

从语义标注文本中推定普罗普的功能项

[美]马克·阿兰·芬雷森 著 张瑞娇 李扬 译

摘要:弗拉基米尔·普罗普的《故事形态学》是民俗学的一项开创性研究,也是计算研究中一个引人注目的课题。我探讨了一种从语义标注文本中学习普罗普功能项的技术,并对普罗普语料库中的15个民间故事进行了语义角色、语义同指、时间结构、事件情感和角色的标注。之后我从普罗普提供的描述中导出了一组合并规则。当这些规则与模型融合学习框架的改进版本相结合时,能够很好地重现普罗普的功能项。三个重要的功能组——即A/a(加害/缺失),H/I(交锋/战胜),以及W(回报)——被精准地识别出来。这是计算系统学习一种真实的叙事结构理论的首次论证。

关键词:人工智能;计算语言学;民间故事;功能主义;语义学

DOI:10.13370/j.cnki.fs.2019.04.012

引言

弗拉基米尔·普罗普(Vladimir Propp)的《故事形态学》一书出版于1928年,1958年首次被翻译成英文。^①这是民俗学的一部开创性著作,引领了结构主义时代,为后来的民间故事叙事结构研究提供了范例,也启迪了一代又一代的民俗学家。普罗普的形态学是迄今为止对叙事结构最精确的表述之一,它提出了一个引人注目的机器学习课题。如果能够从一组给定的民间故事中自动地、可靠地提取形态,这将会引起广泛的兴趣。对民俗学家和文学理论家而言,这种工具将会是进行比较、索引和分类的无价之宝。对文化人类学家而言,它将为研究文化及其跨时空变化提供一种新技术。对文化心理学家而言,它将为探究文化及其对思想的影响的新实验指明方向。对认知科学家而言,它可以作为理解文本抽象和叙事理解本质的模型。对计算语言学家而言,它将推进对自然语言更高层次意义的理解。对研究人工智能和机器学习的人而言,它代表了我们从复杂数据集当中提取深层结构的能力的进步。当然,在每个领域中也都可以发现其他领域取得的相关进展。

然而,直到现在,形态的提取仍旧依靠人工,这类学者如A. J. 格雷马斯(A. J. Greimas),克洛德·列维-斯特劳斯(Claude Lévi-Strauss),阿兰·邓迪斯(Alan Dundes),以及弗拉基米尔·普

作者简介:马克·阿兰·芬雷森(Mark Alan Finlayson),美国佛罗里达国际大学计算机科学助理教授;张瑞娇,中国海洋大学文学与新闻传播学院硕士研究生(山东青岛266100);李扬,中国海洋大学文学与新闻传播学院教授(山东青岛266100)。

基金项目:本文系国家社科基金重大项目“海外藏珍稀中国民俗文献与文物资料整理、研究暨数据库建设”(项目编号:16ZDA163)子课题的阶段性成果。

原文载《美国民俗学刊》(AFS)2016年129期。此译文的翻译和发表得到作者授权,但作者本人并未核查译文准确性,特此说明并向作者致谢。

① 弗拉基米尔·普罗普:《故事形态学(第2版)》,劳伦斯·斯科特译,奥斯汀:得克萨斯大学,1968年。

罗普。^①为一组特定的民间故事构建形态需要多年的阅读与分析。目前还不确定已经完成的形态研究中,有多少是源于民俗学家的个人偏好或其他现存形态的熟悉,而不是通过调查对故事性质做出的正确反映。此外,盲目地对形态分析进行再现或验证是一项异常艰巨的工作,这需要具备必要技能的学者,来回溯人工生成故事形态所需的长达数年的阅读、分析与合成的过程。

我展示了一种技术,可以用计算方式解决从一组给定的故事中识别出形态的问题。该算法是被称为模型融合^②的机器学习技术的改进版,该算法还使用了一组规则,源自普罗普对自身寻找故事间相似性的过程的阐述。在这项技术中,算法将语义标注文本(semantically annotated texts)作为数据运行,并将民间故事的表面语义以计算机可读的表达加以编码。在这个特殊的论证中,数据是普罗普分析的单一回合的(single-move)俄罗斯神奇故事里的一部分,并将之翻译成了英语。值得注意的是,文本表面语义的编码是人工辅助的;而对普罗普功能项特征的实际学习则是由计算机完成的。

本文主要内容如下:第一,我解释了当前机器学习的问题,指出了普罗普理论中我将要重点学习的部分。第二,我描述了所使用的学习技术的结构,以及它与正则模型融合的不同。第三,我阐释了实验中使用的数据,包括文本、语义标注方案以及测量算法性能的黄金标准数据(普罗普的分析)。第四,我列出了一组源于普罗普的描述的合并规则,它在模型融合框架内工作,以重现普罗普的大部分功能项。最后,我阐释了该算法在提取普罗普的功能项指征方面的表现。

学习目标

普罗普的形态学中包括一组人物类别和三级情节结构:总体结构(回合),中级结构(功能)和精细结构(我在本文中称之为亚型;普罗普本人没有给出特定的术语)。登场人物的类别被称为角色,普罗普确定了七种:主人公,对头,公主,差遣者,赠与者,相助者和假冒主人公。由功能项组成的单一故事是不成熟的,一个标准的故事往往是由一个或多个回合组成,它们可能还会以复杂的方式相互交织。功能是一种情节元素,是“从其对于行动过程意义角度定义的角色行为”^③。每个功能都属于一种主要的类型,这由它在一个回合中的位置、情节的目的、以及所涉及的角色来确定。普罗普识别出了31种不同的功能项。每个功能项对应正在发生的事情,但是不一定能指出事情是如何发生的——也就是说,功能项可以通过许多不同的方式例示,这就是我所说的功能项的亚型。

在单词的形式及计算的意义上,普罗普的情节结构定义了一种语法。在这项研究中,我努力从文本本身学习这种语法的某些部分。正如我们从文法推理^④中所知道的,语法力量影响了语法学习的难度。那么普罗普的语法有多强呢?

普罗普将故事的最高级结构定义为可选择的先在序列,其后是一些可能相互交织的回合。这个级别的语法复杂性至少是上下文无关的(context-free),这与拉科夫的分析一致^⑤,自然是一种相当强大的语法。中级结构是一种正则文法,其中功能项要以受到限制的顺序出现,它比上下文

① 阿尔吉尔达斯·朱利安·格雷马斯:《结构语义学:方法研究》,巴黎:拉鲁斯,1966年;克洛德·列维-斯特劳斯:《神话与意义》,纽约:劳特利奇,1978年;阿兰·邓迪斯:《北美印第安民间故事形态论》,《民俗学者通讯》第195期,赫尔辛基:芬兰科学院,1964年。弗拉基米尔·普罗普:《故事形态学(第2版)》,劳伦斯·斯科特译,奥斯汀:得克萨斯大学出版社,1968年。

② 安德烈亚斯·斯托克、斯蒂芬·奥莫亨德罗:《由贝叶斯模型融合推导概率文法》,拉斐尔·C·卡拉斯科、何塞·翁西纳:《文法推理与应用》,柏林:斯普林格,1994年,第106-118页。

③ 弗拉基米尔·普罗普:《故事形态学(第2版)》,劳伦斯·斯科特译,奥斯汀:得克萨斯大学,1968年,第21页。

④ 伊格拉·科林德拉:《文法推理:学习自动化与语法》,剑桥:剑桥大学出版社,2010年。

⑤ 乔治·拉科夫:《神奇故事的结构复杂性》,《人的研究》第1卷,加利福尼亚州尔湾:加利福尼亚大学社会科学学院出版,1972年,第128-150页。<https://georgelakoff.files.wordpress.com/2010/12/structural-complexity-in-fairy-tales-lakoff-1972.pdf>

无关文法弱,因此更容易学习。亚型级则可以在故事弧内产生长期影响,因为在一个故事中,早期对特定亚型的选择(例如 A,加害行为是绑架)会影响后来对特定亚型的选择(例如 K,解决方案是对被绑架者的救助)。这种亚型的影响增加了额外的复杂性,但可以采取特征文法^①或广义短语结构语法^②的形式并入到功能级正则文法(或回合级上下文无关文法)中。因此,抛开角色不谈,普罗普理论的整体文法,至少是其广义短语结构语法(GPSG),确实有很高程度的复杂性。

目前,我们还没有可以同时学习普罗普 GPSG 的字母表、转换及角色类别的计算技术。即使给出了角色,学习 GPSG 也仍旧十分困难。因此,在本文中,我只集中学习普罗普功能项的指征,并指出可以被看作普罗普最突出贡献的功能项类别。几乎所有其他内容都是参考功能进行定义的:回合是功能的复合体,亚型是对功能的调整,角色也部分地由其所参与的功能来定义。大多数以普罗普为基础的民俗学和计算工作都集中在功能层面上。^③

