# 面向意图性的篇章话题结构分析研究与展望

奚雪峰[1),2),3)] 孙庆英[1)] 周国栋[1)]

[1)](苏州大学计算机科学与技术学院 江苏 苏州 215000)

[2)](苏州科技大学电子与信息工程学院 江苏 苏州 215009)

[3)](苏州市虚拟现实智能交互及应用技术重点实验室 江苏 苏州 215009)

**摘 要** 篇章话题结构分析主要针对篇章的意图性,是篇章语义分析的基础,其主要任务是从整体层次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.篇章分析既需要研究篇章的基本构成单元,更需要研究基本构成单元之间的篇章关系.然而当前自然语言处理的研究重心大都集中在词法和句法领域,而忽略了对篇章内在规律的研究,缺乏对篇章话题结构展开有效分析的系统理论方法,这就极大阻碍了基于篇章语义分析的相关应用.本文首先从篇章衔接性和连贯性两个基本特征入手,讨论了篇章话题结构分析的国内外研究现状,从理论体系探索、语料库构建和计算模型三方面展开详细综述,分析对比了各类理论、资源及其模型的特点.其中,理论部分代表性的工作包括语域加衔接理论,Hobbs 模型,修辞结构理论,PDTB 体系,意图结构理论,宏观结构理论等;资源部分主要工作有修辞结构篇章树库、宾州篇章树库、MUC 语料、ACE 评测语料、ARRAU、OntoNotes 和篇章图库等;在计算模型方面,主要围绕上述理论和技术资源展开相关研究;随后,特别讨论了汉语篇章话题结构的最新研究进展.基于上述讨论,本文分析探索了基于主述位理论的篇章微观话题结构表示体系,并描述了相应语料库资源的构建及其一致性检验;篇章微观话题结构形式化表示为一个三元组,其主要特征是一种链式结构,链结点为篇章基本话题(子句),其内部的主位或述位为连接端,连接端之间通过微观话题联接建立起连接关系,其实质是一种语义关联,体现篇章之间的衔接关系.最后,本文还对篇章话题结构研究的未来发展方向进行了总结展望.

**关键词** 篇章话题结构;篇章理论;语料库标注;计算模型;篇章意图性;篇章语义分析

**中图法分类号** TP391 **DOI 号** 10.11897/SP.J.1016.2019.02769

# Research and Prospect of Discourse Topic Structure Analysis for Discourse Intentionality

XI Xue-Feng[1),2),3)] SUN Qing-Ying[1)] ZHOU-Guo-Dong[1)]

[1)](*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215000)

[2)](*School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, Jiangsu* 215009)

[3)](*Virtual Reality Key Laboratory of Intelligent Interaction and Application Technology of Suzhou, Suzhou, Jiangsu* 215009)

**Abstract** The analysis of discourse topic structure which focused on the intension of the discourse is the basis of semantic analysis of discourse level. Its main task is to analyze the semantic relations between the discourse structure and its constituent units from the overall layout, and use the context to understand the discourse. The discourse analysis needs to study the basic constituent units of the discourse, and it is necessary to study the discourse relationship between the basic constituent units. However, most of the researches in Natural Language Processing are focusing on the lexical and syntactic aspects, and the research on the internal law of the discourse is relatively few. The lack of theoretical method system for the effective analysis of the discourse topic has seriously restricted the related application based on the semantic analysis of the discourse.

收稿日期:2016-10-21;在线出版日期:2017-05-31.本课题得到国家自然科学基金(61331011,61472264)、苏州市科技发展计划(重点实验室 SZS201609)项目资助.**奚雪峰**,博士研究生,副教授,中国计算机学会(CCF)会员,主要研究方向为自然语言理解、机器学习.E-mail:xfxi2009@qq.com.**孙庆英**,博士研究生,主要研究方向为自然语言处理.**周国栋**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为自然语言理解、中文计算、信息抽取.E-mail:gdzhou@suda.edu.cn.

This paper begins with the two basic characteristics of discourse coherence and cohesion, discusses the current situation of domestic and foreign research on the discourse structure analysis. It summarizes the three aspects of theoretical system exploration, corpus construction and calculation model, analyzes and compares the characteristics of various theories, resources and models. In the theoretical part, the representative work includes Cohesion and Register Theory (CRT), Hobbs Model, Rhetorical Structure Theory (RST), Architecture of Penn Discourse TreeBank, Intentional Structure Theory (IST), Macro-Structure Theory (MST). In the corpus construction part, the main work of the resource part is Rhetorical Structure Theory Discourse Treebank (RST-DT), Penn Discourse Treebank (PDTB), Evaluation Corpus of Message Understanding Conference (MUC), Evaluation Corpus of Automatic Content Extraction (ACE), ARRAU Corpus, OntoNotes Corpus and Discourse GraphBank. In the calculation model part, the paper mainly focuses on the above theoretical and technical resources to carry out research. Subsequently, we discuss the latest research progress of the topic structure of Chinese discourse. According to the above discussion, this paper analyzes and explores the representation system of the micro-topic structure based on the theme-rheme theory, and describes the construction of the corresponding corpus resources and its consistency test. The discourse micro-topic scheme is expressed as a triple, which main characteristic is a chain structure. The chain node is an EDTU (clause), the internal theme or rheme of EDTU is the end of connection. The end-to-end connections are built by micro-topic link, its essence is a kind of semantic relation which reflecting the relationship between discourses. Finally, this paper also gives a brief introduction to the future research in discourse topic structure.

**Keywords** discourse topic structure; discourse theory; corpus annotation; computational modeling; discourse intentionality; discourse semantic analysis

## 1 引 言

能够让机器自动理解篇章含义是自然语言处理的终极目标.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.概念依存理论(Concept Dependency)[1]开启了篇章理解研究的先河,脚本方法(Script)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是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,因而进一步需要更为通用及开放的结构来表示篇章.为达到此目的,探寻篇章的基本特征来寻求解决之道不失为可行方法.

篇章的七个基本特征[2]已经被自然语言处理领域的研究者广为接受,其中前四个基本特征,即连贯性(Coherence)、衔接性(Cohesion)、意图性(Intentionality)及信息性(Informativity)更是有力地促进了自然语言处理研究的发展[3-10].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,也可以与前两者关联起来综合考虑.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,篇章意图性则更加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,篇章的意图性与篇章理解存在密切关系.

篇章的话题结构充当了篇章意图性的表示形式;而话题结构本身,又通过篇章的连贯性和篇章的衔接性,分别实现了表达和内容两个层面的表示形式.篇章连贯性和衔接性这两个基本特性,对于以中文为代表的汉藏语系,以及以英文为代表的印欧语系的大多数语言,都是值得重点研究的问题.

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[11-12]. 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. 本文结合篇章衔接性和篇章连贯性研究，从篇章意图性角度，重点探讨篇章话题结构.

## 2　问题与挑战

篇章分析（Discourse Analysis）是自然语言处理的核心问题之一，其主要任务是从整体层次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. 篇章分析既需要研究篇章的基本构成单元，更需要研究基本构成单元之间的篇章关系. 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，如篇章修辞结构、篇章话题结构等.

作为篇章分析的基本概念，Beaugrande 和 Dressler[2]将篇章定义为连续的系列语段或句子构成的语言整体单位. 通常，自然语言中有意义的最小单位是词；词可以构成短语、语块和句子；句子又可以构成段落，并最终形成篇章. 这里需要特别强调的是，篇章不是构成单元的无序堆砌；只有当构建的整体单位具有相对完整的意义，表达完整的思想和意图时，才能称之为篇章. 如下给出例 1 和例 2 加以对比说明. 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表示篇章. 其原因在于，这些子句所表达的意义彼此没有关联，难以形成一个整体，也无法表达明确的中心话题（或称为主题）. 与此相比，例 2 中的句子，尽管有些句子的句法成分缺失（图中用【】符号表示缺失的成分），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”作为核心主题的语言整体，因而构成了一个篇章.

**例 1.**　尽管安娜出生在加拿大，但今天天气不好，并且苏州的枇杷上市了，因此，自然语言处理研究迎来了发展高峰期.

**例 2.**　a 李四比较年轻，‖b【】〈而且〉工作经验也不足，|||c【】学历又不高，|d 但是【】不论做啥事情，|||e 他都认真负责，‖f 所以，领导非常器重他.

上述例 2 所表示的篇章，能够表达完整的思想和意图. 通常思想和意图的整体性表现为一个话题，该话题体现了思维的放射性与表达的线性之间的有机联系. “思维的放射性”指的是一个话题（或称主题）由若干子话题（或称基本话题）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.

在篇章分析中，话题在形式上往往表示为某种篇章的基本构成单元，并通过递归组合，基于不同层面的逻辑关系联接，形成不同层次的结构实例. 如图 1 所示，基于篇章修辞结构关系，可以得到与例 2 相对应的篇章层次化结构.

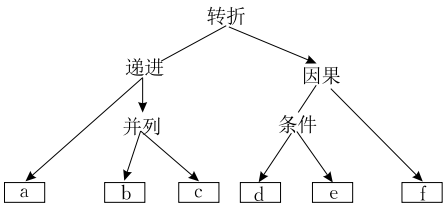


图 1　例 2 的篇章修辞结构实例

如例 2 所示，基本篇章单元采用字母加以表示，其中各个单元之间的关系层次由不同个数的“|”表示. 如“|”代表最高层组合关系，“‖”表示次一级层次. 具体而言，图 1 中的最高层是转折关系，由基本篇章单元 abc 和 def 两个组合构成；随后，abc 单元和 def 单元递归分层又分别得到篇章的次一级关系，即递进和因果关系；最后，再次分层组合，最终形成并列和条件两种最底层篇章关系. 如上例子可见，篇章单位是相对分层的：最底层由各基本篇章单元组成；集成底层的篇章基本单位，构建次高级的；重复组合，不断产生更高级的篇章单位，最终表示成一种树形结构[2]. 在此组合过程中，关键是多种连接关系的存在，其形式上可以表示为某种可见的连接词，如”既…又…”；也可以直接缺省，如例 2 中的“〈而且〉”就是这种省略情况.

上述篇章基本结构（也称为篇章修辞结构）的分析结果对篇章话题理解非常重要. 只有很好地分析出篇章的修辞结构关系，才能更好地理解篇章话题. 例如，在问答系统（Question Answering）中，通过例 2 中的因果关系，不难自动抽取得出相关问题的答案：“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（Document Summarization）而言，根据图 1 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.

综合篇章分析的主要任务(即篇章语义分析)以及篇章的定义来看,篇章的话题理解实质上是篇章分析的核心任务. 从形式上,篇章话题理解又表现为篇章基本结构、及其基本单元之间关系的篇章话题结构分析,两者的研究交叉在一起,密不可分.

篇章话题存在完整性,通常一个话题(或称主题)由若干子话题(或称小主题)构成. 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. 根据图 1 所示例 2 的篇章修辞结构,我们尽管能够在问答系统中提供为什么“领导非常信任他”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),那么图 1 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 2 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”.

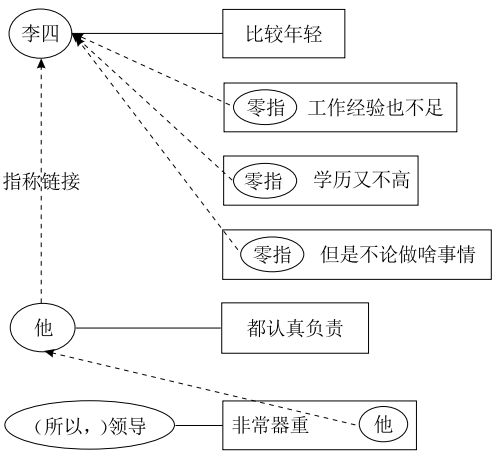


图 2 例 2 的篇章指称结构图

进一步分析可以发现,与上述篇章修辞结构类似,单一的指称结构也只能够解决“Who”这一类问题,对“Why”问题无能为力. 显然,对于需要解决包含 5W 问题(Who,Why,Where,When,What)的篇章话题理解而言,迫切需要联合不同类型(不同粒度/不同层次)的篇章结构(如篇章修辞结构、篇章话题结构等)共同来解决不同类型的篇章理解问题. 其中,一个核心问题是:是否存在一个通用的基本结构可以解决所有篇章话题理解问题?如果没有通用结构,那么对于不同类型篇章话题理解,是否分别存在不同粒度的基本篇章结构可以解决问题?不同粒度的篇章话题结构分别对应可以解决哪些类型的篇章理解问题?不同粒度的篇章话题结构之间的关系又如何?如果存在通用结构,那这种通用结构如何表示?如何实现针对不同类型问题的篇章理解?诸如此类问题,都是篇章话题结构分析的主要困惑,需要进一步研究解决.

## 3 研究现状分析

篇章话题结构分析归属于篇章分析的研究范畴. 下面结合篇章分析,分别从篇章话题结构的理论研究、资源建设、计算模型等方面,分析国内外研究现状.

### 3.1 理论研究

西方语言篇章分析理论研究主要聚焦于篇章连贯性和篇章衔接性的研究,代表性的工作包括:语域加衔接理论(Cohesion and Register Theory)[3],Hobbs 模型[4-5],修辞结构理论(Rhetorical Structure Theory,RST)[6-7],PDTB 体系(Penn Discourse TreeBank)[13-14],意图结构理论(Intentional Structure Theory,IST)[8],宏观结构理论(Macro-Structure Theory)[15],言外行为理论(Illocutionary Act Theory)[16],心理框架理论(Frame Theory of Coherence)[17],主位推进理论(Thematic Progression Theory)[18-19].

实际上,篇章衔接性和篇章连贯性的研究往往是融合在一起的,无法完全独立分开研究. 例如,上述主要侧重篇章衔接性研究的语域加衔接理论,同时也系统解释了篇章连贯性,为后来者对篇章连贯性进一步开展研究提供了帮助,譬如 Widdowson 即是通过深化与补偿语域加衔接理论,提出了“言外行为理论”. 同样得益于语域加衔接理论中的层次和级阶理论,Mann 和 Thompson 在他们提出的用于研究篇章连贯性的修辞结构理论中说到,不同层次的功能块组成的篇章,其功能上的统一性反映了篇章的连贯性. 因此,不同理论在研究过程中,可能存在不同的侧重点,例如有的主要偏重篇章连贯性,有的比较偏重衔接性,还有的则关注上述两者融合后表现出的篇章意图性特点. 下面分别从理论研究的不同侧重点分类加以分析说明.

