稿日期：2012-01-13　**定稿日期**：2012-04-11

基金项目：国家自然科学基金资助项目(61070123,61003155)；江苏省自然科学基金资助项目(BK2011282)；江苏省高校自然科学基金重大研究资助项目(11KIJ520003)；教育部科技发展中心网络时代的科技论文快速共享专项研究资助项目；江苏省普通高校研究生科研创新计划资助项目(CXZZ11_0101)

作者简介：徐凡(1979—)，男，博士研究生，主要研究方向为中文信息处理和自然语言处理；朱巧明(1963—)，男，教授，博士生导师，主要研究方向为自然语言处理，中文信息处理，Web 信息处理和嵌入式系统；周国栋(1967—)，男，教授，博士生导师，主要研究方向为自然语言处理，信息抽取，统计机器翻译和机器学习。

篇章理论主要有句群理论[6]和复句理论[7]。其二是基于篇章理论之上的篇章分析器(Discourse Parsing)的自动构建问题。在英文方面，篇章分析器的代表性成果主要有基于 RST-DT(Rhetorical Structure Theory-Discourse Treebank，简称 RST-DT)和基于 PDTB(Penn Discourse TreeBank，简称 PDTB)风格的篇章分析器。在中文方面，目前的工作主要是在模拟英文篇章分析器的基础之上展开的。其三是基于篇章分析技术的与自然语言处理相关的上层应用，即通过使用篇章分析技术直接或间接地提升上层 NLP(Natural Language Processing)系统的性能。

由于篇章分析技术的应用范围非常广泛，所以它受到了学术界和产业界的高度重视。各大高校和科研院所都从不同角度从事篇章分析技术方面的研究。近 10 年来，在 ACL、EMNLP、COLING、《软件学报》、《计算机研究与发展》、《中文信息学报》等相关的自然语言处理国际顶级会议和国内外核心期刊上都发表了很多高质量的篇章分析方面的研究论文。但是到目前为止，并没有文献对篇章分析技术的综合研究成果进行整体上的介绍，而且近年来关于篇章分析的研究仍有很多高质量的研究成果出现。鉴于此，综述这方面的工作有重要意义。

本文对主流的中英文篇章分析技术工作进行了分类、对比和综述。第 2 节阐述了中英文篇章分析技术的应用；第 3 节介绍了主流的英文篇章分析理论、英文篇章语料库及评测；第 4 节分别针对 PDTB 和 RST-DT 篇章语料库详细分析了完整的英文篇章分析器的自动构建过程；第 5 节阐述了与中文篇章分析有关的篇章理论、篇章语料库和篇章分析器的自动构建等内容。最后总结全文，并展望未来的研究工作。

2　中英文篇章分析技术的应用

据引言所述，篇章分析技术具有重要意义，在 NLP 各传统领域和新型领域都具有相关应用①，我们以下逐一介绍。

统计机器翻译(Statistical Machine Translation，简称 SMT)是自然语言处理最直接的上层应用，篇章分析技术在此起到关键作用[8-12]。现有研究主要从篇章连接词的翻译角度[8-9]、从采用中心理论和指代消解等篇章理论或技术角度[10-11]、从修辞关系角度[12]等来提升 SMT 的性能。

自动文摘(Text Summarization，简称 TS)的主要任务是对给定的一篇或多篇文档，由计算机自动生成相应文档或文档集对应的摘要。传统的自动文摘技术主要采用词串等方法，如考虑词的 TF-IDF(Term Frequency-Inverse Document Frequency)特性和命名实体等信息来抽取相关的句子，但用这些方法生成的文摘质量通常不太高。相比较而言，篇章分析技术可以发挥重要作用[13-15]。文献[13]提出了一种基于有向图的篇章多级依存结构的机内表示法。文献[14-15]分析了篇章结构和篇章的意义表示，通过基于篇章理解的技术达到消除句子歧义的目的，并探索了句子级别和上下文级别两个层次的自动文摘问题。

自动问答系统(Question Answering，简称 QA)的主要任务是用计算机对人们提出的问句自动生成答案的过程，它有两个步骤：其一是问句的理解；其二是答案的抽取。篇章分析技术对此两个步骤都具有重要应用[16-20]。文献[16]提出了一个富于语义的有向无环图篇章表示模型，在问句理解步骤，作者将每个问题和对应的答案都对应为一个篇章状态，然后采用图模型中的优化算法去求解。文献[17]在研究问题构成序列时扩充了中心理论的参照、前向和转换等模型。文献[18]针对阅读理解的 Why 型问题提出了基于话题和修辞识别的方法，其核心思想是先利用基于倒文档频率和基于语义角色的两种相似度计算方法识别出对应问题话题的句子，然后进一步识别出这些句子中与问题话题存在因果关系的句子或短语作为返回答案。文献[19]研究了一种包括用户目的、用户可能性、用户态度和用户知识四个方面在内的用户模型，基于 Schema 和 Process 两种生成策略探索了其对问答系统生成内容和风格的影响。文献[20]研究了汉语篇章理解时以事实—事件网络为基础的知识表示和知识库模型，并将其应用于关于鸟类的问答系统中。

信息抽取(Information Extraction，简称 IE)的主要任务是把文本里包含的信息抽取出来并形成结构化的组织形式。同样，篇章分析技术在信息抽取的模板生成阶段将发挥重要作用[21-22]。这些文献通过引入事件个数、事件在文档中的位置等篇章特征来提升信息抽取系统的性能。文献[23-24]分别探索了信息抽取领域的中文地名识别和中文模板约

① 由于中文篇章分析技术的应用相对较少，本文不专门区分中英文情况。

束问题。其中文献[23]认为出现在同一个篇章中的地名在语义上必然存在一定的关联性，作者利用地名之间的同指关系、静态和动态地理关系对篇章地名进行扩展，以此来提升识别性能。文献[24]分析了语句的逻辑结构和篇章结构对信息模板类型的约束作用，并利用篇章结构中的话题或先行语等元素来找回部分缺失的模板元素等信息。

信息检索(Information Retrieval，简称IR)有两方面的任务：其一是对海量信息的组织和存储；其二是根据用户的需求快速查找出相关的信息。传统的IR系统仅仅关注文本的形态或句法分析，对于语义歧义或篇章一致性考虑比较少。文献[25]扩展了已有的检索模型，对文档的上下文信息建立图模型，并对这种图模型引入重排序策略。文献[26]讨论了篇章分析对于信息检索和分类算法性能的影响问题，利用索引(N-gram过滤)和分类(K-means等)两种算法对不同篇章类型的作用进行了评估。

计算机辅助评估(Computer-Assisted Assessment，简称CAA)的主要任务是利用计算机来辅助个人的学习，它可以对人们提交的小段文本进行自动评分，需要利用NLP相关技术对提交的文本进行语法或语义等方面的分析[27-28]。其中，文献[27]提出了一种语义相似度计算和依存图对齐相结合的方法，而文献[28]将CAA问题转化为排序问题，采用N元语言模型和词性等浅层句法特征利用机器学习方法求解。

情感分析(Sentiment Analysis，简称SA)的主要任务是对给定一个待检测项(可以是词、句子、段落或篇章)，由计算机自动分析其具有的正、负和中性情感。传统的情感分析方法绝大多数没有考虑篇章内部的上下文信息，导致算法效率上相对不高。签于此，文献[29]探索了全局推导范型在篇章关系建模中的作用。

