软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2023,34(2):690−711 [doi: 10.13328/j.cnki.jos.006709]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

属性抽取研究综述*

徐庆婷, 洪 宇, 潘雨晨, 姚建民, 周国栋

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

通信作者: 洪宇, E-mail: hongy@suda.edu.cn

摘 要: 属性抽取是一种自动识别和提取属性表述文字的自然语言处理任务. 首先重温了属性抽取的基本任务、权威数据资源和通用评测规范, 并在此基础上全面回顾了现有前沿技术, 包括基于统计策略和特征工程的传统抽取技术以及利用深度学习的神经抽取技术. 特别地, 以属性表述语言的本质为出发点, 结合现有技术暴露出的不足, 对该领域的技术难点和推演方向给出了详细解释.

关键词: 自然语言处理; 属性抽取; 深度学习

中图法分类号: TP18

中文引用格式: 徐庆婷, 洪宇, 潘雨晨, 姚建民, 周国栋. 属性抽取研究综述. 软件学报, 2023, 34(2): 690–711. http://www.jos.org.cn/1000-9825/6709.htm

英文引用格式: Xu QT, Hong Y, Pan YC, Yao JM, Zhou GD. Survey on Aspect Term Extraction. Ruan Jian Xue Bao/Journal of Software, 2023, 34(2): 690−711 (in Chinese). http://www.jos.org.cn/1000-9825/6709.htm

Survey on Aspect Term Extraction

XU Qing-Ting, HONG Yu, PAN Yu-Chen, YAO Jian-Min, ZHOU Guo-Dong

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: Aspect term extraction is a natural language processing task that automatically recognizes and extracts aspect term from free expression text. The study first goes over the basic task of aspect term extraction, the authoritative datasets of it and general evaluation specifications on it. Based on these, the study comprehensively reviews on the state-of-the-art techniques for the task, including traditional extraction techniques based on statistical strategies and feature engineering, and the neural extraction techniques using deep learning. In particular, the study takes the essence of expression language as the starting point, combines with the limitations of existing techniques and gives an elaboration of the technical difficulties and the future development prospects of aspect term extraction.

Key words: natural language processing; aspect term extraction; deep learning

属性抽取的核心任务是从自由文本中自动识别和提取描述特定属性的语言片段. 比如: 在语句“句法分析鲁棒性优化策略”中, 文字片段“鲁棒性”即为属性表述. 在特定任务场景中, 属性抽取可细分为属性槽(aspect trigger)检测和属性值(aspect value)识别两个部分, 形成一种成对(pairwise)提取的任务模式. 比如: 针对“联想 ThinkPad-E15 的处理器主频是 1.5 GHz”这一语句, 属性抽取器需要将属性槽“主频”和属性值“1.5 GHz”一并正确提取. 此外, 属性抽取既可作为独立的研究任务, 也可以成为其他上游任务的分支. 较为常见的案例是属性抽取与情感分析的组合, 其可形成以情感为中心的属性抽取任务[1], 反之亦然(例: 从语句“1.5 GHz 主频略低, 2.6 GHz 主频够用”中挖掘正向情感“够用”对应的属性“2.6 GHz 主频”).

属性抽取有助于推动非结构化数据向结构化数据(图谱、知识卡片、Profile 等等)转化, 且其在简化信息组成形式和突出关键性质中发挥了重要作用. 学术和工业界在多领域(电商、百科、旅游、政务)开展了相关研究,

* 基金项目: 国家重点研发计划(2020YFB1313601); 国家自然科学基金(62076174, 62076175); 江苏省研究生科研与实践创新计划(KYCX21_2955)

收稿时间: 2021-08-12; 修改时间: 2022-04-14; 采用时间: 2022-05-22; jos 在线出版时间: 2022-07-22

并取得了瞩目的技术成果.

然而, 由于属性表述语言的多样性、灵活性和开放性, 精准且全面地实现属性抽取仍具有较高难度. 针对这一问题, 相关研究在传统计算语言学基础之上, 充分利用神经网络的语义感知和泛化学习能力, 提出并实现了一系列“老道”的神经属性抽取技术. 然而, 现有技术在深层语义理解(如隐式属性感知)、跨领域迁移学习和开放域适应性优化方面仍然面临着较高的挑战. 此外, 单语种和跨语言任务场景下的标记数据匮乏以及常识知识的不足, 都极大地掣肘了现有技术在更为广阔的实用环境下产生突破.

本文集中在属性抽取的方法学部分进行回顾与分析, 包含了早期基于启发式规则和统计技术的传统方法以及近期利用神经网络架构和深度表示学习模型的前沿方法. 重点分析了科学问题、发展现状和尚存的技术挑战. 本文第 1 节、第 2 节结合现有权威数据和评测任务, 对国内外相关研究的整体状态进行分析(含研究机构和发展水平). 第 3 节全面揭示面向属性抽取的技术投入和方法创新, 不仅按启发式规则、统计策略和深度学习的顺序由浅入深地讲解技术发展路线, 也围绕神经网络的整体研究趋势建立相关技术方法的图谱. 第 4 节从不同的“维度”对相关研究的技术细节进行讲解和对比分析, 并就每个方法的设计特色、优势和实验环境加以罗列. 第 5 节总结全文, 分析现有技术的不足并挖掘有待探索的课题, 展望未来研究.

1 背景及现状

属性抽取任务是自由文本分析领域的重要课题, 其建立得益于学术界及工业界在知识获取与结构化数据整合过程中的现实需求. 以知识图谱或知识卡片的自动构建为例, 属性抽取能够提供大规模知识的词级和短语级表示. 换言之, 其支撑了知识节点的初始化. 此外, 属性往往是其他自然语言处理技术瞄准的关键对象之一. 其中, 最具代表性的技术是以属性为中心的情感分析(aspect-based sentiment analysis), 即专门针对特定语句中的商品、实体和技术的属性进行情感极性的自动识别. 比如, 如下两个例子来自商品属性的情感分析数据集(国际语义学评测会议的属性及情感分析语料 SemEval[2]和亚马逊评论语料 Amazon Electronics[3]). 其中, 例 1 给出了显式属性标记(即可见的属性表述文本), 而例 2 的隐式属性词“尺寸”则处于缺省状态.

例 1: This laptop meets every expectation and **Windows 7** is great! (译文: 这台笔记本电脑符合所有预期, 并且 Windows7 很棒！)

Aspect Term (属性项): “Windows 7”

Sentiment Polarity (情感极性): Positive (正面)

例 2: This camera will not easily fit in a pocket. (译文: 这台相机不易装在口袋里.)

Aspect Term (属性项): Default (缺省“尺寸”)

Sentiment Polarity (情感极性): Negative (负面)

属性抽取研究的高地之一是国际语义学评测会议 SemEval 驱动的公开评测任务, 其数据发布和开放实验起始于 2014 年, 并延续至 2016 年[4−6]. 虽然数据集的释出停滞于 2016 年(第 2 节将给出数据集的介绍), 但其研究热度一直持续至今. 本文前期调研的主要对象集中在 SemEval 相关的技术前沿和产出, 综述过程中给出的发展趋势数据和枚举的关键方法学也以这类技术前沿为参照背景. 针对其他同类任务架构下的数据和技术, 本文将给出简介.

- 研究热度(年增文献数量)

图 1 给出了自 2012−2020 年主流国际会议上发表的属性抽取类研究论文数量, 该数据的统计过程以中国计算机学会 CCF 认定的自然语言处理和人工智能领域 A-C 类国际会议论文(比如 ACL、EMNLP、IJCAI、AAAI 等)、主流评测会议 SemEval 论文和高引用率的 arXiv 论文作为信息来源. 统计人员针对历年相关会议中的属性抽取研究, 在谷歌学术上通过关键词搜索的方式捕捉遗漏或错分的文献(但仍不排除图 1 所示的数值与真实数据具有微小的差异). 根据图 1 给出的统计数据说明, 属性抽取的研究热度正逐年上升.

