汉语语篇修辞结构标注实验

邱武松
南京师范大学教育技术系 南京 210097
E-mail:hkqws@163.com

摘 要：本文介绍了我们利用修辞结构理论的理论框架进行汉语语篇修辞结构标注的实验，论述了汉语语篇结构标注的基础理论选择理由、基本语篇单元的界定、修辞结构和关系的标注方法、标注质量的控制和评价方法、标注结果。本文还对基本语篇单元的边界符的分布特征进行了分析讨论。

关键词：语篇结构，修辞结构，语篇分析，语篇结构标注

Experiments in annotating rhetorical discourse structure of Chinese Texts

Qiu Wusong
Department of Educational Technology, Nanjing Normal University, Nanjing 210097
E-mail:hkqws@163.com

Abstract: We introduce our experiments in annotating rhetorical discourse structure of Chinese texts in the framework of Rhetorical Structure Theory. We discuss the reason why we selected the RST, the confining method of elementary discourse units, the tagging style, and how to control and assess the annotating quality. We also discuss the distribution characteristics of boundaries of EDUs.

Keywords: discourse structure, rhetorical structure, discourse analyzing, discourse structure annotation

1 引言

低成本、高效地处理自然语言是人类应对信息急剧增长的一个必然选择，语篇结构的自动分析是海量文本处理系统的重要组成部分。要研究语篇结构的自动分析，必须以一定的实证研究作为指导和支撑，而且自动分析的评测需要以一定的语料作为基准，基于机器学习方法的语篇计算也需要大规模语料库作为基础。为此，我们开展了以面向语篇结构自动分析的手工标注实验，通过实证研究来探索自动分析的策略。

2 标注实验

2.1 理论基础

不管是自动分析还是手工分析，首先需要解决的是语篇结构的形式化问题。为了对语篇的整体结构进行描述，人们先后提出了一些形式化的表征方法。我们采用修辞结构理论（Rhetorical Structure Theory，简称 RST）的基本理论框架作为语篇分析的理论基础。

RST 是 20 世纪 80 年代在美国南加利福利亚大学信息科学研究所由 William Mann、Christian Matthiessen 和 Sandra Thompson 创建并发展起来的。RST 最初是为了实现语篇自动生成（Natural Language Generation）而创建[1]，后来被广泛应用于语篇分析、自然语言处理、语言教学等方向，在计算语言学界和传统语言学界都得到了广泛的关注和讨论，成为影响最大、受关注最高的语篇

结构形式化理论之一。

2.2 标注工具

我们使用了 Mick O'Donnell 开发的开源可使化标注工具 RST-Tool[2]，该工具使用 TCL/TK 进行编写，具有可视化的单元切分、结构关系标注功能，还具有一定的数据统计功能。在使用过程中，我们发现 RST-Tool 的功能并不能完全满足我们的研究需要，而且还存在 bug，因此我们对 RST-Tool 的源码进行修改，主要包括统计功能的扩充、bug 修正和汉语增强三个方面。

2.3 语料选择

RST 是一个对自然文本的语篇结构的描述框架，在理论创建之初，很多的研究者都对 RST 的适用范围做了大量的研究，并证明 RST 适用于包括议论性、说明性、记叙性文本在内的多种类型[1]。一般说来，由于研究精力有限，语篇分析一般都是有选择性地区分文体类型和语料来源。Marcu 等人在开展英文修辞结构树库的标注项目中，选择了三种类型的语料：30 篇 MUC7 参考语料（MUC7 co-reference corpus），主要包括一些关于企业管理的新闻故事，平均每篇 405 个单词；30 篇布朗学术语料（Brown-Learned corpus），主要包括一些很长的、高水平的学术文章，平均每篇 2029 个单词；30 篇华尔街日报语料（Wall Street Journal corpus，WCJ），主要是社论文章，平均每篇 878 个单词[3]。乐明在汉语 CJPL 项目中，选择了 2005 年 4 月 12 日在人民网《主要媒体财经评论》栏目上转载的 395 篇文章，题材包括社会问题、财经政策、证券、汇率、国际贸易、会议导报、上市公司年报分析等；体裁包括财经消息、内外刊文章编译、杂文、社评、学术论文摘要、访谈综述、述评等 [4]。由于时间和精力等因素，我们最后选择了新华网（www.xinhuanet.com）和人民网（www.people.com.cn）上的社论类文章，共计 30 篇，但目前我们仅分析完成了其中的 10 篇。由于我们目前的研究主要在于探索汉语的语篇结构特征和自动分析的策略思路等，而且语篇修辞结构的研究主要考察语篇单元的数量，所以本文所选的语料已能满足研究的需要。