作者身份识别(Authorship Attribution，简称AA)的主要任务是利用计算机自动判别给定文档的作者。文献[30]提出了基于字符的局部直方图的方法，认为文学作品的作者在选词时具有一定的篇章分布相关性，实验结果优于传统的BOW(Bag-Of-Words)模型和全局直方图方法。

综上所述，篇章结构分析在上层自然语言处理系统中均存在着广泛的应用，从而使其成为一个非常重要的研究课题。这些研究从侧面也充分说明了只有对文档进行深层次的语义信息挖掘，即只有基于篇章理解技术，才能在现有统计方法的基础上，取得突破性的进展。

3 英文篇章理论、篇章语料库及评测

本节首先介绍两个主流的英文篇章分析理论D-LTAG和RST，然后阐述篇章语料库PDTB[31]和RST-DT[32]的标注体系，最后介绍篇章分析的评测。

3.1 篇章分析理论

3.1.1 D-LTAG理论

D-LTAG是将传统的词汇化树邻接语法(Lexicalized Tree Adjunct Grammars，简称LTAG)应用于篇章层。

面上而取得的谓词—论元(Predicate-Argument)结构，是一种可以表达句法—语义(Syntactic-Semantic)意义和范围的篇章模型[33]。D-LTAG将LTAG扩展至篇章层面，主要有两点改动：(1)将LTAG中的词汇锚用篇章连接词替代；(2)将部分LTAG中的辅助树结构修改为初始树结构。

图1显示了从属连接句、并列连接句和篇章状语对应的D-LTAG树结构，其中D_c代表篇章从句，“↓”代表可执行替换操作，“*”代表可执行连接操作，subconj代表从属连接词，conn代表显式并列连接词或空连接词。这里，图1(a)对应从属连接句的初始树结构，图1(b)对应并列连接词“so”的初始树结构，图1(c)对应除连接词“so”之外的其他并列连接句的辅助树结构，图1(d)对应篇章状态的辅助树结构。

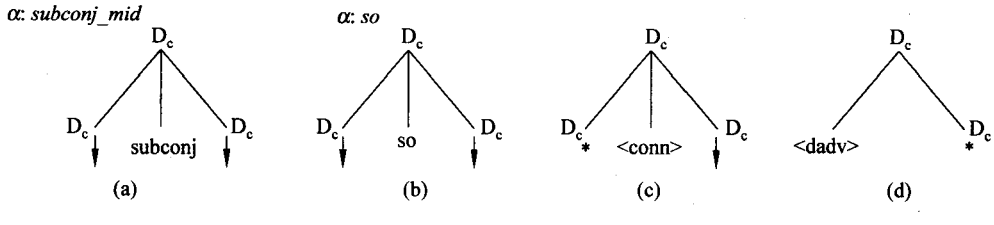

图1 从属连接句、并列连接句和篇章状语对应的D-LTAG结构

为清晰起见，我们采用D-LTAG理论描述带有显式篇章连接词(例1)的D-LTAG的推导过程，限

于篇幅，隐式篇章连接词的情况可以参考文献[34]。

例 1　带有显式篇章连接词实例的 D-LTAG 推导过程

(1. a) John went to the zoo.

(1. b) **However**, he took his cellphone with him.

例 1 中“However”作为篇章连词，从句“John went to the zoo”和从句“he took his cellphone with him”分别作为辅助树上的两个节点。图 2 显示了其对应的 D-LTAG 推导过程。其中，T1 代表从句 1. a 对应的 LTAG 树，T2 代表从句 1. b 对应的 LTAG 树，γ 代表篇章连接词“However”的辅助树，& 代表描述扩展辅助树，“↓”下方代表推导过程：γ 附加到 T1 的根节点，T2 替换成 &，并且 & 附加到 T2 的根节点，推导过程中的实线代表附加操作，虚线代表替换操作，τ1 代表 T1 的推导树，τ2 代表 T2 的推导树。

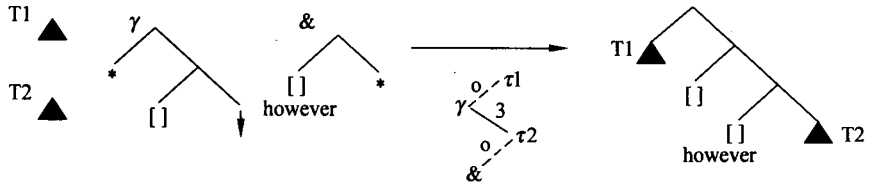

图 2　例 1 的 D-LTAG 推导过程图

3.1.2　RST 理论

RST 的最初目的是用来研究基于计算机的文本生成的，但后来被广泛应用于篇章的功能和结构描述方面[35]①。RST 认为小句（Elemental Discourse Units，简称 EDUs）是最基本的篇章单位，功能语句（Span）是篇章中功能明显的组成部分。其中功能语句间存在各种关系，这种关系集合是开放并且可扩充的，常见的关系有：Circumstance，Elaboration，Background，Enablement and Motivation，Other Relations 等。同时，RST 定义了三个基本概念：(1)核心性：它是指语篇的不对称性，即语篇由核心和辅助部分组成；(2)制约因素：对核心、辅助的分别或同时制约，从而指出命题存在的必要性；(3)效果：使用关系达到的效果解释。RST 采用“核心—卫星”表示结构，其中“核心”在文本中起到更重要的作用，而“卫星”语句却从属于“核心”，每种关系由 5 种要素组成，分别为：(1)核心语句的制约因素；(2)卫星语句的制约因素；(3)单个制约或共同制约因素；(4)结果：即读者使用这种关系后所产生的预期结果；(5)结果的位置：指明的是核心语句、多核心语句或核心语句与卫星语句的共同语句。RST 采用叶子节点代表 EDUs，内部节点代表连续的文本跨度，弧上指明了具体的修辞关系，水平线表明文本跨度，垂直线表明此文本跨度为“核心”。

为清晰起见，我们采用 RST 来描述例 2 的篇章分析过程，其来自《香港城市理工学院的诞生》一文的第一段，读者可参阅文献[36]以了解《香港城市理工学院的诞生》全文对应的 RST 篇章树。

例 2　《香港城市理工学院的诞生》第一段对应的 RST 树分析过程

(2. a) The Genesis of the City Polytechnic of Hong Kong lay in the report of a committee appointed by the Governor in November 1980 to review the scope of post-secondary and technical education to Hong Kong. (2. b1) In its report of June 1981 it recommended (2. b2) that there should be a general and substantial increase in the number of places for tertiary education; (2. b3) one of the measures for achieving this was the establishment of a second polytechnic.

