

中英文篇章依存树库构建与分析*

吴永芃, 李素建, 秦沐坤, 杨安, 王厚峰

(北京大学计算语言学教育部重点实验室, 北京市, 100871)

摘要: 本文在篇章依存关系的基础上, 建立了小规模中英文篇章依存树库, 并针对多核心关系问题、依存关系的选择、长篇章与复杂篇章的标注、层次结构信息的损失等标注过程中遇到的困难进行了分析研究, 给出了解决方案。并对篇章依存树库进行了简单的统计分析, 对中英文篇章中的异同进行了初步探索。

关键词: 篇章依存关系; 篇章依存树库; 篇章结构

中图分类号: TP391

文献标识码: A

Exploring Chinese and English Discourse Dependency Treebanks

Yongpeng Wu, Sujian Li, Mukun Qin, An Yang, Houfeng Wang

Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, Beijing, 100871

Abstract: Based on the idea of Discourse Dependency Relations, a small-scale Chinese and English Discourse Dependency Treebank is constructed in this paper. Difficulties faced during the annotation process, such as multi-nucleus relation problem, selection of relations, annotation of long and complicated discourses, loss of information in hierarchical structure are analyzed and we present solutions to them. We conduct some simple statistical analysis on the Discourse Dependency Treebank and explore the similarity and difference between Chinese and English Discourse.

Key words: Discourse dependency relation; Discourse Dependency Treebank; Discourse structure

1 引言

在自然语言处理领域中, 在词与句子的层次上, 目前已有较多的研究, 并取得了丰硕的成果。随着研究的深入, 人们开始着眼于更高层次的自然语言分析——篇章层次。众所周知, 篇章所独具的完整性和连贯性使得一个篇章与一段由若干句子随机组合而成的文本具有显著的不同。我们可以将篇章视为一系列连续的文本单元(如子句、句子或语段)构成的语言整体单位。任何文本单元都不可以被孤立地进行解读, 而是需要根据其上下文来理解。篇章分析与标注, 旨在对篇章内部的结构和关系进行分析, 并在分析的基础上对其进行相应标注。篇章分析技术在自动文摘^[1]、自动问答^[2]、指代消解^[3]等自然语言处理领域中, 具有重要的意义。

当前两个有代表性的英语篇章树库为宾州篇章树库(Penn Discourse Treebank, PDTB)和RST树库(Rhetorical Structure Theory-Discourse Treebank, RST-DT)。PDTB由美国宾夕法尼亚大学创建, 标记了约100万字的华尔街日报文章, 最新版本为PDTB 2.0^[4]。PDTB将语句看作论元(Argument), 主要标注论元对之间的篇章语义关系和可能的连接词, 把一个大的篇章分解成平面化的论元对, 篇章标注层次较浅, 为浅层篇章标注^[5]。RST-DT建立在修辞结构理论(Rhetorical Structure Theory, RST)之上, 由美国南加州大学和美国国防部共同创建, 共计标记了385篇华尔街日报的文章, 总字数超过176000个^[6]。RST理论通过修辞关系对语篇结构进行描写, 将整个篇章构建成一棵有层次的RST树^[7]。然而, 有层次的RST树结构较为复杂, 节点数目较多, 不同层次的篇章单元有包含关系, 难以构建一个统一的用于篇

* 收稿日期: 定稿日期:

基金项目: 国家自然科学基金项目(61572049和61333018)

作者简介: 吴永芃(1995—), 男, 本科生, 自然语言处理; 李素建(1975—), 通讯作者, 女, 副教授, 自然语言处理; 秦沐坤(1994—), 男, 本科生, 自然语言处理; 杨安(1994—), 男, 硕士研究生, 自然语言处理; 王厚峰(1965—), 男, 教授, 自然语言处理。

章分析的架构,给机器自动分析带来了困难。汉语篇章语料库的建设也取得了一些进展,哈工大参考 PDTB 的标准,并结合中文的特点,从分句、复句和句群三个层次标注显式和隐式关系,构建了篇章语料库 HIT-CDTB^[8]。苏州大学利用汉语依存句法分析技术构造了篇章结构语料库 CDTB^[9],该结构融合了 PDTB 和 RST 的优点,对每篇文档构成一棵篇章树,虽然篇章结构的信息量更加丰富,但是也加剧了自动分析的困难。