#### 3.1.1 篇章连贯性理论研究

(1) Hobbs 模型. 主要研究连贯性,提出篇章单元及其之间的连接关系是组成篇章结构的基本部分. 其中,子句、句子、句群和篇章本身,都能够构建

篇章单元；而连接关系是指篇章单元间的语义关联性，它定义了并列、结果、阐述和时机在内的 12 类关系. 在此基础上，篇章表示理论(Discourse Representation Theory)和分段篇章表示理论(Segmented Discourse Representation Theory)对 Hobbs 模型中定义的篇章关系进行了扩充[20-21].

(2) 修辞结构理论(RST). 是一种基于树状结构的模型，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. 同 Hobbs 模型相比，RST 主要也是侧重篇章连贯性的研究，它定义了 4 大类、25 小类用来连接篇章基本单元的修辞关系. 其中，两个被连接的篇章基本单元中主要的那个被命名为“Nucleus”，次要的那个被命名为“Satellite”；如果两个连接的篇章单元存在并列关系，没有主次之分，则这种关系被命名为“多核(Multinuclear)”关系.

为了有利于篇章计算，RST 定义的句子内部结构篇章单元(Elemental Discourse Units，EDU)最小可以由短语或语块(Span)构成. EDU 间的语义关系具有开放性和良好的扩展性. 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 中的三个基本概念之一——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[11].

(3) PDTB 体系. 将源自修辞结构理论的篇章修辞关系作了改进，把它划分为首层、第二、第三层，每层又分别定义 4、16、23 类关系. 其中，主要的核心是篇章修辞关系连接词. 例如，区分篇章关系是否为显式或隐式的依据，即为是否包含显式的连接词. 又如，PDTB 在标注包含隐式篇章关系的篇章单元之前，首先人工添加连接词用来表示当前隐式的篇章语义关系. 此外，PDTB 没有把短语，而是把从句作为最小篇章单元，也是实用性大大增强的一个重要因素[11].

(4) 宏观结构理论. 强调要从微观结构连贯和宏观结构连贯两个方面表示篇章的连贯性. 微观结构连贯体现篇章中单个句子或多个句子所表达的话题存在彼此关联，形成同一的体系结构[21]. 宏观结构连贯表明篇章总体话题与低一级的话题两者在语义结构上趋向一致. 宏观结构理论认为，微观结构连贯是从句子内部角度表示句子与句子之间的顺序关联；而宏观结构是从篇章全局角度表示上下句子分层后的总体关联；当一个篇章同时满足上述两种结构的连贯时，才算是一个连贯的篇章.

(5) 心理框架理论. 强调连贯的心理属性. 认为连贯是篇章使用者利用背景或百科知识对语篇进行阐释的结果，是篇章的语言形式衔接所带来的一种感觉. 心理框架理论同时认为采用一些知识模型也可以表示背景知识的存在，如框架(Frame)、情节(Scenario)、计划(Plan)等. 那么此时，只要这些背景知识模型，能够在篇章中找到相一致的表达意义，就可以形成统一的相互联系，这就构成了连贯的篇章.

(6) 主位推进理论. 认为连续的主位推进有助于构建连贯的篇章. 这种连续表现为相关联的单位之间都通过相似的成分连接起来[21]. 如果篇章缺乏这一成分，主位推进链就会中断，导致语篇衔接上的缺口，造成不连贯现象发生[22]. 篇章的连贯性依赖连续的主位推进；而关联单元通过相似成分连接则反映了篇章的衔接性，因此，主位推进理论实际上包含了对篇章连贯性及篇章衔接性的融合分析研究.

#### 3.1.2　篇章衔接性理论研究

(1) 语域加衔接理论. 主要研究篇章的衔接性，提出语域(情景语境)可以是衔接含义的有效补充，两者联合构成一个篇章概念[3]. 具体来说，一个合理的篇章必须同时满足两方面的条件，一是篇章的各个组成部分必须具有语域一致性，即有与篇章外部的情景相联系的连续的意义；二是篇章内部的各个组成部分之间必须是粘结在一起的，即由衔接纽带连接起来. 语域加衔接理论提出判断一个篇章之所以是篇章的关键，是要看这个篇章是否一致连贯；而衔接则是保障篇章一致连贯的重要方法. 衔接在篇章与语域上的一致性，可以表示出篇章内部从句关系的连贯性.

衔接被定义为词汇和语法两类，第一类包含搭配、重复、上下义词等；第二类包含省略、替代、照应(指代)、连接词等. 此外，他们把语域概念分为语场、语旨和语态，并特别提出语域能够对衔接起到有益补充. 在实际研究过程中，衔接的形式表示，即形成衔接纽带的衔接词，是研究的重点.

(2) 言外行为理论. 提出衔接是句子及其所表达话题意义之间的连接关系，篇章的连贯是这些话题言外功能的表现. 这一理论是在 Halliday 和 Hasan 的理论基础上进一步地深化与补偿；其依据

的另一个前提是 Austin 和 Searl 创立的言语行为理论[23-24]. 言语行为理论认为言语是一种行为活动,我们使用语言开展交流活动就是一个以言行事的过程. 该理论认为篇章是交际行为的方法,是由多个部分行为组成的序列;连续的、有意义的言语行为序列所构成的篇章就是连贯的篇章.

3.1.3 篇章意图性及其他理论研究

(1)意图结构理论(Intentional Structure Theory). 认为话题序列、目的结构和焦点状态是组成篇章结构的三大重要模块.

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. Moser 和 Moore 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(Dominance)和修辞结构理论中的核(Nucleus)相对应[25].

(2)其他理论. 语言篇章模型(Linguistic Discourse Model,LDM) 则以句法分析为指导,通过使用重写规则建立篇章结构分析模型[26].

## 3.2 资源建设

目前篇章话题结构的资源建设主要与篇章连贯性和衔接性相关,包括修辞结构篇章树库(Rhetorical Structure Theory Discourse Treebank,RST-DT)[27]、宾州篇章树库(Penn Discourse Treebank,PDTB)[28]、MUC①(Message Understanding Conference)语料、ACE②(Automatic Content Extraction)评测语料、ARRAU[29]、OntoNotes[30]和篇章图库(GraphBank)[31].

3.2.1 篇章连贯性资源建设

(1)修辞结构篇章树库 RST-DT. 共包含 385 篇文章,由美国南加州大学标注,于 2002 年经 Linguistic Data Consortium(LDC)正式发布. RST-DT 严格定义了篇章基本单元 EDU,其主要包含:担任状语成分的从句、定语从句、后置的名词修辞短语、或将另外 EDU 分割开的从句及非谓语动词短语;剩余作为主语、宾语以及主要动词补语的从句,全都在 EDU 范畴之外.

(2)宾州篇章树库 PDTB. 由 University of Pennsylvania(美国)、University of Torino(意大利)和 University of Edinburgh(英国)联合标注,于 2006 年由 LDC 正式发布初版,并于 2008 年发布 PDTB 2.0 版,包括了华尔街日报的 2304 篇文章. PDTB 具体标注四类篇章关系:一是显式/隐式连接关系(Explicit/Implicit Relation);二是基于实体的关系(Entity-based Relation);三是替代词(Alternative Lexicalization);四是无连接关系(No Relation). 对于显式、隐式篇章连接关系和替代词类别,PDTB 标注的信息包括篇章连接关系、关系连接词及其论元结构、属性信息等. 关系连接词包括主从连接词(Subordinating Conjunction, 如"because, when"等)、并列连接词(Coordinating Conjunction, 如"and, or")和篇章副词(Discourse Adverbial, 如"however, previously")三大类. 篇章连接关系出现在两个论元之间;与连接词关系密切的称为 Arg2,另一个称为 Arg1;两个论元出现的顺序,与连接词间的距离都比较灵活,可在两个句子(子句)之间出现,或者跨越多个句子(子句)等.

3.2.2 篇章衔接性资源建设

指代消解是自然语言处理在篇章衔接性方面的主要研究问题."指代"是一种存在于篇章中前后两个语言单位之间的特殊语义连接关系;而确定两者的过程即称为指代消解[32]. 下面介绍用于指代消解的主要语料资源.

(1) MUC 语料. 是一个由美国政府赞助,以实际文本理解为目标的国际会议. 该会议的第六和第七届,都把指代消解作为评测任务之一. 相应评测采用的语料即为 MUC 语料,该语料通过指代链方式形成指向特征.

(2) ACE 评测语料. ACE 也是美国政府支持的自然语言处理重要会议. 语料评测起始于 2000 年,自 2004 年开始引入中文语料. 其中的指代信息也是采用指代链的方式加以标注而成,每个指代链独立编号并被记录在文件中,而同一个指代链上的实体,它们的指代关系都相同. 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,但是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.

(3) 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

① MUC. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

② ACE. http://www.itl.nist.gov/iad/mig//tests/ace/

包含汉语部分.

(4) OntoNotes 项目. 由 BBN Technologies,、University of Colorado(美国)、University of Pennsylvania(美国)和 University of Southern California's Information Sciences Institute(美国)相互合作创立的. OntoNotes Corpus 5.0 版是 OntoNotes 项目的最新版本,除了包含早先发布版本的所有内容(之前分别发布了 1.0,2.0,3.0,4.0 多个版本)之外,还添加英文核心文本(基督教的《旧约全书》和《新约全书》). 根据来源,语料可以分为来自英语通讯社、中国通讯社、中国广播新闻、英语广播新闻等,累计包含 290 多万个词. 其中英语通讯社以华尔街日报为主,中国通讯社以新华社为主,中国广播新闻主要包括中国中央电视台、中央人民广播电台、中国电视系统等,英语广播新闻也是主流的如美国广播公司、CNN、NBC 的公共国际广播电台和美国之音等,因此能够确保标注语料的权威性.

OntoNotes Corpus 5.0 是一个标注三种语言(中文、英语和阿拉伯文),同时包含结构信息(语法特征及谓词结构)和浅层语义谓词(实体指代信息和共指语义特征信息)的权威语料库. 其中的语法结构来源于宾州树库(Penn TreeBank)中的语法结构部分,而谓词结构则来自于 Penn PropBank 的谓词论元结构. 可以用来表征的语义形式包括名词和动词的词义消歧、连接本体的词义和共指关系. 汉语部分还标注了部分零指代信息,但仅标注了主语位置的零指代;而汉语的零指代种类很多,且每一类别都有其自身的特点,这就制约了汉语零指代消解的研究.

3.2.3　篇章意图性理论研究

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的缺陷,Wolf 和 Gibson 提出了通过图结构表示篇章的方法[31],并研究了篇章图库(Discourse Graph Bank,DGB)的构建问题;同时以该结构标注了 135 篇文章. 他们主要分为三步:首先,根据标点符号将篇章分为基本单元(句子/子句),称之为篇章段(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).

### 3.3　计算模型

基于不同理论体系和相应语料库,近年来国际上开展了很多研究工作,下面我们就按研究的不同角度分别展开介绍.

3.3.1　篇章连贯性计算模型

(1) 基于 RST-DT 的研究

立足 RST-DT 理论开展的篇章基本结构研究主要包含 EDU 识别和篇章连接关系生成两个子任务[33]. 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.

EDU 识别方面,已有较多的研究工作,取得很好性能. 代表性的工作包括:Soricut 和 Marcus[34] 最早采用概率模型 $p(b|w,t)$(这里的 $w$ 表示为单词,$t$ 代表句法树,$b$ 为变量,取值为 $b\in$ {边界,非边界}),基于最大似然估计法,并采用数据平滑算法对文本进行切分,在自动句法分析树上的实验获得了 $F$ 值为 83.1%的 EDU 识别效果,采用标准句法树则识别效果 $F$ 值达到 84.7%;但是,他们的方法没有包含线索词,因此还无法对复杂句子的边界展开准确辨识[35]. Hernault 等人将文本切分问题看成序列化标注问题,基于条件随机场(Conditional Random Field,CRF)模型,采用文本中的单词词义、词性和词汇中心词等词汇和句法信息特征,实验结果 $F$ 值达到 94%[36]. 受此启发,Vanessa 提出一种基于 linear-chain CRFs 的两阶段 EDU 分割模型,实验结果相比文献[36]高出 1 个百分点,取得当前最好结果 95%[37],与人工篇章分割的效果($F$ 值为 98%)较为接近. 总体而言,目前在基于 RST-DT 上进行的 EDU 识别效果,准确率较高,进一步提升空间有限.

篇章连接关系的生成方面,代表性的工作包括:Soricut 和 Marcus 采用概率模型生成句子级别的篇章结构[34]. 基于全自动方法获得的基本篇章单元和句法分析树生成无关系类别标记和有关系类别标记(18 类)的篇章结构树的 $F$ 值分别为 70.5%和 49.0%;而基于正确的 EDU 和标准的句法分析树,生成无关系类别标记和有 18 类关系标记的篇章结构树的 $F$ 值分别为 96.2%和 75.5%,与人工评测的性能非常接近. LeThanh 等人同时研究了文本级别和句子级别的篇章结构的生成[38];针对句子级别的篇章结构生成,借助标准句法信息和线索词对句子篇章单元进行切分,并生成句子级篇章结构,最终获得含 14 类关系的句子结构树的 $F$ 值为 53.0%,生成含 14 类关系的篇章结构树的 $F$ 值为 39.9%;针对文本级的篇章结构,他们将文本的相邻句子和

文本的组织信息融入搜索算法中，借助搜索方法完成文本级篇章结构的生成. DuVerle 和 Helmut 给出了一种基于 SVM 的篇章结构分析方法[39]，采用较大的特征空间，生成的篇章结构树与文献[38]包含的关系类别一样，但 $F$ 值大幅度提高，达到了 48.1%. 文献[40]联合传统的特征向量以及来自非标注数据中的特征共现向量，综合应用于 RST-DT 下非频繁出现的篇章关系的识别任务，取得了 18.9%的宏平均 $F$ 值；后续有些研究引入了更为丰富的语义特征[41-44].

(2) 基于 PDTB 的研究

作为目前规模最大的篇章语料，PDTB 语料积极推动了篇章分析研究，产生了深远影响. 例如 CoNLL-2015 Shared Task 也选择基于 PDTB 开展浅层篇章分析任务[42]. 采用 PDTB 开展的篇章分析任务主要包括显式和隐式两种篇章关系识别，具体可分成针对显式篇章关系的篇章连接词识别、连接关系识别和关系论元抽取；针对隐式篇章关系和 AltLex 关系的连接关系识别；针对篇章关系的属性识别.

