编号：

****

毕业设计(论文)外文翻译

（译文）

学 院：计算机与信息安全学院

专 业： 软件工程

学生姓名： 薛天杭

学 号： 1400330129

指导教师单位： 计算机与信息安全学院

姓 名： 黄桂敏

职 称： 教授

2020年6月20日

RST语篇分析及其在语篇分析中的应用

# 摘 要

话语解析是确定文本中各话语单位之间的关联性和特定话语关系的任务。特别是，在各种话语分析的理论框架中，我对修辞结构理论感兴趣。我假设，鉴于其最终的成功，话语解析可以提供一般的解决方案，用于许多下游应用程序。本文由两个主要部分组成。首先，我概述了我关于话语化与话语树建设的工作，这是RST式话语解析的两个主要组成部分。在话语树库（RST-DT）上进行评估，我的话语分段人和树构建者都实现了最先进的表演。后来，我讨论了话语关系在话语分析中某些具体任务中的应用，包括评价一致性、确定作者身份和消除欺骗。特别是，我建议使用一组应用程序中性功能，这些功能源自我的话语解析器提取的话语关系，并将the 这些应用程序中性功能的性能与特定于应用程序的经典与每个任务进行比较。在前两项任务中，实验结果表明，话语关系特征本身往往与经典应用特定的fea-tures- 一样表现，这两种特征的组合通常会产生进一步的改进。这些结果为我的假设提供了有力的证据，即话语解析能够 pro-

致 谢

我衷心感谢多伦多大学科普科学系的主管格雷姆赫斯特教授。自从五年前我作为硕士生在多伦多 Toronto开始生活，然后攻读博士学位以来，我非常高兴能和他一起工作。年的学生。格雷姆是个很有幽默感的绅士。在我的整个研究中，他始终为我提供有见地的想法、建议和灵感。他是一位真正的导师，他总是尊敬和善待学生。此外，如果没有他耐心和仔细编辑我所有的研究论文，我would 仍将是一个科学写作新手，为我需要给出的每一个演讲而奋斗。我要感谢我的委员会成员苏珊娜史蒂文森教授和杰拉尔德佩恩教授，在我转过重新搜索的方向时，他们提供了有益的建议和批评。虽然我通常只与我的主管一起研究我的研究项目，但定期与委员会成员的检查站会议为从非诚角度学习有趣和有用的想法提供了机会。我也感谢来自德国的迈克尔斯特鲁贝教授，他非常愿意同意担任我的外部前矿工感谢集团的弗兰克鲁齐茨教授和语言学系的杰克钱伯斯教授，他同意成为我最后一次辩护的新委员会成员。如果没有他们宝贵的建议和有见地的批评，我最后一次的评书工作就会have 少得多。我也感谢我的父母孙华英和冯汉杰，他们留在我的家乡上海，中国，当我在国外攻读博士学位。没有他们的支持，我会在多伦多的日子更艰难。我要感谢多伦多大学小组的所有同事，他们是如此惊人的才能，以及我在多伦多的所有朋友，他们花时间与我共度时光，阻止我成为一个沉闷的博士。最后，我要感谢加拿大自然科学和工程研究理事会和多伦多大学对我的研究提供财政支持。

# 1 导言

写入良好的文本的单元没有完全隔离解释需要了解单位和上下文之间的关系。大多数修辞理论都假定话语的层次结构，其中几个小单元的文本相互关联，形成一个更大的单元，然后可以与其他单元相关。从这个角度看，为给定文本构建分层话语结构类似于句法解析，其目的是建立给定句子相对于其文本单元之间的语法关系的分层结构。因此，在文本中发现层次性话语关系被称为话语解析。我在本论中的终极假设是，话语解析可以成功地以高准确性自动完成，而且，鉴于它的成功，话语解析将能够为话语结构分析中的各种问题提供一般的解决方案。在本论中，为了评价我的假设，我将首先介绍我们关于去解构一个自动话语解析器的工作，并将其表现与人类判断进行比较。此外，基于我们的解析器，我将对话语分析的三个特定应用应用话语解析，并观察从话语中衍生的特征如何解析这些应用。话语解析的一般性是双重的：首先，它可以处理从句子到段落，最后整个文档的粒度。其次，话语解析的目的不仅在于两个给定文本单元的相关性，例如它们是否属于同一个子主题，而且还发现它们之间的确切一致性关系，例如，对比度、因果和解释，它们可以，但通常不必have 依赖于任何特定的目标应用。因此，话语解析能够提供有关文本内容和话语结构的丰富信息，这显然是许多话语分析应用的有力工具。在本章中，我将首先介绍修辞结构理论，这是最广泛接受的话语分析框架之一。此外，我还将简要介绍宾夕法尼亚话语树银行，一个根据另一个流行的话语框架工作而开发的语库，其相关工作，以阐明其他理论和哲学对话语分析的讨论。该则按如下方式组织。在第[一部分中，](#_bookmark17)我将讨论RST式话语解析中的两大任务，即话语分割和话语树建设，以及围绕这两项任务开展的相关工作。到第章结束时，式话语解析器[3,](#_bookmark39)的所有必要组件都将呈现。在本论第[二](#_bookmark87)部分中，我们将在话语分析中看到几个具体的应用，我将在此评估我关于话语解析一般用处的假设。这些申请包括评价一致性（第[6章）、](#_bookmark88)确定作者身份（第[7章）](#_bookmark129)和欺骗行为（第章[）。](#_bookmark169)我将首先介绍针对每个问题的应用程序特定方法，这些方法是每个具体问题的成熟和经典解决方案。之后，我将讨论如何将来自我们应用中立的话语解析器的信息纳入每个问题，并增强整体的形态，从而提供证据来支持话语解析的假定一般性。