篇章依存分析结构的引入一定程度上兼顾了深层篇章结构的标注和降低自动分析的难度。2014年,李素建等^[10]首次提出利用依存结构进行篇章分析,不同于 CDTB 采用的句法依存,这儿的依存指的是篇章层面的依存关系,并在已有的 RST-DT 的基础上进行转换,建立了一个英语篇章依存树库。二元非对称的依存结构解释了篇章的深层关系,保留了 RST 树中的大部分信息,又因相对简单的结构,可以直接分析各个单元之间的关系,使机器自动分析工作能更容易地开展。但 RST 树转换而来的篇章依存树,可能存在一定的问题。例如,无法展现篇章依存树特有的非投射关系。到目前为止,较少有人工构建篇章依存树库的工作。基于这一背景,本文在篇章依存关系的基础上,建立了小规模中英文篇章依存树库,并针对多核心问题、依存关系的选择、长篇章与复杂篇章的标注、层次结构信息的损失等标注过程中遇到的困难进行了分析研究,给出了解决方案。同时,对篇章依存树库进行了简单的统计分析,针对中英文篇章的异同做了简单探索。

2 篇章依存树

2.1. 篇章依存树的形式化表示

篇章依存分析思想认为,篇章由篇章单元(Elementary Discourse Unit, EDU)构成。篇章单元之间由被称为依存关系的二元非对称关系连接。其中,我们称附属(subordinate)篇章单元为“附属单元”(dependent),称被依靠的篇章单元为“头部单元”(head)。利用篇章依存树表示篇章依存结构时,我们需要在篇章依存树起始位置插入一个人工篇章单元,称之为 e_0 ,并视之为该篇章的根(Root),以此简化定义与计算过程。

记一含有 $n+1$ 个篇章单元的篇章 T 为: $T = e_0 e_1 e_2 \dots e_n$,其中 e_0 为根。记该篇章依存关系集为 R , R 为有限的功能关系集合且 R 中的关系存在于两个篇章单元之间。记 V 为篇章的一系列节点, A 为篇章的一系列有向标记弧。记篇章依存图为 G ,则: $G = \langle V, A \rangle$ 。则有:

$$(1) V = \{e_0, e_1, e_2, \dots, e_n\}$$

(2) $A \subseteq V \times R \times V$,其中 $\langle e_i, r, e_j \rangle \in A$ 代表从头部单元 e_i 到附属单元 e_j 被关系 r 标注的一条弧

$$(3) \text{若} \langle e_i, r, e_j \rangle \in A, \text{则对所有} k \neq i, \langle e_k, r', e_j \rangle \notin A$$

$$(4) \text{若} \langle e_i, r, e_j \rangle \in A, \text{则对所有} r' \neq r, \langle e_i, r', e_j \rangle \notin A$$

其中,条件(3)确保了每个篇章单元有且仅有一个头部单元。条件(4)确保了两个篇章单元之间,不能有多于1种的依存关系。

一般而言,对同一个篇章,篇章依存树的结构比 RST 树更加简单,节点数更少,复杂度更低。例如,3个单元的核心-辅助结构 RST 树新增了1个中间节点和1个根结点,而篇章依存树仅新增了1个根结点。可以看出,含有 n 个文本单元的核心-辅助结构 RST 树,共包含 $2n-1$ 个节点;而含有 n 个文本篇章单元的篇章依存树,仅包含了 $n+1$ 个节点。

2.2. 依存关系

本文参考了 1988 年 RST 理论提出的修辞关系^[7]。根据所标注中英文语言特点，删减与合并了部分关系，根据标注时遇到的情况新增了部分篇章关系，最终确定了如表 1 所示的 26 个篇章依存关系，这儿由于篇幅限制，不再展开篇章关系的介绍。

表 1 本文采用的篇章依存关系

temporal	comparison	contrast
elab-aspect	elab-example	joint
elab-addition	condition	elab-enum_member
bg-general	bg-compare	elab-process_step
bg-goal	exp-reason	progression
exp-evidence	evaluation	manner-means
cause	same-unit	attribution
enablement	summary	continuation
elab-definition	ROOT	

3 篇章依存树库的构建

3.1. 语料库的选择

由于短小但完整、连贯的篇章，可以推动篇章依存库的构建和分析，因此我们选择了科技论文摘要和新闻作为标注文本。科技论文摘要的写作一般结构清晰，逻辑性强，容易进行标注和分析。而我们选择人民日报的时政要闻快讯作为语料库，是考虑到人民日报作为国家权威媒体网站，发布的新闻遣词造句较为严谨，结构清晰，逻辑性强，质量较高，不会给篇章依存分析带来困扰。该语料库新闻的平均字数较少，但每一篇新闻，都保持了其作为篇章的完整性和连贯性。

基于以上考虑，我们选用了 ACL 2014 会议的 50 篇英文论文摘要和 EMNLP 2014 会议的 40 篇论文摘要进行篇章依存关系标注，同时还标注了 15 篇英文经济短新闻，中文方面标注了 33 篇人民日报新闻。文献^[10]中篇章依存库是由 RST-DT 语料自动转换而成，依存关系的标注并不够准确，本文中的 138 篇文档均为人工标注和校对。