① 显式篇章关系研究

在连接词识别领域，已有的研究工作包括：Pitler 等人基于最大熵模型分析了在连接词消歧中句法特征的有效性[45]. 在此基础上，Lin 等人额外补充了相关句法信息，进一步提高了连接词消歧性能[43]. 当前连接词识别性能整体较高，在标准句法树和自动句法树上分别取得 95%和 93%的 $F$ 值.

在连接关系识别领域，已有的相关工作包括：基于朴素贝叶斯方法，Pitler 等人[45]融合连接词和句法信息用于识别第一层的显式连接关系，取得了 94.2%的精准率(Accuracy). 随后，基于最大熵模型，Lin 等人[43]又采用连接词的上下文特征开展了第二层的显式连接关系类型识别，在标准句法树和自动句法树上分别取得 86%和 80%的 $F$ 值.

关系论元抽取领域，已有的相关工作包括：面向 Subordinate 类型的连接词所构成的句内论元，Dinesh 等人提出了一种基于规则的树形减法(Tree Subtraction)算法实现自动抽取[46]，但该算法的缺陷在于仅适用 Subordinate 类型的连接词，其它类型效果不佳. 受到该算法的启发，Lin 等人首先利用机器学习方法完成覆盖论元最小子树的识别，然后借助 Dinesh 的树形减法算法在识别出的子树中完成论元抽取[43]. Lin 等人的这一方法为关系论元抽取技术提供了一种新思路，然而该系统的抽取性能比较依赖于第一步最小子树的识别；如果最小子树所覆盖的论元中含有非论元，则会影响到最终论元抽取的质量.

② 隐式篇章关系和 AltLex 连接关系研究

相比显式篇章关系分析，隐式关系和 AltLex 关系的识别更具挑战性. Pitler 等人对 PDTB 语料的分析发现，显式篇章和隐式篇章关系的数量基本上属于各占半壁江山，但显式篇章关系所包含的歧义不多，大约仅占 2%，而且存在确定的连接词，因此相对较为容易识别[47]. 某种角度来看，对隐式篇章关系识别的效果决定了整个篇章结构关系分析的成败.

这方面的研究思路基本可以分成两类：一类是从隐式关系涉及的论元中直接获取相关信息进行关系类别的判定；一类是考虑到显式关系识别中连接词的决定性地位，借助各类外部资源首先进行隐式关系的连接词恢复，再结合连接词进行关系类别的判定.

第一类方法的代表性工作包括：针对隐式篇章关系的识别，Pitler 等人研究了多种语言信息的有效性[45]，实验发现，动词类型、动词短语的长度、情感倾向、情态动词、上下文语境和词法信息等特征对识别篇章关系都有一定的影响；他们的方法在 PDTB 语料库上识别第一层 4 种隐式关系，即 Comparison、Contingency、Expansion、Temporal 的准确率分别为 56.59%、67.1%、60.28%、61.98%. Lin 等人[48]没有对词进行分类，而是根据统计信息获取词对特征，并利用成分结构信息、依存句法信息和上下文信息，以联接篇章关系的词语作为谓词，把谓词所引领的子句、句子等单位，一并归为待识别论元，在 PDTB 语料库上识别第二层隐式关系获得了 40.2%的准确率(隐式关系中出现频度较高的 11 类). 与文献[45]相比，Biran 和 Mckeown 通过引入词对(Word Pair)特征[49]，在 PDTB 篇章语料库上对第一层 4 类隐式关系的识别准确率分别取得 61.72%、61.52%、60.93%、68.09%，除第 2 类 Contingency 关系外，其余均超出文献[45].

第二类方法的代表性工作包括：Zhou 等人提出一个无指导的方法，基于语言模型首先恢复了隐式关系的篇章连接词，之后统计连接词对应的篇章关系，从而直接判定隐式关系[50]. Hong 等人提出一个用于隐式关系识别的无监督的跨论元的推理机制[51]. 首先借助 Web 中的海量信息构建大量可比较的论元对，然后借助这些可比较的论元对进行连

接词的恢复，最后再根据连接词完成隐式关系类别的判定. Xu 等人研究分析了篇章浅层语义信息和基于态度韵的句子级情感信息对隐式篇章关系识别的贡献，在此基础上采用树核方法利用 PDTB2.0 语料库开展的实验表明显著提升了隐式篇章关系的识别性能[33]. Attapol 和 Nianwen 发现[42]，在通过删除显式连接词的方式构建隐式连接词训练数据时，并不是所有显式连接词都能被直接删除；某些显式连接词如果武断地被删除，反而影响隐式篇章关系的识别性能；在此基础上，他们提出两条规则用于实现可删除显式连接词训练数据的筛选，提高了系统性能.

③ 篇章关系的属性识别

篇章关系属性是指文本中与篇章关系对应的提供额外信息的从句. Lin 等人把篇章关系属性的跨度问题作为分类问题加以研究[43]：首先从文本中利用句法和标点符号特征提取小句；然后利用当前小句及其前后小句的词汇化特征进行属性的精确识别；最终在标准句法树和自动句法树上分别取得了 79.68%和 42.59%的 $F$ 值. Kong 等人结合线性标注和子树抽取技术，提出一种新方法(Constituent-based Approach)[52]，相比于 Lin 等人的系统[43]，在标准句法树上的 $F$ 值提高了 1.8%；而在自动句法树上的 $F$ 值则扩大了 6.7%. Wang 和 Lan[53] 使用 Skadhauge 和 Hardt 提出的模型[54] 列举出所有候选属性，并在 Lin 等人的基础上构建篇章关系属性识别系统[43]，在自动句法树上取得 $F$ 值为 68.27%.

3.3.2　篇章衔接性计算模型

(1) 指代的相关研究

在英语指代消解研究方面，由于自然语言处理的应用需要，特别是 MUC 和 ACE 的推动，指代消解问题引起了广泛地关注，研究方法也从基于规则的方法过渡到了基于机器学习的方法. 主要方法有：① 配对方法，根据各种词汇、上下文、位置等特征，构建指定实体是否指向另一实体的分类器[55-59]. 由于这种方法忽略了指代的传递性和先行词候选间的竞争信息，性能受到了一定限制；② 实体-表述方法，根据上下文信息判断候选词与某类表述集是否表示现实世界的同一个实体. 这一模型将当前的候选词与已经归好类的表述集合配对，不仅根据单个先行词候选来选取多种特征，还考虑整个已归类集合的相关特征[40,60-61]；③ 排序方法，依据多种特征计算所有候选词与指示词指向同一实体的可能性，选择得分最高的作为先行词[62-65].

与英文指代消解相比，汉语指代消解起步较晚，目前尚处于跟进发展阶段. 早期张威等人[66] 采用优先和过滤算法，基于句焦点集实现了汉语元指代消解；鉴于元指代形式标记比较明显，机器识别准确率较高，其中元指代词识别的查全率($R$ 值)达到 92.5%，准确率($P$ 值)达到 89%. 王厚峰等人[67] 基于规则方法，分别从语义及领域知识角度开展汉语人称代词的指代消解研究；针对汉语中“他”、“他们”和“她”三类人称代词的消解性能分别达到了 91.7%、71.8%和 88.5%的准确率. 周俊生等人[68] 提出一种基于图划分的无监督指代消解方法，用于汉语名词短语消解及人称代词消解，其在 ACE 语料上的实验结果 $F$ 值分别达到 62.05%、79.36%，与有监督方法性能相当. Ngai 等人[69] 给出的在 ACE 2005 BNEWS 语料上的汉语名词短语消解最佳 $F$ 值性能为 77.2%. 孔芳等人[70] 提出基于树核函数，使用中心理论、融入语义角色和集成竞争者相关信息三个方面，动态扩展结构化句法树来提升中英文代词消解性能的方法，在 ACE2005 BNEWS 语料上的最佳 $F$ 值性能达到 79.3%，优于相同语料上的同类系统[68-69].

上述研究都是针对面向实体的指代，目前中文事件指代相关研究很少. 主要参考实体指代的方法，采用多种特征，针对事件代词和事件名词进行消解[71].

(2) 省略的相关研究

国内外关于省略的研究大致可以分为空语类的识别恢复以及零指代消解. 在英语空语类识别恢复方面，将空语类的识别和恢复看作句法分析的一个子任务，将它集成到句法分析中或看作句法分析的后处理工作[72-74].

汉语空语类的识别和恢复工作有：Yang 和 Xue 提出组合词汇和句法信息进行汉语的空语类恢复，在标准句法树上性能较好，$F$ 值达到 89%，但在自动句法树上[75]，汉语空语类的恢复性能下降至 63.2%. Cai 等人[76] 将空语类的恢复集成到汉语句法分析中，在自动句法树上的性能较文献[75]有了一定的提升，$F$ 值达到 67.0%. Kong 和 Zhou 提出了基于小句的汉语空语类识别方案，使得汉语空语类识别在自动句法树上的性能 $F$ 值提升至 74.6%[77]. 零指代现象是汉语的显著特征之一，但各种语言中零指代现象有着很大的差异. 汉语指示词的省略可发生在主语、宾语等任意位置，而且没有任何标志，因此省略的指示词的识别成了汉语零指代研究的

一个瓶颈.关于汉语零指代消解的探索较少,代表工作有:Converse 在中文树库 CBT 语料上针对"-NONE- *pro*"这种类型的空语类进行了零元素的标注[71],并探讨已知零元素情况下,在标准句法树上的指代消解方法.Kong 和 Zhou 给出了一个完整的汉语零指代消解框架[78],包括零元素识别、零元素待消解项识别和零元素消歧,并针对三个子任务分别给出了结构化特征方法的解决方案;其在 CTB6.0 中文语料上构建的基准系统自动实现三个子任务的性能 $F$ 值分别为 63.78%、59.54%、45.06%.宋洋等人采用全局规划刻画了面向零指代项的识别和消解这两个子任务的关联关系,并基于马尔可夫逻辑网构建了统一处理模型,用来进行两个子任务的联合推断与学习;在给定标准零元素的情况下,零元素消歧性能 $F$ 值达到 65.96%[79].Zhou 和 Li 基于成分候选和块依层方法研究汉语零元素的恢复[80],在 CTB 5.1 语料库上比当前最好系统 $F$ 值性能提高了 1.29%.

3.3.3 篇章意图性计算模型

这部分的研究极少[81-82],代表性工作是Pustejovsky 等人在 GraphBank 上的相关工作[82],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.

### 3.4 汉语篇章话题结构分析

当前自然语言处理研究一般以句子为单位,主要聚焦在词法、句法以及句子内部的语义分析层面,对篇章内在结构的研究相对较少.虽然英语篇章话题结构分析理论已经取得了一定成果,但是由于汉语的特殊性,难以直接应用于汉语篇章话题结构分析.中英文语言特点的最大不同在于,中文重意合,省略多;英文重形合,结构相对完整.有学者认为这两种语言的不同特点源于东西方不同文化的影响.西方文化崇尚逻辑规范,注重主体和客体的独立性,因而在语言上就表示为讲究形式严密,自成系统.而东方文化信奉和谐相处,强调主体和客体的融合,因此在语言上就表现为讲究意会贯通,忽略了过多的形式桎梏.所以,这也就带来了中英文篇章话题结构分析中的策略有所不同.例如刘礼进标注了小规模汉英篇章平行语料库[83],在此基础上研究对比分析了采用话题结构描述汉英宏观语义结构功能时所存在的差异性.

根据中英文语言特点的最大不同,我们可以将英文篇章分析方法分成"结构完整敏感型"和"非结构完整敏感型"两类方法.前者假定的分析策略,是认为所处理的每个英文篇章,结构是否完整对性能具有很大影响,因此在分析过程中,假定篇章逻辑结构上都是完整的,语言成分没有缺失;而后者,则假定篇章结构是否完整对分析性能没有影响或影响很小.我们在借鉴英文分析方法的时候,直接继承第一类方法所带来的性能并不一定理想,例如在标注中文篇章语料库时,由于汉语中大量连接词的缺省,完全采用 PDTB 标注体系就表现为不太适应[63].另一方面,继承第二类"非结构完整敏感型"的分析策略及方法开展篇章结构分析的结果较好,因而比重较高.篇章结构分析的重要任务之一指代消解的研究就属于这种情况.众所周知,英文指代消解的研究经历了基于规则的方法到基于机器学习方法的过渡,目前机器学习中采用的配对方法、实体-表述方法以及排序方法,都属于"非结构完整敏感型",因而在汉语指代消解中也被广泛沿用[68-70].

尽管我们把英文篇章话题结构分析方法分成了两类典型方法,并且也强调,我们的继承主要是面向第二类继承,但是这个继承策略也并非一成不变.或者换句话说,某种情况下,当满足假定策略时,第一类英文篇章分析方法也是可以借鉴继承的.

从中英文语言现象来说,目前中文存在的非完整性结构特点是在自然语言发展过程中自然形成的,具有其"存在合理性";但是从机器计算的角度来看,语言的完整性对语义理解具有重要作用.因此,实质上,中文中存在的缺省、零指代等非完整结构的情况,非常有必要将其补充完整.显然,结构完整的篇章更有利于计算机开展分析处理.

如果中文的篇章结构能够补充完整,则上述第一类"结构完整敏感型"的英文篇章结构的分析方法也就能够借鉴继承应用到中文分析中来.

但是,这种继承的难点也就产生了,即如何补充完整中文篇章结构的缺失部分?中文篇章中的这种缺失,如省略、零指代等,更多依赖于上下文环境、极为泛化(Generally)的常识、约定俗成的规则等,完全是从深层语义角度来加以关联.并且这种依赖并非静止不变,随着社会文化发展,不断有新的热词、新的典型社会背景知识加入,这是一个动态变化的过程.因此,要把"重意合"的、成分缺省的中文篇章结构补充完整,是一项非常有挑战性的工作.

从语言学角度,在汉语篇章话题研究方面,为了阐释汉语中的主语和谓语结构,Chao(赵元任)最早引入说明和话题两个概念[84].随后,Tsao(曹逢甫)

的研究发现：由话题构成的话题链(Topic Chain)[85]在控制小句连接方面有着重要作用.不同小句内话题的代词化和省略的过程，可以通过外延的话题语义来加以控制连接，从而构成链式结构.这种语义上的连接，主要包含零指代(Zero Anaphor，ZA)、代词指代(Pronoun Anaphora，PA)和名词指代(Nominal Anaphor，NA)等形式.其中，以零指代构建话题链的过程，通常认为具有最好的可计算性[86].近年来，还有研究者把话题链从句子拓展到句群和篇章层面，使得话题链由表示相同话题的系列语句构成[87].周强构建了多种类型的话题说明关系集，提出一种基于话题链的汉语复句连贯性描述机制[88].