例 2 共有 4 个小句（2. a，2. b1，2. b2 和 2. b3）。其中，2. a 和 2. b1-b3 构成阐述关系，2. a 为“核心”，2. b1-b3 为“卫星”；2. b1 和 2. b2-b3 构成了背景关系，2. b2-b3 为“核心”，2. b1 为“卫星”；2. b2 和 2. b3 构成了条件关系，2. b2 为“核心”，2. b3 为“卫星”，其完整 RST 篇章树如图 3 所示。

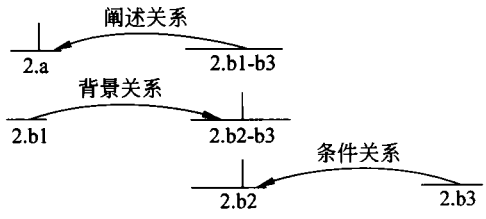

图 3　例 2 对应的 RST 篇章树

3.2　英文篇章语料库

PDTB 是由美国宾西法尼亚大学、意大利托里

① http://www.sfu.ca/rst/05bibliographies/bib_downloads.html (RST Bibliographies)

诺大学和英国爱丁堡大学联合标注，由 LDC(Linguistic Data Consortium)[①]于 2008 年发布。它是目前规模最大的英文篇章级别的语料库，其标注了以下几种类型：(1)显式和隐式篇章连接词；(2)Alternative Lexicalization(AltLex)：篇章关系可以被推导，但加入篇章连接词后会造成表达上的冗余；(3)Entity-based Coherence Relation(EntRel)：不能推导出篇章关系，它是指第二个句子仅仅提供了第一个句子中相关实体的进一步信息；(4)No Relation(NoRel)：既不存在篇章关系又不存在基于实体的一致性。PDTB 对显式和隐式篇章连接词和 AltLex 篇章关系定义了一个三级层次的语义结构：种类—类型—子类型。其中，第一层包括 Temporal、Contingency、Comparison 和 Expansion 在内的 4 类语义，第二层包括 16 类语义，第三层包括 23 类语义。另外，PDTB 还标注了属性，它反映的是显式连接词、隐式连接词和 AltLex 关系的内部单个对象和抽象对象间以及它们参数的“拥有关系”，同时对属性的类型、范围极性、确定性和跨度进行了标记。

RST-DT 由美国南加利福尼亚大学和华盛顿国防部联合标注，由 LDC 于 2002 年发布。它先利用 RST-Tool 工具对文本进行预标注，主要包括文本的切割(生成小句)和初始修辞关系的生成，然后通过人工方式验证预标注的结果，判断文本切分是否正确，并为功能语句对标注一个最可能的修辞关系。

为了清晰起见，表 1 列举了两个篇章语料库的相同点和不同点。

表 1 PDTB 和 RST-DT 比较

<table>
<tr><th>相同点</th><th colspan="6">不同点</th></tr>
<tr><td></td><td colspan="2">采用的理论</td><td colspan="2">标注方法</td><td colspan="2">主要统计数据</td></tr>
<tr><td></td><td>PDTB</td><td>RST-DT</td><td>PDTB</td><td>RST-DT</td><td>PDTB</td><td>RST-DT</td></tr>
<tr><td>(1) 均标注新闻语料
(2) 均采用 WSJ 文章</td><td>采用 D-LTAG 理论</td><td>采用 RST 理论</td><td>(1) 先识别篇章连接词(确定显式、隐式)；(2)然后按不同类型再识别出连接词的两个论元(Arguments)；(3)标记显式、隐式和 AltLex 的篇章关系语义类别(采用类别—类型—子类型层次结构)；(4)对于相邻句对确定除隐式篇章系外的 AltLex、EntRel 和 NoRel 类型；(5)标记属性。</td><td>(1) 先将篇章分割成 EDUs；
(2) 再标注每对 EDUs 可能存在的修辞关系。</td><td># Docs=2 159
Relations=40 600</td><td># Docs=385
Corse-relations=18
Fine-relations=78
EDUs=21 789</td></tr>
</table>

备注：# Docs 代表标注的篇章数；# Relations 代表显式、隐式、AltLex、EntRel 和 NoRel 几大类型的总数目；# Corse-relations 代表粗粒度修辞关系数；# Fine-relations 代表细粒度修辞关系数；# EDUs 代表 EDUs 的总数。

3.3 评测

目前，篇章分析的评测主要考虑算法的正确度和 F1 值两个性能指标。正确度采用式(1)进行度量。

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{All}} \quad (1)$$

这里，*TruePositive* 代表本来是正样例，同时分类成正样例的个数；*TrueNegative* 代表本来是负样例，同时分类成负样例的个数；*All* 代表样例总个数。

F1 值由准确率(Precision)和召回率(Recall)共同体现，采用式(2)进行度量。

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

其中，

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \quad (3)$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \quad (4)$$

这里，*FalsePositive* 代表本来是负样例，但被分类成正样例的个数(通常叫误报)；*FalseNegative* 代表本来是正样例，但被分类成负样例的个数(通常叫漏报)。

4 英文篇章分析器的自动构建

本节分别针对 PDTB 和 RST-DT 介绍如何实现

① http://www.ldc.upenn.edu/

一个完整的英文篇章分析器，并分析和对比了不同的方法体系并报告自然语言处理国际顶级会议上的最新评测性能。

4.1 针对 PDTB 的篇章分析器的自动构建过程

PDTB 作为目前最大的篇章语料库，在其上进行篇章分析的研究者也相对较多。为清晰起见，我们先给出 PDTB 的一些标注实例，然后分析篇章分析器的算法细节。例 3 和例 4 是文献[31]中介绍的显式和隐式篇章关系实例，括号中对应了三层次的篇章关系语义以及实例对应的 WSJ 文章编号①。

例 3　显式篇章连接词实例

In addition, its machines are typically easier to operate, so **customers require less assistance from software.** (CONTINGENCY: Cause: result) (WSJ 1887)

例 4　隐式篇章连接词实例

Mrs Yeargin is lying. Implicit = BECAUSE **They found students in an advanced class a year earlier who said she gave them similar help.** (CONTINGENCY: PragmaticCause: justification) (WSJ 0044)

文献[37]实现了第一个 PDTB 风格的端对端的完整英文篇章分析器，主要完成了四个子任务：(1)篇章连接词分类；(2)论元(Argument)标记；(3)显式和隐式连接词以及 AltLex 关系的语义识别；(4)属性标记。下面我们分别阐述这四个子任务，同时综述每个子任务的相关文献。

4.1.1　篇章连接词分类

此步骤的主要任务就是确定待输入文本中的连接词是否充当篇章连接词的角色。如果此连接词充当篇章连接词角色，则进行论元(Argument)的定位和抽取，否则将判断两个相邻句子的篇章关系语义。文献[38]采用完全监督的机器学习方法，利用连接词本身、Self category、Parent category 等句法特征，取得了 96.26%的 Accuracy 和 94.19%的 F1 性能。文献[39]仅使用篇章连接词作为特征取得了 93%以上的 Accuracy。文献[37]除了利用文献[38]提到的一些特征外，另外加入了连接词的上下文信息、相应的词性(Part of Speech，简称 POS)以及从连接词到根节点的路径等相关特征，利用正确句法树下取得了 97.34%的 Accuracy 和 95.76%的 F1 性能，利用自动生成句法树取得了 96.02%的 Accuracy 和93.62%的 F1 性能。由于此子任务相对简单，其所取得的性能已达到实用阶段，所以它将不再是篇章分析器的研究重点。