3.2. 确定篇章单元的划分方式

确定划分篇章单元的标准是进行篇章分析的先决条件，也是一项较为独立的工作。在英文篇章的篇章单元确定中，涌现出了多种划分方法。修辞结构理论认为，除个别情况外，从句是最基本的单位^[7]。Polanyi 坚持自然句应为最基本的单元^[11]。Grosz 和 Sindner^[12]认为，篇章单元的确定应考虑到该单位在上下文中的位置，且能反映事物的一定状态。

本文英文语料库的篇章划分参考修辞结构理论，以从句层面的结构为最基本单位，因此 to、that、since 等引导从句的介词会成为划分标记，包括一些动词的现在分词作后置定语也会被划分为单独的篇章单元。

中文语料库的篇章单元，与英语中选择从句较为不同，汉语篇章的从句多为隐性。此外，在汉语中，逗号常常起着单句切分的作用，被隔开的单元通常以单句或类似于单句的结构出现。因此，本文选择使用标点符号作为划分的依据。本文汉语篇章的篇章单元由逗号、句号、分号、冒号、问号、叹号、破折号与省略号划分。括号、顿号、引号、连接号、书名号、间

隔号等不作为划分依据。对于新闻语料，可能由多段组成，每个篇章单元起始位置用数字标识其段号和句号，表明其在文本中的位置。例如，“2.3”表示篇章第二自然段的第3个篇章单元。科技论文摘要通常由一段构成，不再区分其段落。

3.3. 标注工具

本文使用了篇章依存关系标注工具¹，为文本标注篇章依存关系。标注工具中，我们用白色文本框表示篇章单元；起始于头部单元并终止于相应附属单元的有向箭头表示依存关系；附属单元左侧写明该依存关系的种类；蓝色方框内的数字为相应篇章单元对应的头部单元编号。通过这种形式，我们将篇章依存关系表示成一个带标注的有向图。

具体使用时，我们首先对篇章进行篇章单元的划分，再使用该标注工具载入已完成篇章单元划分的文档和自定义标签，开始进行依存关系标注。在添加依存关系时，我们依次点击头部单元、附属单元，并在弹出的对话框中选择依存关系的种类，即可标注一个篇章依存关系。若出现关系标注错误，点击依存关系的附属单元后，通过“删除”功能可以删除该关系。标注工具还可以通过“撤销”功能取消之前的添加、删除与标注操作。“加标签”、“删标签”功能则可增、删依存关系的种类。整篇标注完成后，点击“保存”，即可将结果存为后缀名为.dep的文档，以供进一步的分析。

4 标注的困难与解决

篇章依存关系的直接标注需要为每个篇章单元选择其头部单元，并确定关系的种类。表面看来，只需要逐一对每个篇章单元进行分析即可完成，但真正标注时并不容易。在分析每个篇章单元的时候，需要从全文去理解，帮助确定每个依存关系。下面，我们将介绍标注中遇到的问题以及我们的解决方案。

4.1. 多核心关系的处理

篇章结构中存在涉及两个或多个单元、且各单元重要程度相等的多核心关系，如 comparison、joint、same-unit 等。然而，使用篇章依存关系表示多核心关系存在一定的困难：多核心关系连接两个或多个单元，而依存关系仅存在于两个篇章单元之间；多核心关系的各个单元应当同等重要，而依存关系连接的篇章单元重要程度却不同。这使得多核心关系必须进行变换才能在篇章依存树中得以表示。

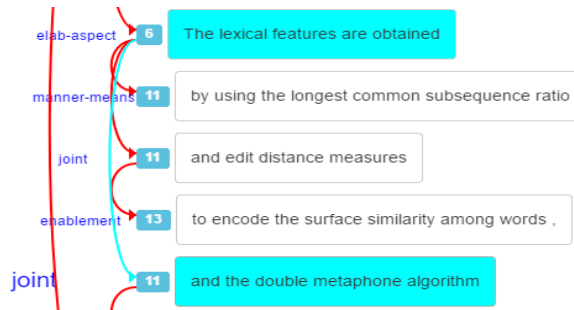


图1 多核心关系（joint）的处理示例

我们的处理方式为：选择多核心关系内部或附近的某一篇章单元作为其余篇章单元的头部

¹ <http://123.56.88.210/demo/depannotate/>

部单元，以表达多核心关系。头部单元的选择视情况而定——可以是多核心关系中相对较重要的一个篇章单元，可以是多核心关系中的第一个篇章单元，也可以是多核心关系前紧邻的一个篇章单元。如图 1 中，我们选取了第一个篇章单元即图中第一句作为头部单元，和其他两个篇章单元（第三句和第五句）构成 joint 关系。这种方式很好地克服了多核心关系与依存结构的矛盾，使多核心关系得以在篇章依存树中得到表示。