我把以下内容留待将来研究:角色类别,功能亚型类别,回合级文法,以及功能级正则文法的转换结构。在本文中,我的关注点仅在研究功能项类别上,相当于只是学习功能级正则文法的字母表。由于使用已知的字母表学习正则文法是一个颇具吸引力的问题,我利用这项工作为学习正则文法的字母表构建了一种新算法。

学习技术

模型融合是一种从正例中学习正则文法的自动化技术^④,这是我的研究方法的概念基础。我的技术采用了模型融合,并扩充了两个关键性内容。第一,虽然模型融合假设语法的字母表是已知的,但学习普罗普形态学的一个主要挑战在于学习功能项本身的指征。为了达到这个目的,我从一个非常大的可能性字母表开始,并在最后加入一个筛选阶段,用以从最终模型中识别真正的字母。第二,尽管模型融合认为模型状态(model states)是相对微小的,且模型状态发出的符号只有一种概率分布,但是我的技术在进行融合时,考虑到了每个模型状态丰富的内部结构(源于文本上的语义标注)。

模型融合可用于从一组正例中导出正则文法。如,两个字符序列的集合{ab, abab},最简明地描述这两个序列的模式是什么?一种猜测是正则文法(ab|abab),确切地说,是第一个或第二个字符串。然而我们觉得这种猜测并不令人满意,因为它没有超出所提供例子的范围。大家都能发现,更合理的猜测是子字符串 ab 重复了一次或多次,或者写成一个正则文法表达式:(ab)+。模型融合是一个框架,它能让我们找到这种模式的良好近似值;我们所需要的只是一种搜索可能的语法空间的方法。

模型融合遵循文法推理范式,该范式始于一个模型,其建构目的在于接受由观察而来的正例

① 乔舒亚·古德曼:《概率特征文法》,哈里·邦特,安东·尼霍特:《概率论和其他解析技术的进展》,多德雷赫特:斯普林格,2000年,第63—84页。

② 加兹达尔·杰拉尔德、伊万·克莱因、杰弗里·K·普卢姆、伊万·A·萨格:《广义短语结构语法》,牛津:巴兹尔·布莱克韦尔出版公司,1985年。

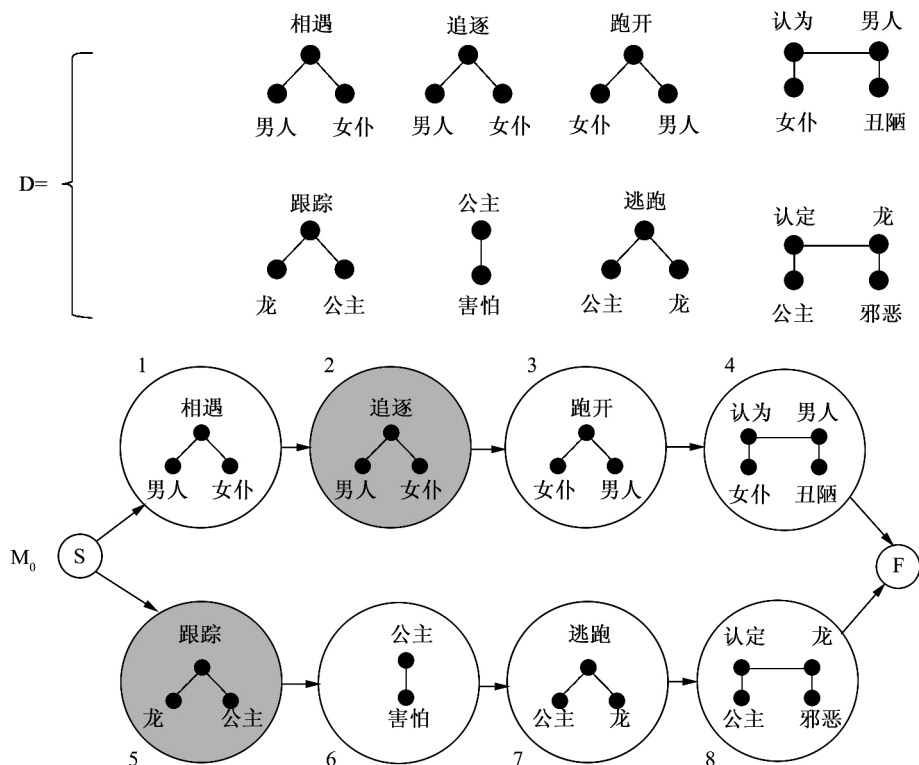
③ 例如,阿兰·邓迪斯:《北美印第安民间故事形态论》,《民俗学者通讯》第195期,赫尔辛基:芬兰科学院,1964年;本杰明·科尔比:《爱斯基摩民间故事的部分语法》,《美国人类学家》1973年第75卷第3期,第645—662页;贝伦·迪亚斯—阿古多、巴勃罗·赫瓦斯、费德里科·佩纳多:《基于案例推理的故事情节生成方法》,《案例推理欧洲会议(ECCBR)论文集》,马德里,2004年,第142—156页;哈里·哈尔平、约翰娜·穆尔、朱迪·罗伯逊:《故事改写情节的自动分析》,《自然语言处理实验方法会议(EMNLP)论文集》,巴塞罗那,2004年,第127—133页。

④ 斯蒂芬·M·奥莫亨德罗:《动态学习与识别的首个最佳模型融合》,约翰·E·穆迪、斯蒂芬·J·韩森、理查德·P·李普曼:《神经信息处理系统研究进展5》,加利福尼亚(圣马特奥):摩根考夫曼,1992年,第958—965页;安德烈亚斯·斯托克、斯蒂芬·奥莫亨德罗:《由贝叶斯模型融合推导概率文法》,拉斐尔·C·卡拉斯科、何塞·翁西纳:《文法推理与应用》,柏林:斯普林格,1994年,第106—118页。

组成的有限语言。^① 通过对模型中的状态进行合并操作以实现一般化,其中两种状态从模型中被移除并被替换为单一状态,后者会继承前者转换与发出的内容。这种合并操作催生了一个很大的模型搜索空间。

为了说明我的技术,图1展示了如何从两个非常短的故事中提取一个简单的形态。编写这些故事也是为了说明该技术。第一个故事是关于一个老人和女仆:他们在路上相遇,他追逐她,她跑开了,最后她认为他是一个丑陋的男人。第二个故事是关于一条龙和一位公主:龙跟踪公主,这让她感到害怕,所以她逃跑并躲了起来,最后她认定龙是邪恶的生物。在某种抽象层面,这两个故事是相似的。追逐与跟踪事件相似,因为它们涉及一个参与者跟随另一个参与者;跑开与逃跑事件相似,因为它们涉及一个参与者远离另一个参与者的行动;认为和认定则都是涉及评估的心理事件。通过这些事件的语义表示,人们可以使用语义距离度量和类比映射算法以发现语义和结构的相似之处。在图1所示的一组合并中,首先被合并的是追逐和跟踪事件,而后是跑开和逃跑项,最后是认为和认定事件。最终的故事形态,可以被看作是一个泛化的故事,故事开头是一个可选择的玩闹事件、一个追逐事件,接着是一个可选择的惊吓事件,而后是一个逃跑和评估事件。一旦最终模型被筛选,只剩下三个状态,它们可能被命名为:追寻(Pursuit),逃离(Flee)和评价(Judgment)。

- (1) 一个老人和女仆在路上相遇。他追逐她,她跑开了。她认为他是一个丑陋的男人。
- (2) 龙跟踪公主,这让她感到害怕。她逃跑,躲了起来。她认定他是邪恶的。