4.1.2　论元(Argument)标记

此步骤的主要任务就是要在输入文本中抽取第一步识别出来的篇章连接词的两个论元(Argument 1 和 Argument 2)。其具有两个子任务，其一是 Argument 1 和 Argument 2 的定位，其二是确定 Argument 1 和 Argument 2 对应的文本跨度。文献[37]仅考虑 Argument 1 与 Argument 2 出现在同一句或在 Argument 2 的前面句子中这两种情况来完成 Argument 1 的定位和跨度识别工作。对于 Argument 的定位，仍将其看成分类问题，采用了连接词、连接词的上下文及词性等特征，利用正确的句法树下取得了 97.95%的 F1 性能，利用自动句法树下取得了 91.44%的 F1 性能；对于 Argument 的跨度确定任务，通过计算句法树上的每个内部节点所具有的概率值，最后将最高概率作为分类结果。对于 Argument 的部分匹配和精确匹配情况下，利用正确句法树分别取得了 86.24%和 53.85%的 F1 性能，利用自动句法树分别取得了 80.96%和 40.37%的 F1 性能。文献[40]把 Argument 的跨度确定任务看成是 Argument 的中心词识别任务，其不识别完整的文本跨度，仅识别每个 Argument 的中心词(Head word)。作者采用对数线性重排序模型，考虑连接词本身、Argument 中的单词、成分路径等在内的平面特征，对于 Argument1 标记，利用自动句法树取得了 69.8%的 Accuracy。但是，这类研究中潜在的问题是 PDTB 中并没有标注 Argument 对应的中心词。对于此任务，我们可以明确 Argument 的抽取和标记任务，尤其是 Argument 的精确匹配问题，仍然是一个很有挑战性的后续研究问题。

4.1.3　显式和隐式连接词以及 AltLex 关系的语义

篇章关系语义识别的主要任务就是对文本中显式篇章连接词、隐式篇章连接词和 AltLex 关系分别指定相应的语义(第一层次、第二层次或第三层次的语义类别)。由于显式连接词的语义识别任务比较简单，所以目前的篇章关系语义识别的研究主要集中在 AltLex 和隐式连接词关系的语义关系识别上。

文献[41]详细介绍了 AltLex 的概念，同时分析了 PDTB 的标注方法，作者建议标注工作应该采用开放类项目对待，而不应受到句法概率等方面的约

① 按照 PDTB 的标注风格，Argument 1 采用斜体表示，Argument 2 采用粗体表示，连接词采用下划线表示。

束。文献[37]采用完全监督的机器学习方法，考虑篇章关系上下文、成分和依存句法等在内的平面特征，将AltLex和隐式关系统一看成非显式关系，利用正确句法树下仅取得39.63%的F1性能，利用自动句法树仅取得25.46%的F1性能。此实验结果从另一侧面也反映了AltLex和隐式连接词的语义识别将是又一大挑战。

文献[42-47]研究了隐式篇章关系识别子任务，分别采用了全监督[42-46]、无监督[44]和半监督方法[47]。文献[42]提出了一种复核树核方法，将平面特征与结构化特征相结合，同时探索了与时态相关的语言学特征，取得了40%的Accuracy。文献[43-44]探索了连接词在隐式篇章关系识别中的作用问题，利用语言模型对无连接词的相邻两个句子预先恢复最可能的连接词，然后把它看作分类问题，针对PDTB第一层语义，取得了49.95%的Accuracy和35.10%的F1性能。文献[45]针对PDTB的第二层语义进行识别，提出了成分句法树产生规则和依存句法树规则等有效特征，取得了40.2%的Accuracy。文献[46]提出了丰富的语言学特征，对PDTB第一层语义进行识别，取得了44.58%的Accuracy。文献[47]考虑了非频繁篇章关系语义识别问题，核心思想为首先在非标注数据下取得特征共现向量，然后将其扩充传统的特征向量，取得了21.3%的Accuracy。文献[42-47]作为隐式篇章关系的语义识别主流方法，虽然都是采用PDTB篇章语料库，但是各种方法之间的直接可比性程度仍然不够。主要原因在于：(1)由于有些文献对语料库做了局限的预处理，如文献[46]把EntRel和NoRel两种类型的实例归为隐式关系，而其他文献不考虑这两类语义对应的实例；(2)文献[42-47]中对PDTB的训练和测试数据的划分也不完全一致，并没有采用PDTB建议的类别(训练集：Section2-21；开发集：Section22；测试集：Section23)。例如，文献[43,44,46]采用的训练集为Section2-20，开发集为Section0-1，测试集为Section21-22；文献[42]采用的训练集为Section2-22，测试集为Section23-24；而文献[45,47]采用的训练集为Section2-21，测试集为Section23。

4.1.4 属性标记

此步骤的主要任务是针对PDTB中显式、隐式和AltLex三种篇章关系，确定输入文本中的哪些从句为其对应的属性。它又可以分为四个子问题：属性的类型、范围极性、确定性和跨度确定。

文献[37]仅考虑了属性的跨度确定问题，并将其看成是分类问题，首先根据句法和标点符号特征将文本分割成从句，考虑当前从句、前一个从句、下一个从句的单词、小写单词和词干化动词等在内的平面特征。对于Attribute的部分匹配和精确匹配问题，利用正确句法树下分别取得了79.68%和65.95%的F1性能，利用自动句法树下分别取得了57.34%和42.59%的F1性能。实验结果表明Attribute的跨度标记工作又将是一个富有挑战性的后续研究工作。

4.2 针对RST-DT的篇章分析器的自动构建过程

相对于PDTB风格的篇章分析器而言，RST-DT风格的篇章分析研究文献相对较少。笔者认为潜在的原因可能在于RST-DT在规模上不及PDTB语料。基于RST风格的篇章分析器自动构建过程主要有以下两个子任务：(1)EDUs的生成，即对文本进行正确切割；(2)修辞关系的确定，即对第一个子过程的输出采用自底向上方法，为功能子句对确定一个最可能的修辞关系。

4.2.1 EDUs的生成

文献[48]综合考虑了句法和词汇等特征对文本进行分割，并取得了84%的F1性能。文献[49]研究了句子级的篇章分析器构建任务，但其不足之处在于其仅生成一个句子内部的篇章结构。对于EDUs的生成，其采用概率模型 $p(b|w,t)$(w为文本中的每个单词，t为句法树，b为二元变量{边界，非边界})，结合最大似然估计和相应的数据平滑算法进行文本切分，取得了84.7%的F1性能。文献[50]采用句法特征，结合分割规则和线索短语对文本进行分割，并取得了86.9%的F1性能。文献[51]将篇章分割问题看成序列化标注问题，抽取出文本中的单词、POS标记、词汇中心词等在内的平面特征，并取得了94%的F1性能。

4.2.2 修辞关系的确定

文献[47]探索了RST-DT下的非频繁篇章关系语义识别问题，其核心思想为：首先在非标注数据下取得特征共现向量，然后将其扩充传统的特征向量，并取得了18.9%的宏平均F1性能。文献[49]采用概率模型生成句子级的篇章结构，首先利用结构函数和关系函数计算出篇章树的概率，然后将控制集作为过滤的条件参数，分别在18种和110种修辞关系下取得了接近于人工评测的性能值。文献[50]同时考虑了句子级和文本级的两种篇章分析

器构建问题，对于句子级情况，其首先采用句法信息和线索短语生成 EDUs，然后生成句子级的篇章结构；对于文本级情况，其融合文本的相邻句子和文本的组织信息至集束搜索算法中，以期望生成最好的篇章结构。文献[52]是一种完全监督的机器学习方法，考虑了潜层词汇、结构化成份句法等在内的平面特征，取得了 48.1% 的 F1 性能。文献[49]与文献[51-52]的区别在于：文献[49]仅考虑了句子级的篇章树构建算法，其提出的基于句子内部的一些特征不能直接应用于跨句子的篇章分析器构建情况，反之，文献[50,52]考虑了跨句子的篇章分析器构建算法，应用范围相对更广。