4.2. 依存标注中的话题链问题

本文选择了 26 种依存关系用于篇章依存树库的标注。大多数关系都容易区分，但是也有一些特殊情况需要单独处理，比如汉语标注中遇到 elab-addition 和 elab-aspect 这两种关系，前者是对核心句主要内容的进一步阐述，后者则是对核心句提到的不同方面进行阐述。由于英文中存在从句结构，elab-addition 和 elab-aspect 较为容易区分，但在汉语中，这两个标签有时难以区分，考虑表达的“重心”和篇章性，我们选择用话题链^[13]来解决这个问题。关于话题的说明可以参考赵元任的《汉语口语语法》^[14]。

若附属单元是头部单元的进一步阐述且属于同一话题链，则关系标注为 elab-addition，若不属于同一话题链则标注为 elab-aspect。

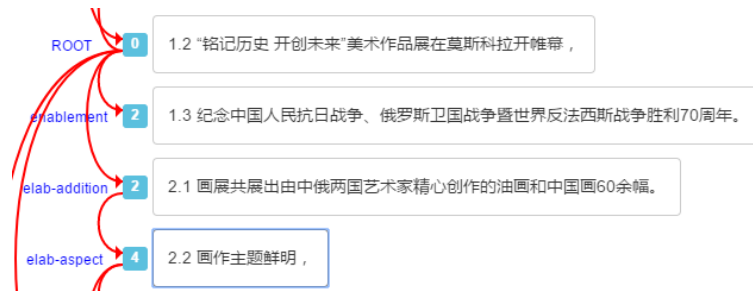


图 2 话题链区别 elab-addition 和 elab-aspect 示例

如图 2 所示，1.2 句的话题是作品展，2.1 句的画展是作品展的回指，则 1.2 和 2.1 两句之间的关系属于同一话题，因而标注为 elab-addition。而 2.2 句的话题是画作，则和 2.1 的关系标注为 elab-aspect。

需要说明的是，我们起初试图用汉语研究中的语句重心^[15]来进行区别以上两种关系的标注，但当前汉语中语句重心的研究认为，类似 1.2 句的陈述句，其重心靠后，也就是“拉开帷幕”这一事实，但联系上下文不难发现，“作品展”才是串起篇章，联结上下文的关键，因此最后采用了篇章性更强的话题的概念对关系进行区分。

4.3. 长篇章及复杂篇章的标注

篇章依存结构相对 RST 结构没有中间节点，复杂度相对较低。但本质上它仍然是一种层次结构，对每一个篇章关系的判断，都需要考虑到上下文信息，也就是其他篇章依存关系对它的影响。因此不能简单地从每个篇章单元独立地去考虑，也就不容易按线性顺序给每个篇章单元确定头部单元及关系。尤其是规模大、复杂度高的篇章，标注过程更为艰难。

对此，本文采用了“自顶向下”和“自底向上”相结合，并兼顾考虑篇章自然段划分的方法，标记长篇章及复杂篇章。在“自顶向下”的过程中，我们首先找到包含篇章中心思想的、最重要的篇章单元，令其头部单元为根结点。然后，找到包含各自然段中心思想的重要篇章单元，并标注它们之间的依存关系。接着，我们再去寻找包含更小范围篇章片段的中心

思想的篇章单元，层层向下进行标注。在“自底向上”的过程中，我们运用层次结构的思想，从篇章结构底层的某一篇章单元着手，不断将其周围更多的篇章单元纳入考虑，层层向上进行标注。

标注过程中，篇章的自然段划分是重要的辅助参考指标。多数情况下，篇章是在自然段内部先形成依存关系后，再与自然段外部形成依存关系的。两个自然段间通常只存在一个依存关系。两种过程可以交替进行，在某一过程中遇到瓶颈难以继续时，则切换到另一标注过程中，继续标注。二者交替进行有助于对篇章进行分块、分层，能够增进对篇章的宏观把握，使依存关系标注更加快速、准确。

4.4. 层次结构信息的损失

篇章依存树一定程度上对 RST 结构作了简化，这虽然降低了标注的难度，但也使其可能缺失了部分篇章层次信息。我们发现，当头部单元两侧各有一个附属单元时，依存树反映了两种可能的 RST 结构，参考论文^[10]，无法判断篇章单元 e_2 与 e_1 、 e_3 中哪个的关系更为密切，一定程度上损失了篇章层次信息。