2.4 标注方法

2.4.1 标注方式

语篇结构的标注通常分为两项工作，一是基本语篇单元（Elementary Discourse Units，简称 EDUs）的识别与切分，二是单元间结构和关系的识别和标注。Marcu 等研究者在英语修辞结构树库的建设过程中，为了给计算机处理提供指导意见，使用了递增式的标注方式，即识别出一个 EDU，立即进行结构关系的识别并将其加入结构树中，实验结果显示，这种标注方式效果很不理想，经常不知道新加入的单元该放入到树中的哪个位置，因此，他们得出结论，认为采用递增式分析方法人类尚不能较好地进行语篇结构分析何况是计算机[3]，我们在前期的语篇分析过程也显示递增式的标注方式难以构建出高质量的修辞结构树。因此，在我们对所选预料的正式标注中，采取了自底向上的标注方式——即先将语篇切分为若干个 EDUs，然后在整体地确定单元间的结构和关系——并对每一步都进行了一致性的评价。

2.4.2 基本语篇单元的识别和切分

在语篇分析中，基本分析单元还无统一的界定标准。在基于 RST 的语篇分析中，理论创建者 Mann & Thompson 认为，EDUs 的切分是任意的，但是也要遵循一定的理论方法，即 EDUs

应该具有独立的功能完整性[5]。在他们的语篇分析中，EDUs 基本是小句（clauses），但是不包括从句主语（clausal subjects）、补语（complements）和限制性关系从句。Marcu 在[6]在英语修辞结构树库标注项目中制定了具体的 EDUs 切分细则，基本是将 EDUs 切分到小句层次，但比 Mann 和 Thompson 切分得更细。乐明在汉语 CJPL 修辞结构树库标注项目中借助了一套选定的标点符号作为 EDUs 的形式切分依据以方便计算机进行自动切分[4]。乐选定的标点符号主要包括：分号、冒号、省略号、破折号、空格、句号、问号、叹号；没有被作为切分依据的标点有：逗号、括号、引号、其他标点符号[5, 8]。乐明利用部分标点符号进行切分的方法比较容易实现，但缺陷也是很明显的，一是很多具有独立功能完整性的 EDUs 没有被切分出来，导致遗漏了很多的修辞结构信息；二是有些切分后的单元本身缺乏独立的功能整体性，分裂为几个部分与前后单元形成修辞关系，乐的切分方法没办法对这种情况进行识别。

在参考上述研究工作的基础上，我们对汉语中基本语篇分析单元的语言范畴进行了考察，发现汉语言学研究中“小句中枢说”的小句基本与基本语篇分析单元具有较好的对应关系。郑贵友从语篇语言学的角度分析“小句中枢”学说在汉语语篇分析实践中的价值，并认为汉语句子的实义切分、汉语语篇主位推进模式的归结、汉语语篇微观结构分析与归纳等很多问题,都必须以“小句”为基本的立足点才能得到合理而充分的观察[8]。经过实际语料观察，我们发现汉语中的“小句”可合理地当作语篇结构分析的基本单元。对于小句的界定方面，我们采取邢福义对小句的定义，即小句主要指单句和结构上相当于或大体相当于单句的分句，不包括充当句子成分的主谓短语[9]。这样做的原因是，如果将充当句子成分的主谓短语作为分析单元，一方面会造成切分过细，很多的单句会被拆分为几部分，还需要采取 Macru 的做法，即利用内嵌结构和伪关系联系起来；另一方面，经过实际的语料分析，发现充当句子成分的主谓结构所表现出的修辞关系非常单一，不需要切分到这么细。

2.4.3 语篇修辞结构的标注

语篇结构的标注一般需要完成结构和关系识别两部分工作。结构的识别主要就是确定单元间的关系是主从还是并列关系，关系的识别需要依据关系的定义进行。RST 对单元间的各种修辞关系规定了具体的定义，以此来准确说明语篇单元间的关系。所有的修辞关系组成一个关系集，Mann 和 Thompson 通过对大量的真实语料进行研究的基础上，总结了 25 种修辞关系[5]，被称为经典的修辞关系集。但他们一直认为修辞关系是一个开放的集合[1,6]。因此，各国的研究者通常都会依据定义格式对修辞关系集进行一些修改和扩充[3, 5, 11]。目前，已有两个面向汉语修辞关系的集合[5, 8, 12]，我们对这些修辞关系集进行了详细的对比分析，并尝试使用它们进行标注，但遗憾的是，不同研究者对修辞关系的定义有着较大的不同。因此，我们最后决定以经典修辞关系集为基础，参考其他的关系集合，边标注边完善修辞关系集合。