5　中文篇章分析技术

相对于英文篇章分析技术的长期研究而言，中文篇章分析技术研究才刚刚起步。本节将围绕中文篇章理论、中文篇章语料库和中文篇章分析器的自动构建三个方面分别阐述。

5.1　中文篇章理论

文献[53-55]对中文篇章研究进行了较深入的综述，它们认为当前中文篇章理论还处于内省阶段，可操作性不强，具有“本土特征”的句群理论和复句理论创立的出发点也不是着眼于篇章理论，而是更偏重于汉语语法方面的研究。然而，文献[56]却认为句群理论可以经过修改后作为切实可行的中文篇章分析理论。其认为句群理论和 RST 在研究对象、研究内容、研究方法和呈现形式等方面都极其相似，但由于句群理论根植于句子层面的定位模式直接导致了其没有发挥应有的价值，句群理论本身及其发展和应用可以借鉴 RST 的发展和应用模式。文献[57]认为复句和 RST 在超句结构、研究对象、语义关系、标记和图式等方面都极其相似，复句理论本身的可操作性比较强，经过略微修改后同样可以作为切实可行的中文篇章理论。基于此，我们分别介绍句群理论和复句理论的定义、分类、实际操作和两者的区别与联系等内容。

5.1.1　句群理论

文献[6]认为句群是语义上有逻辑关系，语法上有结构关系，语流中衔接连贯的一群句子的组合，它是介于句子和段落之间的，或者说是大于句子、小于段落的语言表达单位。其将句群按如下体系进行分类：(1)按照结构上，将其分为并列关系、连贯关系、递进关系、选择关系等 12 大类；(2)按照功能上，将其分为主体句群(包括记叙句群、描写句群、说明句群、议论句群、抒情句群和对话句群)、过渡句群和插入句群；(3)按照形式上，将其分为一重句群和多重句群。在实际操作层面上，可以按句群的内部和外部接应对其进行组合和切分。其中，内部接应是指句子和句子的组合，主要有词语接应、句式接应、辞格接应等类型。外部接应指句群和句群，或句群和句子组合成为段落的手段，可以是时间词语、处所词语、同义词词语等类型。

文献[58]把句群定义为一些句子结合而成的单位，这种结合具有条件性，主要体现在句子都是前后相连的、各句子都围绕一个基本意思进行表述、内部不能分出比句子大的单位、且所有句子紧密地结合成一个比句子大一级的单位等方面。并从构成方式上把句群分为词语的关联、句式的重复、总括性提示和说明等五种类型。文献[59]其把句群定义为一组有明晰的中心意思的、前后衔接连贯的句子，也称为句组或语段，句子间有语义上的联系、逻辑事理上的联系和语法上的联系三种类型。句群中句子和句子的组合方式和词与词组合成短语、分句与分句组合成复句有相同之处，即也有两种形式：一是句子和句子直接组合，靠语序来表示句与句之间的关系；一是借助虚词(关联词语等)来组合。同时，把句群分为并列、承接、递进、选择、转折等几大类。

5.1.2　复句理论

文献[7]认为复句是包含两个或两个以上分句的句子，它包括三个方面的诠解。其一，凡是复句，都包含两个或两个以上的分句；其二，任何一个复句，在口头上都具有“句”的基本特征；其三，复句的构成单位，从构成的基础看是小句，从构成的结果看是分句。其将复句按如下体系进行分类：(1)按照关系上，将其分为联合复句(包括并列复句、连贯复句、选择复句、解说复句和递进复句)和偏正复句(包括假设复句、转折复句、条件复句、因果复句和目的复句)；(2)按照非关系上，将其分为单重与多重，有间与紧缩，有标与无标，陈述和非陈述等几大类。在实际操作层面上，目前主要还是基于关联词上的操作，例如：“因为…所以…”、“如果…就…”等句式可以表示因果关系；“既…又…”、“不但…而且…”等句式可以表示并列关系；“…但是…”、“…否则…”等句式可以表示转折关系等。

文献[59]认为复句是由两个或两个以上的单句所构成的，同时分句是构成单句的单位。其中分句

可以是主谓句也可以是非主谓句。在形式上，分句和分句间由逗号或分号隔开，可以直接组合，也可以借助虚词构成。其按照分句和分句间的关系，把复句分为并列、承接、递进、选择等类型，同时也指出了复句具有一定的层次性。文献[60]从句子的分类、结构和句法变化上详细讨论了汉语句子。从结构上把句子分为简句和繁句，并将复句归为繁句当中，研究了复句在诸如数量、指称、方所、时间、正反等范畴上的语义内容表达手段和形容事情之间的诸如离合、向背、异同、高下等语义关系。同时，其指出单句和复句的划分是非常困难的问题，涉及到句子中主谓结构的个数、句子中是否存在关联词语、句子中有无停顿等三个相互交错的因素。文献[61]把复句定义为可以用语音停顿隔断的两个句子形式的构成者。其着重强调了两种语言现象：其一是句子形式，它是指一个连系式。其二是复句中的语间停顿现象。其把复句进一步区分为等立复句和主从复句两种类型，其中等立复句中所包含的句子形式具有平等价值，而主从复句所包含的句子形式具有“主要”和“从属”分别。文献[62]提出了与传统的复句概念极为类似的整句概念，把整句定义为一个前后都各有一个全停顿的主谓形式。在结构上，整句仅指前后有全停顿的主谓形式的语言片段。文献[63]把复句分为包孕复句、等立复句和主从复句三种类型。其中，包孕复句由两个以上的单句构成，且“母句”包孕着其余的“子句”；等立复句由两个以上单句构成，且构成彼此接近或互相联络却都是平等而并立的关系；主从复句是由两个以上的单句构成，不能平等而并立，具有主从性质。

5.1.3 句群与复句的比较

为清晰起见，表2列出了句群理论和复句理论之的相同点和不同点。

表2 句群和复句比较

相同点	不同点			
（1）均采用意合法或关联法进行组合； （2）结构类型基本相同。	构成单位		关联词语使用情况	
	句群	复句	句群	复句
	由句子构成	由分句构成	一般不成对使用关联词语，必要时只使用其中的一个	经常使用成对关联词语

备注：常用的关联词语主要有：“一边…一边”、“之所以…是因为”、“与其…不如”、“因为…所以”、“虽然…但是”等。

除了句群和复句中文篇章理论之外，文献[64]提出了一种混合确定性中文篇章分析方法，它是RST分析、主位模式分析、向量空间模型等方法的混合，利用主述位分析、平行句式分析等多种方法来推测输入文本最可能的篇章结构。作为一种混合方法，此方法较比传统的向量空间模型等方法在适用范围上也相对更广。文献[65]探索了中文篇章理解的元指代消解问题，提出了句焦点概念，采用相应的规则过滤算法生成较为连贯的语篇。文献[66]深入分析了指代消解问题，研究了基于树核函数的指代消解技术，并在中心理论的指导下，采用平面特征与结构化特征相结合的方法，较大程度地提升了中文篇章指代消解的性能。