- 分布和影响力

图 2 显示了全球不同国家在 2012−2020 年之间, 面向属性抽取研究的前沿科技文献数量. 其中, 总量排名

前三的国家包括中国、新加坡和美国. 尤其, 国内前沿技术产出的数量在各级别会议的文献数量(蓝色标记条对应 A 类、绿色对应 B 类、黄色对应 C 类)均远高于其他国家.

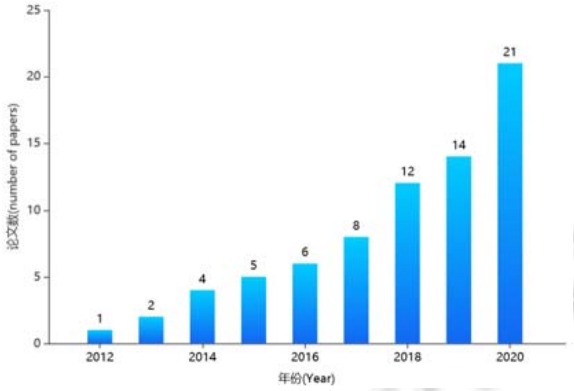


图 1　属性抽取相关的主流国际会议文献历年数量统计(2012–2020)

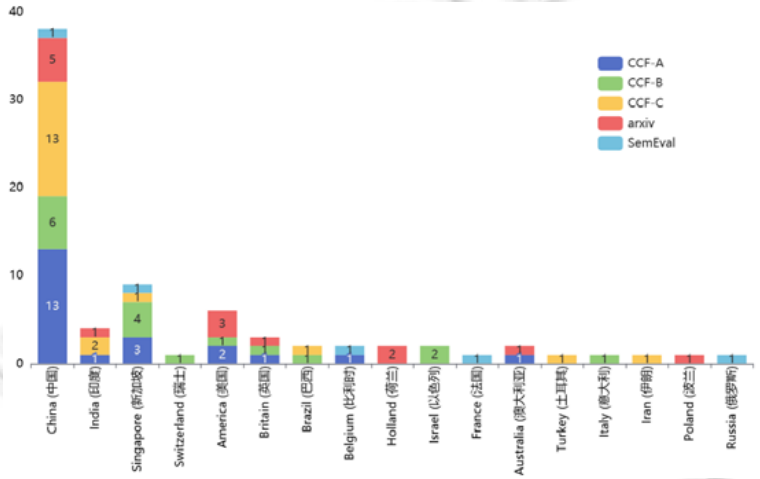


图 2　不同国家在属性抽取领域产生的前沿科技文献统计(2012–2020)

尽管如此, 技术文献的影响力却呈现了不同趋势. 图 3 给出了影响力的分布情况.

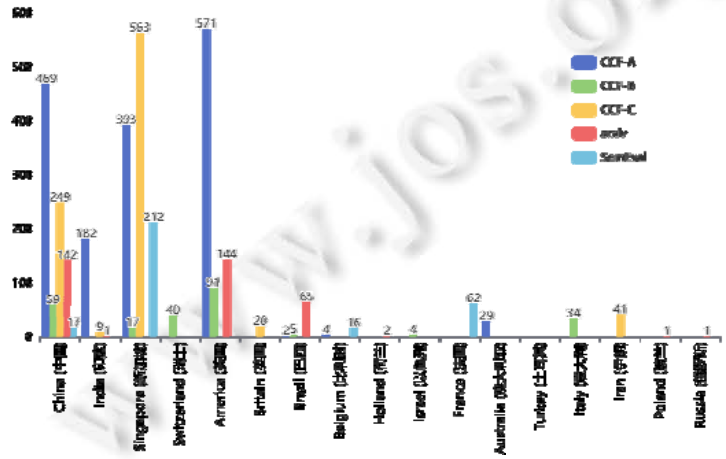


图 3　不同国家属性抽取发表论文的“他引”情况(影响力分布)

本文利用科技文献的“他引(other-citation)”数量作为影响力的评价指标. 其中, 国内同行的 CCF-A 类技术文献的“他引”总量(如图 3 中蓝色柱状标记所示)为 469 次, 低于美国同级别文献的“他引”总量(563 次). 国内所有 CCF-A 类文献中, “他引”次数最多的技术来自东南大学 Liu 等人[7]在 2015 年 IJCAI 国际会议上发表的“抽取规则排序与选择算法”(截止目前, “他引数”为 133 次). 国际范围内, 美国伊利诺伊大学芝加哥分校的 Mukherjee 等人[8]和 Chen 等人[9], 先后在 2012 年和 2014 年 ACL 国际会议上发表的“半监督抽取模型”和“先验知识学习”方法, 其“他引”次数分别为 391 和 180. 上述 3 项工作是属性抽取领域中早期的代表性研究, 起到了重要的引领作用.

总体而言, 属性抽取研究有着较高的应用价值, 其研究热度正逐年升温. 但是相较于其他信息抽取领域, 面向属性抽取的高显示度研究成果并不多见. 因此, 其仍然是一项处于发展期的“年轻”的研究课题.

2 任务特色、数据及评价指标

2.1 任务特色

属性抽取任务的特色是“以词汇和短语为分析对象, 通过语义级的表示学习和句子级上下文的信息融合, 实现针对属性表述文字的自动标记以及属性类别的自动判别”. 事实上, 属性抽取对象的颗粒度(词、短语、语块 Chunk 或子句 Clause)是界定任务难易程度的重要标志. 词汇和短语级的属性表述较易, 语块和子句级较难. 其原因在于, 前者表述边界的确定性高于后者. 比如: 词汇级属性“性能”的前后边界即为其自身, 相比而言, 语块级属性“归一化折损累积增益 NDCG 指标高低”中, 大部分词汇都不是前后边界, 但被误判为边界或独立属性表述的概率极高. 尽管如此, 学术界尚未有针对性地释出和共享语块和子句级属性 Benchmark 数据集(某些词和短语级属性语料偶然地包含了个别语块级属性特例). 从而截至目前, “长条”属性表述抽取任务并无明确定义, 相应研究的技术报告和文献也乏善可陈.

词汇级和短语级属性抽取往往被视作逐词分类标记任务的一种特例[10]. 具体而言, 给定一条自然语句 $S=\{w_1,w_2,\ldots,w_{n-1},w_n\}$, 抽取过程被事实上指定为针对每个词汇 w_i 的 BIO 标记[11]过程, 即为 w_i 唯一地标记上 B、I 或 O 的标签. 其中, B (Begin)表示属性表述的开始, I (Inside)对应属性表述中非起始位置的其他词(包括中间和结尾的词), O (Outside)表示词 w_i 并不在属性表述中(与属性无关的语句中的其他词). 如果仅仅使用 BIO 这 3 个标签进行标记, 其效果等同于区分属性表述片段和非属性表述片段的二元分类(通常可称为“aspect identification”, 即属性判别). 更为复杂的情况下, 属性抽取任务需要达成多种不同类型属性的分类标记, 即在判别出属性表述片段的基础上, 对其额外赋予类型的判定结果, 比如“性能”“质量”和“效率”类属性(此时的任务形式可称为“aspect recognition”, 即属性识别). 针对这一情况, BIO 标记类型需要相应地给予扩充, 形成每一种属性类别的 BI 标记(O 标记不变), 比如“质量”(“quality”)属性的 BI 标记将被具体化为 QB 和 QI(即“质量”类属性表述的开始词 QB 和其他词 QI). 由此可以发现: 属性抽取是一种二元或多元的词一级标记分类任务, 词义的表示学习和上下文语义的融合, 是实现正确标记分类的重要前提.

值得指出的是: 标记语言的定义较为灵活, 可以根据主观的需求自由更改. 原因在于: 实验的评测阶段仅考虑机器抽取的属性与标准属性(ground truth)的文字一致性以及类型的一致性, 至于机器利用何种标签实现抽取并不在评测范围之内. 目前除了 BIO 外, 较为通用的标记语言还包括 BMESO[12].

2.2 数据集

面向属性抽取任务构建的评测数据集来源广泛, 除了国际语义评测会议提供的英文数据, 也包含不同科研单位和组织提供的其他语种数据. 值得注意的是: 随着该领域研究的逐步深入, 近期的数据建设已将“隐式属性”这一特色语言现象纳入标注范围. 下面按照语种和显隐性对现有数据进行划分, 并分别作简要介绍.