这种结构在各种篇章中十分常见。例如，图 3 所示 4.1、4.2、4.3 之间的依存结构，可能对应着图 4 和图 4 所示的两种 RST 结构 A 和 B。我们知道，因为 4.1 说明了 4.2 及 4.3 的引用来源，结构 A 反映了作者的写作意图和该篇章的篇章结构。然而，在没有额外知识和信息的情况下，我们根本无法做出上述判断。也就是说，用这种篇章依存树表示篇章结构出现了信息损失。在未来的工作中，我们将考虑如何处理这种情况。

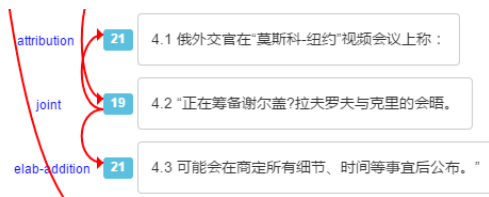


图 3 层次结构信息损失示例

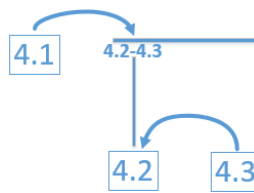


图 4a RST 结构 A

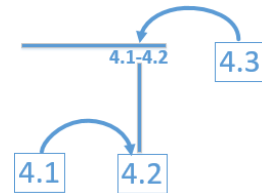


图 4b RST 结构 B

4.5. 篇章单元划分错误

在使用标点符号作为划分篇章单元的标志时，若不进行特殊规定，在部分情况下会出现错误。篇章中出现的汉语冒号，在提起下文、引用话语或总结上文时，可作为篇章单元的划分标志。但在篇章撰写过程中，可能出现将冒号用为“比号”的情况。例如，“尼泊尔制宪会议 16 日晚以 507:”和“25 的压倒性票数表决通过了新宪法草案”，由于根据标点符号划分篇章单元，则被错误地划分为了两个篇章单元。再例如，“9 月 16 日 21 时 08 分在台湾宜兰县附近海域（北纬 24.3 度，”和“东经 121.9 度）发生 5.4 级地震”，括号中的逗号导致篇章单元划分的不合理，括号内的内容被分拆进了 2 个篇章单元中。更好的划分方式为分成 1 个篇章单元（即不划分）或 3 个篇章单元（括号前、括号及括号内、括号后）。

在现有的划分标准下，篇章单元的自动划分存在错误。本文选择在篇章单元自动划分后，再进行一遍人工修正与校对，以排除自动划分造成的问题。

4.6. 一致性问题

本文篇章依存树库的构建及篇章依存关系的标注均只由一人完成。为了提高标注语料的一致性，标注者对每一篇语料均进行了两次标注。两次标注有一定的时间差。最后，再对两

次的标注结果进行对比和分析，对不一致的标注进行修改，得到最终的标注结果。这一方法一定程度上弥补了单人标注的缺陷，提高了标注语料的一致性与篇章依存树库的质量。

5 篇章依存树库的统计

此次标注的语料库总计 138 篇文献，共对 2044 个篇章单元进行了依存关系标注。其中英文文献中最长的篇章单元有 39 个单词，平均长度为 9 个单词；中文新闻中最长的篇章单元有 70 个汉字，平均长度为 13 个汉字。由于每个依存树有且只有一个 ROOT 节点，且只有 ROOT 节点与文章核心句之间的关系被标注为 ROOT，故表中 ROOT 标注的数量等同于依存树的数量。