## 1.1 修辞结构理论

修辞结构理论（）（ [(Mann和Thompson，1988年）](#_bookmark262)是话语分析最广泛交流的框架 [Thompson,](#_bookmark262)之一，在（[年）](#_bookmark263)的话语解析[Marcu](#_bookmark263) 开创性工作中被采用。在的框架下，一个连贯的文本或相当独立的文本片段可以表示为话语树。在RST 风格的话语树中，叶节点是非重叠文本范围，称为*基本话语单元*（）这些是话语树的最小文本单元（参见第 [1.1.1](#_bookmark3)节），而内部节点是连续的串联国家。相邻节点通过特定的话语关系（详情请参阅第[1.1.2](#_bookmark5)节）形成话语子树，然后与other 树结构中的其他相邻节点相关。通过这种方式，建立了分层树结构。正如[Taboada和Mann（2006年）](#_bookmark289)在其最初提案中详细讨论的那样，被设计为一个开放的系统，为从事二等领域和应用的研究人员提供了灵活性。在的最初设计中，只有少数固定部分强制执行，包括将文本划分为一组非重叠的话语单元，以及话语关系和文本一致性之间的紧密性。因此，为了继续介绍该理论的细明细节，下面，我将连接到特定的注释方案及其生成的语料库，话语树库（RST-DT），并侧重于本语集中注释指南提供的相应定义。话语树库 （RST-DT） [（Carlson等人，2001 年），](#_bookmark211)是一个在RST 框架中批文，由语言数据联盟（）发布，目录号和[1](#_bookmark0)。它由《华尔街日报》的份文件（份用于培训，份用于测试）组成*Wall Street Journal*。RST-DT被广泛用作RST式话语解析研究的标准基准，因为它为定义曼恩和汤普森在最初开发中基于直觉的概念提供了系统指南and ，包括的定义和若干话语关系。在本 this 篇全则中，术语*RST式话语解析*将引用特定类型的话语解析根据中的注释框架。

## 1.2 基本话语单元

正如曼恩和汤普森（年，第页）所述，提供了一种描述 the 文本中条款之间的关系的一般方法，无论这些条款是语法的还是用词法表示的。因此，基本话语单元（）是最小的话语单元，不一定是句法句，也没有明确的词法线索来表示边界。*can*在RST-DT 中，为了在注释的一致性和粒度之间提供平衡，开发人员选择子句作为的一般基础，并采用以下一组异常。因此，仅仅依靠句法信息并不是分割的，需要采取更复杂的。在第章[中，](#_bookmark18)我将介绍我关于开发用于确定边界的话语分割模型的工作。

1.3 话语关系清单

根据RST，有两种类型的话语关系，假设（单核）和辅助战术（多核）。在单核关系中，其中一个文本跨度，*即原子核*，is 比卫星更突出*satellite*，而在多核关系中，所有文本跨度对于解释都同样重要。在RST-DT中，[曼恩和汤普森（1988年）](#_bookmark262)定义的原有的种话语关系进一步分为个细粒度修辞关系（个单核和个多核），提供了高度的表达性。个关系可以分为个关系类，如表[1.1](#_bookmark7)所示。例如，类原因是关系类型*的原因*、*结果*和*结果*的粗粒度聚类。此外，三个关系用于在树上施加结构：文本组织、跨值和同一单位（用于链接由嵌入单位或跨值分隔的单位部分）。附于核性后，有种不同类型的话语关系类，如表[1.2所示。](#_bookmark9)例如，对比度关系可以是三种不同类型的对比度：对比度（两个范围都是原子核）、对比度（第一个跨度是原子，the 另一个是卫星），另一种是卫星，另一种是对比度这种不同类型的关系类是当前对RST式话语关系 is 进行分类的细粒度层次。每个特定关系的定义基于四个要素：（）核约束（）核心约束（）核心关系。（）卫星的限制（）核与卫星组合的限制和（）在文本接收器上实现的等。例如，表 [1.3](#_bookmark10)说明了条件类的定义，相对于上述四个定义元素[2](#_bookmark0)。