^① 马修·扬-拉伊:《文法推理》,刘玲(音译)、M·塔梅尔·厄兹叙:《数据库系统百科全书》,柏林:斯普林格,2009年,第1256-1260页。

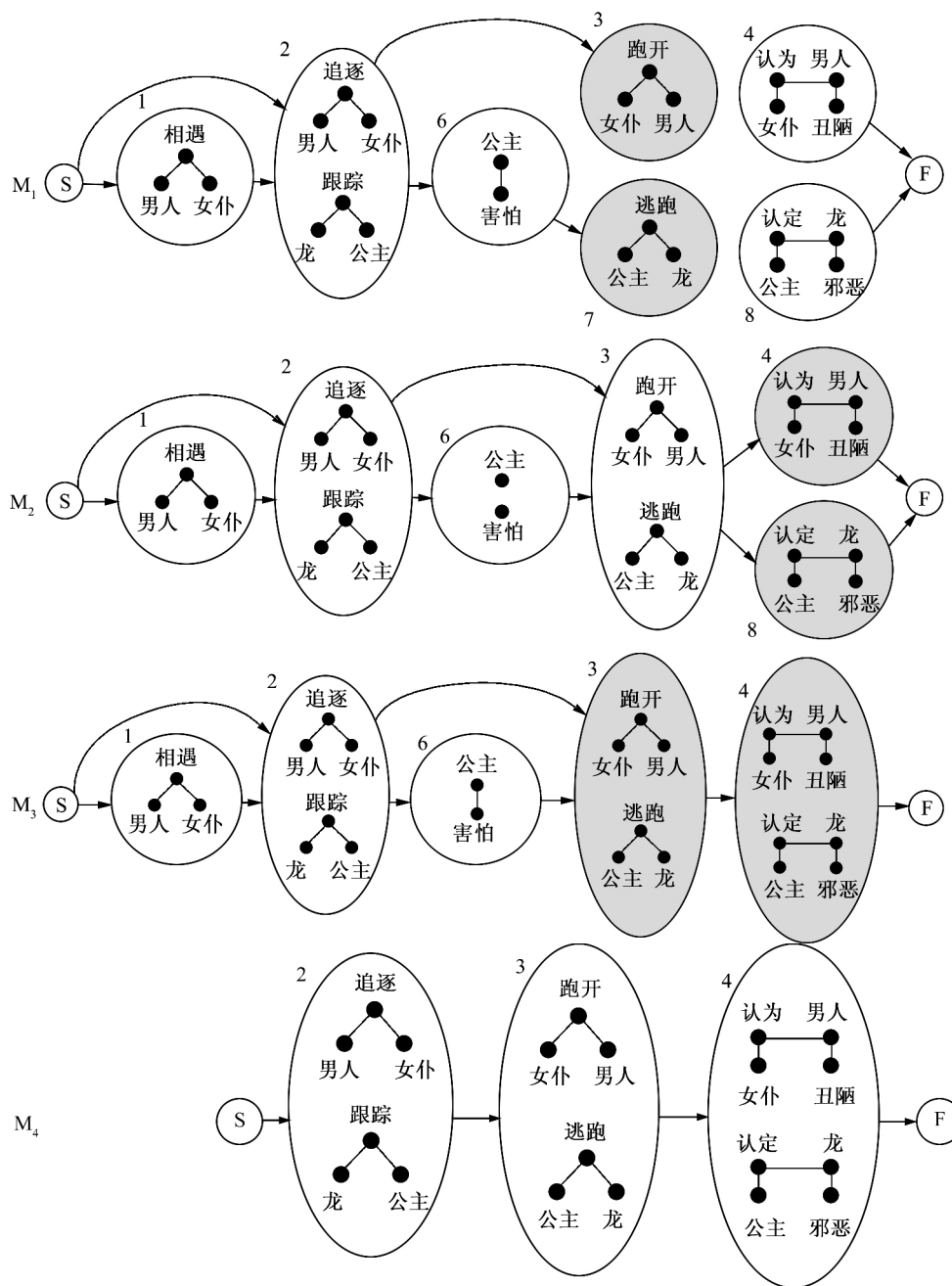


图1 两个简单故事的合并示例。模型 M₃ 不仅描述了两个被输入的故事,而且还增加了另外两个可以包含或排除节点 1 和 6 的故事。因此,这一模型已经在两个输入示例之外实现了一般化。由筛选步骤产生的模型 M₄ 即代表最终形态。

初始模型是从故事世界本身的事件时间线导出的。模型中的每个初始状态都来自各个故事的单一事件,当它们在故事时间线中出现时会被排序。然后,每个单独的故事时间线会作为一条单独的支线并入初始形态中。在图 1 中有一个被标为 M₀ 的初始模型示例,其中,两个简单故事及其各自的四个构成事件都被转换成了一种包含四种状态的序列。

有许多方法来驱动搜索合适的合并集。我曾经在其他研究中探索过一种常见的方式,是由贝

叶斯法则(Bayes's rule)得出的概率来驱动搜索。^① 相比之下,此处描述的工作使用了一组源于普罗普专著中的语义和结构合并规则来驱动搜索。我将在解释了实验运行的数据之后,在标题为“合并规则”的部分对这些规则加以概述。但显而易见的是,我们需要一些规则、启示或偏好来发现一个好的模型:在大多数情况下,穷举搜索是不可能的。^②

筛选阶段

如图1所示,倒数第二个模型(M_3)尚且不是一个形态:它包含的状态与两个故事之间的抽象相似性(即状态2和3)并不对应。这是因为初始模型会以包含各种可能符号的字母表开始。使用筛选步骤则可以从融合模型转变为表现实际形态的模型。筛选过程会在最终的融合模型中构造另一个模型,从中移除所有不符合特定条件的状态。筛选后剩下的状态成为普罗普的语法或功能项的字母表。有关此筛选过程的详细信息,请参阅下文的“合并规则”部分。

数据

普罗普选择了一组特定的故事来分析并导出了他的形态学:亚历山大·阿法纳西耶夫俄罗斯神奇故事集的前一百个故事。^③ 普罗普在他的附录Ⅲ中,提供了他所分析的约一半故事的功能图式:在普罗普作品的英译本中,功能表内有45个故事,整个文本中散布着少量的附加分析。在本文中,我不打算学习回合级语法、亚型语法以及角色类别。这种范围限制了数据准备的特定方法。首先,回合级上的异文被筛选,只留下普罗普认定为只包含单一回合的故事。其次,学习数据明确包含对人物角色类别的识别。

由于我将范围限定在单一故事中,所以普罗普分析的45个故事中可用的故事减少了一些;在普罗普形态学的几个译本中,我发现总共有21个单一故事包含了功能分析。我的研究预算进一步限制了我对这一组故事的详尽语义标注。最后,我留下了共计18862个单词的15个单一故事,对此我完全能够进行详细标注。

此外,虽然普罗普因现实原因在研究中采用了故事的原始语言(俄语,有时是白俄罗斯语或乌克兰语),但我使用英文翻译进行了我的分析。民俗学家有时也会研究被翻译过的故事,并且大家的共识是,对最初的结构语义分析而言,故事的重要信息应保留在一个良好的译文中。正如J. L. 费希尔(J. L. Fischer)所说:“如果一个人将故事翻译成另一种语言,那故事的结构和故事图像的基本特征应该保持原貌。”^④

语义标注

我在这里所使用的“标注”一词与语料库语言学相同,它涵盖了“所有应用于原始语言数据的描述性或分析性标记”^⑤。自动生成本文所需的多方面高质量语义标注超出了当前自然语言处理(NLP)的技术范围。因此,为了实现高质量、低误差的语义标注,我需要雇用人力,来更正自动生成的标注(即所谓的半自动标注)或从一开始就提供完全的人工标注。虽然这很慢而且花费不菲,但进行半自动或人工标注的好处是,我们可以获得尚且无法自动创建的高质量标注。因此,虽然

① 马克·阿兰·芬雷森:《从被标注的民间故事中学习叙事结构》,麻省理工学院博士学位论文,2011年。

② 对于非平凡起始故事,模型融合搜索空间变得太大而难以管理:它相当于贝尔数 B_n ,其中 n 是模型中的初始状态数(罗塔,1964)。当 n 增大时,贝尔数也会迅速变大。例如,当 $B_2=2$ 时, $B_3=5$,当 $B_{10}=115975$ 时, $B_{35}\approx 3.59e+31$ 。

③ 亚历山大·N·阿法纳西耶夫:《俄罗斯民间故事》3卷本,莫斯科,国家艺术出版社,1957年。请注意,普罗普使用的是阿法纳西耶夫故事集的旧版。为方便起见,我们在文中提供了更现代的引文。

④ J. L. 费希尔:《民间故事的社会心理分析》,《现代人类学》1963年第4卷第3期,第249页。

⑤ 史蒂文·伯德、马克·利伯曼:《语言标注的形式框架》,《语言通信》2001年第33卷第1—2期,第23—60页。

对普罗普功能的学习是通过机器完成的,但研究的原始数据“文本的形式化语义”基本上是由人工产生的。

标注者为这项工作进行的所有自动、半自动或人工标注都是使用 Story Workbench 标注工具完成的。^① Story Workbench 是一种通用的文本标注工具,支持多层语义标注,提供容易操作的图形用户界面,并支持对任意文本的标注。表 1 中列出了标注层次。普罗普的形态建立在人物和事件结构之上,即什么时候谁在对谁做什么:我称之为文本的“表面语义”。每个列出的层次都是从每个文本中提取表面语义的关键。

表 1 本文中使用的标注。一致性被同时表示为 F_1 度量或一个偶然性校正兰德指数 (chance-adjusted rand index)。 F_1 度量的范围从 0(无一致性)到 1(完全一致)。兰德指数范围从 -1(完全不一致)到 1(完全一致)。

层次	语义捕捉	标注方式	一致性
指称表达式	世界上的事物	人工	0.91
语义同指	可语义同指的指称表达式	人工	0.85*
时间表达式	时间和日期	人工	0.66
事件	发生的事情与状态	半自动	0.69
时间连接词	文本时序	人工	0.59
语义角色	动词论元	半自动	0.60 [†]
Wordnet 意义	字典定义	人工	0.78
事件效价	事件对主人公的影响	半自动	0.78
角色	普罗普的人物类型	人工	0.70
功能	普罗普的功能项	人工	0.71 [‡]