2.5 标注质量控制和评价

2.5.1 标注质量控制

为了保证本研究的标注质量，我们依据 RST 理论框架制定了基本原则和校验方法。所有标注的工作都按照基本原则进行标注，然后使用校验方法进行标注校验。

标注基本原则主要是根据 RST 的理论框架来制定的，主要有以下四点：1) 立足原文，尊重作者的写作意图和交际意图；2) 严格遵守修辞关系的定义进行标注；3) 保证核心单元保留作者所想要表达的核心意思；4) 着眼全局，不要只在局部范围考虑单元间的关系。

本文所采用校验方法主要来自 RST 理论框架中对核心单元的核心性的论述，本文称之为单元删除法。对于待校验的单核关系，单元删除法的操作步骤如表 1 所述。

单核关系	Do：删除核心单元
	if 卫星单元反映作者所要表达的主要意思 **then**
	　Result：标注错误
	else
	　Do：删除卫星单元
	　if 核心单元不能反映作者所要表达的意思
	　then
	　　Result：标注错误
	　else
	　　Result：标注正确
	　end if
	end if
		多核关系	**for each** 单元-k **in** 所有单元
			　if 单元-k 不是作者所要表达的核心意思
			　then
			　　Result：标注错误
			　end if
			end for
			Result：标注正确

表 1　单元删除法操作步骤

2.5.2　标注结果评价

语篇结构标注的质量评价主要是对实验误差大小的评定，是语篇结构标注研究中一项必不可少的环节，常常利用重复标注然后评价结果的一致性（agreement）来反映误差情况。一致性通常包括标注者间的一致性（inter-annotator agreement）和标注者自身的一致性（intra-annotator agreement）两种类型。为了全面评价对语料的标注质量，我们结合了两种一致性方法进行了语料的标注实验。大致过程如下：笔者和一名研究生同学独立完成了所有语篇的 EDUs 切分任务，检验标注者间的一致性；然后笔者对所有语篇进行了三遍独立的语篇结构和关系的标注，计算标注者自身的一致性；然后笔者请两位研究生同学对其中两篇文章进行了心理实验以评价标注质量。

一致性的评估通常使用卡帕（kappa）系数来衡量，卡帕系数反映了评分者实际评定一致的次数百分比与评分者理论上评定一致的最大可能次数百分比的比率，其计算公式如公式 1 所示，其中，P(A)指 K 位评分者评定一致的百分比，P(E)指 K 位评分者理论上可能评定一致的百分比。

$$K = \frac{P(A) - P(E)}{1 - P(E)} \qquad \text{公式 1}$$

对于语篇修辞结构来说，确定统计类别和样本的考察范围是 kappa 值计算中非常关键的两个方面。对于 EDU 切分一致性的 kappa 值计算，最简单的办法是假设所有的符号（包括标点、汉字、英文单词）后面均可以插入 EDU 切分标记，统计不同标注者在所有插入位置上 EDU 切分标记的一致性，计算 kappa 值为 K_w。但是，由于大部分的字符后面不是 EDU 切分符应插入的位置，所以这种计算方法会导致 kappa 值偏高。偏高的原因是大量字符被判定为非 EDU 而拉高了统计者间的一致性。因此，我们使用了另外一种统计方式，假设 EDU 切分符只能插入所有的标点后面或者是任一标注者认为是 EDU 分界的位置上，这样就大大减小了非 EDU 分界位置的影响，从而使得一致性的统计更趋真实。这两种方式统计的 kappa 值如表 2 所示。从表 2 中可知，在 EDU 分界点的判定上，标注者间的一致性是非常高的，Marcu 等在英语语料 WSJ 的语篇分析中在相同评价方法下 K_w 值为 0.919[3]，与本文的差值较大，这一方面与我们对 EDU 切分的粒度有差异有关，另一方面也可能与汉语和英语两种语言之间的差异有关。