### 1.1.1 RST 风格话语树表示示例

图[1.1](#_bookmark8)所示的示例文本片段由四个*（e*1*e*4）组成，按方括号段分。其话语树表示图如下所示。

RST的表示法约定。两个*e*1和*e*2由单核关系归因相关，其中*e*1是更突出的跨度，由指向*e*1的箭头表示。跨度（e*e*1-*e*2）和*e*3由多核关系相关，其中它们同样突出，由连接（e*e*1-*e*2）和*e*3的两条直线表示。最后，跨度（e*e*1-*e*3）与*e*4相关的单核关系条件，形成句子的完整偏离树。这样，我们有have 一个树结构的分层表示对应于整个句子。请注意，风格的话语树表示范围没有约束，因为树结构表示可用于描述文本的表达结构：从句子到段落，最后到整个文本。由于具有如此的能力来表示关于等粒度水平的话语关系，对话语分析领域的许多研究人员特别感兴趣。更im- 令人不解的，它很好地符合本章开头概述的目标，即为分析话语结构中的各种问题提供一般解决方案。正如我们在后面的章节中看到的，在文本中确定RST式话语关系确实有利于一些话语分析问题。

### 1.1.2 RST 风格话语解析管道

由于话语关系的树结构表示的性质，RST 风格的不课程解析通常采用由两个单独阶段组成的管道框架：

话语分割：将原始文本分割为非重叠的，这是文本级话语树表示的底层话语单位。

话语树构建：鉴于第阶段的分段集，采用策略来构建与全文对应的话语树，例如图[1.1](#_bookmark8)所示的示例话语树。在第一部分[，](#_bookmark17)第[2](#_bookmark18)章和第[3](#_bookmark39)章将详细讨论相关工作和我自己在这两个阶段的工作。

### 1.1.3 RST 和RST-DT 的问题

在近三十年的历史中，在各种话语理论中获得了无与伦比的人气，并应用于各种应用，不仅适用于文本生成其原始动机和设计目的也用于大量的文本理解任务。并非巧合的是，也有许多文献专门质疑或批评RST的几个方面。然而，正如前面提到的，根据[塔波阿达和曼恩（2006年），](#_bookmark289)这些批评大多源于对原始设计的误解或偏离。相反，应被视为具有高度灵活性的开放系统，并鼓励针对特定应用和领域进行创新和适应。事实上，在应用RST 风格的话语分析the 时，仅执行以下一般规则：

文本分析是通过应用遵循完成性约束的架构（一个架构应用程序包含整个文本）来执行的;连接性（每个范围（每个范围（除了包含整个文本的跨度外，是最小单位或另一个架构应用程序的组成）;唯一性（每个架构应用程序包含一组文本跨度）;和邻接（每个架构应用程序的跨度构成一个连续的跨距文本）。然而，就目前RST式话语分析的计算方法而言，特别是由于使用RST-DT作为基准数据集，确实有几种普遍接受的公式，实际上值得怀疑。在这里，我简要地谈到了一些最突出的问题，有关RST-DT和一般。

首先，基于条款的分段规则被批评为过于粗糙，无法捕捉一些语言现象。例如，如RST-DT所指定的，作为主动词主题或对象的子句不被视为（参见第[1.1.1](#_bookmark3)节）因此，以下句子被视为单个。他努力学习使他通过了考试。

然而，这种分割并不擅长细粒度，因为它排除了两个动作之间难以*学习和通过考试*之间潜在的因果关系的任何表现。此外，还有人担心，通过树形话语结构表示文本是否可行，以及这种树形表示形式是否是给定文本的唯一有效表示形式。诚然，这可能是一个过于强烈的假设，即一棵树能够捕捉整个文本中的话语结构：对于一个普通作家所写的文本，偶尔从主要主题偏离或思想的逐渐发展是正常的，这样小文本片段中有一定程度的一致性，而二等片段之间的关系则相当松散。因此，为了处理真实文本中的这些复杂问题，[和 Gibson （2005）](#_bookmark291) 建议使用基于图形的替代数据结构进行分析，该结构允许具有多个父项的跨依赖项和节点。然而，尽管它们具有更大的表达性，但基于图形的表示也强制自动话语解析面临更大的挑战。最后，RST中的邻接约束，即每个话语关系的跨度构成一个连续文本跨度，也不是完全合理的，其微妙之处在于嵌入的话语单元的存在。根据RST-DT中的定义，嵌入式话语单元具有以下一个或两个属性：（）它单独分解了一个合法为的单位（）它单独分解了一个单元，即（）它只修改的一部分，而不是整个。例如，图[1.2](#_bookmark13) 显示了包含三个的文本片段，其中第二个是嵌入的。嵌入式*e*2打破了*e*1和*e3，*3, 当串联时，这是一个合法的本身。因此，为了描述*e*1和*e*3之间的一致性，这基本上是一种延续，的开发者RST-DT 必须发明一种伪关系，称为但维持其战略的关键组成部分*e*稳定的汇率和高水平的进口*[]e*12将消耗大量外汇。*e*3同一单位。但是，这样，嵌入*e*2的存在就违反了邻接约束。