* 偶然性校正兰德指数

[†] 仅核心论元

[‡] 部分重叠

指称表达式与语义同指

用于计算故事人物的原始信息由“指称表达”和语义同指标注给出。^② 指称表达的表示 (representation) 标注了指代某些事物的词语集合,其中的单词集合连续与否都可以。这种表示是人工标注的。例(1)展示了指称表达式的三个示例,以下划线标出。

(1) 伊万 有 一把剑。它是锋利的。

在这句话中,指称对象是人和事,是故事世界中的具体事物,但它们并非总是如此。指称表达式还可以指代抽象对象(如想法)、事件、时间、动作、情感和许多其他事物。

例(1)也说明了一个显而易见的要点,即一个单一的指称对象可以在文本中被多次提到。在本例中,一个单一的指称对象(剑)有两个指称表达式(短语“剑”和“它”)。句中的后两个指称表达式是语义同指的,因为它们指代的对象相同。为了使用指称表达式来标注指称对象,同指性指称的表达式集合被汇集在了语义同指的集合之中。因此,语义同指集是一个指代同一类事物的指称表

① 马克·阿兰·芬雷森:《在自然环境中收集语义:Story Workbench》,《人工智能协会秋季学术研讨会论文集:关于自然灵感人工智能》,华盛顿特区,2008年,第46-53页;马克·阿兰·芬雷森:《从被标注的民间故事中学习叙事结构》,麻省理工学院博士学位论文,2011年。

② 拉克尔·埃尔瓦斯、马克·阿兰·芬雷森:《新闻和叙事中描述性指称表达的盛行》,《第48届计算语言学协会年会论文集》,乌普萨拉,2010年,第49-54页。

达式列表。这种表示是人工标注的。

被标注的语义同指集的第二个方面是集合内成员的关系。下面的例(2)展示了一种简单形式,其中的指称表达式“杰克和吉尔”指的是包括杰克和吉尔的集合。该信息对于确定哪些个体角色实际参与了哪些事件非常重要。

(2)杰克和吉尔去了山上。他们取来一桶水。

时间表达式,事件,时间连接词

为了构建故事的时间线,我使用了 TimeML 标注方案。^① TimeML 包含三种表示:时间表达式,事件和时间连接词。前两者会标记居于时间线上的对象,最后一个则定义时间线上各对象的顺序。本节中的示例来自 TimeML 标注指南。^②

时间表达式会标记时态表达式的位置、类型和值。每个表达式都是一个可能不连续的事件符号序列,表明时间或日期、某事持续多长时间或某事发生的频率。时态表达式可以是日期、一天的时间,也可以是持续的一段时间,例如几个小时、几天、甚至几个世纪。时态表达式可以精确,也可以模糊。

(3)龙在中午来了。(时间)

(4)龙在春天的最后一日来了。(日期)

(5)他在地下世界住了将近一年。(一段时间)

有趣的是,在本项研究分析的神奇故事中,时间表达式非常稀少,在 18862 个单词的整个语料库中只有 142 个实例,平均每 1000 个单词只有 7.5 个时间表达式。事实上,大多数故事的时间表达式都不到 10 个,甚至有两个故事都只有一个时间表达式。这可能是因为民间故事通常发生在不确定的日子,或完全在历史之外。不管原因是什么,都说明时间表达式对于整体的时间线并不重要。

事件是居于时间线上的第二类对象。事件被定义为发生的事情或状态。它们可以如例(6)所示立即发生,也可以如例(7)所示持续一段时间。在大多数情况下,达到或适用某些事物的状况被视为事件,如例(8)中的“短缺”。

(6)伊万迅速击中了龙的头。

(7)英雄们前往遥远的国度。

(8)整个国家食物短缺。

事件和时间通过表示时序的连接词衔接在一起。时间连接词分为三大类,包括对两个时间、两个事件、或时间和事件之间的排序,如例(9)和例(10)所示。

(9)伊万的兄弟们在战斗结束之后才到达。(时间——之后)

(10)他在底下住了将近一年。(时间——期间)^③

体连接词(aspectual links)表明了一个事件与它的某个组成部分之间的关系,如例(11)所示。从属性连接词(subordinating links)表明了带论元的事件的关系,如例(12)所示。对从属性连接词出现在开头的事件而言,好的例子是在其论元中加入部分真值条件,或是暗指其论元与未来或可能世界有关。

① 詹姆斯·普斯捷约夫斯基、何塞·卡斯塔尼奥、罗伯特·因格里亚、罗泽·绍里、罗伯特·盖佐斯卡斯、安德烈·塞策、格雷厄姆·卡茨:《TimeML:文本中事件和时间表达式的稳定规范》,《第五届计算语义学国际研讨会(IWCS-5)论文集》,蒂尔伯格,2003年,第193页。

② 罗泽·绍里、杰西卡·利特曼、鲍勃·克尼彭、罗伯特·盖佐斯卡斯、安德烈·塞策、詹姆斯·普斯捷约夫斯基:《TimeML 标注指南》(1.2.1版),2006,http://www.timeml.org/site/publications/timemldocs/annguide_1.2.1.pdf.

③ 例(10)原句为:He lived in the underworld for almost a year.其时间连接词为 for。——译者注

- (11)伊万开始 寻找他的妻子。(体——开始)
 (12)伊万忘了 带上咒语。(从属——叙实性的)

单词意义

词义消歧(WSD)是众所周知的自然语言处理任务,其中每个开放类符号或多词表达(即,每个名词、动词、形容词或副词)会从词义清单中被指定一个单一的意义,这为我们提供了每个词实际意义的指标。^① 为了本项研究,标注者使用电子词典 Wordnet3.0 对每个单词进行了词义消歧。^② 由于大多数 WSD 算法并不比默认的高频词义基准好多少,所以这一标注完全由标注者人工完成。当他们指定单词意义时,还更正了多词表达边界、词性标记、以及词干。虽然 Wordnet 的覆盖面非常广,但有时它也会缺乏一个适当的词义。在这类情况下,标注者会用一个合理的同义词取代原来的词义。在极少数情况下,标注者找不到合适的替代词,则被允许将之标记为“没有可用的适当意义”。

语义角色

标注者还捕捉了文本中所有动词的论元结构,这一任务被称为语义角色标注。具体而言,我们使用了 PropBank 体系。^③ 本项标注是由一个统计语义角色的初级标注器以半自动方式完成的,该标注器的建模基于研究者的文献描述。^④ 这个标注器在文本上运行,为每个动词创建论元边界和语义角色标签。每个动词也被分配了一个 PropBank 框架,它是被承认的语义角色及其描述的列表。这个框架本身是唯一一则未被自动标注的信息,标注者需要添加其框架、所有缺少的论元、语义角色标注,并更正已有的论元边界和标注。与单词意义的情况一样,有时,PropBank 的框架集内并没有适当的框架。这可能在每个文本中发生一两次,在这类情况下,标注者会找到最相近的匹配框架,并以之取代原来的框架。

事件效价

每个 TimeML 事件也因其效价而被标注,旨在获取事件对主人公的正面或负面影响。其标度与温迪·莱纳特(Wendy Lehnert)的积极或消极心理状态类似。^⑤ 但我的标度数值是从-3到+3,并以0(中性)作为潜在效价(potential valence),而不是像莱纳特的表述那样,仅限于正或负。表2中列出了标度范围内每个效价的重要性。这一表示是人工标注的。

表2 效价标度,描述了每种影响的级别,并列举了一些例子

效价	描述	例子
+3	对主人公或其盟友立即有利	主人公与公主成婚;主人公被赠送黄金
+2	可能直接导致一个+3事件	某人在主人公被追捕时将他隐藏起来
+1	某人承诺某件事将会+2或+3	一位老人承诺当主人公最需要的时候他会提供帮助

① 埃内科·阿吉尔、菲利普·埃德蒙兹编:《词义消歧》,多德雷赫特:斯普林格,2007年。

② 克里斯蒂亚娜·费尔鲍姆编:《WordNet:电子词汇数据库》,剑桥麻省理工学院:麻省理工出版社,1998年。

③ 马莎·帕尔默、保罗·金斯伯里、丹尼尔·吉尔德:命题库:《命题库:语义角色标注的语料库》,《计算语言学》2005年第31卷第1期,第71-105页。

④ 萨米尔·普拉丹、卡的里·哈吉奥卢、瓦莱丽·克鲁格勒、韦恩·沃德、詹姆斯·H·马丁、丹尼尔·尤尔基:《支持向量学习在语义论元分类中的应用》,《机器学习》2005年第60卷第1-3期,第11-39页。丹尼尔·吉尔德、丹尼尔·尤拉夫斯基:《语义角色的自动标记》,《计算语言学》2002年第28卷第3期,第245-288页。