5.2 中文篇章语料库

由于中文篇章理论的不成熟性直接导致了中文篇章语料库的缺乏。根据我们的调研，目前的中文篇章语料库方面的工作都比较初步，可以大致分为三大类：其一，以“本土”句群和复句理论为代表的中文篇章语料库；其二，以借鉴西方RST为代表的中文篇章语料库；其三，以借鉴西方PDTB体系为代表的中文篇章语料库。下面我们分别介绍。

5.2.1 “本土”句群和复句理论为代表的中文篇章语料库

文献[67]对国内外几个主流的汉语树库的建设过程和主要特点进行了综述，其中清华汉语树库作为国内第一个大规模汉语短语结构树库，已经标注的复杂句子比例为56.8%，说明清华汉语树库已经成为中文篇章语料库的雏形。据文献[68]介绍，作者已经开发了100万词规模的汉语句法树库，标注体系中采用的标记组{单句句型、复句句型、整句、句群}较好地体现了句子间的组合关系，其中单句句型和复句句型既可以灵活地充当句子特定的成分又可以构成整句，但整句则不充当句子中的句法成分，多个整句便构成句群。另外，华中师范大学语言与语言教育研究中心开发了一个面向汉语复句研究的专用语料库，采用《人民日报》和《长江日报》作为语料来源，已收有标复句658 447句，约44 395 000字，

收录了各种句式的现代汉语有标复句①。

5.2.2　借鉴西方 RST 为代表的中文篇章语料库

文献[69]是迄今为止较完整的中文篇章语料库，其采用 RST 预计对 395 篇财经评论文章进行标注。第一阶段已经完成了 97 篇中文文章的标注，在句子切分时考虑句号、问号、叹号、分号、冒号等进行自然切分，对有主次重要成分的单模型采用二叉树结构进行标注，对多个同等重要成分的多模型采用多叉树结构进行标注。其首先利用 RST-Tool 工具对文本预标注，然后人工验证和修改，但初步取得的人工标注一致性程度不是很高（Kappa 系数为 0.638）。实验结果初步说明了将 RST 不加修改地直接应用于中文篇章是否切实可行，其需要篇章研究人员进一步探索。

5.2.3　借鉴西方 PDTB 体系为代表的中文篇章语料库

文献[70]分析了中文树库上的篇章连接词标注工作，其采用类似 PDTB 的标注标准对中文树库中的显式连接词进行标注，分析了中文篇章连接词的分布情况，探索了中文篇章连接词的意义消歧和中文篇章连接词的变形等问题。文献[71]是中国台湾大学在中文篇章关系识别方面的最新工作，其主要贡献之一是基于 Sinica Treebank3.1 之上手工标注了 81 篇中文文章，完成了 3 081 个句对的小规模的中文篇章树库。但笔者认为其当前的版本主要存在以下两个问题：其一是在篇章连接词的参数定位上，他们以句子作为基本单位，然而实际情况却更加复杂，这种参数单位可以是从句、一个句子或多个句子；其二是标注一致性有待于验证，文献作者并没有给出标注一致性 Kappa 值，但笔者认为 Kappa 值是任何语料不可缺少的一部分，因为 Kappa 值可以从另一侧面反映出中文篇章语料库标注工作的难度和语料本身的质量。

5.3　中文篇章分析器的自动构建

据我们调研，目前仅有文献[71]提到中文篇章分析器的自动构建工作，总体来说，它采用类似英文篇章分析器的构建思路，将其看成分类问题，利用了句子长度、标点符号、连接词、词性、上位词等在内的平面特征，采用完全监督的机器学习方法，对于显式和隐式篇章关系取得了 88.28% 的 Accuracy 和 63.69% 的 F1 性能。

6　总结和展望

多年来，篇章分析的研究工作主要围绕篇章建模、篇章分析器自动构建、基于篇章分析的上层应用三个子方向进行。篇章分析的这三个子方向形成了一个自底向上的关系，其中篇章理论是基石，它的成熟性和完备性将直接影响到篇章语料库的质量，进而影响篇章分析器的性能，最终影响到基于篇章分析的自然语言处理的上层应用的性能。通过对这些已有研究的分析和总结，我们可以归结出篇章分析的后续研究的几个方向：

（1）篇章分析器的自动构建

通过上文分析，我们可以明确到，目前不管是基于 PDTB 风格的还是基于 RST-DT 风格的篇章分析器的整体性能都不高。鉴于此，仍有以下子问题需要探索：其一，隐式篇章关系识别问题：当前隐式篇章关系的第一、二层次语义的识别性能均仅在 40% 左右，从目前的研究看来，词汇级、短语级等平面特征相对较多，但结构化特征如树核、复合树核相对较少。文本中较深层次的语言学特征的挖掘、提取和选择需要篇章分析研究者和语言学工作者长期不懈的努力。其二，论元识别问题：虽然 Subordinating 和 Coordinating 类别的 Argument 识别达到 82% 左右的性能，但它考虑的是句内的情况。随着时间的推移，笔者相信跨句的 Argument 的定位和识别以及 Discourse adverbial 的 Argument 的识别将逐渐会成为研究热点；其三，除了显示和隐式篇章关系之外，如何确定 AltLex、EntRel 和 NoRel 几大类型将是篇章分析器的又一难点；其四，篇章分析器的整体性能提升问题：自动生成一个离适用阶段性能要求相近的且较为完整的篇章分析器仍需要很长一段时间的努力。

（2）篇章级的语义分析

传统的文本语义分析大部分都是建立在单个句子层面上的，没有综合考虑句子所处的上下文方面的信息，而这些信息对于文本的较深层次的语义挖掘起到非常关键的作用。我们认为篇章级的语义分析至少存在以下三个研究点，其一，篇章语义树模型。对文本建立类似 RST 的篇章树，同时对篇章树中的节点引入对应 WordNet 语义信息（同义词、上位词、下位词、反义词等），可以避免传统 BOW 模型

① http://ling.ccnu.edu.cn:8089/jiansuo/TestFuju.jsp

缺乏深层次语义信息的缺陷;其二,篇章中实体间的语义关系网络模型。通过对篇章中的句子进行浅层语义分析(Semantic Role Labeling,简称 SRL),然后对 SRL 标记的每个 Argument 建立类似中心理论的实体关系网络,并结合实体间的指代消解技术,可以生成较丰富的篇章实体关系网络模型;其三,基于篇章级语义模型的上层应用。上述两个篇章模型都可以应用在篇章的一致性和连贯性评估、文本的相似度检测、科技论文的复制检测、计算机辅助评估等领域。

(3) 篇章级的"话题"结构分析

"话题"是现代语言学的一个重要概念,它是指文本中被讨论的对象。它包括针对单个句子的句内话题和针对整个语篇的篇章话题两种形式。"话题"对于篇章的一致性和连贯性非常重要,有良好延续性的"话题"可以使整个篇章更易于理解,相反,频繁转换"话题"的篇章则不易于理解。关于篇章级的"话题"结构研究方面,我们认为至少存在以下三个研究点,其一,语篇"话题"的识别问题。随着自然语言处理的句法分析和浅层语义分析等工具性能的不断提高,我们可以借助这些工具对文本预先进行句法和语义分析,抽取出文本中的名词短语、动词短语、时间词、地点词和小句等这些潜在的"话题"对象,然后采用中心理论或指代消解技术,利用回指频度和回指方式等特征(例如,我们可以把回指次数相对比较多和具有大量回指的实体看成整个语篇的"话题")进行"话题"的识别工作;其二,"话题链"识别问题。"话题链"具有强大的篇章组织功能。通过对自然语言文本抽取出不同的"话题链",可以分析出"话题链"内部和"话题链"之间的延续或跳转关系,对于探索语篇的组织形式和篇章意识的培养具有关键作用;其三,基于篇章级"话题"结构分析的上层 NLP 应用。通过确定整个篇章的"话题"和"话题链"后,可以在"话题"的指导下探索基于"话题链"的教学、基于篇章"话题"结构的机器翻译等等。总之,针对整个语篇的篇章级的"话题"结构研究将逐渐成为研究热点。