## 1.4 宾夕法尼亚话语树银行和PDTB风格的话语解析

宾夕法尼亚话语树银行（ [(Prasad等人 al.,，2008年）](#_bookmark278)是另一个带注的论述其文本是RST-DT（2159篇《华尔街日报》文章）的超集。与RST-DT不同，不遵循RST的框架 相反，它遵循话语词化树毗邻语法（D-LTAG）（[韦伯，2004年），](#_bookmark290)这是一个词汇基础，谓词论证方法与一套预定义的话语关系。在此框架中，话语连接（例如，*因为*）被视为以两个文本跨度作为其参数的谓词。话语在结构上附着的论点称为**Arg2，**另一个论点称为**Arg1;**与RST不同，这两种论点不是因其突出性而被解释所迷惑的。示例[1.1](#_bookmark15)中显示了来自的示例注释，其中显式共和（*当*）下划线，两个参数和**Arg2**分别以*斜体*和**粗体**显示。该示例使用其三级层次结构关系类型进行批带：它是 is 应急类、原因类型和原因子类型。

示例 1.1.当第三天的测试显示一些阳性结果时，分散剂的使用被批准

在这中，关系类型是分层组织的：有*个类*：扩展、帕里斯、原因和暂时性，可进一步分为*种*和*个子类型*。在发布后，已尝试多次识别风格的重新化。皮特勒等人进行的语料库研究[。（2008）](#_bookmark276)表明，整体话语构成词组大多明确，允许对话语关系进行高精度分类：通过简单地将每个连接体映射到其最频繁的感官，它们实现了以上的准确率。因此，话语解析的真正挑战在于隐性关系（没有由显式连接者指示的话语关系），最近的研究重点是on 承认这些隐含的话语关系。特别是[林等人。（2009）](#_bookmark257)尝试通过使用四类要素（上下文要素、组成解析特征、依赖项解析特征和词法特征）来识别中的这种隐式话语关系，并探讨了它们对性能的个体影响。它们表明，从组成解析树中提取的生产规则是最具有特征的，而上下文特征是最弱的。随后，他们全面实施了端到端风格的话语解析器[（Lin 等人，2014 年）。](#_bookmark259)[皮特勒等人（2009）](#_bookmark277)采用了一套类似的语言动机特征，并进行了一系列与他人的分类，以识别各种类型的隐含话语关系。后来，根据[皮特勒等人](#_bookmark276)的见解。[（2008）](#_bookmark276) 以上所述，[周等人。（2010）](#_bookmark295)通过首先预测适当的话语连接，然后将预测的连接与最常见的话语感映射，从而帮助解决识别隐式关系的问题。具体来说，周等人训练了一种语言模型来评估一组合成文本的困惑性，这些文本通过将每一种可能的话语连接插入the 兴趣的隐含话语关系中而形成。最可能的连接是从最令人费解的合成文本中选择的。然而，这种方法并没有取得多大成功。主要原因是，合成文本以这种方式由插入的连接体形成因此，计算困惑性将考虑连接附近非常有限的上下文单词（通常在图中使用三图序列）。事实上，正确解释特定的隐式关系通常需要更大比例的文本。最近一项认识隐性话语关系的研究重点是特征再细腻。[Park 和 Cardie （2012）](#_bookmark273)在[Pitler等人](#_bookmark277)以前使用的要素集上应用了一个简单的贪婪特征选择[。（2009）](#_bookmark277)提高隐式关系识别的性能。最近，[卢瑟福和薛（2014年）](#_bookmark282)认为，单词对，这是识别隐含关系最的生态特征，当可用的训练样本有限时，从稀疏问题。因此，他们建议通过用布朗词簇对[3](#_bookmark0)和共引用模式表示关系来克服这种稀疏问题。[Ruther-ford和 Xue](#_bookmark282)在中实现了识别级隐式关系的当前最先进的一对一分类性能，从*F*1得分（时间对其他人）到（扩展对他人）。