- 通用数据(英语)

通用数据来自国际语义评测会议(SemEval)的 4 个基准数据集[13], 包括 SemEval-L14 (http://alt.qcri.org/semeval2014/task4) 和 SemEval-R14 (http://alt.qcri.org/semeval2014/task4)、SemEval-R15 (http://alt.qcri.org/

semeval2015/task12)和 SemEval-R16 (http://alt.qcri.org/semeval2016/task5). 数据集名称中的“L”代表其领域为笔记本电脑(Laptop), “R”代表餐馆(Restaurant)领域; 数字指明了数据集释出的年份, 比如, L14 和 R14 分别表示 2014 年的笔记本电脑领域和餐馆领域的数据. 该数据集不仅提供了商品类(电脑和餐馆)属性的标注结果, 也给予了每个属性的本地(句子内)情感词和情感极性的人工标注. 从而, 该数据集不仅支撑属性抽取研究, 也能支持情感分析研究. 尤其, 该数据对于挖掘属性和情感之间的内在联系并形成协作和互助的识别技术(比如利用情感特征辅助分析属性值的优劣, 反之亦然)具有重要的应用价值. 其中, L14 训练集有 3 045 条评论文本、2 342 个属性词, 测试集有 800 条评论文本、650 个属性词. R14 训练集有 3 041 条评论文本、3 686 个属性词, 测试集有 800 条评论文本、1 134 个属性词. R15 训练集有 1 315 条评论文本、1 209 个属性词, 测试集有 685 条评论文本、547 个属性词. R16 训练集有 2 000 条评论文本、1 757 个属性词, 测试集有 676 条评论文本、622 个属性词.

SemEval 数据的主要来源包括 Yelp (https://www.yelp.com/dataset/challenge)和 Amazon Electronics (http://jmcauley.ucsd.edu/data/amazon/). 其中, Yelp 是 Yelp 官方挑战赛数据集, 包含了来自 4 个国家 10 个城市总共 80 000 多家商铺的 270 万条评论, 以及针对其中商品属性和图片的标记数据. Yelp 为 SemEval-R14、SemEval-R15 和 SemEval-R16 提供了辅助数据[3]. Amazon Electronics 包含了 60 000 多种商品的 160 万条评论, 它是 SemEval-L14 的辅助数据[3]. 此外, IMDB (Internet movie database) (https://www.imdb.com/)是建立在亚马逊电商资源上的同类数据. IMDB 是包含电影、电视节目、明星、演员和电影制作的在线数据, 其规模始终处于动态增加的过程中. 截至 2018 年 6 月份, IMDB 共收录了 400 多万部作品资料和 800 多万名人物资料. IMDB 中的属性标注工作亦是一项持续性的工作. 常用的基础版 IMDB 数据集由 Pang 等人[14]提供, 它包含 13 841 条评论语句和 538 个已标注的属性词. 伊利诺伊大学芝加哥分校的 Hu 等人(2004)[15]提出了 CRD (customer review dataset, https://github.com/yangheng95/LCF-ATEPC)数据集. 该数据集同样来源于亚马逊电商平台, 它包含了相机、DVD、MP3 等 5 种电子产品的用户评论, 经过人工筛选保留了 500 条评论数据和 348 个属性词, 该数据集也是最早应用于属性抽取任务的数据集之一.

- 其他语种数据(阿拉伯语、汉语和印度语)

面向属性抽取的其他语种数据集包括 AOTC (http://www.cs.columbia.edu/~noura/Resources.html)、CPRO 和 Hindi. 其中, AOTC (arabic opinion target corpus)由 Farra 等人(2015)[16]提供, 是一套阿拉伯语观点分析数据. 数据来源是卡塔尔半岛电视台新闻网站的在线评论, 包含文化、体育、政治这 3 个主题的 1 177 条评论以及其中 5 437 个属性词. 香港中文大学的 Xu 等人(2008)[17]提供了中文数据集 CPRO (Chinese product review opinion). 该数据集包含了 7 538 条数字照相机的评论(www.xitek.com)和 3 397 条移动电话的评论(www.soit.com.cn). 该数据集不仅标注了显式属性词(比如评论“按键和按钮都很粗糙”中的“按键”和“按钮”), 也标注了隐式属性(比如“Canon 30D 太贵了”中隐式存在的属性是“价格”). Akhtar 等人(2016)[18]遵循 SemEval 2014 的标注规则, 构建了印度语 Hindi 数据集, 其来源于 12 个领域(比如笔记本电脑、智能手表、旅游等), 包含 5 417 个句子和 4 509 个属性词.

特别地, Bhattacharya 等人(2021)[19]将英语数据集 L-14 和 R-14 翻译成印度语, 并进行了严密的人工校验, 形成了一个包含 5 989 个句子和 5 864 个属性词的印度语数据集 Nhindi. 该数据集保留了原始数据集语法结构的多样性, 使定量比较更有代表性. Bhattacharya 等人(2021)使用当下主流语言模型对 Hindi、L-14、R-14 和 Nhindi 数据集分别进行实验验证. 实验结果显示, 面向 Nhindi 的属性抽取性能($F1$)介于同类模型在 Hindi 和 L-14、R-14 上取得的性能之间.

- 隐式属性数据集

CRD 数据集在属性抽取任务中扮演了很重要的角色, 从最初(2004 年)的提出到现今(2021 年)[20], 仍有人在使用. 此外, 该数据集的存在, 也为隐式属性抽取提供了重要素材. 墨西哥国立理工大学的 Cruz 等人(2014)[21]提供的 AEAIE (implicit aspect extraction and implicit aspect indicator extraction, http://www.gelbukh.com/resources/implicit-aspect-extraction-corpus/)是对 CRD 数据集的重新再标注. Hu 等人对 CRD 数据集仅标注

了显式属性词, 并未对隐式属性做标注. 在 CRD 数据集的基础上, Cruz 等人保留了原数据集中的评论文本内容, 删除其原有的显式属性标注以及不含有隐式属性的句子, 重新人工标注了隐式属性词(比如评论“This MP3 player is really expensive.”中, 隐式属性词是处于缺省状态的“price”)和隐式属性词的指示词(如 small [size]、light [weight]、slick [appearance]中括号里面的词语即为隐式属性词), 为隐式属性抽取提供了数据. AEAIE 数据集包含 314 条评论数据、4 025 个句子和 445 个隐式属性词.

2.3 评价指标

如前所述, 绝大部分现有研究将属性抽取认定为词一级的序列标注任务, 并利用 BIO 标记逐词进行标记分类. 因此, 适用于分类方法的 PRF 评价测度被广泛作为检验属性抽取性能的手段. 在评价过程中, 是属性词判定为真, 不是属性词判定为假. 在此基础上, 准确率 Precision (P 值)的计算方法为“预测为真且正确预测的数量与所有预测为真(包含实际预测结果正确和错误两种情况)的属性总数的比值”. 召回率 Recall (R 值)的计算方法为“预测为真且正确预测的数量与标准属性(人工抽取的属性)总数的比值”. F 测度是对准确率和召回率的权衡指标, 其通用计算方法如下:

$$F=\frac{(a^2+1)\times P\times R}{a^2(P+R)} \tag{1}$$

其中, 超参 a 为人工可调权衡系数, 其以数值 1 为界调整准确率 P 和召回率 R 在整体 F 测度中的重要程度(也可认为是贡献度). 在属性抽取领域的评测中, 超参 a 被设置为 1, 表示准确率 P 和召回率在 F 测度中同等重要, 此时的评价指标也称为 $F1$ 测度.