⑤ 温迪·G·莱纳特:《情节单元和叙事概述》,《认知科学》1981年第5卷第4期,第293-331页。

续表

效价	描述	例子
0	不好也不坏	
-1	某人威胁称某件事将会-2或-3	女巫以死亡威胁主人公
-2	可能直接导致一个-3事件	主人公与龙交锋
-3	对主人公或其盟友立即不利	公主被绑架;主人公被放逐

角色

普罗普从其民间故事人物中识别出了七种类型,这些人物类型在他的理论中非常重要。如前所述,我打算将角色学习留待将来研究。因此,被标注的角色信息被用来帮助获得形态结构。这种表示包括七个标签:主人公,对头,公主,差遣者,赠与者,相助者和假冒主人公。不论多少,它们都可以附在文本中特定的指称对象上。正如普罗普所指出的那样,在某些情况下,某个人物会扮演多个角色。这一表示是人工标注的。

功能项

最终的标注获取了普罗普的功能项。该标注用作度量学习算法结果的标准。标注普罗普的功能项是一项精细的任务。虽然普罗普非常详细地描述了他的形态,但仍未能在文本中以一种清晰标注的方式加以明确表示。普罗普的专著富有启发性,但并不是一本有效的标注指南。普罗普描述的方案中至少有四个主要问题:布局不清晰;功能项隐含;多重标记(连续重复两次、三次或四次的功能组)不一致;而且,在少数情况下,普罗普自己的分类方案与故事内容之间存在明显分歧。

关于布局不清晰,可以参考下文摘录的阿法纳西耶夫第148号故事:

沙皇亲自去乞求硝皮匠尼基塔(Nikita),希望他能使沙皇的疆域摆脱恶龙的威胁,并能够把公主拯救出来。当时尼基塔正在揉搓皮子,他手里拿着十二块生皮。当他见到沙皇亲自朝他走来,他胆战心惊,双手颤抖起来,把那十二块皮子都扯破了。但是不管沙皇和皇后怎样恳求(entreated)他,他都不肯去对付龙。于是他们召集了五千个小孩子,并派他们去哀求尼基塔,希望孩子们的眼泪会让他产生怜悯之心。孩子们来到尼基塔身边,流着泪乞求(begged)他去和那条龙战斗。尼基塔看到孩子们的泪水,也开始流下(shed)眼泪。他弄来一万二千磅大麻,浇上树脂,一下子全裹在身上,以防止自己被龙吞下,就找龙去了。^①

普罗普表示,在这个故事中存在功能项B和C。普罗普称B为“调停,承上启下的环节”,其定义扩展为:“灾难或缺失被告知,向主人公提出请求或发出命令;派遣他或允许他出发。”^②。他称C为“最初的反抗”,其定义扩展为:“寻找者应允或决定反抗。”^③大体而言,这两个功能项是向主人公呈现任务(B),以及接受任务(C)。

在这段故事中找到这两个功能项并非易事。B到底在哪里?是整段内容吗?是从“恳求”(entreated)一词到“乞讨”(begged)一词之间吗?功能边界应该与句子或段落边界对应吗?小孩的哀求可以看作是B的一部分吗?在识别功能项时,标注者标记了两组符号。首先,他们标记了一个区域,该区域捕捉了一个功能项的大部分意义及范围。这通常是一个句子,但在某些情况下会扩

① 亚历山大·N·阿法纳西耶夫:《俄罗斯民间故事》3卷本,莫斯科:国家艺术出版社,1957年;亚历山大·N·阿法纳西耶夫:《俄罗斯神奇故事》,诺伯特·古特曼译,纽约:帕特农丛书,1975年,第310—311页。

② 弗拉基米尔·普罗普:《故事形态学(第2版)》,劳伦斯·斯科特译,奥斯汀:得克萨斯大学,1968年,第36页。

③ 弗拉基米尔·普罗普:《故事形态学(第2版)》,劳伦斯·斯科特译,奥斯汀:得克萨斯大学,1968年,第38页。

展到一个或更多段落。其次,他们标记了该功能项的定义词,通常是单个的动词形式。如果单个动词或其同义词在紧邻第一个标记的地方重复了,并且指代相同的动作,则这些重复词也会被标记。在上面的例子中,标注者将“不管沙皇和皇后怎样恳求……流着泪乞求他去和那条龙战斗”的部分标记为 B,并将动词“恳求”和“乞求”选为定义词。

C 又究竟在哪里呢? C 是指前往对抗龙的决定。它似乎发生在尼基塔流泪和他获取大麻为战斗做准备之间的某个地方,但这并没有直接用文字表达;也就是说,功能项是隐含的。普罗普提及了发生在故事中的特定功能,但是当标注者无法找到其明确体现时,便会酌情选择逻辑上与之关联最密切的事件并将其标记为前因(Antecedent)或后续(Subsequent)。引文中 C 的区域是句子“尼基塔看到孩子们的泪水,也开始流下(shed)眼泪”,并且“流下”被标记为定义动词。这个隐含的功能项被标记为前因。

当多重标记不一致时,或者当所指示的功能似乎与故事本身不匹配时,标注者会尽力确定正确的标记。幸运的是,普罗普表中的大多数印刷错误仅限于功能亚型的不一致,对这些结果并没有直接影响。

一致性

度量标注者之间的一致性可以对标注质量做出评估。在已建立的层次被标注的情况下,我从可用材料中为标注团队准备了一份标注指南。一个标注团队由两个标注者和一个裁定者组成。裁定者要么是对这种工作已有经验的标注者,要么是我自己(如果没有其他裁定者可用的话)。在两个标注者对相同的几千个单词(两到三个文本)进行标注之后,整个标注团队会面,将标注合并为一个单独的文档,然后在裁定者的指导下进行讨论更正。重复该过程直到所有文本都被标注。

对不同层次间一致性的度量,最统一的方式是统计学家所熟悉的 F_1 度量,它以标准方式计算,并提供了查准率和查全率的调和平均值。^① 我采用 F_1 度量而不是更常见的 Kappa 统计^②,后者用以评估去除偶然性后的一致性,是因为计算大多数层次一致性的偶然性(chance-level)是很困难的。 F_1 度量是合并过程的自然产物,它对数据有明确的解释,并且允许直接比较不同的层次。表 1 概括了人工或半自动标注的不同层次的一致性。总体而言,一致性的值是很好的。

初始模型构建

有了人工标注数据之后,我们便可以进入自动化研究部分了。构建合并算法的初始模型需要以下步骤:首先,从标注中自动提取每个故事的事件时间线。其次,每个事件都自动与一组施事和受事字符相关联。图 2 简要展示了初始模型中包含的信息。

TimeML 标注允许提取每个故事的事件时间线。语料库中的神奇故事在时间结构上非常简单;除了一个之外,其他所有的都可以用线性时间线加以描述。为了给每个故事构建线性时间线,我首先删除了所有从属事件。仅由从属连接词衔接的事件表示的是在时间线上实际不发生的事件。其次,我使用时间连接词(之前,之后,同时等)的直接定义,写了一个按照起点顺序排列事件的简单算法。^③

应该注意的是,时间线上很多事件是通用的,并且仅依据表面语义是无法与其他非功能性事件区分开来的。这些事件最终被过滤;这将在“合并规则”一节中进行更多讨论。表 3 展示了被标