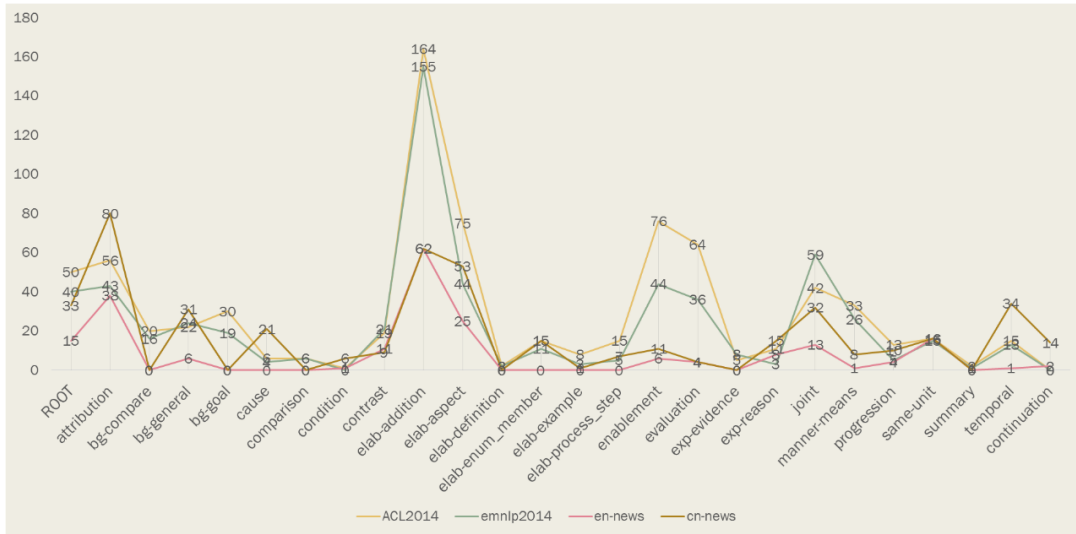


图 5 篇章依存关系统计

图 5 为篇章关系在中英文和新闻、科技论文摘要两个领域上分布的折线图，虽然不同语料上的标注基数不同，但其起伏的趋势是类似的。只有个别依存关系的标注上差别较大，接下来会就这些有代表性的关系标注频率情况做一些对比分析。由于目前篇章数量较少，和大规模数据统计相比，以上数据可能会出现一些偏差。

5.1. 中英文新闻标注

中文新闻来自人民网，总长度 6006 字，平均长度为 182 字。英文新闻数据则来自华尔街见闻，全部都是经济领域，共计 1656 个单词，平均长度为 110 个单词，且标注难度较大。

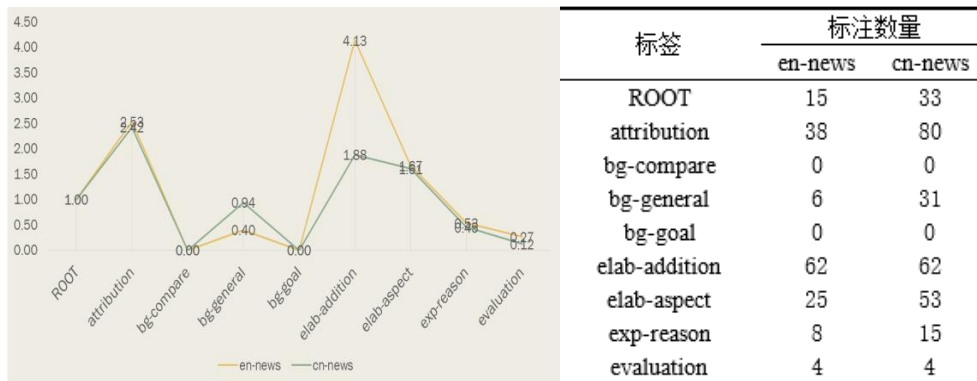


图 6 中英文新闻部分篇章依存关系分布

图 6 给出了中英文新闻上的篇章依存关系分布,其中折线图是依存关系的平均出现次数,在新闻标注中,中文新闻中一般性背景介绍的 **bg-general** 关系使用频率比英文新闻高;而英文的 **elaboration** 关系使用得较多,其中比例上更侧重于 **elab-addition**,中文的 **elab-addition** 和 **elab-aspect** 比例较为均匀,也就是说在对核心的阐述方面,英文更倾向于深入说明,中文则更倾向于泛泛说明。整个篇章来说,英文新闻的核心与附属单元之间联系相对更紧密。另外中英文新闻中出现了较多的 **attribution** 关系,对此我们在标注过程中也有直观感受,新闻常常会引用或者转述权威的分析、看法或者报道。

尽管使用频率有区别,但可以看到不管中文还是英文新闻,其整体的篇章结构,还是以 **bg-general**—**attribution**—**elaboration** 这样的结构为主。