总体而言，相对于西方语言(特别是英语)篇章分析的长期研究，汉语篇章话题分析的研究刚刚起步，目前主要处于理论体系探索和语料库资源建设阶段.

(1)汉语篇章理论

就篇章结构而言，从基本篇章单元、篇章结构的组织、篇章关系的分类，连接词及其分布样式等，汉语与英语相比均有不同.因此围绕汉语篇章结构分析，就不能直接或完全使用面向英语篇章结构分析的篇章理论，如 RST 和 PDTB 体系等.

就汉语而言，复句句群理论来源于本土研究，最早用于汉语句子层面的分析；但是徐赳赳[89]对比研究了复句句群理论和 RST 理论，发现两者研究的对象、内容、方法及其表现形式等都有相通之处，因而推断在篇章分析层面，复句句群理论应该还有很大的潜在应用价值.

此外，宋柔等人针对汉语篇章话题结构进行了比较深入的研究，提出了广义话题结构的概念和相应的表示方法[90-92]；并提出了“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.依据这一理论，他们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.这一研究成果是汉语篇章分析领域的一项开创性工作.

(2)汉语篇章语料库及计算模型

与英语相比，由于大规模高质量的适于汉语篇章分析的标注语料严重缺乏，制约了相关篇章话题计算模型的研究，近年来建立汉语篇章语料库资源日益成为研究者关注的焦点.相关工作主要包括两类：

一是针对汉语话题结构的篇章标注.宋柔课题组基于他们提出的广义话题结构的概念，把标点句看作基本篇章单位，开展了汉语篇章的话题结构标注工作，已标注了《围城》、《鹿鼎记》和其它语料(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[91].其数据仍在修订整理中.其中，《鹿鼎记》第一回的广义话题结构标注及其说明已经在网上公开发布(http://clip.blcu.edu.cn/).

二是在参考 RST 和 PDTB 体系的基础上，结合汉语复句和句群理论的研究成果，对汉语篇章结构的标注体系进行探索.代表性工作有：乐明[93]根据 RST 理论，面向汉语篇章结构，结合汉语句群和复句理论开展了标注探索，主要工作包括以标点符号为边界，定义了篇章修辞结构分析的基本单元；定义了 47 种汉语修辞关系集用于区分核心单元；定义了篇章结构标注的具体规则；在此基础上，选取来自大陆主要媒体中的 97 篇财经评论文章开展了修辞结构标注，探索了中文篇章分析中采用 RST 的可行性.Xue 分析了汉语中篇章连接词的分布情况，并对汉语篇章连接词的意义消歧和变形等问题进行了探讨；他采用类 PDTB 标注体系，面向中文树库篇章连接词标注问题，尤其是显式连接词标注开展了标注实践及相关研究[94].在此基础上，Zhou 和 Xue 在 PDTB 标注系统下对来自中文树库的 98 个文件进行了标注，并对汉语和英语在该体系下的差异进行了分析[95].汉语句子中由于缺省较多连接词，因此无法直接采用面向英语的 PDTB 体系开展相关研究.Huang 和 Chen 提出一种弹性的汉语篇章结构标注框架并完成了网上标注系统的开发；结合此标注框架及 PDTB 体系，标注者完成了隐式或显式、跨句及句内等篇章关系，以及情感信息的标注[63].哈尔滨工业大学刘挺教授的课题组选取 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、复句和句群三个层次的篇章关系的标注[96].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、扩展和并列等六类，标注的关系连接词共有 1472 类.

苏州大学自然语言处理课题组结合 PDTB 和 RST 体系的优势，在充分考虑汉语篇章特点的基础上[97]，将基本篇章单位和连接词分别采用树结构的叶子节点及中间节点的形式加以表示.高级篇章单位相对分层：最底层由各基本篇章单元组成；组合底层的篇章单位，构建次高级的篇章；重复组合，不断

产生更高级的篇章单位，最终将汉语篇章修辞结构表示成一棵篇章结构树[11-12]. 在此方案的指导下，该课题组标注完成了中文树库上 500 篇文章的篇章修辞结构，其中涉及基本篇章单位、篇章结构边界、篇章连接词、篇章分层关系及主次篇章单位等.

## 4 研究展望

### 4.1 主要问题及研究趋势

从上述对国内外研究现状的分析中我们可以看到，相对西方语言，汉语的篇章研究刚刚起步，目前主要存在如下一些问题和研究趋势.

(1) 针对汉语篇章话题结构分析的理论指导体系还不够完善. 如何更好地参考英语篇章理论(如 RST 和 PDTB)，并融合汉语的复句、句群、广义话题结构等本土理论，是逐步建立汉语篇章话题结构分析理论的有效途径.

(2) 适用于汉语篇章话题结构分析的大规模标注资源非常缺乏. 虽然有一些研究者，或基于英语篇章分析理论体系，或基于汉语的复句、句群理论和广义话题结构理论，对标注汉语篇章话题结构分析资源库展开了研究，但有待进一步深入系统研究.

(3) 面向汉语篇章话题结构分析的关键研究十分匮乏. 适用于汉语篇章话题结构分析的理论体系尚未完全建立，相关的标注资源缺乏，因此很难大规模有效地进行相关关键技术研究.

### 4.2 我们的探索

值得注意的是，正如本文第 2 节所述，篇章基本结构的表示粒度对篇章话题结构分析至关重要. 鉴于此，我们在前期构建的基于连接依存树表示汉语篇章修辞结构的基础上，又借鉴主述位理论、英语修辞结构理论和宾州篇章树库体系的优点，参考汉语复句和句群的研究成果，结合汉语本身特点，分析了汉语篇章话题结构的表示模型：首先基于主述位理论构建篇章微观话题结构实体的形式化表示模式；其次基于主位推进理论，搭建话题上下文关联模式；最后，融合上述实体/关联构建了汉语篇章话题结构的表示体系. 其中，根据汉语特点定义了汉语篇章话题结构表示模型的关键元素：篇章基本话题(子句)、篇章微观话题结构(Micro-Topic Scheme)、篇章微观话题结构的主位(Theme)和述位(Rheme)、篇章微观话题联接(Micro-Topic Link)、篇章微观话题链(Micro-Topic Chain).

综合基于连接依存树表示的汉语篇章修辞结构和基于微观话题链表示的汉语篇章话题结构，我们构建了面向意图性的篇章话题结构表示体系，如图 3 所示. 针对此种篇章话题结构表示形式，从形式化表示及可计算角度，给出如下定义.

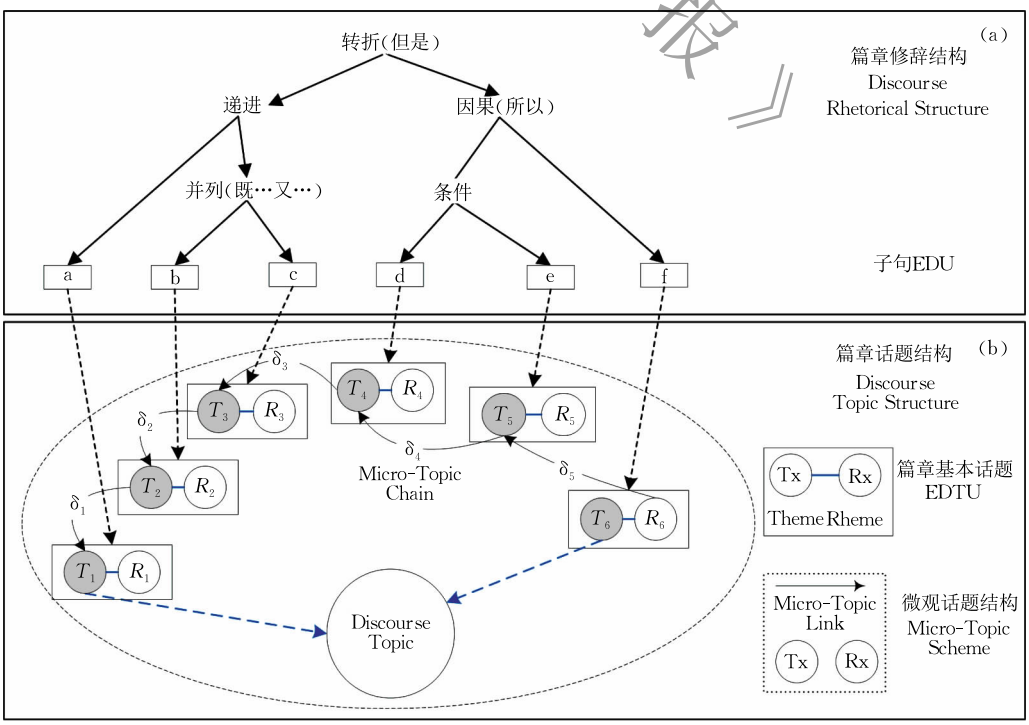


图 3 例 2 的汉语篇章话题结构表示体系

4.2.1　汉语篇章微观话题结构形式化表示

(1)基于主述位理论的汉语篇章话题结构形式化表示

篇章微观话题结构形式化表示为一个三元组；其主要特征是一种链式结构，链结点为篇章基本话题(子句)，其内部的主位或述位为连接端；连接端之间通过微观话题联接建立起连接关系，其实质是一种语义关联，体现篇章之间的衔接关系.

**定义 1.**　篇章基本话题(Elemental Discourse Topic Unit，EDTU)是最小的独立表达意图性的单位，通常是一个含主谓的独立句子.

**例 3.**　(a)两名巴基斯坦飞行员登上飞机.

(b)两名巴基斯坦飞行员登上飞机开始准备起飞.

上述例 3(a)表示一个基本话题，主语是“两名巴基斯坦飞行员”，谓语是“登上(飞机)”；而例 3(b)则包含了两个基本话题，主语相同，谓语分别是“登上(飞机)”和“开始(准备起飞)”.

在图 3 表示的例 2 的汉语篇章话题结构中，篇章基本话题共有 6 个，分别以 a～f 标注. 这里所提到的篇章基本话题结构，从形式上与我们前期有关篇章修辞结构中所标注的篇章子句是一致的，这也有利于开展篇章修辞结构与篇章话题结构的联合研究.

**定义 2.**　篇章微观话题结构(Micro-Topic Scheme，MTS)是一个三元组，

$$MTS=(S_n,S_{n+1},\delta_n),$$

其中，$S_n\in\{T\cup R\}$，$S_{n+1}\in\{T\cup R\}$，$T$ 为一个篇章中的篇章基本话题(EDTU)的主位(Theme)集合；$R$ 为同一个篇章中的篇章基本话题(EDTU)的述位(Rheme)集合；$\delta_n\in\Gamma$，$\Gamma$ 为同一个篇章中的微观话题联接(Micro-Topic link)的集合.

有关主位、述位以及微观话题联接的定义见下述定义 3 和定义 4.

**例 4.**　(a)这张条子是安娜留的，(b)她刚才来过.

上述例 4 中，篇章基本话题(a)与篇章基本话题(b)之间即通过微观话题联接构成一个篇章微观话题结构. 其中，“是安娜留的”是基本话题(a)中的述位，而“她”则是基本话题(b)中的主位. 有关微观话题结构中的主位和述位的概念，见如下定义 3.

**定义 3.**　篇章微观话题结构中的主位(Theme)，是指包含在一个篇章基本话题(EDTU)之中的谓词前面的成分，一般包含主语；谓词及其后的剩余部分，即为述位(Rheme).

上述例 4 中，基本话题(b)的主位“她”与基本话题(a)的述位中的“安娜”形成回指照应. 这里的回指照应即为一种语义关联，形成微观话题联接(Micro-Topic link)，见如下定义 4.

**定义 4.**　微观话题联接(Micro-Topic Link)是一种上下文篇章基本话题(EDTU)内主述位之间语义关联的表示，体现篇章之间的衔接特性，主要包含照应、省略、替代、重复、同义/反义、上下义(具体与抽象)、局部/整体、搭配.

上述定义 4 所述词语间的联接，实质是篇章内衔接的一种表示方式，即词汇衔接. 功能语言学创始人 Halliday 采用两类手段定义篇章衔接性：一为词汇衔接，二为语法衔接. 同时还认为后者也是实现篇章连贯性的重要机制之一. 我们同意 Halliday 的观点，并且认为，篇章话题的联接，需要体现语篇的连贯性，因此，本文所定义的基于词汇衔接的微观话题联接适用于篇章微观话题结构联接.

**照应.** 指的是一个主述位作为另一个基本话题中主述位的参照点，如例 5 中的人称代词“他”指前面出现的“彼得”.

**例 5.**　彼得有一个妻子，非常爱他.

**省略.** 指的是把一个基本话题中的主述位省去不提，是一种避免重复，突出新信息，并使语篇上下紧凑的一种语法手段. 如例 6 中，“看到一只猫”前省略了“我”.

**例 6.**　我早上出门，看到一只猫.

**替代.** 指的是用替代词去取代基本话题中的主述位，替代词的语义来自于所替代的成分.

**重复.** 指的是基本话题中的主述位多次出现，如例 7 中的“熊”.

**例 7.**　安哥拉碰到了一只熊，这只熊显然非常饥饿.

**同义/反义.** 指的是关联上下两个基本话题结构中的主述位是一对同义词/反义词.

**上下义.** 指的是表示抽象和具体关系的两个基本话题中的主述位. 如例 8 中，“生物”对“动物”的界定.

**例 8.**　动物/是生物的一大类，这一类生物/多以有机物为食料，有神经，有感觉，能运动.(《现代汉语词典》第 260 页)

**局部/整体.** 指的是一个基本话题中的主述位是另一基本话题主述位的局部表示. 如下例 9 中的“面”，“身”，“头发”，“嘴”和“手”，可以同上文提到的

“中年男子”形成局部与整体的语义关系.

**例 9.** 前面走来一名中年男子,他满面红光,一身名牌,头发又光又亮,嘴里叼着“红塔山”,手中拿着大哥大.(顾文绮《找头》)

**搭配.** 指的是词汇同现,即一组语义上有联系的词汇关联上下基本话题结构中的主述位.例如下面两组词:(冰,雪,白色)和(夜晚,星星).

在图 3 表示的例 2 的汉语篇章话题结构中,篇章微观话题结构共有 5 个,分别以微观话题联接(图中箭头)相关联,可以表示为($T_1$,$T_2$,$\delta_1$),($T_2$,$T_3$,$\delta_2$),($T_3$,$T_4$,$\delta_3$),($T_4$,$T_5$,$\delta_4$),($T_5$,$R_6$,$\delta_5$).从研究内容来看,首先,微观话题联接涉及多种语义级关联,体现篇章衔接性,有待重点研究关联的可计算性;其次,英语属于语法制导的印欧语系,语法与语义直接对应,而汉语属于汉藏语系,语法与语义相距较远,与英语体系不同,存在较大差别,因此,也有必要开展针对性的汉语篇章微观话题研究.