## 1.5 两种话语框架的差异

作为话语解析研究中最流行的两个框架，和有几个固有的区别，这使得这两个框架可能对各种应用有用。在第二部分[中，](#_bookmark87)我们将看到话语分析的几个具体应用，以及两个框架在应用程序上生成的分析的。两个框架之间最重要的矛盾是，在风格的解析中，文本最终表现为一个话语树，因此话语结构完全用文本的二等粒度来修饰然而，在中，不一定存在涵盖全文的树结构，即式话语关系只存在于文本中。棕色词类聚类是一种基于前字类的层次类词聚类，由[Brown等人（1992年）提出。](#_bookmark210)非常本地的上下文窗口。如[第6.3节所示，](#_bookmark111)话语结构的完整层次对于某些特定应用非常有用。此外，由于在 RST 风格的解析中，文本首先被分割为非重叠的，这是最终话语树表示中最小的单位，因此文本中的任何给定有效话语单元都参与至少一个话语关系。换句话说，RST 式解析中的话语关系涵盖了整个文本。但是，在风格的话语解析中，这通常不是事实。因此，RST式话语关系对文本的覆盖优于式的话语关系。 这种覆盖范围更好的属性对于某些特定应用程序也很有用。最后，一般来说，RST式话语关系比式关系更为受限：RST风格的关系只能存在于相邻的文本跨度（单个或多个连续的串联）之间，文本中的两个RST 风格的话语关系只能是两种情况之一：对应于两个关系的文本彼此完全不协调，或一个关系的文本范围是另一个关系文本范围的适当子序列，即两个文本范围不能部分重叠。但是，在风格的话语关系中找不到这种约束，因此在样式关系的注释中具有更大的灵活性。上文讨论的二等内容不一定导致明确声明，即一个话语框架优于另一个话语框架相反，它们说明了 the 两个框架的基本哲学之间的矛盾，因此，我们应该根据我们感兴趣的特定应用程序选择更合适的。例如，由于RST式话语重传中存在分层结构和完整覆盖where ，RST式话语解析可能更适合需要全局理解文本的应用，例如本文后面部分要讨论的应用程序later 。相反，由于式的话语解析是有词法的，在 have相当局部的语境窗口中代表话语关系，因此，对于我们希望查明相关信息的应用程序来说，它更具有相关性，在文本的其余部分中可能没有什么内容。此类应用程序的示例包括信息检索和问答。

# 2 话语分割

如[第 1.1](#_bookmark2)节所述，对于RST 风格的话语解析，识别of 话语单元的边界是管道工作流的第一阶段因此，其性能对整体准确性至关重要。在本章中，我将首先介绍一些以前关于RST风格的话语分割的工作，然后讨论我自己的基于的话语分段。

## 2.1 以前的工作

通常，自动分段的任务被表述为：给定一个句子，分段模型通过预测a 是否应在句子中的每个特定标记之前插入边界来标识复合的边界。特别是，以前关于话语分割的工作通常分为两个主要框架。第一种是按顺序和独立地考虑句子中的每个标记。在此框架中，分段模型按令牌扫描句子令牌，并使用二进制（如支持向量机或逻辑回归）来预测在检查令牌之前插入边界是否合适。这一框架之后的例子包括[索里库特和马库（2003年）、](#_bookmark283)[苏巴和迪尤金尼奥（2007年）、](#_bookmark286)[费舍尔和罗克（2007年）](#_bookmark229)和[乔蒂等人。（2012）。](#_bookmark243)给定句子被视为一个整体，模型为每个标记分配一个标签，指示此令牌是否是的开头。通常，类标签*B* 分配给作为开头的令牌，并将标签*C beginning* 分配给其他令牌。由于句子的开头是的开始，因此在此标记过程中排除了句子中的第一个标记。例如，图[2.1](#_bookmark20) 说明了此顺序标记过程。示例句由个标记组成，由空格分隔，最后个标记在顺序标记过程中被考虑。每个标记由标签模型分配一个标签或*C。* 如果标记标记为*B*，例如，*以*粗体显示的标记和*标记*，则边界将放置在它之前。因此，该句子分为三个，由方括号对指示。在此顺序标签框架之后的一项具有代表性的工作是[Hernault 等人。（2010a），](#_bookmark233)其中the使用条件随机字段（）实现顺序标记。上述两个主要框架的一个有趣的例外是[Bach](#_bookmark203)s [et等人（2012 年）](#_bookmark203)重新排名模型，该模型获得了迄今为止报告的最佳分段性能：对于*B* 类，*F*1得分为，*B* 类和*C*类的宏观平均值为。其想法是训练一个排名函数，其输入是基段的*N-*最佳输出，并输出这些*N* 候选项的重新排序。在他们的作品中，巴赫等人使用一个类似的基于的分段器，以等人al.’s 为基础分段。由于重新排序过程几乎与基段器的实现是正交的，因此有必要探索基段器的增强，以进一步提高每个形式。关于基本分段（通常采用之前介绍的两个框架[），Fisher 和 Roark （2007）](#_bookmark229) 报告了最佳性能*，F*1得分为，用于识别句内边界（类），使用三个单独的功能集：基本有限状态功能、完全有限状态功能和无上下文功能。本部分开头介绍的现有基本分段模型具有某些限制。首先，采用的功能集都围绕单个令牌（如令牌的语音部分）或语法树中最高节点的生产规则（特定令牌是词义头）为中心。尽管上下文信息可以通过或部分语音*n-g*等功能进行部分捕获，但这些上下文要素的表示能力可能有限。*n*相反，我们假设，与其利用以单个令牌为中心的功能，不如平等地考虑来自相邻令牌对的信息，因为分段模型的基本输入单位是一对令牌，其中每个令牌由其自己的一组要素表示。此外，现有模型从不重新考虑其以前的分段决策，即话语边界是通过只运行一次分段算法获得的。但是，由于单个决策彼此相互关联，通过执行第二次分段，其中集成了编码分段全局特征的功能，我们也许能够纠正初始运行的一些不正确的分段。因此，在这项工作中，我们建议通过我们的配对功能和两路分段过程来克服这两个限制，该过程将在[第 2.2](#_bookmark21)节中介绍。

