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[]、基于图的篇章模型[]、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

2.1 篇章理论

（1）修辞结构理论RST

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。同时，每对子句或者文本跨度之间都由特定的篇章关系进行连接，整个篇章关系集合是由人工标注的结果确定的，因而它是一个可扩充的开放性集合。RST的结构属于树状图，它用叶节点代表EDUs，内部节点作为Span,

2.2 语料库标注体系

3 可计算模型

4 重要评测和比赛

5 state-of-the-art

参考文献