**定义 5.** 篇章话题结构(Discourse Topic)由 $n$($n\geq 1$)个篇章微观话题结构(MTS)组成,且篇章微观话题结构(MTS)之间也通过篇章微观话题联接(MTLink)相关联.

上述定义 5 中所指篇章话题结构(DT)与篇章微观话题结构(MTS)构成一类整体与部分的组合关系;它们两者之间通过 MTLink 联接起来.实质上,篇章话题结构是一种递归定义,可以表示为:

① 篇章微观话题结构是篇章话题结构;

② 通过篇章微观话题联接(MTLink)相关联的两个篇章话题结构也是篇章话题结构;

③ 篇章话题结构,当且仅当有限次地使用上述规则①和规则②所构成.

**定义 6.** 在一个篇章话题结构(DT)中,多个相关联的篇章微观话题联接(MTLink)构成了一个篇章微观话题链(Micro-Topic Chain).

在图 3 表示的例 2 的汉语篇章话题结构中,$\delta_1$,$\delta_2$,$\delta_3$,$\delta_4$,$\delta_5$构成了一个篇章微观话题链.

(2)基于主位推进模式的汉语篇章话题链体系构建

主位推进模式直观地反映了篇章话题演变关系,将其应用于汉语篇章话题链的识别即可构建一个完整的篇章话题结构体系.我们把篇章话题演变关系作为一个独立模块进行研究,资源标注部分同时也涉及话题结构体系.两者的关系在于,在标注资源上对篇章基本话题进行话题链的识别,即可获得篇章话题动态演变模型.如何运用这个篇章话题结构体系,取决于不同使用者对标注资源的具体应用目的.这里需要研究的问题是如何对话题链进行形式化表示,从而获得一个逻辑性强,并且应用广泛的篇章话题结构体系.

图 4 提出了我们采用的四种主位推进模式.与传统主位推进模式表示不同的是,我们在判断上下子句之间的主述位关系时,不仅包含传统主述位语义相等关系,而且还提出了包含其它具有篇章之间衔接关系的微观话题联接(MTLink)的概念(详见上文定义 4),即主述位之间只要形成微观话题联接,上下句之间的关联关系就能成立.

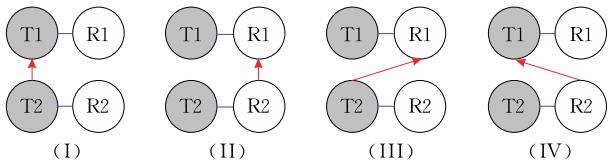


图 4 我们拟采用的四种常用主位推进模式

(I)放射型(平行型或主位相同型).各个子句的主位相关联,而述位各不相关,T2→T1.

**例 10.** 两个绑匪(T1)躲藏了起来(R1),他们(T2=T1)捂住了哈利的嘴巴(R2).

上述例 10 中,第二句中的主位,即人称代词“他们”与上一句中的主位“两个绑匪”,存在回指照应关系,构成一个微观话题联接.

(II)集中型(述位相同型).后一句的述位和前一句的述位相关联,R2→R1.

**例 11.** 孩子们(T1)笑了(R1),然后他们的母亲(T2)也笑了(R2=R1).

上述例 11 中,上下句述位包含“笑了”,存在重复关系,构成一个微观话题联接.

(III)延续型(主位线型发展型).前一句的述位或述位的一部分与后一句的主位形成关联关系,T2→R1.

**例 12.** 我们的小区(T1)是一个大公园(R1),公园里(T2=R1)长满了各种花草(R2).

上述例 12 中,后一句的主位中核心词“公园”,包含在前一句的述位中,构成一个微观话题联接.

(IV)交叉型.后一句的述位与前一句的主位形成关联,R2→T1.

**例 13.** 这只小狗(T1)非常可爱(R1),小朋友们(T2)都非常喜欢它(R2=T1).

上述例 13 中,后一句的述位中核心词“它”,与

前一句主位“小狗”存在回指关系，R2→T1，构成一个微观话题联接.

在图 3 表示的例 2 的汉语篇章话题结构中，5 个篇章微观话题联接，其中 4 个联接采用了第(I)类主位推进模式，即主位同一型；1 个联接采用了第(IV)类主位推进模式，即交叉型.

4.2.2　篇章微观话题资源库构建

根据上述篇章话题结构体系，我们在项目组前期标注的来源于中文宾州树库新闻类文档的篇章修辞结构文本基础上，基于一种自顶向下和后向搜索联合的标注指导思想，借助人工标注和机器辅助相结合的方式，追加与篇章衔接性相关的照应、省略和替代等标注信息并进行微观话题结构及微观话题链的识别标注，构建了汉语篇章话题结构语料库(Chinese Discourse Topic Corpus，CDTC).

(1) 标注策略

对于 CDTC 的标注策略，总体指导原则是：一切从便于篇章理解的角度出发，制定相应的标注规范；充分利用自动标注缩小标注范围，采用手工标注提高标注准确度. 根据主述位篇章微观话题结构和基于主位推进模式的微观话题联接机制，在一定规模的语料上试标注，针对包含照应、省略、替代等微观话题联接的标注及微观话题链识别提出具体的标注规范. 标注规范注重可操作性，分别从判定原则、动态联接方法等方面入手制定，并给出例子详细说明，初步制定标注规范. 进一步在较大规模语料上，实施和验证标注规范的科学性，适当做出调整，最终形成一套完整的汉语篇章话题结构标注规范.

为保证标注质量的同时，又能迅速扩大标注规模，我们结合自举学习(Bootstrapping)和主动学习(Active Learning)方法，机器推荐人工修正，半自动构建标注资源.

主动学习有两种主流方法：不确定性方法(Uncertainty Sampling，US)、专家询问委员会方法(Query-By-Committee，QBC). 其中，不确定性方法利用一种度量机制来评估学习器输出结果的置信度，并选择置信度低(或者叫做不确定)的样本加入训练集. 我们采用基于不确定性的主动学习方法，并结合自举学习方法进行语料的构建. 同时，我们结合基于多视图的半监督学习方法，充分利用单视图分类置信度高的样本. 算法 1 给出了我们的基本算法流程.

**算法 1.**　半自动资源标注算法.

输入：少量标注语料 RL

输出：新标注的语料 NL

步骤：

1. 设置迭代次数为 $N$.
2. 利用 RL 训练学习器 F1.
3. 利用 F1 测试非标注语料 U.
4. 分别选择打分最高的 $N1$ 个正、负样本构成样本集 $A1$.
5. 分别选择打分最低的 $M1$ 个正、负样本构成样本集 $B1$.
6. 将样本集 $A1$ 加入到新标注的语料 NL.
7. 人工标注样本集 $B1$，并加入新标注语料.
8. 返回步骤 2，重复执行上述步骤，直至达到迭代次数 $N$.
9. 输出标注的语料 NL.

虽然自举学习和主动学习方法分别在机器学习领域有广泛研究，但如何将这两种方法有效结合的研究极少. 这或许是未来研究的方向之一.

(2) 标注流程

在 CDTC 语料库标注时，首先导入我们前期已经完成的篇章修辞结构标注处理的语料，作为需要话题结构标注的生语料，然后利用计算机辅助工具生成语料的可视化篇章结构，以辅助人工分析话题结构；通过人工分析识别主述位，寻找候选主述位，建立话题链接关系. 为评估多人标注完成的语料是否达到一致性要求，我们利用一致性检验方法完成了相应的一致性计算，并统计分析了所完成的标注语料结果. 此外，为了克服手工标注生文本费时费力，且容易出错的问题，我们设计开发了汉语篇章微观话题结构计算机辅助标注系统(如图 5 所示)，功能模块包含有篇章结构预处理、计算机辅助可视化结构生成、语料半自动标注、标注结果生成、语料自动统计和一致性自动计算等. 其中在核心功能语料半自动标注模块中，还细分为微观话题结构中主述位标注、微观话题链识别标注等操作.

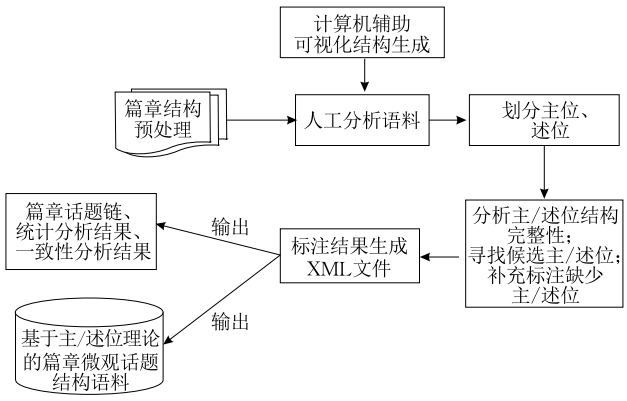


图 5　汉语篇章微观话题结构标注平台处理流程

(3) 质量评估

在语料标注过程中，尽管不同标注者遵循同一

标注规范,但依然存在由于个体主观性差异而导致标注语料结果的不一致.一致性检验即用来验证这种差异程度,并反映问题的本质难易程度.常用的一致性检验方法是 Kappa 检验[98].

Kappa 检验借助观察一致率(observed agreement)和偶然一致率(agreement by chance)两个参数来计算 *Kappa* 值,用来反映标注语料的一致性.*Kappa* 值的计算公式如下:

$$Kappa=\frac{p_0-p_c}{1-p_c},$$

其中,$P_0$ 表示观察一致率,$P_c$ 表示偶然一致率.*Kappa* 值$\in[-1,1]$.在评估一致性时,如果 *Kappa* 值超过 0.75,一般认为标注一致性较好;如果 *Kappa* 值不大于 0.4,则表明一致性较差.为符合常规要求,我们采用 Kappa 方法来检验语料标注质量.

我们以篇章基本话题(子句)为单位,当微观话题结构中的链式结构,即链式结构两端的主位或述位完全相同时,认为微观话题结构的标注结果一致.在语料上分别计算篇章基本话题(EDTU)、主/述位(Theme/Rheme)以及微观话题结构(MTS)的 *Kappa* 值.表 1 给出了语料库的标注一致性检验.所有识别项目的 *Kappa* 值均大于 0.75,因此,我们认为该语料的标注结果是可靠的.

**表 1 标注一致性检验**

| 识别项目 | *Kappa* |
|---|---|
| 篇章基本话题(EDTU)识别 | 0.91 |
| 主/述位(Theme/Rheme)识别 | 0.83 |
| 微观话题结构(MTS)识别 | 0.81 |

目前,CDTC 共包含 500 个文档,2342 个自然段落,6648 个自然句子,10 640 个子句构成篇章基本话题(EDTU);一致性检验表明 CDTC 能够充分体现汉语篇章话题分析问题本身的难度,并为相关研究提供语料资源支持.

4.2.3 汉语篇章话题结构计算模型关键技术研究

篇章话题存在完整性,通常一个话题(或主题)可能由若干子话题(或称基本话题)构成.如果希望准确理解作者所要表达的意图,就必须从上到下整体把握篇章话题.为了达到此目的,清楚地识别出篇章层次结构组成,并找出其中组成结构的关联关系,是一种比较可行的思路.

汉语篇章话题结构分析的主要任务是:首先划分出篇章基本话题(话题句),识别出基本话题中的主位和述位,然后识别出微观话题联接,最后构造出篇章话题结构链.在此过程中,需要结合篇章话题语义和上下文信息,判断微观话题之间的衔接关系类型,并根据篇章中主位推进模式,得到篇章的主要话题.因此,需要研究:

(1)基本话题的边界识别

篇章基本话题结构的边界识别对于篇章话题结构分析非常重要.面向英语的基本篇章单元识别研究工作开展的较早,也取得了较好效果.尽管在篇章层面,英语和汉语存在较大的差别,但在基本话题识别这一层面存在较大的共性.未来或许可以借鉴英语基本篇章识别方法,结合汉语基本话题的特点,采用词法、句法、谓词等信息进行基本话题的边界识别.

(2)篇章微观话题结构的识别

篇章微观话题结构是一个三元组,主要包括篇章基本话题中的主位和述位的识别、以及前后基本话题之间的联接识别.汉语重意合,在子句中会大量出现缺省主语(或宾语等)的情况,因此也带来了包含主语的主位(或包含宾语的述位)的缺省.我们把这种情况称为隐式主位或隐式述位现象;对应则称为显式主位或显式述位.显式主位或述位的识别比较容易确定,问题是隐式主位或述位的确定.由于汉语中此类隐式现象所占比例较大,因此,对于隐式主述位的确定及其微观话题联接关系的确定就成为未来研究的重点和难点.为便于后文展开说明,这里我们特别将包含隐式主述位的微观话题联接关系命名为"篇章微观话题隐式联接关系".

(3)篇章话题的链式结构自动分析

篇章话题的链式结构自动分析的主要任务是:基于主位推进模式,分析出篇章中话题结构及各微观话题结构之间的动态演变规律,重点研究主述位理论在篇章话题结构分析中的作用,解决篇章话题结构全局性优化问题.需要指出的是,基于主述位理论的微观话题结构与传统主题模型不同的是:传统主题模型从词频角度入手分析篇章整体话题,而基于主述位理论的微观话题结构分析的是基本话题句.两者对话题的分析粒度不同.

(4)篇章微观话题联接关系的判断

篇章微观话题联接关系的判断,需要结合关系本体和上下文信息,判断联接关系的类型,如照应、省略、替代等.联接关系的类型体现话题之间的衔接性.未来可以研究如何将主述位推进模式中的主/述

位等价机制，推广到微观话题结构的语义关联性，体现话题结构本体的关联关系.