5.2. 英文科技文献摘要标注

英文论文摘要包括 ACL2014 的 50 篇和 Emnlp2014 的 40 篇,共计 12300 个单词,平均每个文档的长度约为 130 个单词。

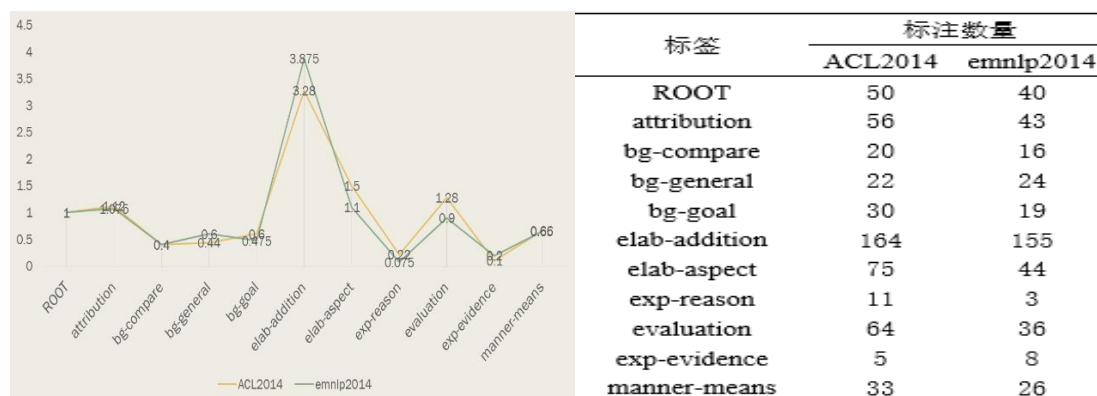


图 7 中英文科技文献篇章依存关系分布

图 7 给出了中英文科技文献篇章依存关系的分布情况,其中折线图是依存关系的平均出现次数,可以看到这些标签的使用情况基本相同,说明英文科技文献摘要在写作上基本有比较统一的规范。其中三个背景关系使用比较平均,还可以看到大量的 **elab-addition** 关系,用于说明其算法或是方案。**evaluation** 关系用于标示出摘要的评测部分,这些依存关系的使用情况也符合论文摘要的写作目的和要求。

5.3. 新闻与科技文献摘要对比

综合考虑新闻和科技文献摘要的依存关系出现频率可以发现一些明显的区别,图 8 是所有科技文献摘要和中英文新闻中依存关系平均出现次数的对比。

首先是新闻中出现的 **attribution** 要远多于论文摘要,正如之前提到的,一篇新闻中常有多次转述和引用,而论文摘要的 **attribution** 多数情况下只会在 **evaluation** 部分出现一次。在科技文献中 **background** 关系的使用更为丰富,可能为了某种目的提出全新的方法则用 **bg-goal** 关系,或者和旧方法进行比较则用 **bg-compare** 关系,而新闻中的背景内容则相对单一一些,主要采用 **bg-general** 关系,目的或比较的背景信息较少。我们还可以看到英文论文摘要和英文新闻的 **elaboration** 关系出现频率相当一致,而与在中文新闻中的使用存在区别,这意味着该关系的使用差别并非文体原因而是语言原因。大量的 **elab-addition** 表示解释或说明一个话题,而英文又相当依赖其从句结构。论文摘要与新闻还有一个显著区别在 **enablement** 和 **evaluation** 关系的频率上,科技文献摘要中告诉读者某项措施、某种方案的采

用，其目的是什么，而之后基本也一定会对文献提出的方法进行一个评价或是评测，因此科技文献中对于 enablement 和 evaluation 这两种关系的使用更为频繁。

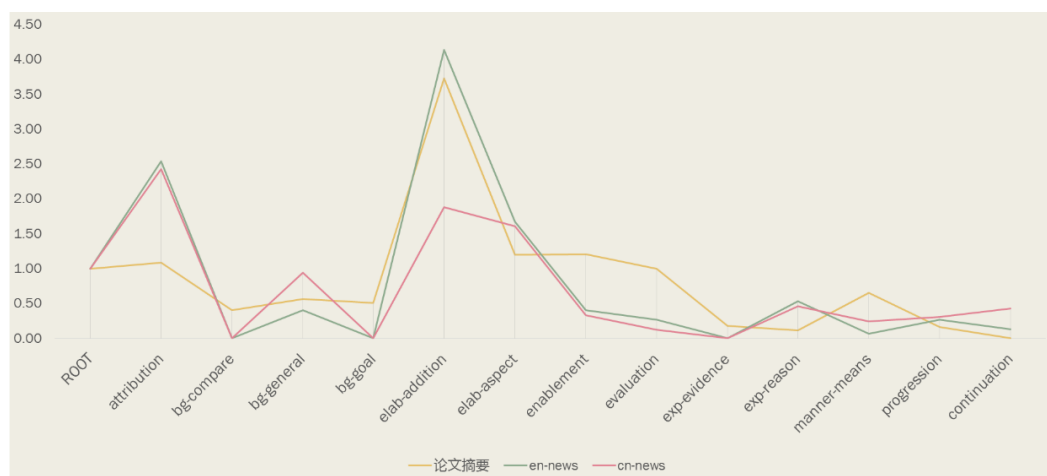


图 8 新闻与论文摘要部分篇章依存关系分布

综合来说，英文科技文献摘要内容更丰富，逻辑性也比新闻更好，其基本上有一个结构 bg-goal/bg-compare—(elaboration—enablement)—evaluation，即大致分成三个部分：背景，说明和评测。结合 5.1 对新闻的分析来看的话，英文科技文献摘要更倾向“线性”叙述，即围绕说明部分，说清其前因（背景）和后果（评测），中文新闻则倾向于叙述事件所涉及的多个方面。