## 2.2 方法

图[2.2](#_bookmark22) 以线性链条件随机字段的形式显示了我们的分段模型。每个句子由单个线性链表示。对于句子中的每个相邻标记，即*Ti*=1和*Ti，*有一个关联的二进制节点*Li*来确定对的标签，即两者之间存在边界：如果*LiB，*则插入边界

*在 Ti*之前如果*LiC，*则两个相邻的令牌被视为中的连续部分。We effective [Hernault et al.（2010a）](#_bookmark233) 和[巴赫等人。（2012）。](#_bookmark203)此顺序标记框架也有利于培训过程，也就是说，无需进行额外的来处理数据中边界的稀疏性，这通常是传统二进制分类器的问题。如前所述，我们的分段模型从以往关于RST式话语分割的工作中，从两个重要方面进行了细分。首先，我们的约束元节点使用以单个令牌为中心的要素表示形式（可能带有一些专门设计的要素以部分合并上下文信息），而是从一对相邻令牌中获取输入，以完全合并上下文输入，从而允许相邻令牌之间的竞争。其次，我们采用双通分割算法，采用如下方法，而不是通过一次模型来生成边界的预测。 我们首先 apply 为每个句子应用一次分割模型。然后，我们通过考虑从初始分段派生的一些**全局特征**（将在[第 2.3](#_bookmark23)节中描述）来执行第二次。这些新颖的全局特征背后的直觉是，给定 given令牌是否应标记为边界有时取决于相邻的边界。例如，正如[Joty等人](#_bookmark243)所建议的那样。[（2012），](#_bookmark243)由于通常是多字表达式，因此当前令牌与相邻边界之间的距离可能是一个有用的指示。此外，了解令牌和相邻边界之间的令牌是否构成有效的语法组成也很有帮助。由于这些全局指标只有在我们对have 边界进行初始猜测时才可用，因此有必要进行第二次of 分割。

## 2.3 特征

如图[2.2](#_bookmark22)所示，线性链中的每个边界节点*Bi*在句子中采用一对相邻的标记*Ti*和*Ti*=1的输入。每个这样的对都使用曲面词法和句法特征列表进行编码，如下所示。这些要素分为三个子集：基本要素、全局要素和上下文要素，其中基本和上下文要素适用于第一次和第二次传递，全局要素仅适用于第二次传递。与其他模型的比较

我们首先研究我们基于每种形式的配对功能和现有分段模型提出的双通话语分段。在本实验中，我们在话语树库（RST-DT）（Carlson[等人，2001年）](#_bookmark211)上训练我们的线性链模型，这是一个根据RST进行加带的一个大型反课程序列。按照惯例，将语料库分为 into 一组份文件和一组个文档的测试组。图Table[2.1](#_bookmark24)显示了公司的详细特性。数据使用夏尼亚克和约翰逊的重新排序解析器[（Charniak 和](#_bookmark213)，）进行预处理 [Johnson, 2005)](#_bookmark213) ，以获得句法结构。我们的线性链采用（，）设计[(Okazaki, ，](#_bookmark270)这是线性链的快速实现。为了应用我们的双通分段策略（在第[2.2](#_bookmark21)节中介绍），我们首先使用第[2.3](#_bookmark23)节所示的基本特征和上下文特征，使用单个线性链表示每个句子，从而训练模型。对于训练和测试集，我们应用经过训练的单通道模型，以获得每个句子的初始分段。We 然后，我们从此初始分段派生全局要素，并训练我们的二通道模型以及基本和上下文要素。此任务使用了两种评估方法：第一种是评估检索句内边界（*B* 类）的精度、调用和*F*1分数，这是我们更关心的类。第二种是根据检索每个类的宏平均精度、撤回和*F*1分数来评估*B* 和*C*这两个类的性能。表[2.2](#_bookmark26)演示了在*B*类上评估的性能。我们比较了几个现有的模型。在第一部分，（[等人，2010a）](#_bookmark233)是一个模型，采用类似的基于的序贴标框架作为我们的，但没有涉及配对和全球功能。第二部分列出了以前四部遵循每个令牌的独立二进制分类框架的作品，包括[SPADE（Soricut和Marcu，2003年](#_bookmark283)）、（SPADE [和Di Eugenio，2007年](#_bookmark286)）、等人。[（Joty等人，2012年）](#_bookmark243)和（[费舍尔和罗克，2007年）。](#_bookmark229)最后一个模型，重新排序[（Bach等人 al.,，2012年），](#_bookmark203)利用子树特征来重新排列基础分段器的*N-*最佳输出，实现一个歧视性的重新排列模型，并获得了迄今为止报告的最佳分段性能[报告1](#_bookmark0)。可以看出，与所有以前的模型相比，我们的双通道模型在两个类的所有三个指标上获得最佳性能。事实上，我们得到的召回与等人相同，但其精度明显低于我们。关于*F*1分数，我们的模型实现了的错误率降低超过最佳基线，即，重新排序，并接近水平的人类性能在这个任务[2。](#_bookmark0)此外，由于等人的重新排序框架几乎与我们的双通方法有正交，因此，我们的双通分段模型可以作为更强大的基段，因此通过将我们的双通模型插入重新排列框架，可以预期进一步的改进。表[2.3](#_bookmark27)演示了对两个类及其宏观平均值的评估性能。只有前两个模型，和排位，根据此评估报告了其性能，因此此比较中不包括其他以前的模型。可以看出，在这里考虑的三个模型中，我们的双通分割模型与配对fea- 不仅在*B*类上执行得最好，而且在*C* 类上表现最佳，因此宏平均*F*1得分为。