$F1$ 测度建立在较为严格的字符串比对标准之上, 该标准是“当且仅当系统抽取出的属性字符串与人工标记的标准属性字符串完全一致(前后边界及内容都一致)的情况下, 该抽取结果可作为正例予以认定; 否则皆为负例”. 然而在实用过程中, 人们对属性表述(字符串)的边界有着较为模糊的认定, 换言之, 对确切边界的人为感知并不统一. 比如: 出生日期“1997 年”和“1997”是两种不同的字符串, 差异仅在于后边界是否包含“年”字, 部分标注人员认定前者为正例, 另一部分标注人员认定后者也成立. 在这一情况下, 过于严格的边界约定往往使得 $F1$ 测度并不能恰如其分地传递出真实应用场景下的系统性能. 因此, 现有研究的实验建设中引入了一种较为宽松的评测方法, 称为覆盖率 c[10]. 单一候选样本的覆盖率计算方法如下:

$$c(s_1,s_2)=\frac{|s_1\cap s_2|}{|s_2|} \tag{2}$$

其中, s_1 表示抽取出的属性词的起止位置区间, s_2 表示标准属性词答案的起止位置区间, $\cap$表示交集, $s_1\cap s_2$ 表示两个属性词交集的长度, $|s_2|$表示标准属性词答案的区间长度. 样本集合整体的覆盖率 C 计算如下:

$$C(S_1,S_2)=\sum_{s_1\in S_1}\sum_{s_2\in S_2}c(s_1,s_2) \tag{3}$$

其中, S_1 为属性抽取结果集, S_2 为标准答案集. 借助覆盖率的定义, 可对现有准确率 P 和召回率 R 的计量方法进行改进, 并形成具有边界模糊性判别的新型 $F1$ 测度. 计算方法如公式(4)和公式(5)所示:

$$P=\frac{C(S_1,S_2)}{|S_1|} \tag{4}$$

$$R=\frac{C(S_1,S_2)}{|S_2|} \tag{5}$$

3 前沿技术回顾

本节首先概述属性抽取领域的技术发展概况, 在给出 3 项传统方法的基础上, 侧重汇总了利用神经网络的建模架构, 并对比了测试性能(对比环节的技术皆来自基于通用数据 SemEval 的测试场景, 其性能为相关工作报告中的最优指标); 其次, 本节细致地对各项关键技术进行剖析, 从设计动机、建模原理和优劣势方面逐项介绍.

本文聚焦属性抽取领域的“主阵地——SemEval 评测”, 对其中显露头角的主要方法进行汇总, 并按照其采用或集成的核心技术进行了分析, 见表 1. 表 1 中: 每行对应一种属性抽取方法; 每列对应一种通用的关键技术, 包括早期基于规则的技术、条件随机场(CRF)[22]、卷积神经网络(CNN)[23]、循环神经网络(RNN)[24]及其变体(如长短期记忆网络 LSTM[25]和门控循环单元 GRU[26])、注意力机制(attention)[27]以及预训练语言模型 BERT[28]. 根据技术应用的分布, 本文总结出如下发展概况.

- 总体上, 基于规则的属性抽取研究主要集中在 21 世纪初, 其在后期较少被涉及, 研究主流逐步集中到基于神经网络的深度学习技术之上, 并于近期引入了预训练语言模型(如 BERT);
- 条件随机场是基于动态规划的择优策略, 其对于 BIO 标记分类场景有着天然的适用性, 尤其在不同时刻的最优化判别过程, 依赖于前期不同时刻的最优解, 由此与循环神经网络有着较高的兼容性, 见表 1, 2019 年出现的前沿模型 TOWE[29]和 RINANTE[3]都利用 CRF 和 LSTM 进行建模;
- 相较于基于卷积计算的语义编码方式, 现有属性抽取方法更多采用基于序列化循环计算的编码技术. 尤其, 集合长短期记忆单元 LSTM 的循环神经网络是最常见的核心编码步骤. 原因在于: 属性抽取的处理对象是自然语句中的词项, 对其中任意候选词项进行编码(及属性判别)都对上下文语境信息有着较强的依赖性. LSTM 能在一定程度上缓解长程依赖问题(即融合上下文语义信息的记忆), 从而成为近期研究采用的必要手段之一;
- 注意力机制[30−32]也是相应方法经常使用到的关键技术, 其在依据上下文进行高注意力信息的识别和加权方面有着重要的贡献. 值得注意的是: 诸如 BERT 等预训练语言模型中的 Transformer 架构[33−35]集成了多层多头注意力机制, 可以认为是预训练语言模型在属性抽取任务中的成功应用, 注意力机制在编码和表示学习过程中凸显重要信息的能力.

表 1 现有属性抽取方法及其内含的关键技术

Model	Rule-based	CRF	RNN	CNN	LSTM	Attention	GRU	BERT
FBS (Hu, *et al*., 2004)	▲	–	–	–	–	–	–	–
MKB (Zhuang, *et al*., 2006)	▲	–	–	–	–	–	–	–
IHS-RD (Chernyshevich, 2014)	–	♠	–	–	–	–	–	–
LSTM (Liu, *et al*., 2015)	–	–	◆	–	♥	–	–	–
RNCRF (Wang, *et al*., 2016)	–	♠	–	–	–	–	–	–
MIN (Li, *et al*., 2017)	–	–	–	–	♥	–	–	–
CMLA (Wang, *et al*., 2017)	–	–	◆	–	–	♣	–	–
HAST (Li, *et al*., 2018)	–	–	–	–	♥	♣	–	–
DE-CNN (Xu, *et al*., 2018)	–	–	–	●	–	–	–	–
TOWE (Fan, *et al*., 2019)	–	♠	–	–	♥	–	–	–
Seq2Seq (Ma, *et al*., 2019)	–	–	–	–	–	♣	★	–
RINANTE (Dai, *et al*., 2019)	▲	♠		–	♥	–	–	–
SpanMlt (Zhao, *et al*., 2020)	–	–	–	–	♥	♣		▼
LOTN (Wu, *et al*., 2020)	–	–	–	–	♥	♣	–	–
BBCR (Wei, *et al*., 2020a)	–	♠	–	–	♥	–	–	▼

表 2 给出了上述属性抽取方法的性能(*F*1 测度指标), 所有性能皆来自通用评测场景 SemEval 的测试结果. 总共有 2 个领域的 3 期数据用于 SemEval 测试, 包括 2014–2016 年采集和标注的餐馆领域属性数据(即 R14–R16)以及 2014 年便携式电脑领域的数据(L14).

由表 2 可见: 相较于传统基于规则和模板的方法, 利用神经网络的属性抽取模型取得了显著的性能优势; 此外, 引入预训练语言模型的方法进一步推动了性能优化. 上述方法的性能变化趋势可以借助图 4 中的柱状指标分布更为清晰地得以观测. 根据表 2 和图 4 的性能分布, 我们能够窥见现有方法设计过程中暴露的如下问题.

- 基于序列标注的属性抽取方法对条件随机场 CRF 有着普遍的依赖性, 以同时引入预训练语言模型的 SpanMlt[36]和 BBCR[37]为例, 两者的主要差异在于“是否装配了 CRF 的动态规划过程”. 如表 2 和图 4 所示, 结合 CRF 进行属性判别的 BBCR 具有更高的性能优势(注意: 虽然 SpanMlt 额外部署了注意力机制, 但鉴于其使用的 BERT 内含多层多头自注意力机制, 本文认为, 其额外增设的注意力计算对整

体性能的贡献并未起到主导作用, 从而在刨除 CRF 的情况下, 基于 BERT 的 SpanMlt 和 BBCR 具有极为近似的编码结构). 此外, 在未结合预训练语言模型的其他神经网络方法中, 弃用 CRF 的 MIN[38]、CMLA[11]和 HAST[39], 相比于启用 CRF 的 TOWE[29], 取得了略低的性能指标;

- 基于神经网络进行语义编码的属性抽取方法虽然在判别性能上取得了显著优势, 但其并未彻底扭转鲁棒性不佳的局面. 通过观测图 4 中的柱状指标可以发现: 所有方法在不同数据集(R14–R16 和 L14)上的表现并不稳定, 最大的性能落差达到近 18%的 *F*1 值, 最小的落差不低于 3.6%的 *F*1 值.