6 结语和未来工作

本文建立了一个小规模的中英文篇章依存树库，并针对标注过程中遇到的困难进行了分析研究，给出了解决方案。其中对于多核心关系在篇章依存树中表示的问题，设立了依存关系选择的规范，采用了“自顶向下”与“自底向上”相结合的标注方法，研究了层次结构信息损失和非投射结构，对篇章单元的错误划分进行了人工修正与校对，并提高了单人标注语料的一致性。同时通过统计已构建的中英文篇章依存树库中的关系分布，简单分析中英文在科技文献摘要和新闻两个领域上的篇章现象

本文的探索，为未来篇章依存关系分析与标记的研究，指出了一些可供研究或改进的方向，其中包括：(1) 扩充语料库的篇章数及篇章平均字数。现有的语料库规模较小，得到的结果在统计学视角下意义有限。(2) 扩充语料库的种类。标注除新闻外的其他语料类型，例如书信、小说、广告、剧本等。(3) 研究如何解决篇章依存树略简化的层次结构带来的层次结构信息损失。(4) 增强标注的一致性。标注数据的人员由一人改为两人或多人，相互对照与校对，提高标注的一致性与准确率。

感谢

感谢三名匿名评审专家的宝贵建议。感谢王亮同学开发了自动篇章标注的工具，为本工作的进行提供了支持。

参考文献

- [1] Louis A, Joshi A, Nenkova A. Discourse indicators for content selection in summarization[A]. Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on

-
- Discourse and Dialogue[C]. Tokyo, Japan: Association for Computational Linguistics, 2010, 147-156.
- [2] Verberne S, Boves L, Oostdijk N, et al. Discourse-based answering of why-questions[J]. *Traitement Automatique des Langues, Discours et document: traitements automatiques*, 2007, 47(2): 21-41.
- [3] Webber B, Stone M, Joshi A, et al. Anaphora and discourse structure[J]. *Computational linguistics*, 2003, 29(4): 545-587.
- [4] Prasad R, Dinesh N, Lee A, et al. The Penn Discourse TreeBank 2.0[A]. *Proceedings of the International Conference on Language Resources & Evaluation[C]*. Marrakech, Morocco: LREC, 2008, 2961-2968
- [5] Miltsakaki E, Prasad R, Joshi A K, et al. The Penn Discourse Treebank[A]. *Proceedings of the International Conference on Language Resources & Evaluation[C]*. Lisbon, Portugal: LREC, 2004.
- [6] Carlson L, Marcu D, Okurowski M. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory[J]. *Sigdial Workshop on Discourse & Dialogue*, 2001, 18(18):1-10.
- [7] Mann W C, Thompson S A. Rhetorical structure theory: Toward a functional theory of text organization[J]. *Text-Interdisciplinary Journal for the Study of Discourse*, 1988, 8(3): 243-281.
- [8] 张牧宇, 秦兵, 刘挺. 中文篇章级关系体系及类型标注[J]. *中文信息学报*, 2014, 28(2): 28-36.
- [9] Li Y., Feng W., Kong F., Sun J. and Zhou G. Building Chinese discourse corpus with connective-driven dependency tree structure[A]. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing [C]*. Doha, Qatar: EMNLP, 2014, 2105-2114.
- [10] Li S, Wang L, Cao Z, et al. Text-level Discourse Dependency Parsing[A]. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)[C]*. Baltimore, USA: Association for Computational Linguistics, 2014, 25-35.
- [11] Polanyi L. A formal model of the structure of discourse[J]. *Journal of pragmatics*, 1988, 12(5): 601-638.
- [12] Grosz B J, Sidner C L. Attention, intentions, and the structure of discourse[J]. *Computational linguistics*, 1986, 12(3): 175-204.
- [13] 屈承熹. 汉语篇章语法. 潘文国等译[M]. 北京: 北京语言大学出版社, 2006: 248-249.
- [14] 赵元任. 汉语口语语法. 吕叔湘译[M]. 北京: 商务印书馆, 1979: 45-47.
- [15] 杨晓宇. 句子的表达重心及其与相关概念的关联[J]. *宁夏大学学报(人文社会科学版)*, 2015, 37(4): 8-13.

吴永芑(1995—), 本科生, 主要研究领域为自然语言处理。Email: 1300012963@pku.edu.cn
李素建(1975—), 通讯作者, 副教授, 主要研究领域为自然语言处理。电话: 62753081-105,
email: lisujian@pku.edu.cn

秦沐坤 (1994—), 本科生, 主要研究领域为自然语言处理。Email: qmk@pku.edu.cn

杨安 (1994—), 硕士研究生, 主要研究领域为自然语言处理。Email: yangan@pku.edu.cn

王厚峰 (1965—), 教授, 主要研究领域为自然语言处理。Email: wanghf@pku.edu.cn