## 2.4 错误传播到话语解析

如前所述，话语分割是式反课程解析器的第一阶段。因此，评价话语解析的整体性能如何受二等分割模型结果的影响是很有帮助的。为了评估性能，我们使用标准未标记和标记的分数的跨度，核度，和关系，定义 [Marcu （2000）.](#_bookmark264)此外，为了进一步说明文本的二等级自动分段的等，我们分别对the 发送内、多发送和文本级别进行了评价。在发送内层面，评价单位是不跨越句子界限的话语子树。在多发送层次上，考虑所有跨越至少两个句子的话语子树。在 the 文本级别上，评估所有话语子树。结果如表[2.4](#_bookmark28)所示。可以看出，在发送内水平上，分段的内向是显著的。在跨度、核度和关系上进行评估，使用我们自己的细分结果在得分（p）中产生的。在所有情况下）[4，](#_bookmark0)而使用等人的分段时，这种花边甚至更大尽管如此，整体分析性能明显更好（p）使用我们的分段模型时，使用using 等人。然而，在多发送级别评估时，使用手动和自动分段之间的二分数几乎脱节梨。事实上，所有指标的绝对严重性都小于，而且微不足道。事实上，这不是一个令人惊讶的the 发现：RST式话语解析树中的大多数话语成分都符合句子边界，即很少跨越多个句子。此外，本实验采用的目标话语解析器采用两阶段解析策略：在第一阶段，将句子处理成句子级话语子树，进而作为第二解析阶段的基本过程单元。因此，由于RST式话语树的性质和目标话语解析器中特定的解析算法，假分割的影响在很大程度上局限在每个句子中，因此对在此发布版本中，句子拆分作为软件的预处理过程的一部分合并。为了进行公平的比较，为了排除不同句子分裂软件和我们自己的模型的复杂性，我们修改了代码，以确保所有分段与输入相同的句子集进行馈送。所有显著性测试均使用威尔科森签名等级测试执行。较高的树水平。根据分析，分段对文本水平的影响几乎完全归因于其对发送内水平的影响。特征分析

在本节中，我们研究了配对和全球特征的等，我们的双通道分段模型的两种截然不同的参数，即整体性能及其跨分段框架的通用性。首先，从我们的完整模型开始，我们执行一系列特征消融实验。在每次这些实验中，我们从训练中的功能集中 the 删除其中一个组件特征或其组合，并评估结果模型的性能。

删除配对功能（+*p*） 通过删除配对功能，图[2.2](#_bookmark22) 中所示的基于的段段模型将减小到图[2.3](#_bookmark30)所示的模型，其中每个标签节点*Li*的输入是单个标记*Ti，*, 而不是一对相邻的令牌*Ti*=1和*Ti*。请注意，第一个令牌 *T*1在序列中被排除在外，因为 *T*.1之前始终存在边界（句子的开头）。根据第 [2.3](#_bookmark23)节中列出的功能，现在减少到描述每个令牌 *Ti*的功能。