(4) 中文篇章分析技术

通过本文的分析,当前篇章分析技术的研究者主要集中在欧美国家,中文篇章分析技术研究相对较少。虽然部分英文篇章分析技术可以直接移植到中文篇章环境,但我们仍需要针对中文所具有的特点,专门设计和完善相应的中文篇章理论、中文篇章语料库和中文篇章分析器。这些工作都将切实推动中文篇章分析技术的前进。

致谢 在此,我们向对本研究工作提供帮助的老师和同学表示感谢。

参考文献

[1] Grosz B J, Joshi A K, Weinstein S. Centering: A Framework for Modeling the Local Coherence of Discourse[J]. Computational Linguistics, 1995, 21(2): 203-225.

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

[3] Webber B. D-LTAG: extending lexicalized TAG to discourse[J]. Cognitive Science, 2004, 28(5): 751-779.

[4] Jerry R H. On the coherence and structure of discourse[R]. USA: Stanford CA, 1985.

[5] Wolf F, Gibson E. Representing discourse coherence: a corpus-based analysis[C]//Proceedings of the 20th International Conference on Computational Linguistics. Morristown: Association for Computational Linguistics, 2004: 134-140.

[6] 吴为章, 田小琳. 汉语句群[M]. 北京: 商务印书馆, 2000: 1-246.

[7] 邢福义. 汉语复句研究[M]. 北京: 商务印书馆, 2001: 1-693.

[8] Meyer T. Disambiguating Temporal-Contrastive Discourse Connectives for Machine Translation[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Morristown: Association for Computational Linguiscs, 2011: 46-51.

[9] Meyer T, Belis A P. Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation[C]//Proceedings of the 10th Annual Sigdial Meeting on Discourse and Dialogue. Morristown: Association for Computational Linguiscs, 2011: 194-203.

[10] Nagard R L, Koehn P. Aiding Pronoun Translation with Co-Reference Resolution[C]//Proceedings of Workshop on SMT and MetricsMATR. Morristown: Association for Computational Linguiscs, 2010: 252-261.

[11] Haenelt K. Towards a Quality Improvement in Machine Translation: Modelling Discourse Structure and Including Discourse Development in the Determination of Translation Equivalents[C]//Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation. Mor-

ristown: Association for Computational Linguiscs, 1992: 205-212.

[12] Mitkov R. How could rhetorical relations be used in machine translation (and at least two open questions)? [C]//Proceedings of ACL Workshop on Intentionality and Structure in Discourse Relations. Morristown: Association for Computational Linguiscs, 1993: 86-89.

[13] 刘挺，王开铸. 基于篇章多级依存结构的自动文摘研究[J]. 计算机研究与发展，1999，36(4)：479-488.

[14] 王建波，王开铸. 自然语言篇章理解及基于理解的自动文摘研究[J]. 中文信息学报，1992，6(2)：1-7.

[15] 王建波，杜春玲，王开铸. 基于篇章理解的自动文摘研究[J]. 中文信息学报，1995，9(3)：33-42.

[16] Chai J, Jing R. Discourse Structure for Context Question Answering[C]//Proceedings of the Workshop on Pragmatics of Question Answering at the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2004: 23-30.

[17] Sun M, Chai J Y. Discourse processing for context question answering based on linguistic knowledge[J]. Knowledge-based Systems, 2007, 20(6): 511-526.

[18] 张志昌，张宇，刘挺，等. 基于话题和修辞识别的阅读理解 Why 型问题回答[J]. 计算机研究与发展，2011，48(2)：216-223.

[19] 吴华，黄泰翼. 问答篇章生成系统中的用户模型和文本规划[J]. 中文信息学报，2001，15(4)：28-34.

[20] 崔耀，陈永明. 一个实验性的汉语篇章理解系统[J]. 中文信息学报，1994，8(3)：24-34.

[21] Huttunen S, Vihavainen A, Etter P V, et al. Relevance Prediction in Information Extraction using Discourse and Lexical Features[C]//Proceedings of the 18th Nordic Conference of Computational Linguistics. Latvia, 2011: 114-121.

[22] Cimiano P, Reyle U, Saric J. Ontology-driven discourse analysis for information extraction[J]. Data & Knowledge Engineering, 2005(55): 59-83.

[23] 唐旭日，陈小荷，许超，等. 基于篇章的中文地名识别研究[J]. 中文信息学报，2010，24(2)：24-32.

[24] 袁毓林. 用逻辑和篇章知识来约束模板匹配——逻辑结构和篇章结构知识在信息抽取中的运用[J]. 中文信息学报，2004，19(4)：39-45.

[25] Wang D Y, Luk R W P, Wong K F, et al. An Information Retrieval Approach Based on Discourse Type [C]//Proceedings of the 11th International Conference on Applications of Natural Language to Information System. Springer. 2006: 197-202.

[26] Morato J, Llorens J, Genova G, et al. Experiments in discourse analysis impact on information classification and retrieval algorithms[J]. Information Processing and Management, 2003, 39(6): 825-851.

[27] Mohler M, Bunescu R, Mihalcea R. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2011: 752-762.

[28] Yannakoudakis H, Briscoe T, Medlock B. A New Dataset and Method for Automatically Grading ESOL Texts[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2011: 180-189.

[29] Somasundaran S, Namata G, Wiebe J, et al. Supervised and Unsupervised Methods in Employing Discourse Relations for Improving Opinion Polarity Classification[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Morristown: Association for Computational Linguistics, 2009: 170-179.

[30] Escalante H J, Solorio T. Local Histograms of Character N-grams for Authorship Attribution[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2011: 288-298.

[31] Prasad R, Miltsakaki E, Dinesh N, et al. The Penn Discourse Treebank 2. 0 Annotation Manual [R]. USA: University of Pennsylvania, 2008.

[32] Carlson L, Marcu D, Okurowski M E. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory[C]//Proceedings of the Annual Sigdial Meeting on Discourse and Dialogue, Morristown: Association for Computational Linguiscs, 2001: 30-39.

[33] Forbes K, Miltsakaki E, Prasad R, et al. D-LTAG System: Discourse Parsing with a Lexicalized Tree-adjoining Grammar[J]. Journal of Logic, Language and Information, 2001, 12(3): 261-279.

[34] Joshi A K, Schabes Y. Tree-Adjoing Grammar and Lexicalized Grammars [R]. USA: University of Pennsylvania, 1991.

[35] Taboada M, Mann W C. Applications of Rhetorical Structure Theory[J]. Discourse Studies, 2006, 8(4): 567-588.

[36] 卫真道(著)，徐赳赳(译). 篇章语言学[M]. 北京：中国社会科学出版社，2002：1-171.