语料	方式一		方式二	
	总数	K_w	总数	K_p
所有文章	13417	0.992	879	0.963

表 2 EDU 分界点的标注者间的一致性

对于语篇层级结构、核心单元、修辞关系三个方面的一致性计算，我们参考 Marcu 等人提出的做法[3]，即将层级关系映射到对应一个标有关系和核心性判断的单元集合中。按照这种统计方法，对笔者所进行的三遍进行的独立标注进行了一致性统计，统计结果如表 3 所示。从表 3 中可知，经过三遍的独立标注，标注者内的一致性的已经达到很高的程度。这个结果一方面与我们制定了基本标注原则和校验原则的做法有关，另一方面也可能与同一标注者对同一文章的理解逐渐趋于一致相关。实际上，这种一致性的求解方式会使得 kappa 偏高，因为，判断集合中存在大量一致的不是语篇单元的判断元素。

	语篇层级结构	核心单元	修辞关系
第一遍与第二遍	0.946	0.940	0.898
第二遍与第三遍	0.990	0.989	0.985

表 3 语篇修辞结构的标注者内的一致性

需要说明的是，语篇分析的一致性统计是很难找到一个真正好的评价方法，一方面， kappa 统计虽然能很好地反映一致性，但是，对于语篇标注实验来说，往往在标注前就规定了一些标注方法或手册，这实际上已经破坏了独立性假设，使得 kappa 值的可信度降低；另一方面，正如前面的分析，纳入考察的样本范围对 kappa 值的确定有较大的影响。如何有效地考察语篇分析的一致性，还是一个需要继续深入探讨的课题。

标注者内的一致性并不能全面地评价语篇结构标注的质量，但是，由于我们人力限制，没有采取多人独立标注的实验方式，因此，不能求得标注者间的一致性。为了弥补这方面的不足，我们设计了一个心理分析实验用以评价标注质量，实验由两位硕士研究生同学完成，实验过程，实验过程如图 1 所示。由于这样的评价是在已标注语料上进行，破坏了一致性统计中的独立性假设，因此，我们仅统计了实验结果并未计算一致性，如表 4 所示。

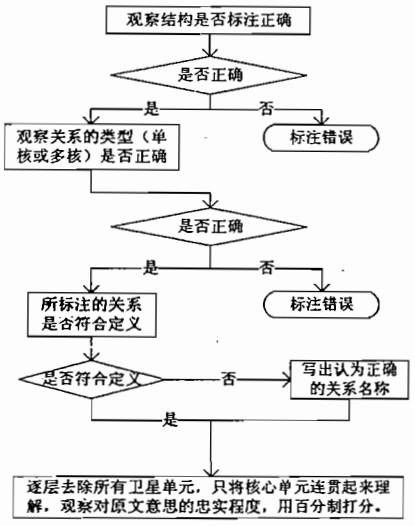

图 1：标注评估流程图

		实验者 1	实验者 2
结构判定错误		0	0
关系类型标注错误		1	0
不符合关系定义的标注		8	3
核心单元对原文意思忠实度	语篇 1	90	70
	语篇 2	95	75

表 4：标注评估心理实验的结果

从表 4 中可以看出，两位心理实验者对几乎所有的结构判定和关系类型都不存在异议，对99.11%的关系判断都保持认同，这说明我们对语料的语篇修辞结构标注具有很高的一致性，具有非常高的可信度。从核心单元对原文意思忠实度的结果来看，两者相差都较大，这与不同的人的判断标准不一样；但是，可以肯定的是，利用 RST 理论框架进行语篇分析，可以较好地反映文章的语篇结构。

2.6 标注结果

经过多月的标注实验，我们最后确定了一个稳定的语篇结构树库。我们对树库的各项指标进行统计，统计结果如表 5 所示。

总字数	EDUs 数量	EUDs 中平均字数	单核关系数量	多核关系数量	S-N 数量	N-S 数量
12538	698	18	400	200	248	152

表 5 语料标注的统计数据表

由表 5 可计算，单核关系与多核关系的比例约为 66.67%：33.33%，这说明汉语中多核结构比较多，但单核关系仍占主导。进一步地，我们对比了语篇单元个数，单核单元与多核单元的数量之比为 61.63%：38.37%，这说明语篇中大部分的语篇单元都是通过单核关系联系的。在所有的单核关系中，卫星-核心结构（S-N）与核心-卫星结构（N-S）的比例为 62%：38%，这说明汉语中前偏后正的结构在我们的语料中是很明显的。这个结论与汉语复句研究中关于“汉语复句多是偏句在前，正句在后”的结论是一致的。