表 2 代表性模型实验结果

Category (类)	Method (方法)	L-14	R-14	R-15	R-16
Rule-based (规则方法)	FBS (Hu, *et al*., 2004)	49.92	40.42	45.97	51.83
	MKB (Li, *et al*., 2006)	58.04	37.14	55.98	64.62
	IHS-RD (Chernyshevich, 2014)	74.55	79.62	–	–
NN (神经网络方法)	LSTM (Liu, *et al*., 2015)	75.00	82.06	64.30	71.26
	RNCRF (Wang, *et al*., 2016)	78.42	84.93	67.47	–
	MIN (Li, *et al*., 2017)	77.58	–	–	73.44
	CMLA (Wang, *et al*., 2017)	77.80	85.29	70.73	–
	HAST (Li, *et al*., 2018)	79.52	85.61	71.46	73.61
	DE-CNN (Xu, *et al*., 2018)	81.59	85.20	68.28	74.37
	TOWE (Fan, *et al*., 2019)	80.24	71.39	73.51	81.84
	Seq2Seq (Ma, *et al*., 2019)	80.31	–	–	75.14
	RINANTE (Dai, *et al*., 2019)	73.47	84.06	66.17	–
	LOTN (Wu, *et al*., 2020)	82.21	72.02	73.29	83.62
NN+Pretrained (+预训练)	SpanMlt (Zhao, *et al*., 2020)	77.87	85.24	71.07	–
	BBCR (Wei, *et al*., 2020a)	82.68	87.11	70.23	77.51

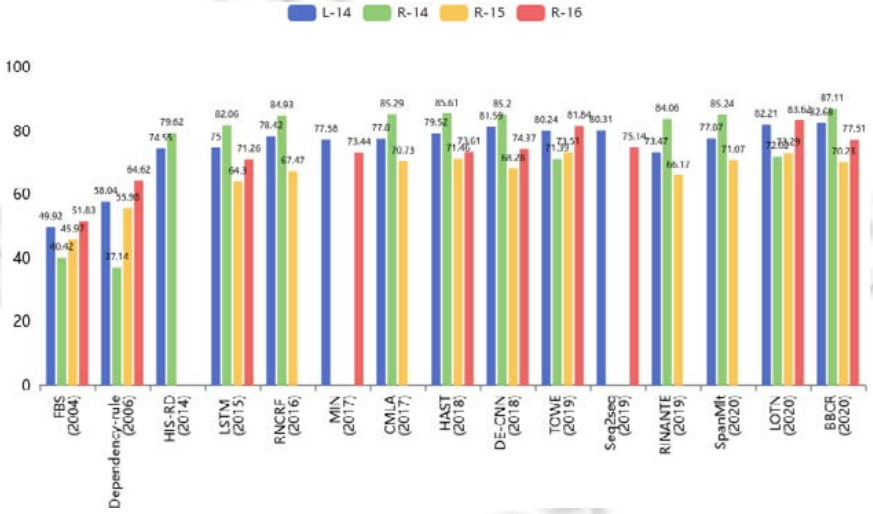


图 4 性能分布趋势(纵轴为 *F*1 测度)

4 属性抽取方法分类解析

如前所述, 属性抽取方法可基本划分为传统基于规则的方法和基于神经网络的方法. 事实上, 在利用神经网络的过程中, 前人分别从无监督、监督学习、半监督学习和自监督学习这 4 种模式入手, 通过结合独特的特征抽取、表示学习和知识运用等方法, 形成了大量各具特色的研究成果. 本节对相关研究的技术细节进行分类讲解和对比分析, 借以辅助读者全面了解属性抽取领域的各项技术(注意: 部分技术并未列入表 1 和表 2, 原因在于其实验数据集或数据划分并不相同). 在特征抽取和表示学习方面, 本节按照两个角度陈述相关技术: 其一(第 4.1 节)是属性抽取模型的设计方式(含基于规则、CRF 和基于神经网络两种设计手段), 其二(第 4.2 节)是机器学习模式(含无监督学习、监督学习、半监督学习和自监督学习). 此外, 在知识的有效运用方面,

本节针对现有数据驱动(第 4.3 节)和情感“加持”(第 4.4 节)的属性抽取方法进行了解析.

4.1 规则、CRF和神经网络的对比分析

基于规则和基于神经网络的属性抽取方法具有截然不同的设计原理与特性: 前者依赖领域专家制定的经验性模板, 具有处理速度快和对同构数据判别精准度高的优势, 但也因为模式匹配过拟合和可观测模式种类的有限性问题, 导致在异构数据上迁移能力差和普适性偏低的缺陷; 后者借助神经网络中庞大的记忆和运算单元, 抽象表示属性及上下文的语言编辑模式, 从而对语义具有更深的感知能力, 往往在大规模且语用模式多样的数据上表现出更为卓越的性能, 即普适性较高.

在早期的基于规则的研究中, Hu 等人(2004)[15]提出了 FBS (feature-based summarization)模型, 其使用关联规则实现属性抽取. 特别地, FBS 基于词性标注、高频特征识别、冗余特征剪枝等基础 NLP 技术建立关联规则, 且利用这类规则识别“常见”的属性表示. 在此基础上, FBS 借助词典 WordNet[40]获取情感词, 并将情感语义的指向性作为判别关联属性的依据, 该方法对识别“少见”的属性表示有着较好的效果. Zhuang 等人(2006)[41]提出了一种基于多知识的 MKB (multi-knowledge based)方法. MKB 能够借助句法分析工具构建依存关系模板, 从电影评论文本 IMDB 中抽取属性词. 具体地, MKB 使用词典 WordNet、电影演员表以及带标签的训练数据构建查询, 目的是获取属性词、情感词和上下文的候选文字片段. 在此基础上, MKB 根据属性词和情感词“组团出现”的语法规则, 实现“属性词-情感词”单元对的“组团抽取”. 近现代, Qiu 等人(2011)[42]提出一种双重传播(double propagation, DP)的方法. DP 通过信息共享获取属性词和情感词之间的信息交互模式, 并利用依存关系构建两者的关系模板. 特别地, DP 使用了情感词极性分配和噪声目标剪枝技术, 并构建了交叉迭代的双重抽取模式(不仅以属性词为依据抽取情感词, 同时反向依据情感词对属性词进行抽取). 近期, Liu 等人(2015)[7]提出了一种规则挑选算法 RS-DP (rule selection-DP). RS-DP 面向不同样本, 自动地从规则集合中选择可靠性高的规则子集, 并利用这些规则组装成特定的抽取模式. RS-DP 能够针对不同的依存关系制定相应的抽取规则, 并使用贪心算法筛选出最佳的规则. 表 3 对比分析了基于规则的属性抽取方法的设计特色、实验数据(含数据集名称及语种)及其呈现的主要优点. 数据集的具体情况如第 2.2 节所示.

表 3 基于规则的属性抽取方法比对

年份	模型	设计特色	数据集	语言	优点	来源
2004	FBS	基于词频	IMDB	英文	采用剪枝策略; 关联规则; 简单易行	Ref.[15]
2006	MKB	WordNet+语法关系+特征关键词表	IMDB	英文	整合 WordNet、统计分析和关键词表, 利用 IMDB 实现属性抽取	Ref.[41]
2011	DP	DP+依存关系	CRD	英文	通过双重传递来获取属性词和情感词之间的信息; 提出了新的意见词极性分配和噪声目标剪枝方法	Ref.[42]
2015	RS-DP	语法抽取规则+贪心选择算法	IMDB	英文	规则挑选算法, 自动地从规则集合中选择出抽取性能更好的规则子集	Ref.[7]

自从 2010 年 Jakob 等人(2010)[43]引入条件随机场(CRF), 属性抽取方法设计中便普遍增加了一项策略性的环节, 即利用 CRF 动态规划句子级属性标记序列的最优组合模式. 具体地, Jakob 等人[43]利用词特征、词性特征、依存句法特征、词间距特征等多种特征构建词项的特征表示, 并利用特征空间的分布模式实现属性 BIO 标记的判别. 在此基础上, CRF 权衡所有 BIO 标记的发射和转移概率, 探寻序列标记的全局最优解. Xu 等人(2011)[44]在 CRF 的基础上引入浅层句法分析和启发式位置特征, 在不增加领域词典的情况下, 有效地提高了属性抽取的性能. Chernyshevich (2014)[45]开发了 HIS Goldfire 语言处理器, 并在 CRF 模型中使用了丰富的词汇、句法和统计特征, 构建了 IHS-RD 模型, 该模型对明显蕴含上述特征的属性词有着较高的抽取准度. 值得指出的是, CRF 在属性抽取领域有着较高的受众群体. 其原因在于 CRF 对序列标注的天然适用性, 以及它与不同判别模型的适配性. 此外, CRF 可以接收灵活的特征设计方式, 且善于动态规划择优策略. 总之, 利用 CRF 实现属性抽取的研究阶段是一个重要的过渡期, 它的出现代表了研究焦点正从“约定确定的模式”向“规划不确定的模式”进行转变. 事实上, 具备动态规则选择能力的 RS-DP 模型[7]也在尝试突破僵化静态的经验性抽取模式. 这类研究激发了同期科研人员“求变”的模型设计思维, 而在不久之后, 对模式变换具有更高感知

和模拟能力的神经网络, 即被快速且大量地应用于属性抽取领域.