[37] Lin ZH, Ng H T, Kan M Y. A PDTB-styled end-to-end discourse parser[R]. Singapore: National Uni-

versity of Singapore, 2010.

[38] Pitler E, Nenkova A. Using Syntax to Disambiguate Explicit Discourse Connectives 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. Morristown: Association for Computational Linguistics, 2009: 13-16.

[39] Pitler E, Raghupathy M, Mehta H, et al. Easily Identifiable Discourse Relations[C]//Proceedings of the 22nd International Conference on Computational Linguistics. Morristown: Association for Computational Linguistics, 2008: 85-88.

[40] Wellner B, Pustejovsky J. Automatically Identifying the Arguments of Discourse Connectives[C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Morristown: Association for Computational Linguistics, 2007: 92-101.

[41] Prasad R, Joshi A, Webber B. Realization of Discourse Relations by Other Means: Alternative Lexicalizations[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Morristown: Association for Computational Linguistics, 2010: 1023-1031.

[42] Wang WT, Su J, Tan C L. Kernel Based Discourse Relation Recognition with Temporal Ordering Information[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2010: 710-719.

[43] Zhou ZM, Xu Y, Niu ZY, et al. Predicting Discourse Connectives for Implicit Discourse Relation Recognition[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Morristown: Association for Computational Linguistics, 2010: 1507-1514.

[44] Zhou ZM, Lan M, Niu ZY, et al. The Effects of Discourse Connectives Prediction on Implicit Discourse Relation Recognition[C]//Proceedings of the 9th Annual Sigdial Meeting on Discourse and Dialogue. Morristown: Association for Computational Linguiscs, 2010: 139-146.

[45] Lin ZH, Kan M Y, Ng H T. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank [C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Morristown: Association for Computational Linguistics, 2009: 343-351.

[46] Pitler E, Louis A, Nenkova A. 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. Morristown: Association for Computational Linguistics, 2009: 683-691.

[47] Hernault H, Bollegala D, Ishizuka M. A Semi-Supervised Approach to Improve Classification of Infrequent Discourse Relations using Feature Vector Extension[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Morristown: Association for Computational Linguistics, 2010: 399-409.

[48] Tofiloski M, Brooke J, Taboada M. A Syntactic and Lexical-Based Discourse Segmenter[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. Morristown: Association for Computational Linguistics, 2009: 77-80.

[49] Soricut R, March D. Sentence Level Discourse Parsing Using Syntactic and Lexical Information[C]// Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2003: 149-156.

[50] LeThanh H, Abeysinghe G, Huyck C. Generating Discourse Structures for Written Texts[C]//Proceedings of the 20th International Conference on Computational Linguistics. Morristown: Association for Computational Linguistics, 2004: 329-335.

[51] Hernault H, Bollegala D, Ishizuka M. A Sequential Model for Discourse Segmentation[C]//Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics. Morristown: Association for Computational Linguistics, 2010: 315-326.

[52] DuVerle D A, Prendinger H. 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. Morristown: Association for Computational Linguistics, 2009: 665-673.

[53] 田然. 近二十年汉语语篇研究述评[J]. 汉语学习, 2005, 1: 51-55.

[54] 郑贵友. 中文篇章分析的兴起与发展[J]. 汉语学习, 2005, 5: 40-48.

(下转第55页)

and optimization of confusion network for LVCSR [J]. 中国科学院电子学报合集 1994-2007.

[11] L Mangu, E Brill, A Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks[J]. In Computer, Speech and Language, 2000,14(4): 373-400.

[12] J Xue, Y Zhao. Improved Confusion Network Algorithm and Shortest Path Search from Word Lattice [J]. ICASSP 2005, 2005: 853-856.

[13] J Zheng, C Huang, M Chu, et al. Generalized segment posterior probability for automatic Mandarin pronunciation evaluation[J]. ICASSP 2007, 2007: 201-204.

[14] 王璐,赵欣如,谢簪,等. 普通话测试信息分析[J]. 中文信息学报,2010,24(4):104-110.

[15] P D Patterson, K Robinson, J Holdsworth, et al. "Complex sounds and auditory images"[C]. in Auditory and Perception. Oxford, UK: Y Cazals, L Demany, and K Horner, (Eds), Pergamon Press, 1992: 429-446.

[16] B C J Moore, B R Glasberg, "A revision of Zwicker's loudness model" [J]. Acustica—Acta. Acustica, 1996. 82: 335-345.

[17] M Slaney, "Auditory Toolbox Version 2" Interval Research Corporation Technical Report, 1998, no. 010, 1998.

[18] 李净,郑方,张继勇,等.汉语连续语音识别中上下文相关的声韵母建模[J]. 清华大学学报(自然科学版),2004, 44(1):61-64.

[19] J Li, F Zheng, W H Wu. Context-independent Chinese initial-final acoustic modeling[J]. ISCSLP'00, Oct. 13-15, 2000: 23-26, Beijing.

[20] 孙景涛. 介音在音节中的地位[J]. 语言科学,2006, 5(2):44-52.

[21] 魏思. 基于统计模式识别的发音错误检测研究[D]. 安徽:中国科学技术大学,2008.

[22] 何珏,刘加. 汉语连续语音中 HMM 模型状态数优化方法研究[J]. 中文信息学报,2006, 20(6):83-88.

(上接第 32 页)

[55] 聂仁发.汉语语篇研究回顾与展望[J].宁波大学学报(人文科学版),2009, 22(3): 40-45.

[56] 陈莉萍.修辞结构理论与句群研究[J].苏州大学学报(哲学社会科学版),2008, 4: 118-121.

[57] 徐赳赳, Webster J J.复句研究与修辞结构理论[J].外语教学与研究, 1999, 4: 16-22.

[58] 曹政.句群初探[M].杭州: 浙江教育出版社, 1984: 1-130.

[59] 张志公.张志公文集①汉语语法[M].上海: 上海教育出版社, 1962: 1-651.

[60] 吕叔湘.中国文法要略[M].北京: 商务印书馆, 1956: 1-463.

[61] 王力.中国现代语法[M].北京: 商务印书馆, 1985: 1-402.

[62] 陆俭明.现代汉语句法[M].北京: 商务印书馆, 1993: 1-235.

[63] 黎锦熙.新著国语文法[M].湖南: 湖南教育出版社, 2007: 1-347.

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

[65] 张威,周昌乐.汉语语篇理解中元指代消解初步[J].软件学报, 2002, 13(4): 732-738.

[66] 孔芳.指代消解关键问题研究[D].苏州: 苏州大学, 2009.

[67] 王跃龙,姬东鸿.汉语树库综述[J].当代语言学, 2009, 11(1): 47-55.

[68] 周强.汉语句法树库标注体系[J].中文信息学报, 2004, 18(3): 1-8.

[69] 乐明.中文篇章修辞结构的标注研究[J].中文信息学报, 2008, 22(4): 19-23.

[70] Xue Nianwen. Annotating Discourse Connectives in the Chinese Treebank [C]//Proceedings of CorpusAnno. Morristown: Association for Computational Linguistics, 2005: 84-91.

[71] Hen-Hsen Huang, Hsin-His Chen. Chinese Discourse Relation Recognition[C]//Proceedings of the 5th International Joint Conference on Natural Language.

[72] Processing. Asian Federation of Natural Language Processing. 2011: 1442-1446.