4.2.4　汉语篇章话题结构标注实例分析

结合主述位理论、RST、PDTB、汉语复句理论、汉语句群理论和广义话题结构理论等的研究，我们提出用主述位构建微观话题链的形式表示汉语的篇章话题结构. 其标注规范如图 6 所示，具体标注说明如下：

(1) TYPE 表示标注对象的类型是实体或事件或联合；

```
<MTS ID=[1..N]      //ID 号
  TYPE=[Entity|Event|Union]
  //实体、事件、联合单元
  POSITION=[Theme|Rheme]
  //主位、述位
  LOCATION=[Root|NotR]
  //初次出现、非初次出现
  KEY=[Complex|Satellite|Nucleus]
  //复合、辅助、核心
  RTYPE=[NotZ|Zero]
  //非零主述位、零主述位
  USETIME=[Numbers]
  //自动计算的标注用时，单位：秒
  {Union=?,?}
  //联合单元指示
  >被标注的主位或述位对象</MTS>
```

图 6　CDTC 标注规范

(2) POSITION 表示当前标注是“主位”或“述位”；

(3) LOCATION 表示当前标注是否初次出现，即是否处于话题链的开始位置；

(4) KEY 表示当前标注类型：复合主/述位、辅助主/述位、核心主/述位；

(5) RTYPE 表示当前标注是否属于“非零”或是“零结构”；

(6) USETIME 表示当前标注所用时间，由标注软件自动计时；

(7) UNION 指示当 TYPE 类型为 union 时，所包含的各个联合指称单元 ID 号.

下面我们通过一个具体的例子介绍我们标注方案中标注的篇章话题结构内容.

例 14 的篇章经标注后，采用 xml 标记对标注后如图 7 中标注实例所示，图 8 表示其结构实例. 其中字母所标记的语段表示篇章基本话题(EDTU)，Tn 前面的语段表示主位，Tn 后面的语段表示述位，用 Rn 表示；各篇章基本话题通过连接主述位的微观话题联接组合后形成微观话题结构，进而再通过组合形成更高一级篇章话题结构(其组合过程也是微观话题联接构建微观话题链的过程)；如此层层组合，最后形成中心篇章话题结构，并且形式上表现为微观话题链. 从图 8 可知，例 14 所示篇章最后可以由两条微观话题链表示(由图中指向中心圆的两条链领衔)，并形成整个篇章的核心话题.

**例 14**　(a)浦东$_{(Satellite(T1))}$开放建设(T1)是一项振兴上海，建设现代化经贸、金融中心的伟大工程(R1)，‖(b)<T2=ZeroA=T1>(因此)大量面对的是以前不曾碰到的新状况、新事务(R2)。| (3)(对此)，浦东$_{(T3=Satellite(T2))}$不是简单的采取“先行动，等克服了各种困难有了经验后再完成法规条例制定”的做法(R3)，‖(4)<T4=ZeroA=T3>而是吸取先进国家及深圳等特区的发展经验(R4)，|||(5)<T5=ZeroA=T4>聘请国内外相关领域的学者专家 (R5)，|||(6)<T6=ZeroA=T5>及时、迅速地制定和推出适合的法规制度文件(R6)，|||(7)<T7=ZeroA=R6>让这些经济活动一旦发生即可迅速被归入合法的处理流程(R7)。

**【标注实例】**

<EDTU><MTS ID=1 TYPE=Event POSITION=Theme LOCATION=NotR KEY=Complex RTYPE=NotZ USETIME=31><MTS ID=1 TYPE=Entity POSITION=Theme LOCATION=NotR KEY=Satellite RTYPE=NotZ USETIME=31>浦东</MTS>开放建设</MTS>是一项振兴上海，建设现代化经贸、金融中心的伟大工程，</EDTU><EDTU><MTS ID=4 TYPE=Entity POSITION=Theme LOCATION=Root KEY=Complex RTYPE=NotZ USETIME=61><MTS ID=1 TYPE=Entity POSITION=Theme LOCATION=NotR KEY=Satellite RTYPE=Zero USETIME=19>null</MTS>因此大量面对的</MTS>是以前不曾碰到的新状况、新事务。</EDTU>

<EDTU><MTS ID=1 TYPE=Entity POSITION=Theme LOCATION=NotR KEY=Complex RTYPE=NotZ USETIME=37>对此，浦东</MTS>不是简单的采取“先行动，等克服了各种困难有了经验后再完成法规条例制定”的做法，</EDTU><EDTU><MTS ID=1 TYPE=Entity POSITION=Theme LOCATION=NotR KEY=Complex RTYPE=Zero USETIME=13>null</MTS>而是吸取先进国家及深圳等特区的发展经验<B TYPE=Y LAYER=6 RID=34>，</EDTU><EDTU><MTS ID=1 TYPE=Entity POSITION=Theme LOCATION=NotR KEY=Complex RTYPE=Zero USETIME=8>null</MTS>聘请国内外相关领域的学者专家，</EDTU><EDTU><MTS ID=1 TYPE=Entity POSITION=Theme LOCATION=NotR KEY=Complex RTYPE=Zero USETIME=10>null</MTS>及时、迅速地制定和推出适合的法规制度<MTS ID=2 TYPE=Entity POSITION=Rheme LOCATION=Root KEY=Nucleus RTYPE=NotZ USETIME=36>文件</MTS>，</EDTU><EDTU><MTS ID=2 TYPE=Entity POSITION=Theme LOCATION=NotR KEY=Complex RTYPE=Zero USETIME=6>null</MTS>让这些经济活动一旦发生即可迅速被归入合法的处理流程。</EDTU>

图 7　例 14 及其微观话题结构标注实例

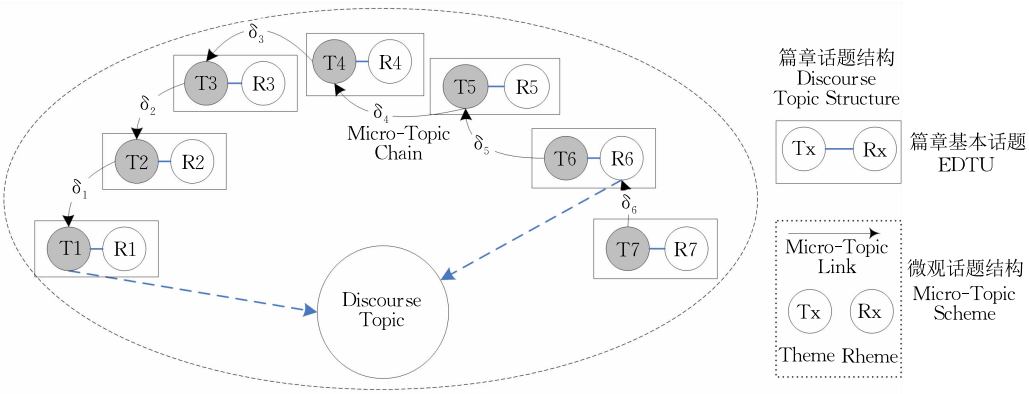


图 8 例 14 的微观话题结构实例图

基于微观话题链的标注方案，我们标注的篇章话题结构信息有：篇章基本话题单元，篇章微观话题结构（含联合主/述位，核心主/述位），篇章微观话题联接及其联接关系类型、篇章话题结构（一种层次化的链式结构）.

4.2.5 汉语篇章微观话题结构分析平台探索

基于上述标注方案和已标注的资源库，提出汉语篇章微观话题结构分析平台主要由以下几个模块构成：

（1）基于标点消歧和成分筛选的篇章基本话题边界识别

关于篇章基本话题边界的识别，研究方向可以从句子分割和主构成成分识别两个视角展开.

① 句子分割视角

相比英语，汉语中的长句比例较高，有时标点符号承担了基本篇章单位的分割功能. 其中，相关统计表明，逗号是分割基本篇章单元最为常用的一种标点符号. 在前期对篇章修辞结构基本单元划分的研究基础上，我们认为可以选择逗号做为篇章基本话题边界识别的主要依据.

根据逗号在汉语中表现出来的功能差异，可以分层表示为不同类型，如图 9 所示. 首先根据是否可以标记篇章基本话题单元，将逗号一分为二，分别表示为篇章基本话题单元可标记逗号（RELATION）和基本话题单元不可标记逗号（OTHER）；其次，根据待分割的篇章单元之间的关系，又将可标记逗号（RELATION）分为两类并列关系篇章单元的分隔逗号（COORD）和从属关系篇章单元的分隔逗号（SUBORD）；再有，基于逗号在句法树上所属的不同层次，又将 COORD 逗号分为用于分隔句子边界的逗号（SB）、用于分隔两个并列 IP 结构的逗号（COIP）、用于分隔两个并列 VP 结构的逗号（COVP）和用于分隔动词与长宾语的逗号（COMP）四类逗号；同时，基于句法角色不同，又将 SUBORD 逗号分为用于分隔附属从句和主句的逗号（ADJ）、用于分隔宾语中并列 IP 结构的逗号（OBJ）两种类型.

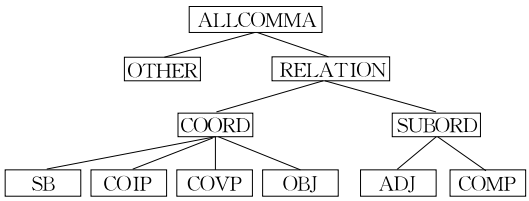


图 9 多类逗号示意图

如上所示，我们所采用的逗号分类体系显然与句法关系密切，所以，可以在完成句法分析的基础上，通过提取逗号所处上下文的词汇、句法等信息进行基于机器学习方法的逗号消歧，并最终根据逗号消歧的结果进行篇章基本话题单元（EDTU）的分割.

② 主构成成分识别视角

篇章基本话题单位识别性能对后续篇章话题结构的分析至关重要. 因为篇章基本单位至少包含一个谓语部分，我们首先可以借助语义角色分析（SRL）提取语句中的多个谓词-论元结构，将每个谓词-论元结构构成的最小单词串看作主构成成分，基于这些成分，探究它们是不是一个 EDTU，最后基于标点信息对生成的 EDTU 进行最终确认.

**例 15.** 尽管浦东开发区制定的法规性文件还不太完善，有些只是试行规定，有待经过实践有效性检验，但这种法制建设为经济和社会活动提供快速保障的做法，得到了国内外投资者的肯定，他们认为，到浦东开发区投资建设有条件，讲法制，利益能得到有效保护.

借助 SRL 提取的谓词-论元结构，我们可以得到下列 8 个成分：

A.［浦东开发区制定的法规性文件］

B.［经过实践有效性检验］

C.［法制建设为经济和社会活动提供快速保障］

D.［他们认为，到浦东开发区投资建设有条件，讲法制，利益能得到有效保护］

E.［［A］还不太完善］

F.［但这种［C］的做法，得到了国内外投资者的肯定］

G.［［E］有些还只是试行规定］

H.［［G］，有待［B］］

基于这 8 个成分提取它们各自的上下文信息，我们可以进行基于机器学习方法的 EDTU 识别.

(2)基于有监督和无监督联合方法开展的篇章微观话题结构识别

自然语言处理中的相关研究表明，零指代识别是一项挑战性极高的任务. 而篇章微观话题结构识别中，因识别的联接关系中也包含指代等衔接，类似零指代识别，同样存在难度. 而且相比英语，汉语中的零指代所占比重更大，因此对整个汉语篇章话题结构分析性能影响也就更大. 所以，对于隐式主/述位的确定及其篇章微观话题隐式联接关系的识别就成为该部分研究的重点和难点. 我们认为可以采取有监督和无监督两种方法，双管齐下开展有关篇章微观话题结构识别的研究.

**有监督方法**. 利用标注资源对显式篇章主/述位开展处理机制研究，为隐式篇章主/述位联接的确定产生正反训练语料，产生隐式篇章主位联接确定模型.

**无监督方法**. 借助搜索引擎确定隐式篇章主位述位联接关系. 给定语段对，分别取一个短语构成一个具有一定语义联系的短语对集合，利用搜索引擎查询相应短语对，获取查询反馈中两者之间可能存在的显式篇章主述位联接分布以及上下文语境，在此基础上建立精准的推理模型，给出相应的隐式篇章主位述位联接关系判断结果. 其本质是利用大规模语料(例如 WEB)，判断给定语段对可能存在的主/述位联接，提高隐式篇章话题结构关系识别的性能.

具体探索高性能汉语篇章微观话题隐式联接关系识别的分析方法可以开展如下探索：

① 基于上下文结构信息和语义相似度方法进行篇章微观话题隐式联接关系的识别

隐式主位之所以省略主位，是因为上下文已经提供了足够的信息，对话题的语义理解不存在歧义. 基于此，我们通过分析语料发现：微观话题隐式联接关系涉及的两个篇章基本话题在构成结构、构词以及语义上有着明显的相似性和可比性. 例如前文给定的例 14 中，片段“吸取先进国家及深圳等特区的发展经验，聘请国内外相关领域的专家学者”包含两个 EDTU，它们都缺少微观话题联接单元(即存在篇章基本话题中的主位或述位)，句法结构存在极高的相似度，而在构词及语义上，“吸取…发展经验”和“聘请专家学者”，这两个表述的语义也关系密切. 因此，可以首先基于上下文结构和语义相似度进行微观话题隐式联接关系的识别.

这一工作可以从以下两方面展开：一方面借鉴前期的研究成果，进行适用于篇章微观话题隐式关系识别的结构化特征树自动获取方案和上下文语义相似度计算方法研究：前期针对指代(实体和事件)消解任务，进行结构化句法信息和语义信息的深入研究(这是笔者所在的课题组已结题的两个国家自然基金的核心工作：基于句法结构和语义信息的指代消解研究，#61003153；篇章衔接性分析：指代、省略及其消歧研究，#61272257)，提出了结合句法和依存两类信息动态获取结构化特征的多种方案；还提出了一个将谓词论元结构看作局部语义信息的表现形式，将实体指代链看作全局语义信息的一种表现形式，借助多种方式将全局和局部信息相集成的基于实体指代链的上下文语义相似度计算方法. 在此基础上，未来可以针对篇章微观话题隐式关系识别任务开展进一步研究.

另一方面针对篇章微观话题隐式联接关系识别任务，探讨高效的上下文表示形式：高效体现在两方面：充足性和低冗余性. 上下文仅考虑逻辑语义关系涉及的两个 EDTU 是否足够？未来的研究可以考虑相邻两个 EDTU 同时构建实体、事件和话题缓冲队列，借助实体链、事件链和话题链来充实上下文，从而更准确地进行隐式联接关系的识别.