近期, 研究人员将基于神经网络的编码模型引入属性抽取研究. Liu 等人(2015)[46]提出一种基于递归神经网络(RNN)和词嵌入的判别模型, 该模型无需手工设计特征类体系, 即可完成上下文依赖的特征表示和编码. 此外, Liu 等人(2015)详细分析了 RNN 不同变体(Elman、Jordan 和 LSTM)在属性编码表示及抽取中的表现. 随后, Toh 等人(2016)[47]将双向循环神经网络(bidirectional recurrent neural network, Bi-RNN)与 CRF 相结合, 其不仅在局部词项的编码方面融合了短期上下文信息, 而且同时具备动态规划全局最优解的能力. 在 RNN 的基础上, Dai 等人(2020)[3]升级了 RNN 的计算模式, 将双向长短期记忆单元 Bi-LSTM 引入每个时刻的词项编码, 如同 Bi-LSTM 在其他 NLP 任务中的效用一样, 结合 Bi-LSTM 的 RNN 能够有效解决长程依赖问题, 以提升属性词与远端上下文的信息融合效率. 此外, Dai 等人利用规则和句法依存规律实现了针对外部数据属性词的挖掘, 并构建了远程监督的学习模式, 有效缓解了数据稀疏问题对属性抽取性能的影响. Li 等人(2017)[38]提出了基于 LSTM 的多任务学习模型 MIN, 它利用双通道分别进行属性抽取和情感分类, 并在双通道的编码过程中实施特征共享, 增强判别模型所依赖的信息总量. 该模型不仅借助 LSTM 解决了 RNN 的长程依赖问题, 并利用双通道多任务架构, 在特征共享与信息交互方面给出了较好的范本. Xu 等人(2018)[48]尝试着将卷积神经网络[49,50]引入属性抽取任务, 其动因是 CNN 能够对非等长的文字片段进行局部的深度抽象与编码. 具体地, Xu 等人以多层 CNN 为基础框架实施编码, 同时开发了双嵌入机制, 其形成的 DE-CNN 模型不仅能够充分利用整个句子的语义信息, 并且缓解了标记的“受限依存”问题. Ma 等人(2019)[51]提出一种端到端的多任务联合学习模型, 该模型具备 3 个编码通道, 分别由基本的端到端 Seq2Seq 模型、门控单元网络 GUN (gated unit networks)和位置预知注意力网络 PAA (position-aware attention)构成, 且不同编码通道分别用于属性识别和情感分析, 形成了多任务驱动的联合学习模式. 该模型与 MIN 有着类似的设计原理, 但其复杂度略有提高, 不仅具备异构的通道模型, 且形成了 3 种通道的特征共享. 表 4 围绕 CRF 和神经网络设计方法进行比对, 具体比对体现在设计特色、实验数据(含数据集名称及语种)及其呈现的主要优点. 数据集的具体情况如第 2.2 节所示.

表 4 基于神经网络或 CRF 的属性抽取方法比对

	年份	模型	设计特色	数据集	语言	优点	来源
CRF	2010	CRF	基于 CRF	IMDB	英文	基于 CRF 的算法在单一领域中取得了优于基线的效果, 加入特征可以捕获更复杂句子中的更多信息, 并且在跨域中也取得了不错的效果	Ref.[43]
	2011	CRF	CRF+ 浅层句法分析+ 启发式位置特征	COAE2008	中文	采用条件随机域模型, 引入多种有效的语言学特征和启发式位置特征, 提高了产品属性抽取的精准率	Ref.[44]
	2014	IHS-RD	CRF+特征	L-14, R-14	英文	基于 CRF, 融入丰富的词法、句法和统计特征, 性能好且可移植性强	Ref.[45]
	2016	CRF	Bi-RNN+CRF+ 各种特征	L-14, R-14, R-15, R-16	英文	提取各种词汇特征、句法特征和集群特征, 并且结合深度学习系统提取额外的神经网络特性	Ref.[47]
神经网络	2015	–	LSTM+ RNN+ 词嵌入	L-14, R-14	英文	将长短期记忆网络应用于属性抽取, 使用 LSTM 对词向量编码, 并通过最后一层全连接获得每个词的概率分布	Ref.[46]
	2017	MIN	LSTM+ 记忆单元	L-14, R-16	英文	使用双通道 LSTM 同时抽取属性词和情感词, 且通过记忆单元进行交互	Ref.[38]
	2018	DE-CNN	CNN+ 词向量	L-14, R-14, R-15, R-16	英文	使用多层 CNN 构建模型, 同时利用通用词向量和领域词向量提升属性抽取的性能	Ref.[48]
	2019	Seq2Seq	Seq2Seq+ GUN+PAA	L-14, R-16	英文	在实现属性词的抽取的同时, 也可以实现属性词和情感词的联合抽取, 增强属性词、属性词和情感词联合抽取时多任务之间的信息传递, 提升各个任务的效果	Ref.[51]
	2020	RINANTE	Bi-LSTM+CRF	L-14, R-14, R-15	英文	使用规则标注数据扩大训练语料, 有效地解决了属性抽取任务缺乏数据的瓶颈问题, 通过 Bi-LSTM 和 CRF 进行属性词和情感词的联合抽取	Ref.[3]

4.2 面向属性抽取的机器学习模式

面向属性抽取的相关研究中, 最为主要的科学问题是“以属性为中心”的表示学习方法设计. 截止目前, 传统的统计学习方法、机器学习方法和近期的深度学习方法, 都在属性抽取研究领域得以应用, 其具体可划分为无监督、监督、半监督以及自监督学习模式.

- 面向属性抽取的无监督学习模式

建立在统计学原理之上的无监督处理方法中, Hasegawa 等人(2004)[52]首先提出了基于聚类的属性抽取方法, 其基本假设是“处于同一聚类簇内的候选属性词项, 具备相同的属性类”(包括非属性词的聚类簇及其对应的非属性类). 由此, 属性抽取转化为词项的聚类问题. 相应地, 研究焦点也集中在无监督的聚类方法以及面向聚类的特征表示方法上. 如同其他领域的聚类应用方法一样, 机器并不进行严格意义上的学习, 也并不会利用可观测样本学习类别的划分模式. 事实上, 划分模式的学习过程已由专家根据经验人工完成. 聚类方法的实现仅仅是人脑处理过程的模拟, 包括了统计特征的组织、特征分布的相似度计算以及聚类算法的流程化. Chen 等人(2005)[53]对 Hasegawa 等人的基准方法进行了改进, 其在不需要手动标注数据的情况下, 实现了关键特征子集和聚类簇的自动构建, 并重点解决了类簇关联计算中聚类中心表示的偏差问题. 类似的研究还包括 Popescu 等人(2007)[54]设计的判别语引导方法, 其中, 判别语是一种表征实体目标和实体属性的自然语言表述模式. Popescu 等人依据经验对判别语及其关联的属性类进行了预定义, 并在属性抽取过程中, 依据候选属性词与判别语之间的互信息权衡前者的归属. 该方法本身也是一种聚类策略, 区别仅在于其利用属性与判别语的关系的紧密程度作为聚类的条件, 而 Hasegawa 等人和 Chen 等人的方法则是利用属性本身的统计关联性来实施聚类.