① C·J·范·里杰斯伯根:《评估》,《信息检索》,伦敦:巴特沃斯,1979,第 112—140 页;另参见本期的尼科利奇、巴卡里奇。

② 琼·卡莱塔:《评估分类任务的一致性:Kappa 统计》,《计算语言学》1996 年第 22 卷第 2 期,第 249—254 页。

③ 马克·阿兰·芬雷森:《从被标注的民间故事中学习叙事结构》,麻省理工学院博士学位论文,2011 年。

注的 15 个故事^①、事件的数量、完整时间线上(不包括从属事件)的事件数量以及在最终实验中使用过的被筛选的时间线上的事件数量。

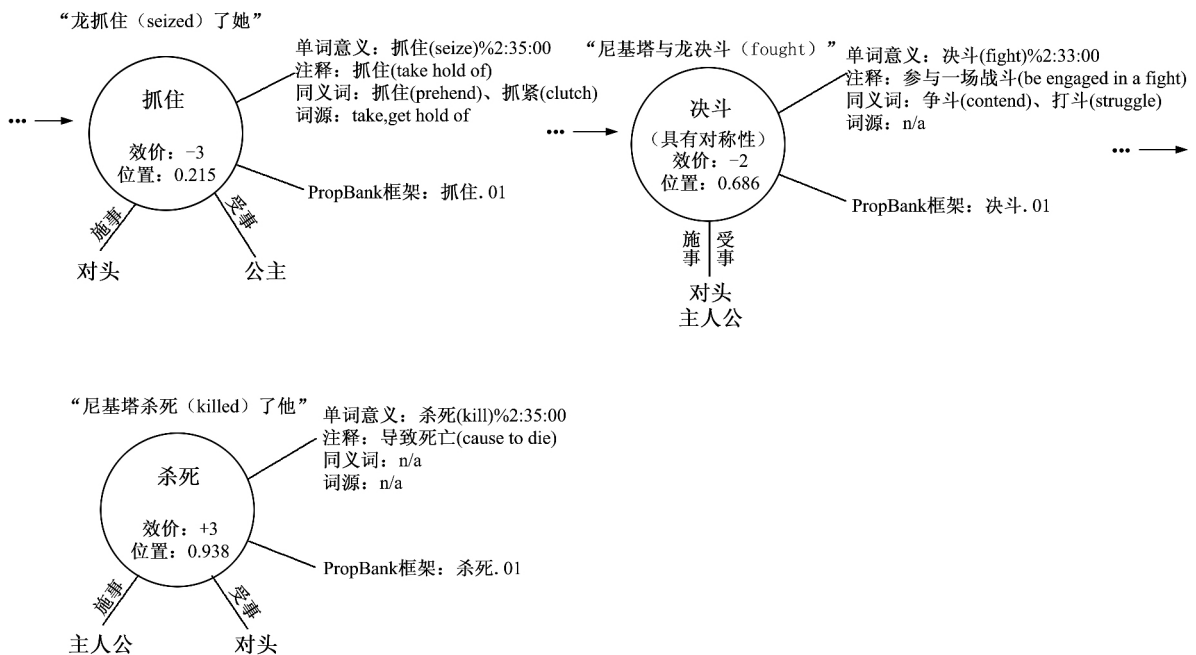


图2 从标注中提取信息的示意图。每个故事由一个有序的事件列表(时间线)表示,它是从TimeML标注中提取的。如果可能的话,为每个事件分配一组施事和受事角色,这些角色从附加到参与指称表达的角色标签中收集,其中的人物群体被替换为个体。每个事件也与一个或多个PropBank框架、一个或多个单词意义、以及一个事件效价相联系。

表3 被分析的文本。所有文本都是单一回合的民间故事,普罗普为之提供了功能分析。表中列出的是:英文翻译中的单词数;每个故事中所标注的TimeML事件的数量;在故事完整时间线中出现的非从属事件数量;以及在学习实验中所使用的“筛选”时间线上出现的事件数量。

故事序号	俄语标题	英语标题	# 字数	# 事件	完整时间线	被“筛选”的时间线
148	Никита кожемяка	Nikita the Tanner	646	104	75	16
113	Гуси—лебеди	The Magic Swan Geese	696	132	94	43
145	Семь симеонов	The Seven Simeons	725	121	87	42
163	Бухтан Бухтанович	Bukhtan Bukhtanovich	888	150	107	62
162	Хрустальная гора	The Crystal Mountain	989	150	104	43
151	Шабарша рабочий	Shabarsha the Laborer	1202	236	122	55
152	Иванко Медведко	Ivanko the Bear's Son	1210	223	143	65
149	Змей и цыган	The Serpent and the Gypsy	1210	250	138	80

① 表3中15个故事的中文名称按表中顺序翻译如下:硝皮匠尼基塔,神奇的天鹅,谢缅七兄弟,布赫坦·布赫坦诺维奇,水晶山,机灵的工人萨巴尔沙,熊之子伊万科,蛇与吉普赛人,伊万·波普洛夫,老坐在那儿的弗罗尔卡,伊瓦什科与巫婆,逃兵与魔鬼,丹尼拉·戈沃里拉王子,商人的女儿和女仆,黎明、黄昏和午夜。——译者注

续表

故事序号	俄语标题	英语标题	#字数	#事件	完整时间线	被“筛选”的时间线
135	<i>Иван Попялов</i>	<i>Ivan Popyalov</i>	1228	220	170	46
131	<i>Фролка — сидень</i>	<i>Frolka Stay-at-Home</i>	1388	248	169	56
108	<i>Ивашко и ведьма</i>	<i>Ivashko and The Witch</i>	1448	276	157	61
154	<i>Беглый солдат и черт</i>	<i>The Runaway Soldier and the Devil</i>	1698	317	190	76
114	<i>Князь Данила — Говорила</i>	<i>Prince Danila Govorila</i>	1774	341	223	92
127	<i>Купеческая дочь и служанка</i>	<i>The Merchant's Daughter and the Housemaid</i>	1794	331	234	89
140	<i>Зорька, вечерка, и полуночка</i>	<i>Dawn, Evening, and Midnight</i>	1934	339	250	78
		平均数	1258	229	151	60
		总计	18862	3438	2253	904

一旦我构建了事件时间线,(如果可能的话)我就会自动为每个事件分配一个施事和一个受事。我从语义角色、指称表达和语义同指的标注中提取了此信息。语料库中的每个动词都标有语义角色,该角色为表现为文本范围的动词提供了一致性。语料库中几乎每个事件都通过其动词表达式与至少一个语义角色相关联。事实上,在故事时间线上的 3438 个事件中,只有两个事件没有语义角色。在后期处理中,我手动指定了这两个事件的施事和受事。当一个事件的语义角色不止一个时,意味着使用某动词多次提到了该事件,我为每个相关联的语义角色合并了其主语和宾语的填充词,在冲突情况下支持首次提到的语义角色。

我使用每个语义角色的相关 PropBank 框架来查找主语和宾语。根据 PropBank 的规则,标记为 ARG0 的动词论元通常是主语,标记为 ARG1 的论元通常是宾语。然而,由于框架定义的特殊性,许多 PropBank 框架没有这种 ARG0—ARG1 的主—宾结构。此外,一些 PropBank 框架可以被认为是**对称的**,其中施事和受事的角色在语义上并不是截然不同的(例如,当动词“结婚”以不及物动词被使用时:“安娜和鲍勃结婚了”)。由于这种信息没有被 PropBank 囊括,所以我对语料库中发现的所有对称类 PropBank 框架及施事受事角色进行了人工分类。

一旦正确的主语和宾语范围被确定后,每个范围内最大的指称表达式将会被自动选择为最合适的语义角色填充词。填充一个事件主宾语角色的指称表达式被确定以后,仍会有一个或多个初级指称来自动替换该指称表达式。有时,这需要用部分指称来替代复合性指称。

合并规则

为了设计在模型融合框架内再现普罗普功能的合并规则,我考虑了三个特征,它们与普罗普本人在其分析中所注意到的相同。普罗普在他的专著中描述了这三个特征:事件语义、涉及的角色、以及事件在回合弧中的位置,通过这些特征他发现了事件之间的相似性。我在一个两阶合并过程中巧妙地利用了这三个方面的相似性。第一阶段将语义相似的事件进行粗略合并。第二阶段仅合并包含多个事件的状态,并在这些状态中合并了附近对主人公具有相同情感效价的状态。

这两个阶段只合并了包含一致角色集合的状态。当两个事件中的角色完全一样或者是彼此的固有子集时,它们被认为是一致的。更具体而言,就是在施事和受事位置上的每个参与者,其角色标签都被添加到了一个施事或受事的标签集合中。如果主人公标签在某个集合中,则相助者标签也会被添加进去,反之亦然。如果一个事件中,角色标签的施事和受事集合与另一个事件的施

事和受事集合相同(或者是其固有子集,反之亦然),则认为两个事件具有一致的角色。如果其中一个事件被标记为对称性事件,其中施事和受事的位置可以互换,则每个事件的角色集合会被合成一个以便进行匹配。

第一阶段:语义

第一阶段的合并规则如下。两种状态会自动合并的条件是:(1)结果状态(resultant state)中所有事件都是非通用的(参见下文),(2)就 Wordnet 意义而言,结果状态中所有成对事件都同义或其上位词同义(hyper-synonymous),(3)结果状态中,附属于所有事件的每个独特的 PropBank 框架都会被至少表示两次。我在下文更详细地定义了这些条件。

通用事件(Generic Events):我识别了一种动词类型,并称之为“通用”动词。它们被自动排除在合并之外,因为无法将这些词的信息性、功能性用法与其通用的填充意义区分开来。动词“说”及其同义词就是一个很好的例子:它们占据了所有事件的近四分之三,而普罗普的每一个功能项都包括至少一个“说”的事件。也就是说,人物可以通过言语行为完成普罗普的所有功能项。角色可以相互威胁(A,加害,或 Pr,追捕),初次见面或提供帮助(D,第一次与赠与者相遇),对其他人的行为做出反应(E,主人公对赠与者的反应),提供某种效劳(C,决定反抗),因某任务而派出主人公(B,派遣),等等。更确切地说,通用事件是指其动词被 Wordnet 标记为归属于词典编纂者档案的交际动词、感知动词或位移动词的事件。这些动词包括“说”“看”或“走”等。