② 基于篇章关系互补性的篇章微观话题隐式联接关系识别研究

尽管篇章微观话题隐式联接关系识别是我们课题组提出的一个新任务，除了笔者所在课题组的前期初步研究外，尚无其他研究工作. 但因为与其它自然语言处理任务存在相似性，可以借鉴其它任务的处理思路. 例如，在篇章关系的识别中，因为语料资源缺乏，一些研究者提出通过去除显式关系中的连接词来构建隐式篇章关系语料的思路，并基于这些语料对隐式关系的识别进行了初步的研究. 某种意义上，若显式关系中的连接词不能去除，则它们恰好构成了隐式关系的负例. 借鉴该方法，未来可以通过区分篇章微观话题显式联接中的主位或述位是否可

删除与不可删除,以及篇章微观话题隐式联接添加的主位或述位是否符合自然语言的表达习惯,从连贯性的视角出发,从正反两方面基于这两种关系间的互补性对篇章微观话题隐式联接关系的识别进行研究.

(3)基于主位推进模式的篇章微观话题联接识别

篇章微观话题联接识别可以作为序列化标注问题加以处理.对于该类序列化标注问题中,首先确定需要使用的标注集合.依据话题联接采用的常用主位推进模式,将话题联接分成四类:主位相同型、述位相同型、主位线型发展型和交叉型.借鉴中文分词及短语识别的标注集合,使用 5 个标注符:T,表示主位相同型,即当前句子的主位与相邻上文句子中的主位或主位的一部分相关联;R,表示述位相同型,即当前句子的述位与相邻上文句子中的述位或述位的一部分相关联;L,表示主位线型发展型,即当前句子的主位与相邻上文句子中的述位或述位的一部分相关联;X,表示交叉型,即当前句子的述位与相邻上文句子中的主位或主位的一部分相关联;O,不属于联接关系.确定了标注集合后,就可以分别从词、句、段落、篇章等方面抽取多种特征信息,借助机器学习方法进行微观话题联接的识别.

(4)基于整数线性规划和结构化感知器的全局优化研究

汉语篇章话题结构分析器由相互作用的多个模块构成,以传统的级联方式进行组合必将产生错误的传播和相互作用利用不足的问题.对此提出的解决方法是首先独立地构建汉语篇章话题结构分析器的各个模块,然后借助整数线性规划(Integer Linear Programming,ILP)和结构化感知器(Structured Perceptron,SP)框架进行平台的全局优化研究.

① 基于 ILP 的全局优化研究

汉语篇章话题结构分析器的各个模块独立地获得各自的结果,并未考虑彼此间的不一致.可以将这个全局优化问题看作一个带约束条件的优化问题,基于 ILP 任务的方式进行建模:输入各个模块获取结果时的置信度,并根据各模块间的相互关系设定一系列约束条件,最终找到符合约束条件的最佳结果,达到全局优化的目标.

针对基于主/述位理论的汉语篇章微观话题结构体系中的显式主/述位及联接关系识别两个子任务,可以首先实现一个基于 ILP 的全局优化方案;之后,借助这一方案实现多系统集成.

② 基于 SP 的全局优化研究

针对基于主述位理论的汉语篇章微观话题结构分析器,进一步提出一种基于结构化感知器的全局优化方案.

Collins(2002)提出的结构化感知器(SP)算法扩展自线性感知器(Linear Perceptron),主要用于结构化预测问题.该算法首先调用 Viterbi 算法求解输出最优标注序列;随后,通过将输出的最优标注序列与标准标注序列对比反馈来训练模型参数.训练结束后,对每次迭代后产生的中间参数求平均值,以缓解过拟合现象.该算法实现难度不高,并且运行速度和准确度也有较好表现,已经应用于当前如句法分析、词性标注的结构化预测等众多自然语言处理问题中.

整个篇章微观话题结构分析平台通过结构化感知器来训练模型.感知器训练的过程是一个在线学习(online learning)的过程,具体见算法 2.

**算法 2.** 结构化感知器学习算法.

输入:训练数据 $D=\{x_i,y_i\}_{i=1}^{n}$;最大迭代次数 $T$

输出:模型参数 $w$

步骤:

1. 设置模型参数 $w=0$.
2. for $t=1:T$ do
3. for $(x,y)$ in $D$ do
4. $z=decode(x,y,w)$
5. if $z!=y$ then
6. $w=w+f(x,y)-f(x,z)$

对于每个实例,感知器算法在每次迭代过程中,通过解码算法确定当前模型参数下最优解 $Z$:

$$Z=\arg\max_{y\in Y(x)} w\cdot f(x,y) \tag{1}$$

其中 $f(x,y)$ 表示实例 $X$ 在解 $y$ 下的特征向量.如果 $Z$ 不正确,模型层数按如下规则进行更新:

$$w=w+f(x,y)-f(x,z) \tag{2}$$

### 4.2.6 同类篇章话题结构模型比较及应用分析

从基本单元、联接词、关系表示结构等方面,将我们提出的基于 MTS 构建的汉语篇章 CDTC 语料库体系与 PDTB 中文标注体系以及汉语广义话题结构体系进行比较,结果表明 CDTC 体系吸收了 PDTB 体系和广义话题结构体系的优势,具有合适的篇章话题结构分析粒度,可以满足篇章话题结构分析的需求.具体结果如表 2 所示.

**表 2　同类汉语篇章话题结构体系比较**

| 项目 | 广义话题结构体系 | PDTB 体系 | CDTC 体系 |
|---|---|---|---|
| 基本单元 | 以标点句为基本分析单元 | 谓词-论元结构 | 含主谓结构的独立标点句，通常是小句. |
| 话题结构表示 | 广义话题和话题小句 | PropBank | 基于主述位理论的微观话题结构 |
| 话题联接表示 | 动态生成堆栈模型 | 基于词义的本体和指代关联 | 基于主位推进模式的微观话题联接 |
| 篇章结构表示 | 线性叠加模式 | 基于连接词和论元的结构树 | 采用自顶向下切分的微观话题链结构 |

当然如果面向应用，目前的语料库规模还有待进一步提高. 但是，如果出于利用语料库分析发现语言现象，总结语言规律这样的目标来看，目前的语料库规模能够达到这个要求. 初步标注的 500 篇文章，共 2342 个自然段落，6648 个自然句子，10 640 个子句构成基本单元.

语料库的研究，一般可以分为三个阶段：第一阶段是利用语料库分析发现语言现象，总结语言规律的过程；在此基础上，第二阶段扩大语料规模，在不同领域验证语言规律的过程；第三阶段，进一步扩大语料规模，为具体应用提供充分的语料资源. 从研究阶段来看，本文综述所讨论的语料库资源建设及其语言模型计算，尚处于第一阶段. 这一方面是遵循语料库研究的基本规律，另一方面也由于篇章话题结构的复杂性及研究难度，难以快速逾越，还需要持续深入一个时期的研究.

同时，从目前实际面向应用的典型语料库建设来看，在语料规模和覆盖领域两个方面都有不同建设特点. 例如语料标注规模并非很大的知名语料库就有修辞结构篇章树库、篇章图库等. 修辞结构篇章树库 RST-DT 共包含 385 篇文章，由美国南加州大学标注，于 2002 年经 Linguistic Data Consortium (LDC)正式发布，为修辞结构理论 RST 研究提供了研究资源. 篇章图库(Discourse Graph Bank，DGB)是根据 Wolf 和 Gibson 提出的图结构表示篇章的方法加以标注的语料库，共标注了 135 篇文章，用作篇章结构分析的语料资源.

相对而言语料规模比较庞大的典型语料库也有，如宾州篇章树库 PDTB，包括了华尔街日报的 2304 篇文章，于 2008 年正式发布，共标注四类篇章关系. OntoNotes 语料库包含广播和脱口秀节目、新闻、网络日志、电话用语等各种体裁的语料；根据来源，语料可以分为来自英语通讯社、中国通讯社、中国广播新闻、英语广播新闻等，累计包含 290 多万个词. 其中英语通讯社以华尔街日报为主，中国通讯社以新华社为主，中国广播新闻主要包括中国中央电视台、中央人民广播电台、中国电视系统等，英语广播新闻也是主流的如美国广播公司、CNN、NBC 的公共国际广播电台和美国之音等，因此能够确保语料来源的权威性.

上述不同规模和覆盖领域的语料库资源，事实上都在自然语言处理的不同研究领域、不同阶段发挥着不同程度的影响和作用. 因此，能否达到应用需求，是衡量所建设的语料库领域和规模大小是否合适的可行标准.

从后续应用来看，基于篇章话题结构的分析结果，在自动摘要、文本分类、信息抽取和机器翻译等领域都有广泛的应用价值. 比如在自动摘要中，通过话题结构的主述位推进，可以反映话题的变化规律，从而推断作者表达的意图及重点内容，为自动摘要提供素材. 又如在文章体裁分类中，不同体裁的文章所采用的篇章话题结构推进模式是不同的，其中蕴含着某种结构规律，这个可以为体裁分类提供新的特征. 又如在机器翻译领域，统计翻译方法可以考虑词对齐、短语对齐，子句对齐，那是否也可以基于主述位结构的对齐方法呢？基于主述位结构的对齐反映话题的变化规律，能够从篇章层面提供更为准确的语义对齐.

随着后续研究的推进，也将逐步开展第二阶段研究. 届时计划首先要扩展不同领域的标注语料；其次，逐步增加语料规模.

## 5　结　论

综上所述，相比于词法分析、句法分析研究，在自然语言处理领域中的篇章分析研究相对滞后. 特别是汉语篇章话题结构分析的研究处于起步阶段，尚未形成有效的理论体系，相应语料库资源建设薄弱，关键技术研究严重滞后.

众所周知，汉语与英语等西方语言相比有很大的不同，无论是篇章结构和意图表达方式，还是句法结构、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汉语篇章话题结构分析的理论体系，建立一定规模的适用于汉语篇章话题结构分析的资源库，并在此基础上建立汉语篇章话题结构分析的计算模型，实现高性能的汉语篇章话题结构分析. 这对面向汉语的自然语言处

理相关的诸多应用如自动文摘、信息抽取、机器翻译等都有重要的应用价值,并为推动自然语言处理进一步向自然语言理解发展做好基础准备.

**致 谢** 本文作者衷心感谢匿名评审专家富有建设性的评审意见!

## 参 考 文 献


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

[2] De Beaugrande R-A, Dressler W U. Introduction to Text Linguistics. London and New York: Longman Paperback, 1981

[3] Halliday M A K, Hasan R. Cohesion in English. Harlow England: Longman Pub Group, 1976

[4] Hobbs J R. Coherence and coreference. Cognitive Science, 1979, 3(1): 67-90

[5] Hobbs J R. Information, intention and structure in discourse //Proceedings of the NATO Workshop on Burning Issues in Discourse. California, USA, 1993: 1-26

[6] Mann W C, Thompson S A. Rhetorical structure theory: Toward a functional theory of text organization. Inss Journal, 1988, 8(3): 243-281

[7] Mann W C, Thompson S A. Relational propositions in discourse. Discourse Processing, 1986, 9(1): 57-90

[8] Grosz B J, Sidner C L. Attention, intentions, and the structure of discourse. Computational Linguistics, 1986, 12(3): 175-204

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

[10] Xu Fan. Research of Key Issues in English Discourse Structure Analysis[Ph. D. dissertation]. Soochow University, Suzhou, 2013(in Chinese)
(徐凡. 英文篇章结构分析关键问题研究[博士学位论文]. 苏州大学, 苏州, 2013)

[11] Sun Jing, Li Yan-Cui, Zhou Guo-Dong, Feng Wen-He. Research of Chinese implicit discourse relation recognition. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1): 111-117(in Chinese)
(孙静, 李艳翠, 周国栋, 冯文贺. 汉语隐式篇章关系识别. 北京大学学报(自然科学版), 2014, 50(1): 111-117)

[12] Li Yan-Cui. Research of Chinese Discourse Structure Representation and Resource Construction [Ph. D. dissertation]. Soochow University, Suzhou, 2015(in Chinese)
(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]. 苏州大学, 苏州, 2015)

[13] Marcu D. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts [Ph. D. dissertation]. Department of Computer Science, University of Toronto, 1997

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

[15] Van Dijk T A. Semantic macro-structures and knowledge frames in discourse comprehension//Just M A, Carpenter P A, eds. Cognitive Processes in Comprehension. Hillsdale, NJ, US: Lawrence Erlbaum Associates Press, 1977: 3-32

[16] Widdowson H G. Teaching Language as Communication. Oxford: Oxford University Press, 1978

[17] Brown G, Yule G. Discourse Analysis. Cambridge: Cambridge University Press, 1983

[18] Daneš F. Functional sentence perspective and the organization of text//Dane F, eds. Functional Sentence Perspective. Prague, Czech Republic: Academica, 1974: 106-128

[19] Fries P H. On the status of theme in English: Arguments from discourse. Forum Linguistica, 1981, 6(1): 1-38

[20] Lascarides A, Asher N. Segmented discourse representation theory: Dynamic semantics with discourse structure. Berlin, Germany: Springer Netherlands Press, 2008

[21] Zhang De-Lu, Liu Ru-Shan. The Development of the Theory of Text Coherence and Cohesion and Its Applications. Shanghai: Shanghai Foreign Language Education Press, 2003(in Chinese)
(张德禄, 刘汝山. 语篇连贯与衔接理论的发展及应用. 上海: 上海外语教育出版社, 2003)

[22] Zhong Mao-Sheng, Wang Xiao-Hu. Method of automatically identifying thematic progress mode in Chinese discourse. Application Research of Computers, 2015, 32(5): 1313-1315(in Chinese)
(钟茂生, 王小虎. 汉语篇章主位推进模式自动识别方法. 计算机应用研究, 2015, 32(5): 1313-1315)

[23] Austin J L. How to do Things with Words. New York: Oxford University Press, 1962

[24] Searle J R. Speech Acts: An Essay in the Philosophy of Language. Cambridge: Cambridge University Press, 1969

[25] Moser M, Moore J D. Toward a synthesis of two accounts of discourse structure. Computational Linguistics, 2010, 22(3): 409-419

[26] Scha R, Polanyi L. An augmented context free grammar// Proceedings of the International Conference on Computational Linguistics. Budapest, Hungary, 1988: 573-577

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

[28] Prasad R, Dinesh N, Lee A, et al. The Penn Discourse Ttreebank 2.0//Processings of the International Conference on Language Resources and Evaluation. Marrakech, Morocco, 2008: 2961-2968

[29] Poesio M, Artstein R. Anaphoric annotation in the ARRAU corpus//Processings of the International Conference on Language Resources and Evaluation. Marrakech, Morocco, 2008: 1170-1174

[30] Pradhan S, Moschitti A, Xue Nian-Wen, et al. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes//Proceedings of the Joint Conference on EMNLP and CoNLL-Shared Task. Jeju Island, Korea, 2012: 1-40

