Survey of English and Chinese Discourse Parsing

1 任务定义

篇章分析是一项旨在研究自然语言文本的内在结构并理解文本单元之间关系的任务，其最终目的是从整体上理解篇章。篇章分析因其能挖掘出文本内部丰富的信息而对很对其他的自然语言处理任务起着重要作用，比如情感分析、文本摘要、问答系统等。篇章分析任务自底向上包含了三个子任务：以语言学为主的篇章理论研究、基于篇章理论的篇章解析器的自动构建、基于篇章分析技术与自然语言处理相关的上层应用。

以语言学为主的篇章理论研究主要解决篇章建模的问题。在英文方面代表性的篇章理论主要有基于实体关系的**修辞结构理论（Rhetorical Structure Theory, RST）、篇章词汇化树型连接语法（Discourse Lexicalized Tree Adjoining Grammar, D-LTAG）、中心理论（Centering）、脉络理论[]、篇章图库理论[]、和篇章表示理论[]**等。在中文方面，代表性的篇章理论主要有**复句理论[]和句群理论[]**。

基于篇章理论的篇章分析器的构建任务是通过自动化的方式构建篇章解析器，该解析器应具备对生文本的内部结构和单元间关系合理解析的能力。在英文方面，篇章解析器的代表性成果主要有基于PDTB和基于RST-DT风格的篇章解析器。在中文方面，目前的工作主要有模拟英文解析器风格的建立在中文语料库之上的解析器。

篇章分析技术的上层应用具体而言就是将篇章分析技术应用于像统计机器翻译、自动文摘、自动问答系统和信息抽取等任务，从而直接或间接地提升上层NLP系统的性能。

本文对主流的中英文篇章解析器的自动构建方法进行了分类、对比和综述。第2节阐述了中英文篇章理论及篇章语料库标注体系；第3节介绍了主流的应用于篇章解析器自动构建的可计算模型；第4节阐述了篇章分析的重要评测方法并列举了相关比赛；第5节总结了当前中英文篇章分析领域所达到的最高水平。

2 篇章理论及篇章语料库标注体系

首先，本节介绍了主流的中英文篇章分析理论修辞结构理论和宾州篇章树库理论等。此外，本节还对英文篇章语料库RST-DT和PDTB的篇章标注体系和中文篇章语料库CDTB的标注体系做了相关阐述。

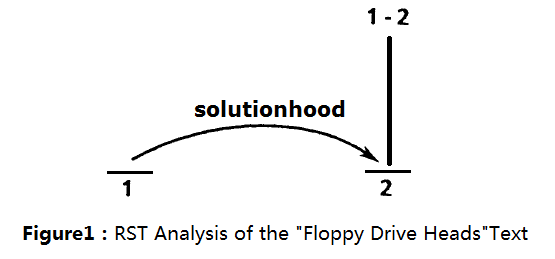
2.1 英文篇章理论

（1）修辞结构理论

修辞结构理论（RST）是由文献[Mann and Thompson, 1987]等提出的有关篇章分析得在RST中，每个句子由若干独立的子句组成，这些字句作为最基本的篇章单位（Elemental Discourse Units，简称EDUs）。由若干连续子句构成的文本跨度（Span）是篇章中功能明显的组成部分。RST采用“核心—卫星”的表示结构来确定一个文本跨度中各个部分之间的主次关系。核心是篇章的最重要的部分，表示中心信息单元，具有相对完整的语义。卫星是传达支撑信息的其他单元，用于补充说明核心部分，脱离核心部分的卫星通常在篇章中是没有实际意义的。

多个子句或者文本跨度之间都由特定的修辞关系进行联结，整个关系集合是一个开放性集合，它会随着研究的。判定修辞关系需要四个因素：对核心的限制条件、对卫星的限制条件、对核心卫星的联合限制条件以及效果。最基本的修辞关系有两种：具有不对称的单核关系和无主次之分的“多核心关系”，其中，单核关系占据了主要部分。