通过删除全局特征，我们的双通道分段模型将简化为一个简单的单通道模型，其中仅使用第[2.3](#_bookmark23)节中的基本和上下文要素来训练模型。删除两个功能 **（***p***）** 在这种情况下，我们的模型将简化为一个简单的单通道模型，其中只使用基本和上下文要素，并且所有要素都基于每个单独的令牌*Ti，*而不是令牌对*Ti*和*Ti*=1。此外，我们希望通过, 评估它们的，探索配对特征和双通层-egy的一般性。特别是，由于我们的双通道分段模型本身是基于的序号标签模型，因此，在本文ex- 中，我们还研究了在独立二进制分类的框架下删除配对和全局特征的等。回想一下，在独立的二进制分类阳离子的框架下，句子中的每个标记（不包括*T*1） 在序列中独立检查，二进制分类器用于预测该令牌的标签。图[2.4](#_bookmark33)显示了我们在独立二进制分类框架内的模型。如果启用配对要素（如图[2.4a 所示，](#_bookmark33)则在每个分类中，将检查一对相邻令牌，而不是单个令牌，分类器预测边界之间是否存在。如果禁用配对功能，则模型将减小图[2.4b](#_bookmark33)所示的配对要素。在本实验中，我们探索了独立二进制分类器中的两个基础分类器：逻辑回归（）和线性内核支持向量机（）。我们爆裂使用学习[（Pedregosa等人，2011年）](#_bookmark274)来分类这两个分类器。对于，所有参数都保留为其默认值，而对于，我们使用自动类权重，这些权重根据训练数据中第二类的分布进行调整，以克服of 数据集中*B* 类的稀疏性。表[2.5](#_bookmark34) 演示了我们的功能分析结果。第一部分列出了我们完整模型在二等分段框架中的每个形式。可以看出，我们的完整模型在框架中的性能类似，其中*F*1中的绝对小于且微不足道。这与[Hernault等人](#_bookmark233)s [（2010a）](#_bookmark233) 发现，当合并大量上下文要素时，等二进制分类器SVM可以实现与的竞争性能。第二部分列出了of 没有配对功能的模型的性能（=*p*）。对于所有三个生成的模型，CRF= p、LR=*p*和SVM*p*=*p，*,性能明显较低（p*p*）比他们在第一节中对应的完整模型。当在第三部分删除全局特征时，观察到类似的趋势。但是，就底层框架本身而言，明显比和LR（p），而当删除配对要素时，这种显著的二等特征是无法观察到的。最后，当两组要素被删除（=*p*g），如上一节所示，我们的模型的性能急剧下降（从以上下降到以下）。 have这表明，配对和全局特征本身对性能有重要的作用，even 它们的组合更为重要。在本实验中，我们论证了配对特性和全球特征在提高整体分段性能方面具有内质性等，这种改进是显著的。此外，我们观察了跨框架的类似，这表明我们细分模型这两个新方面的一般性。现在，我们进行逐令牌错误分析，以研究基于的模型与特征设置所犯的错误分布情况。特别是，我们评估由我们成熟的双通道分段模型或已删除的配对或全局功能训练的模型在测试集中的每个令牌上出现的标签错误。在这里，我们将比较限制在顺序标签框架之后的模型，即基于的模型，表 [2.5](#_bookmark34)中具有+*p*或上文。同样，作为句子开头的所有令牌都不包括在此分析中。结果如表[2.6](#_bookmark36)所示。一个有趣的观察是，如表第二部分所示，在=g之上，通过添加全局功能，我们的完整模型能够纠正=g的个错误，同时在此过程中不会引入其他错误。此外，如表第三部分所示，配对和全局特征几乎相互互补，即=*p* 所犯错误的发生在=正确且对等的情况下，=g所犯的错误的g发生在CRF=*p*正确的情况下。最后，在图[2.5中，](#_bookmark37)我们展示了一些句子示例，我们成熟的双通道分段模型标签正确，而较弱的模型会犯一些错误。

## 2.5 未来工作

在本章中，我们介绍了基于线性链的双通RST式话语分割模型。与在建模中依赖于以令牌为中心的功能的分段的典型方法不同，我们的分段器中的要素以令牌对为中心，以同样地考虑上一个和以下围绕潜在边界位置的信息。此外，还提出了一种新颖的两通分割策略。在初始分段通过后，我们获得了一组全局特征来描述整个分段结果的特征，在第二次传递中考虑这些特征，以获得更好的分段。与几种现有的话语分割模型相比，我们在识别边界和非边界方面达到了最佳的每次形式。此外，我们研究了the 我们新颖的配对和全球特征的等特性，并证明这两组特征对整体性能都很重要，并且这种重要性在分段框架和分类器中是可以观察到的。最后，我们尝试了我们的分段模型作为话语解析器的插件，并评估了它对整体解析精度的影响。我们发现，自动分段在计算在发送内水平时对解析交流性有巨大的影响然而，这种影响在多发送层面上是非常次要的。对于今后的工作，我们希望探讨将我们的双通分割模型纳入[巴赫等人](#_bookmark203)的重新排名框架。[（2012）。](#_bookmark203)由于我们的模型被证明是一个更强的基段，随着重新排列过程，可以预期分段精度的进一步提高。