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

[32] Xi Xue-Feng, Zhou Guo-Dong. Pronoun resolution based on Deep Learning. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1): 100-110(in Chinese)
(奚雪峰, 周国栋. 基于 Deep Learning 的代词指代消解. 北京大学学报：自然科学版, 2014, 50(1): 100-110)

[33] Xu Fan, Zhu Qiao-Ming, Zhou Guo-Dong. Implicit discourse relation recognition based on tree kernel. Journal of Software, 2013, 24(5): 1022-1035(in Chinese)
(徐凡, 朱巧明, 周国栋. 基于树核的隐式篇章关系识别. 软件学报, 2013, 24(5): 1022-1035)

[34] Soricut R, Marcu D. Sentence level discourse parsing using syntactic and lexical information//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Edmonton, Canada, 2003: 149-156

[35] Li Yan-Cui, Zhu Kun-Hua, Zhou Guo-Dong. Summary of research on English discourse parsing. Application Research of Computer, 2012, 29(6): 2018-2023(in Chinese)
(李艳翠, 朱坤华, 周国栋. 英语语篇结构分析研究综述. 计算机应用研究, 2012, 29(6): 2018-2023)

[36] Hernault H, Bollegala D, Ishizuka M. A sequential model for discourse segmentation//Proceedings of the 11th International Conference of Computational Linguistics and Intelligent Text Processing. Tokyo, Japan, 2010: 315-326

[37] Feng V W. RST-Style Discourse Parsing and Its Applications in Discourse Analysis [Ph. D. dissertation]. Department of Computer Science, University of Toronto, 2015: 17-32

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

[39] DuVerle D A, Prendinger H. A novel discourse parser based on support vector machine classification//Proceedings of the International Joint Conference on ACL. Suntec, Singapore, 2009: 665-673

[40] Yang Xiao-Feng, Su Jian, Zhou Guo-Dong, Tan C L. Improving pronoun resolution by incorporating coreferential information of candidates//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain, 2004: 127-134

[41] Feng V W, Hirst G. A linear-time bottom-up discourse parser with constraints and post-editing//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA, 2014: 511-521

[42] Rutherford Attapol T, Xue Nian-Wen. Improving the inference of implicit discourse relations via classifying explicit discourse connectives//Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL. Denver, USA, 2015: 799-808

[43] Lin Zi-Heng, Ng Tou Hwee, Kan Min-Yen. A PDTB-styled end-to-end discourse parser. Natural Language Engineering, 2012, 20(2): 151-184

[44] Li Sheng, Kong Fang, Zhou Guo-Dong. A joint learning approach to explicit discourse parsing via structured perceptron//Proceedings of the Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Suzhou, China, 2014: 70-82

[45] Pitler E, Louis A, Nenkova A. Automatic sense predication for implicit discourse relations in text//Processings of the International Joint Conference on ACL. Singapore, 2009: 683-691

[46] Dinesh N, Lee A, Miltsakaki E, et al. Attribution and the (non-) alignment of the syntactic and discourse arguments of connectives//Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky. Michigan, USA, 2005: 3-6

[47] Pitler E, Raghupathy M, Mehta H, et al. Easily identifiable discourse relations//Proceedings of the International Conference on Computational Linguistics. Manchester, England, 2008: 87-90

[48] Lin Zi-Heng, Kan Min-Yen, Ng H T. Recognizing implicit discourse relations in the Penn Discourse Treebank//Proceedings of the SIGDAT Conference of Empirical Methods on Natural Language Processing. Singapore, 2009: 343-351

[49] Biran O, McKeown K. Aggregated word pair features for implicit discourse relation disambiguation//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 69-73

[50] Zhou Zhi-Min, Xu Yu, Niu Zheng-Yu, et al. Predicting discourse connectives for implicit discourse relation recognition//Proceedings of the International Conference on Computational Linguistics. Beijing, China, 2010: 1507-1514

[51] Hong Yu, Zhou Xiao-Pei, Che Ting-Ting, et al. Cross-argument inference for implicit discourse relation recognition//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Suzhou, China, 2012: 295-304

[52] Kong Fang, Ng H T, Zhou Guo-Dong. A constituent-based approach to argument labeling with joint inference in discourse parsing//Proceedings of the SIGDAT Conference of Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 68-77

[53] Wang Jian-Xiang, Man Lan. A refined end-to-end discourse parser//Proceedings of the 19th Conference on Computational Natural Language Learning- Shared Task. Beijing, China, 2015: 26-31

[54] Skadhauge P R, Hardt D. Syntactic identification of attributions in the RST Treebank//Proceedings of the International Workshop on Linguistically Interpreted Corpora. Jeju Island, Korea, 2005: 57-61

[55] Soon W M, Ng H T, Lim D C Y. A machine learning approach to coreference resolution of noun phrase. Computational Linguistics, 2001, 27(4): 521-544

[56] Ng V, Cardie C. Improving machine learning approaches to coreference resolution//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 104-111

[57] Yang Xiao-Feng, Zhou Guo-Dong, Su Jian, Tan C L. Coreference resolution using competition learning approach//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Sapporo Convention Center, Sapporo, 2003: 176-183

[58] Kong Fang, Zhou Guo-Dong, Zhu Qiao-Ming. Employing the centering theory in pronoun resolution from the semantic perspective//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore, 2009: 987-996

[59] Stoyanov V, Cardie C, Gilbert N, et al. Coreference resolution with reconcile//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, 2010: 156-161

[60] Yang Xiao-Feng, Su Jian, Tan C L. A twin-candidate model for learning-based anaphora resolution. Computational Linguistics, 2008, 34(3): 327-356

[61] Iida R, Poesio M. A cross-lingual ILP solution to zero anaphora resolution//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Porland, USA, 2011: 804-813

[62] Denis P, Baldridge J. A ranking approach to pronoun resolution //Proceedings of the International Joint Conference on Artificial Intelligence. Hyderabad, India. 2007: 1588-1593

[63] Rahman A, Ng V. Coreference resolution with world knowledge //Proceedings of the Annual Meeting of the Association for Computational Linguistics. Portland, USA, 2011: 814-824

[64] Ma Chao, Doppa J R, Orr J W, et al. Prune-and-Score: Learning for greedy coreference resolution//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Doha, Qatar, 2014: 2115-2126

[65] Dutta S, Weikum G. C3EL: A joint model for cross-document co-reference resolution and entity linking//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 846-856

[66] Zhang Wei, Zhou Chang-Le. Chinese yuan to the text comprehension refers to preliminary. Journal of Software, 2002, 13(4): 732-738(in Chinese)
(张威, 周昌乐. 汉语语篇理解中元指代消解初步. 软件学报, 2002, 13(4): 732-738)

[67] Wang Hou-Feng, Mei Zheng. Robustness of Chinese personal pronoun resolution. Journal of Software, 2005, 16(5): 700-707(in Chinese)
(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)

[68] Zhou Jun-Sheng, Huang Shu-Jian, Chen Jia-Jun. A kind of unsupervised Chinese refer to eliminate algorithm based on graph partition. Journal of Chinese Information Processing, 2007, 21(2): 77-82(in Chinese)
(周俊生, 黄书剑, 陈家骏. 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)

[69] Ngai G, Wang C S. A knowledge-based approach for unsupervised Chinese coreference resolution. Computational Linguistics and Chinese Language Processing, 2007, 12(4): 459-484

[70] Kong Fang, Zhou Guo-Dong. Based on tree kernel function study of pronouns in English and Chinese. Journal of Software, 2012, 34(5): 1085-1099(in Chinese)
(孔芳, 周国栋. 基于树核函数的中英文代词消解研究. 软件学报, 2012, 34(5): 1085-1099)

[71] Converse S. Pronominal Anaphora Resolution in Chinese [Ph. D. dissertation]. Department of Computer Information Science, University of Pennsylvania, 2006

[72] Johnson M. A simple pattern-matching algorithm for recovering empty nodes their antecedents//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 136-143

[73] Campbell R. Using linguistic principles to recover empty categories//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain, 2004: 645-652

[74] Gabbard R, Marcus M, Kulick S. Fully parsing the Penn Treebank//Proceedings of the Human Language Technology Conference of the North America. New York, USA, 2006: 184-191

[75] Yang Ya-Qin, Xue Nian-Wen. Chasing the ghost: Recovering empty categories in the Chinese treebank//Proceedings of the International Conference on Computational Linguistics. Beijing, China, 2010: 1382-1390

[76] Cai Shu, Chiang D, Goldberg Y. Language-independent parsing with empty elements//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers. Portland, USA, 2011: 212-216

[77] Kong Fang, Zhou Guo-Dong. A clause-level hybrid approach to Chinese empty element recovery//Proceedings of the International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 2113-2119

[78] Kong Fang, Zhou Guo-Dong. A tree kernel-based unified framework for Chinese zero anaphora resolution//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Massachusetts, USA, 2010: 882-891

[79] Song Yang, Wang Hou-Feng. Chinese zero anaphora resolution with Markov logic. Journal of Computer Research and Development, 2015, 52(9): 2114-2122(in Chinese)
(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)

[80] Zhou Guo-Dong, Li Pei-Feng. Improving syntactic parsing of Chinese with empty element recovery. Journal of Computer Science and Technology, 2013, 28(6): 1106-1116

[81] Borisova I, Redeker G. Same and elaboration relations in the discourse GraphBank//Proceedings of the Annual SIGDIAL Meeting on Discourse and Dialogue. Tokyo, Japan, 2010: 63-66

[82] Wellner B, Pustejovsky J, Havasi C, et al. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources//Proceedings of the 7th SIGDIAL Workshop on Discourse and Dialogue. Sydney, Australia, 2006: 117-125

[83] Liu Li-Jin. Comparative study between English and Chinese discourse structure mode. Guangzhou: Sun Yat-Sen University Press, 2011: 166-178(in Chinese)
(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)

[84] Chao Y R. A Grammar of Spoken Chinese. Berkeley and Los Angeles: University of California Press, 1968

[85] Cao Feng-Fu. Clause and Sentence Structure in Chinese: A Functional Perspective. Taipei, China: Student Book Co, 1990

[86] Qu Cheng-Xi. Chinese Discourse Grammar. Translated by Pan Wen-Guo, et al. Beijing: Beijing Language and Culture University Press, 1998(in Chinese)
(曲承熹. 汉语篇章语法. 潘文国等译. 北京：北京语言大学出版社，1998)

[87] Wang Jian-Guo. A Continuation of the Theory of Topic: Based on the Topic Chain of Chinese-English Discourse Research. Shanghai: Shanghai Jiao Tong University Press, 2013(in Chinese)
(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)

[88] Zhou Qiang, Zhou Xiao-Cong. Based on the topic of Chinese discourse coherence description system. Journal of Chinese Information Processing, 2014, 28(5): 102-110(in Chinese)
(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)：102-110)

[89] Xu Jiu-Jiu. Chapter in Modern Chinese Linguistics. Beijing: The Commercial Press, 2010(in Chinese)
(徐赳赳. 现代汉语篇章语言学. 北京：商务印书馆，2010)

[90] Jiang Yu-Ru, Song Rou. Based on the theory of generalized topic sentence recognition. Journal of Chinese Information Processing, 2012, 26(5): 114-119(in Chinese)
(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)：114-119)

[91] Song Rou. Chinese chapter generalized topic structure model of the water. Chinese Language, 2013, (6): 483-494(in Chinese)
(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)

[92] Shang Ying, Song Rou, Lu Da-Wei. General topic structure theory perspective self-sufficient in topic sentences and study. Journal of Chinese Information Processing, 2014, 28(6): 107-113(in Chinese)
(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)

[93] Le Ming. Chinese discourse rhetoric structure tagging research. Journal of Chinese Information Processing, 2008, 22(4): 19-24(in Chinese)
(乐明. 汉语篇章修辞结构的标注研究. 中文信息学报，2008，22(4)：19-24)

[94] Xue Nian-Wen. Annotating discourse connectives in the Chinese Treebank//Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky. Ann Arbor, USA, 2005: 84-91

[95] Zhou Yu-Ping, Xue Nian-Wen. PDTB-style discourse annotation of Chinese text//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Jeju, Republic of Korea, 2012: 69-77

[96] Zhang Mu-Yu, Li Yao-Bing, Qin Bing, Liu Ting. Based on the center word matching refers to dissolve. Journal of Chinese Information Processing, 2011, 25(3): 3-8(in Chinese)
(张牧宇，黎耀炳，秦兵，刘挺. 基于中心语匹配的共指消解. 中文信息学报，2011，25(3)：3-8)

[97] Li Yan-Cui, Feng Wen-He, Sun Jing, et al. Building Chinese discourse corpus with connective-driven dependency tree structure//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 2105-2114

[98] Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20(1): 37-46

**XI Xue-Feng**, Ph. D. candidate, associate professor. His main research interests include natural language understanding, machine learning.

**SUN Qing-Ying**, Ph. D. candidate. Her current research interests include natural language processing.

**ZHOU Guo-Dong**, professor, Ph. D. supervisor. His research interests include natural language understanding, Chinese computing, and information extraction.

**Background**

As a key technology to explore the discourse intension and play a fundamental role to discourse-level semantic analysis, analysis of discourse topic structure has become one of the core scientific problems in the area of discourse analysis. This paper, based on reviewing traditional theoretical exploration, corpus construction and computational modeling for discourse structure analysis, studies the frontier research and challenges about discourse topic structure analysis for discourse analysis.

This paper is supported by NSFC's Key Program named "Cross-lingual social opinion analysis: fundamental theories and key techniques". Cross-lingual social opinion analysis has been an important and hot research topic in recent years. However, there exist various kinds of problems in current studies, such as passivity, non-objectivity, and lack of deep understanding and analysis. This program attempts to establish a theoretical infrastructure suitable for cross-lingual social opinion analysis, develop a series of key techniques, construct relevant resources, and build a cross-lingual social opinion analysis platform, in addressing various scientific obstacles in front of its wide application.

As one part of the research of NSFC's Key Program, the main task of discourse analysis is to analyze the structure of a discourse and semantic relations between units from the overall level, and the use of context to carry out discourse-level event information extraction.

High level study is carried out on the discourse structure analysis, syntactic parsing and semantic computing and published in respectable international conferences and journals, such as ACL, EMNLP, COLING and IJCAI, *IEEE Transactions on Audio, Speech and Language Processing*, *Information Processing & Management*, *Journal of Computer Science and Technology*, *Journal of Software* (in Chinese), etc.