另一项有代表性的无监督处理模式来自 Garcia-Pablos 等人(2015)[55]的研究. 其基于词性收集候选属性词和情感词集合(名词作为属性词的候选集, 形容词作为情感词的候选集), 并依照依存关系等建立属性词与情感词的关系子图; 同时, 结合语义相关性和情感相关性归并上述子图, 从而集成了属性节点和情感节点的全图. 在此基础上, Garcia-Pablos 等人将信息检索领域的 PageRank[56]算法引入全图节点的权威性(可信度)计算中, 并将其作为权衡词项具备属性特质的概率化标准, 建立了排序和“择优录取”的抽取模式. 该方法考虑了全局关联属性, 并适用于不同领域, 具有较高的实用价值. Garcia-Pablos 等人的研究凸显了上下文情感词及其依存关系的重要性, 揭示了它们对于牟定属性信息的关键参照性(注意: 在缺少情感分析的数据集和应用领域, 该方法并不适用). 由此, Liu 等人(2016)[57]在 2017 年专门研究了属性词和情感词的依存关系, 并经验性地建立了 8 套基于依存的抽取规则. 特别地, Liu 等人建立了依存规则的评估方法, 能够有效地对人工建立的规则进行可靠性评估和筛选. 上述主要无监督方法的设计特色、优点和实验环境见表 5. 数据集的具体情况如第 2.2 节所示.

表 5 无监督方式的属性抽取方法比对

年份	模型	设计特色	数据集	语言	优点	来源
2007	OPINE	PMI	IMDB	英文	在文献[14]上的改进, 引入 PMI, 提出了判别语引导方法	Ref.[54]
2015	V3	PageRank+ 语义关系+ 情感关系+图	L-14, R-14	英文	基于节点和加权边构建图, 依照依存等关系 建立属性与情感词的关系子图, 再经 PageRank 排序, 生成属性词列表, 实现属性抽取	Ref.[55]
2016	Rule- Based	贪心算法	CRD	英文	建立了 8 套基于依存的抽取规则, 并且 建立了依存规则的评估方法, 能够有效地对 人工建立的规则进行可靠性评估和筛选	Ref.[57]

- 面向属性抽取的监督学习模式

如第 2.1 节所述, 属性抽取任务通常被解释为序列化标注问题(自然文本向 BIO 标记序列的转化). 因此, 主要的监督学习模式都是用于服务序列化的语义编码和解码. 常见的序列标注模型包括隐马尔可夫 HMM、长短期记忆神经网络 LSTM 和条件随机场 CRF. 本文第 4.1 节已对基于 LSTM 架构的属性抽取方法进行了介绍, 本节不再赘述. 相对地, 本节重点解释监督学习过程中的特色科学问题(特征工程、终身学习、数据不平衡、

感知深度)及其解决方法.

Jin 等人(2009)[58]首次利用词级的 HMM 模型实现属性抽取, 其主要贡献是设计了词性、短语信息模式、上下文语境线索等语言特征的监督学习模式, 并以此类特征为依照, 形成了基于剪枝技术的属性抽取方法. 在面向属性抽取的神经网络编码模型设计中, Poria 等人(2016)[59]尝试检验学习(感知)深度对属性语义编码的作用. Poria 等人提出了一种基于 7 层深度卷积网络的抽取模型(包括输入层、2 个卷积层、2 个池化层、全连接层和 softmax 层), 侧重利用学习深度更为有效地提取和表示特征.

Shu 等人(2017)[60]结合 CRF 和终身机器学习 LML (lifelong machine learning)[61]提出了 L-CRF 方法. 其有效利用了 LML 学习过程中的知识存储优势, 并借此促进了 CRF 在增量数据集上逐步形成更优的判别模式. LML 的终身学习模式能够在观测样本不断扩充的情况下, 渐进地提高基于 CRF 的属性标记识别能力, 使其在实际应用场景中扮演了极为“稳健”的角色.

近期, Wei 等人(2020)[62]提出了一种基于类卷积交互式注意力机制的属性抽取方法. 类卷积交互式注意力模型的核心思想是: 在卷积层之间植入注意力加权因子, 旨在依照前一层卷积层的聚合信息, 对下一层的卷积层进行注意力加权. 其避免了全局注意力加权的盲目性, 同时保证了局部注意力加权仍能具有一定的波及范围. 该模型不仅利用了更深层的卷积特征学习, 且累加了注意力感知层. 特别地, 其模型的输入来自 LSTM 循环神经网络输出的隐状态, 因此, 该模型具有多样且“深邃”的感知学习通道. 实验结果表明, 其模型的实际性能确实优于同期基于浅层感知学习的神经属性抽取模型.

失衡的观测数据分布和健壮性问题引起了相关学者的关注. 具体地, 观测样本中标记数据的分布不平衡, 往往使得机器学习模型趋向“有偏见”的判别模式, 特别是在过度拟合训练数据的模型中, 这类偏见不可避免地影响到其测试性能, 使其体现出较低的健壮性. 简言之, 由数据分布不平衡导致的低健壮性是大部分监督学习过程中常见的痼疾. 属性抽取领域的现有语料集也存在标记分布失衡的问题(属性词较少, 非属性词较多). 为此, Venugopalan 等人(2021)[63]提出一种结合了数据平衡和鲁棒性特征选择的属性抽取方法. 该方法通过拼写检查、分词、分块、词性标注等方法进行特征提取, 并利用 SMOTE[64]解决数据分布不平衡问题. 特别地, 该方法采用基于关联性的特征子集评估算法和最佳优先搜索方法进行特征选择, 有效地去除了冗余或不相关的特征. Venugopalan 等人在数据平衡和特征优选的基础上实现了基于朴素贝叶斯、支持向量机、随机森林[65]和多层感知机的 4 种属性抽取模型, 并证明随机森林的性能最优. 上述属性抽取模型的设计特色、优势和实验环境见表 6. 数据集的具体情况如第 2.2 节所示.

表 6 监督方式的属性抽取方法比对

年份	模型	设计特色	数据集	语言	优点	来源
2009	HMM	HMM+ Bootsraping	CRD	英文	不需要统计学信息, 加入语言学特征, 并且能够自主学习新词	Ref.[58]
2016	CNN	CNN+POS+词向量	Google+ Amazon	中文/英文	深度神经网络, 精度较高, 可以适用于较复杂的数据集, 网络结构对参数较敏感	Ref.[59]
2017	L-CRF	CRF+LML	IMDB	英文	采用终身机器学习的方法; 传统 CRF 与 LML 结合, 可以更好地实现属性抽取	Ref.[60]
2020	CIA-CRF	CRF+类卷积交互式注意力机制	L-14, R-14, R-15, R-16	英文	类卷积注意力机制可以降噪以及获得重要的全局信息; 融入词的字符级特征, 有助于识别未登录词, 从而有助于属性词的预测	Ref.[62]
2021	SMOTE	随机森林	CRD	英文	一个包含不平衡数据处理的简单分类器, 使用最小的鲁棒特性集, 已经能够实现与属性抽取任务中 SOTA 可比较的结果	Ref.[63]

- 面向属性抽取的半监督学习模式

监督学习需要大量人工标注数据, 从而开发成本偏高. 相对地, 无监督属性抽取依赖专家制定规则或者模板, 虽然开发成本较低, 但可迁移性差. 因此, 结合两者所长的半监督属性抽取逐渐得到研究人员的重视.

Su 等人(2016)[66]提出一种直推式的半监督学习模型 TSVM (transductive support vector machines). 其核心思想是: 利用特征同分布现象(实际为分布近似现象), 借助观测样本的泛化学习, 对无标签的样本数据实现自

动标记, 形成训练集的扩展数据. 事实上, 该方法在思路上近似于主动学习, 但其研究目的仅在于优化训练数据的质量和数量, 借助少量人工标记数据获取大规模可靠的训练数据, 而非直接在测试阶段进行补充学习和推理(即 test-time inference, 基于主动学习和测试推理算法的工作在属性抽取领域并未出现). Su 等人的实验本身也是建立在低资源的模拟场景中, 其将部分训练数据视为真正的可观测样本, 其他训练数据则被视为标记未知的数据. 在此基础上, Su 等人利用 TSVM 实现了 23 122 个样本表示的自动生成, 形成了直推式的学习模式, 并证明利用附加的生成样本的模型能够取得较为明显的性能优化.