同义性:如果两个事件所附带的 Wordnet 意义或这种意义的上位词共享同义词,则认为它们是同义的。这定义了一种宽泛的语义相似性,允许事件以意义为基础进行聚类。

双重 PropBank 框架:如前所述,PropBank 框架通过语义角色标注被附加到事件上。对于要合并的两种状态,在某个状态中某个事件上找到的每个 PropBank 框架,都需要在该状态中其他至少一个事件中被找到。这种更具体的语义相似性能够平衡 Wordnet 同义词所提供的更丰富的相似性。

第二阶段:效价和位置

在合并的第二阶段,两种状态会自动合并的条件是:(1)两状态中都已包含多个事件,(2)状态中的事件效价是相容的,(3)两种状态是故事弧中最密切的事件对。

效价匹配:如果一种状态中的事件效价是相容的,则两个状态在此阶段将会自动合并。如表 2 所示,事件效价是在从+3 到-3 的 7 点标度内测量的。如果两个效价的值相等,则它们是相容的,只有中性效价(值为 0)可以与其他所有效价相匹配。

最密切的一对:这个阶段也按照特定顺序自动合并为状态,其顺序视状态的组成事件在时间线上相隔多远而定。每个状态的位置计算如下:事件的位置被定义为 0 到 1 之间的分数(包括 0 和 1),对应于其在最初的线性时间线中的相对位置。合并节点的位置是其组成事件位置的平均数。然后根据所合并的状态之间的位置差异,对成对合并的状态进行排序,其中最小的被推到搜索队列的前面。

结 果

根据前文描述的普罗普功能标注,我构建了度量最终模型的黄金标准。最终模型中,功能标记的黄金标准集合实际上从普罗普专著中的功能项列表中减少了很多,原因有三个:普罗普的省略,功能项没有在语料库数据中出现或太稀少,以及功能项隐含。

在 31 个功能项中,普罗普没有说明前 7 个功能项的存在(它们是预备功能项,标有希腊字母)。因此,必须将这些功能排除在分析之外。在剩下的 24 个功能项中,J、L、M 和 N 在我的语料库的

15 个故事中没有被找到,因此只剩下了 20 个功能项。它们当中的四个功能项——o, Q, Ex 和 U——只有两个或更少的实例,也都因太稀少无法学习而被排除在外。

在 276 个功能项标记中,有 186 个是显性的(explicit),90 个是隐性的(implicit)。由于我没有进行常识性推断,因此这些隐性功能项或超过 30%的数据在文本中没有实际的事件实例。这个问题在很大程度上被回避了,因为我只注意到大多数隐性功能项是包含在 E—F(反应和获益)和 H—I(交锋和战胜)这两对之中的功能项之一。在这些情况下,如果其中一对是隐性的,则另一对是显性的。例如,当主人公与对头进行战斗时,只有实际的交锋被提到而战胜是隐含的,或是战胜被提到而交锋是隐含的。因此,为了进行度量,我将这两组功能项合并在了一起。这导致 45 个隐性功能标记在合并中转变为显性功能实例,在 276 个中留下了 234 个显性功能项标记;其余 45 个隐性标记被排除在目标之外。这些数据汇总在表 4 中,最右边的一列表示在筛选了通用事件后的功能项数量(参见下一节)。

我使用了三种不同的度量方式来分析学习程序的性能。首先是应用偶然性校正兰德指数以度量在普罗普功能项中事件聚类的总体质量。^①第二个是应用于普罗普每个功能项的 F_1 度量。第三个是交叉验证分析,用以测试该实现(implementation)与少量数据的合作程度。

事件聚类

我使用偶然性校正兰德指数^②来检验普罗普功能项类别中事件聚类的质量。我创造了三种标准,通俗而言,可以从“严格”(strict)到“宽松”(lenient)进行排列。它们是:(1)严格分数,最终模型中的聚类与表 4“筛选前的显性功能”纵列所列举的所有普罗普显性功能标记聚类进行比较;(2)仅交互式分数(an Interactive—Only score),最终模型中的聚类与普罗普的显性功能聚类进行比较,并移除非交互事件;(3)仅交互且非通用(Interactive Non—Generics Only)分数,最终模型中的聚类与普罗普的显性功能聚类进行比较,并移除非交互的、通用的事件。这三个结果列于表 5 中。对于最宽松的度量(仅交互且非通用的)而言,该算法性能相当好,对普罗普最初的功能项获取的偶然性校正兰德指数大致为 0.714。我之所以在这里说“相当好”,是因为实际上我们不清楚这种性能究竟有多好,因为没有先例:有史以来,以计算机方式在民间故事中学习普罗普的功能,这是首次尝试,所以没有以前的技术与之比较。

表 4 时间线筛选前后存在于语料库中的功能

符号	描述	# 筛选前的显性功能	# 筛选后的显性功能
A/a	加害/缺失	18	15
B	调停/派遣	7	7
C	最初的反抗	7	5
up	出发	13	7
D	与赠与者相遇	16	16
EF	反应/获益	30	29
G	转移	4	2 [†]
HI	交锋/战胜	71	66

① 吉安—卡洛·罗塔:《集合的分区数》,《美国数学月刊》1964 年第 71 卷第 5 期,第 498—504 页。

② 劳伦斯·休伯特·菲普斯·阿拉比:《对比分区》,《分类期刊》1985 年第 2 卷第 1 期,第 193—218 页。

续表

符号	描述	# 筛选前的显性功能	# 筛选后的显性功能
K	灾难或缺失的消除	12	9
down	归来	10	2 [†]
Pr	追捕	18	14
Rs	获救	13	10
T	改头换面	3	2 [†]
W/w	回报	12	8
	总计	234	186

† 数据中可被提取的实例太少,不被列入总数。

表 5 关于聚类质量衡量的三种偶然性校正兰德指数。分数从最严格到最宽松。

方法	分数
严格	0.511
仅交互式	0.581
仅交互且非通用	0.714

功能项类别

第二种度量是针对单个功能项类别的 F_1 度量。在最终数据中的 14 个功能类别中,有 8 个被复原。这些结果显示在表 6 中对交互式非通用的功能项 O 的度量中。重要的是,该算法提取了形态最核心的功能:最初的加害(A),遇到赠与者的三重步骤(DEF),与对头的交锋和战胜(HI),灾难的消除(K),追捕—获救双重步骤(Pr—Rs),以及最终的回报(W)。在所分析的故事中乃至普罗普的整个形态学中,这些都是关键功能。

最显著的成功之处是提取了 HI,即交锋—战胜这一组功能。完整的 51 个实例被正确分类,并且,在对被筛选后的时间线进行度量时,这使得整体 F_1 度量值为 0.823。这种成功可能归因于这一特定功能语义的基本一致,因为所有动词都是关于角逐和战斗的。

另一个显著的成功之处是对 A(加害/缺失)和 W(回报)的识别,其 F_1 度量值为 0.8。这是两个关键性功能,因为它们代表着行动的开始与结束。与 HI 类似,这些功能项的语义一致性对于它们的成功提取很重要。在俄罗斯故事中,最常见的加害行为是绑架公主或其他弱势群体,而回报通常是公主获救并与其结婚或得到金钱报酬。

表 6 功能项识别的 F_1 度量

符号	描述	语义	# 假正类	# 真正类	# 假负类	F_1
A/a	加害/缺失	吞噬、拖拽、伤害、抓住	3	12	3	0.8
D	与赠与者相遇	拖曳,击打	0	6	10	0.585
E-F	反应和获益	掩护,吃,做	3	9	20	0.839
H-I	交锋和战胜	攻击,击破,切割,击败,拖拽,战斗,击打,伤害,推动,竞争,密封,抓住,投掷,吹口哨	7	51	15	0.823

续表

符号	描述	语义	#假正类	#真正类	#假负类	F ₁
K	灾难或缺失的消除	充满	0	3	4	0.6
Pr	追捕	追逐, 考虑	0	5	9	0.529
Rs	获救	攻击, 投掷	1	6	4	0.706
W	回报	礼物, 结婚	1	6	2	0.8

交叉验证

第三个成功的度量标准是交叉验证研究。在交叉验证研究中,算法在不同的数据子集上运行,并且在数据量较小的情况下表现出了平稳下降趋势。值得注意的是,在仅有两个故事时,该技术仍获得了偶然性校正兰德指数 0.457。图 3 以语料库不同子集上的最佳参数值展示了这种性能,表 5 中的三个偶然性校正兰德指数对其进行了衡量。图中的每个数据点,是民间故事语料库的所有故事子集的平均数。可以看出,该算法的运行呈现出平稳下降趋势,直到一次只考虑两个故事时,它对非通用类的度量保留了 0.457 的惊人良好值,仅交互式度量的值为 0.360,严格度量的值为 0.325。这一度量方式表明,该工具应对数据变化非常稳定。