修辞结构理论认为，连贯的篇章由不同层次的修辞关系组成，并且可以表示为一种树形结构，用叶节点代表EDUs，内部节点代表Span。图1展示RST-DT中典型的单核关系，其中垂直线指向核心，弧线由卫星单元1指向核心单元2。可以把图中的solutionhood关系理解为，核心单元的文本跨度是对由卫星文本跨度中的问题的解答。



（2）宾州篇章树库理论

宾州篇章树库理论（Penn Discourse Treebank）是宾州大学的研究人员采用的一种以词汇为中心的方法[]，在句子级的Penn Treebank的基础上，以篇章关联词为核心，从语义角度出发构建了篇章关系库[]。该研究检测同一篇章内两个文本单元之间的逻辑语义关联（因果关系、转折关系等），因为文本单元的范围大于句子范围，可以把这种语义关联的检测理解为对句子内部语义关联的扩展。

根据文本单元间是否存在篇章连接词，将篇章句间的关系分为包含连接词的显式篇章关系（Explicit Dsicourse Relation）[]与不包含连接词的隐式篇章关系（Implicit Discourse Relation）[]。隐式关系因缺少连接词而很难推测出两个文本单元之间的语义关系，需要结合上下文进行推测，难道较大。

（3）篇章图树库理论

篇章图树库（Discourse GraphBank）最初由Wolf et al[1]提出。该理论认为，相比于树结构，篇章更适合表示为图。他们提出，图的表示方式允许将文章中不同内容以更自由的形式表达出来，从而能够从中获取更丰富的信息。同时，他们构建了一个由135篇文档构成的篇章树资源。关于Discourse GraphTreebank和RST-DT的区别，文献[]做了更详细的介绍。

2.2 中文篇章理论

文献[]对中文篇章研究进行了较深入的综述，他们认为中文篇章理论更偏重于汉语语法方面的研究，可操作性不强。其中句群理论和复句理论创立的出发点也不是篇章理论，而文献[]和[]认为句群理论和复句理论可以经适当修改而与RST相似并可作为中文篇章分析理论。关于句群和复句理论的相关内容，文献[]中做了比较详细的综述。

RST在汉语中的跨语言可转移性有特殊背景，当前已经有不少对RST的中文介绍和初步应用计划等，但实质性的发展很少。目前，山西大学李茹教授的团队正在尝试进行中文RST数库的构建工作，产出的资源也有一定的规模，但是尚未公开发表。

目前采用PDTB标准构建的篇章预料主要面向英语[]，除此以外在印度语[]、土耳其语[]和阿拉伯语也有相应的研究和资源出现。在中文上，布兰迪大学的Xue教授最早尝试了中文关联词标注于分析工作[]，并尝试按照PDTB体系标准标注中文树库。国内北大语料库、哈工大语料库？

综上所述，中文篇章理论主要还是在模仿英文篇章理论的阶段，旨在产出可操作性较强的适用于中文的篇章分析理论。

2.3 语料库标注体系

RSTDT和PDTB两种语料均选自PTB（Penn Treebank）[]语料，根据各自定义的规则及目的进行标注。PTB语料内容来自美国《华尔街日报》（Wall Street Journal, WSJ）的新闻报道，而PDTB相较于RSTDT来说，规模更大。CDTB?...

（1）RST-DT标注体系

RST-DT是基于修辞结构理论的产物，主要对WSJ中的385篇文章进行标注，其中53篇（13.8%）倍重复标记。该语料库包含了两个部分的内容：切割篇章文本形成的子句群和篇章修辞结构树。联结各个文本单元的篇章修辞关系共有78种（53种单核与25种多核）细粒度篇章修辞关系。根据修辞关系的相近程度将其划分为18种粗粒度关系类。由这18种关系类结合篇章结构关系（NS、SN、NN）共得到41种不同的状态，如表1所示。

关于修辞结构理论中篇章关系的理解可参考文章[]。

（2）PDTB标注体系

（3）CDTB标注体系

3 可计算模型

4 重要评测和比赛

5 state-of-the-art

参考文献