3 EDU边界符的特征分析

EDU 边界符指 EDU 的首部或尾部的字符或词语。在实验中，我们发现汉语中 EDUs 的边界具有很强的规律性，即绝大部分 EDUs 都是通过标点符号分隔的。这可能与汉语比较偏爱使用标点来分隔具有独立功能完整性的语言片段有关。为此，我们统计了作为 EDUs 边界符的标点的出现频率，如图 2 所示。

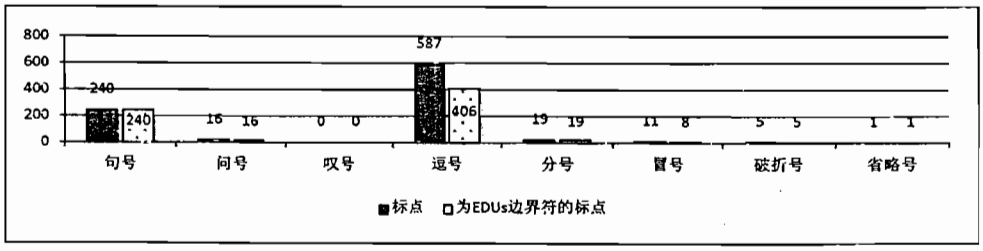

图 2：标点符号为 EDUs 边界符的统计图

从图 2 中可以看出，所有的句号、问号、分号全部都是 EDUs 的边界符，逗号约有 69%为边界符，冒号有 72.73%为边界符。叹号没有出现可能与所选的语料为社论类文章有关。这个结果固然与采取的 EDUs 切分方法有关系，但也非常明显地说明，标点符号是汉语中 EDUs 切分的一个重要的形式依据，而且切分出的 EDUs 满足功能独立完整性的要求。此外，结论也证明，乐明[4]在 CJPL 标注工作直接以部分标点为切分标记的做法的可行性，但不将逗号考虑为切分标记确实损失了较多的语篇结构信息。从图 2 还可以看出逗号作为 EDUs 边界符的歧义是最大的，因此，判断逗号是否为 EDUs 边界符就成为了自动 EDUs 切分的难点和重点。进一步地观察表明，

逗号作为 EDUS 边界符时也表现较强的规律性，如：连词后面的逗号不是边界符，没有动词的语言片段后的逗号往往不是边界符。因此，可以预测，在基于 RST 的汉语语篇自动分析中，利用标点符号进行 EDUs 的切分是可行的，并且利用一定的策略，可以使 EDUs 的自动切分可以达到较高的准确率和召回率，不过这还需要实践来检验。

4 结论与进一步工作

建设一个高质量的汉语语篇结构树库对中文语篇计算具有重大的意义。本文介绍了我们进行汉语新闻评论语料标注实验的实验方法、实验过程、实验评价等；同时对汉语语篇分析的理论基础、EDU 界定、质量控制、一致性评价等方面进行了探讨。目前，我们标注的语料还不多，而且语料类型单一，因此，我们计划在已有经验的基础，继续扩大语料库，以满足后续语篇自动分析研究的需要。从基于语料研究本身来看，标注质量的控制和评价方法需要更加深入地研究，语料的范围和数量需要增大。另外，语料标注的最终目的是为了实现语篇计算，因此，我们也将基于语料研究的成果开展汉语语篇结构自动分析的研究。

参 考 文 献

[1] Maite Taboada, William C. Mann. Rhetorical Structure Theory: Looking Back and Moving Ahead[J]. Discourse Studies, 2006, 8(3), 423-459

[2] Mick O'Donnell. 1997. RST-Tool: An RST analysis tool. In Proceedings of the 6th European Workshop on Natural Language Generation, Duisburg, Germany, March 24–26.

[3] Daniel Marcu, Estibaliz Amorrortu, and Magdalena Romera. Experiments in Constructing a Corpus of Discourse Trees[A]. Marilyn Walker. Proceedings of the ACL'99 Workshop on Standards and Tools for Discourse Tagging[C]. New Brunswick, USA : Association for Computational Linguistics, 1999, 48-57.

[4] 乐明. 汉语财经评论的修辞结构标注及篇章研究[D]. 北京: 中国传媒大学, 2006.

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

[6] Marcu D . Instructions for Manually Annotating the Discourse Structures of Texts[EB/OL]. http://www.isi.edu/~marcu/.

[7] 乐明. 汉语篇章修辞结构的标注研究[J]. 中文信息学报, 2008, 22(04): 19-23.

[8] 郑贵友. "小句中枢说"与汉语的篇章分析[J]. 汉语学报, 2004(01): 61-65.

[9] 邢福义. 小句中枢说[J]. 中国语文, 1995(06): 420-428