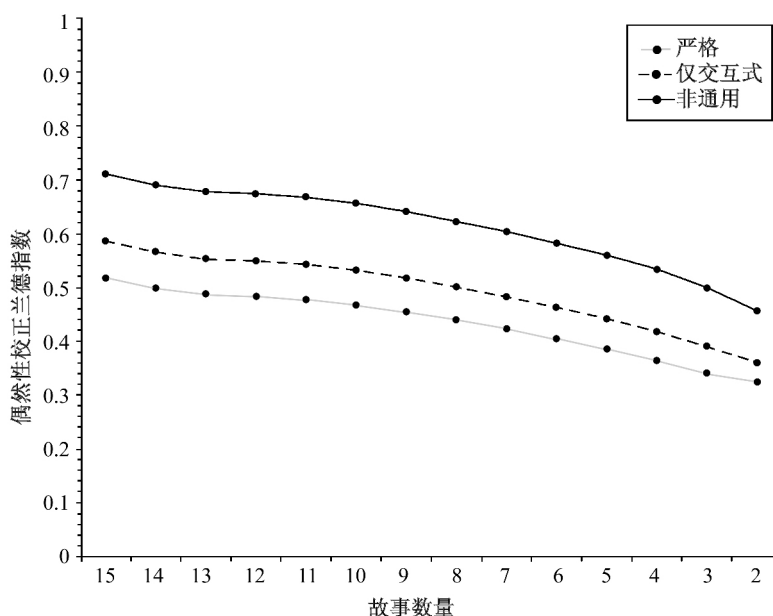


图 3 普罗普 ASM 实现在语料库所有子集上的性能

相关研究

虽然这是第一篇通过计算的方法学习叙事结构实际理论的文章,但最近还有一些关于学习更一般的叙事模式的有趣研究。首先,纳撒内尔·钱伯斯(Nathanael Chambers)和丹·尤拉夫斯基

(Dan Jurafsky)利用对大型语料库的分布式学习来识别常见的事件序列。^① 该技术依赖于动词之间的逐点式交互信息分数,这些动词共享论元以构建公共事件对及其顺序,然后将这些事件对编织在一起形成叙事链。叙事链有几个有趣的共同点,与本文有所不同。钱伯斯、尤拉夫斯基和我都试图识别出各组文本中常见的事件链。此外,他们的研究是另一个数据点,其支持的观点是:明白人物的角色(如,谁是主人公)对识别常见叙事结构的重要性。另一方面,该技术依赖于惊人的文本数量(他们检验了超过100万个文本)来发现相似之处。这种方法与我的算法形成鲜明对比,我的交叉验证研究表明,只剩两个故事时其工作效果可能更好。与我的方法相比,钱伯斯和尤拉夫斯基使用的叙事链模型非常接近文本的含义:共享词根的动词被认为是相同的。而我的技术超越了这种表面意义,我从数据中进行抽象和概括——例如,使用语义知识统一诸如“绑架”和“抓住”之类的事象,然后用诸如导致“伤害”或“加害”的“折磨”之类的动词进一步统一它们。

此外,米夏埃拉·勒涅里(Michaela Regneri)、亚历山大·科勒(Alexander Koller)和曼弗雷德·平克尔(Manfred Pinkal)的研究力图从行动清单中学习事件脚本。^② 该技术是生物信息学中的多序列比对技术的变体。他们能够从数据中提取合理的类似脚本的结构。其数据类型(与自然故事相对,在完成一项任务时关键行动的主题生成列表)与我的工作有所不同,而不能学习循环这一点则与钱伯斯、尤拉夫斯基相同。此外,他们也没有过滤掉不重要的事件,因为其起始数据只包含与特定脚本相关的事件。

结 语

本项研究体现了人工智能领域和民俗学领域的共同进步。对人工智能而言,它展示了一种学习语义级别的技术,这种技术很少被尝试,也从未以这种经过验证的方式被学习。对民俗学而言,它表明计算技术可以为检测民间文学的更深层结构提供重要帮助,而不仅是在词汇或关键词分析的表面水平进行操作。在未来的工作中,还有许多方面可供探索。首先,我们应该继续扩展这些技术,以自动学习其他级别的普罗普理论:回合、亚型和主人公。其次,关于功能项,将这项研究应用于其他形态学分析是很自然的事,如科尔比和邓迪斯的那些形态学分析。^③ 第三,基础技术本身也有很大改进空间:如关于原因、通用类和其他语义的常识性知识的更大整合;学习隐性功能项的尝试;以及通过心理或文化实验,验证形态分析的有效性以结束循环。通过这些努力,人工智能和民俗学可以期待将来诸多令人兴奋的跨学科互动,这将丰富和推进这两个领域的研究。

[责任编辑 龙 圣]

① 纳撒内尔·钱伯斯、丹尼尔·尤拉夫斯基:《叙事事件链的无监督学习》,《计算语言学协会第46届年会论文集》,俄亥俄州哥伦布,2008年,第789—797页。纳撒内尔·钱伯斯、丹尼尔·尤拉夫斯基:《叙事模式及其参与者的无监督学习》,《计算语言学协会第47届年会论文集》,新达城,2009年,第602—610页。

② 米夏埃拉·勒涅里、亚历山大·科勒、曼弗雷德·平克尔:《利用网络实验学习脚本知识》,《计算语言学协会第48届年会论文集》,乌普萨拉,2010年,第979—988页。

③ 本杰明·科尔比:《爱斯基摩民间故事的部分语法》,《美国人类学家》1973年第75卷第3期,第645—662页;阿兰·邓迪斯:《北美印第安人民间故事形态论》,《民俗学者通讯》第195期,赫尔辛基:芬兰科学院,1964年。

mix up ghost legends with ghost stories and rumors, and the differences between ghost legend and urban legend is also unclear, and it makes the study of ghost legend more and more difficult. Although ghost legend, ghost story and rumor share the same narrative theme, ghost legend has a unique narrative structure and narrative function. If we can figure out the definition between contemporary ghost legend, traditional ghost legend, urban legend, ghost story and rumor, we will be able to understand ghost legend accurately.

Key words: urban legend; ghost legend; realness; fright

Inferring Propp's Functions from Semantically Annotated Text

FINLAYSON, Mark Alan, trans. by ZHANG Ruijiao, LI Yang

Vladimir Propp's *Morphology of the Folktale* is a seminal work in folkloristics and a compelling subject of computational study. I demonstrate a technique for learning Propp's functions from semantically annotated text. Fifteen folktales from Propp's corpus were annotated for semantic roles, co-reference, temporal structure, event sentiment, and dramatis personae. I derived a set of merge rules from descriptions given by Propp. These rules, when coupled with a modified version of the model merging learning framework, reproduce Propp's functions well. Three important function groups—namely A/a (villainy/lack), H/I (struggle and victory), and W (reward)—are identified with high accuracies. This is the first demonstration of a computational system learning a real theory of narrative structure.

Key words: artificial intelligence; computational linguistics; folklore; functionalism; semantics

Stratified "Locality": Taking the Legend of Bald-tailed Lao Li as an Example

MA Guangting

Although folklore has a certain relevance to a certain place, it is not only for "one" place, especially for the wide spread legends like that of bald-tailed Lao Li, which itself is a product of the integration of super-local and local. On the one hand, the localization of the legend of bald-tailed Lao Li is a stratified process which is generated by the interaction of "inside and outside" of local community, and the process of dynamic construction of local history. On the other hand, with the local people adapting the identity and ownership of "inside" and "outside", under the background of great history and culture, the area of legendary integration gradually expands, and the scope of "place" gradually expands and changes, which mainly depends on the attributes and regulations of foreign lands. Through oral narration and literature analysis, the "place" presented in layers is intended to extend legends in time and space at the same time.

Key words: the legend of bald-tailed Lao Li; inside and outside; stratification; localization

The Creation and Dissemination of Cai Xiang Gods in the Mountainous Areas of Southern Zhejiang

LI Yang, XU Xiaoqing

Cai Xiang Belief is distributed in the upstream area of Qiantang River in Southern Zhejiang. The legends of its origin express the early memory of the exploitation and settlement of mountainous forests. The traces of the change of Cai Xiang Belief centering from mountains to flatland were preserved in documents. This change was not only caused by the change of population structure brought about by immigrants in the Ming and Qing Dynasties, but also the result of gentry's initiative construction, which reflected the economic changes in the process of regional development. The Lunar July Rally in Shilian was the most important ritual expression of Cai Xiang Belief. At the Lunar July Rally, organizations and institutions such as "Tan", which were formed in the festival, not only maintained the activities of temple fair, but were also the way for local people to constantly integrate social relations, adjust the distribution of benefits and strengthen social identity.

Key words: Southern Zhejiang; mountainous areas; Cai Xiang belief; Lunar July Rally