此外, Qu 等人(2019)[67]提出了一种基于半监督自训练的属性提取方法 AESS (aspect extraction based on semi-supervised self-training), 其核心思想与上述 Su 等人的工作近似, 都是以少量观测样本为参照, 实现近似样本的获取与应用. 两者的差异在于: Qu 等人专注于新型属性词的挖掘和扩充, 并形成了直接可用(look-up 模式)的属性词典. 具体地, Qu 等人首先依照统计指标 TF-IDF (TF-IDF 为检索领域常用的关键词鉴别指标)对可观测样本中的属性词项进行排序, 并选择排序较高的对象作为“黄金属性词(golden aspect term)”. 利用不同属性类中具有代表性的“黄金属性词”, Qu 等人建立了种子集, 并通过对种子属性词的语义表示学习和语义近似计算, 挖掘新的同类属性词, 借此构建了属性词词典. 实验结果表明: 该方法形成的可查属性词表及其表示向量能够用于解决短文本属性抽取结果不稳定的问题, 且在中英文数据集上均表现出了理想的效果.

Ansari 等人(2020)[68]提出了一种基于图的半监督属性抽取方法 GSSL (graph-based semi-supervised learning). 该方法将已标记属性词和未标记词项作为顶点, 将词与词之间的文本编辑距离作为边的权重, 构建了以词为元素的混淆图. 在此基础上, Ansari 等人针对图中每个未标记词项, 利用已标记的属性词形成邻居顶点社区(node community), 根据邻居节点的语义表示和属性类分布, 对未标记词项进行重构表示和类型预测. 该工作可以认为是一种图的补全方法, 利用图中的标记样本和加权边的信息传递, 有效地描述未知类的顶点和类型判别. 其在实际计算环节, 只需构建临界矩阵和参数矩阵, 通过图神经网络的演算模式即可实现顶点的补全(即词的属性类鉴别). 从而, 该方法形成了一种易于实现和操作的半监督学习模式. 实验结果表明, 基于图的半监督属性抽取方法比同期的监督学习方法具有更好的移植性和可扩展性. 上述 3 种半监督属性抽取方法的设计特色、优势和实验环境见表 7.

表 7 半监督方式的属性抽取方法比对

年份	模型	设计特色	数据集	语言	优点	来源
2016	TSVM	SVM+ 直推式学习	百度百科	中文	该模型能够在小语料条件下取得较好的抽取效果, 泛化学习能力较强, 可以节省大量的人力成本	Ref.[66]
2019	AESS	TF-IDF+ 种子词+ 词向量+词典	美团网的评论 数据+Citysearch corpus	中文/ 英文	避免了大量的文本标注, 充分利用 未标签数据的价值, 在中文和英文 数据集上都表现出了理想的效果	Ref.[67]
2020	GSSL	KNN+ 图模型	L-14, R-14, Yelp, Amazon	英文	与监督学习方法相比, 不需要人工标注 大量的数据, 节约了时间和成本. 鲁棒性和 修剪特征集的识别可以更好地表示该模型	Ref.[68]

- 面向属性抽取的自监督学习模式

自监督学习侧重利用生成模型在未标注样本上产生丰富的表示模式, 自助式地为监督学习提供感知学习的隐式样本. 基于自监督学习模式形成的模型中, 较为独特的一员是变分自编码器(variational autoencoder, VAE)[69]. 它能够以同一样本为蓝本, 生成多样的表示形式. 从而, 基于 VAE 的神经网络模型往往能够适应样本表示形式的微妙变化, 提升其自身的泛化能力.

在属性抽取领域, Liao 等人(2019)[70]提出将 VAE 思想结合到词项的编码阶段, 侧重在提升模型的同时感知属性词局部和全局特性的能力(该模型简称 LGC). 具体地, Liao 等人建立了局部语境特征表示模块(local context modeling, LCM)和全局语境表示模块(global context modeling, GCM). LCM 并未使用 VAE 进行编码, 而是利用常见的 LSTM 循环神经网络输出词项的隐状态, 并将其定义为局部语境敏感的词项表示. 相对地, GCM 利用了 VAE 对 LCM 输出的表示进行“改写”(即利用变分原理和随机采样, 重构和推断新的隐状态), 并将“改写”得到的隐状态作为全局语境敏感的词项. 在此基础上, Liao 等人对局部和全局语境敏感的表示作进

一步加工, 包括注意力计算、耦合编码和信息融合, 形成了用于属性类别判别的最终词项表示. 该项研究中, VAE 扮演了极为重要的角色. 在原理上, 经过 VAE“改写”的多样表示, 有助于神经网络感知词项的概念级信息(每个词项的词义丰富性可在“掺和”了VAE扰动的表示中得以呈现, 即概念化的信息感知), 从而, 模型对每个词在不同语境中呈现的差异性概念具有了较高的鉴别能力. 相对地, 在未结合VAE的LCM中, 局部特征的表示尽管更为精确, 但模式显然较为“僵硬”, 使得模型难以适应不同语境中语义表示模式的变化. 较为有形的例子是, 属性词项在某些语境中扮演了“非属性”的角色, 原因在于其概念发生了变化(特别是在一词多义情况较为多见的数据中). LCM 过分依赖训练数据中的样本, 泛化能力不强, VAE 的介入, 可在一定程度上缓解该问题(使模型具有了一定的“柔性”). 值得注意的是: VAE 的自监督学习模式使得 VAE 的利用并不需要外部数据的支持, 从而其可操作性更强.

4.3 数据驱动的属性抽取方法

数据驱动的属性抽取方法研究主要涉及两类: 其一是从数据的本质特性入手, 结合人工观测的经验和模型运行时(run-time)的响应特点, 对样本的处理和应用模式进行调整; 其二是瞄准样本数量和质量导致的弊端, 通过扩展和增强数据的方式, 实现抽取模型的优化. 表 8 对上述两类数据驱动方法的设计特色、优势和实验环境进行了罗列, 方法解析如下文所示.

表 8 数据驱动的属性抽取方法

	年份	模型	设计特色	数据集	语言	优点	来源
样本重构与回收	2011	NSSR	CRF+核心句+句法结构	COAE 2009	中文	将核心句和句法结构相结合, 提高CRF的标注效能; 单一领域表现更优	Ref.[71]
	2020	BBCR	BERT+Pointer net-work+BiSELF-CRF	L-14, R-14, R-15, R-16	英文	利用指针网络改善边界定位问题, 单独训练的指针网络作为后处理器, 可以很容易地与不同的属性抽取耦合	Ref.[37]
	2019	TOWE	LSTM+CRF+IOG	Lap-14, Res-14, Res-15, Res-16	英文	IOG可以有效地将目标信息分别编码成左右上下文. 然后结合目标的左右上下文和全局上下文, 提取解码器中相应的属性词	Ref.[29]
数据扩展与增强	2020	MASS	MaskFrag+Encoder+Decoder	Lap-14, Res-14, Res-15, Res-16	英文	可控性强, 可以产生更多样化的数据, 亦适用于分块、命名实体识别等任务	Ref.[72]
	2021	BRIDGE	SynBridge+SemBridge	L-14, R-14, R-15	英文	提出了一种新的主动域自适应方法, 通过构建句法和语义桥来积极增强属性词的可转移性, 并且设计了一个轻量级的端到端标记器实现序列标记	Ref.[73]

注: Lap-14、Res-14、Res-15、Res-16 分别表示重新标注之后的 L-14、R-14、R-15、R-16

- 样本重构与回收

南京大学的 Zhang 等人(2011)[71]提出了一种基于核心句的属性抽取方法 NSSR (nuclear sentences and syntactic relations). 核心句是指原句中围绕属性进行表述的核心文字片段(含属性表述及其密切相关的上下文). Zhang 等人的研究动机在于“相比于原句, 其内含的核心句更便于针对属性表述进行聚焦”(事实上, 核心句含有较少的迷惑性上下文, 有助于模式回避不必要的“排他”处理). 为此, Zhang 等人结合句法关系特征和标记信息, 总结了 10 种常见的核心句句法模式, 并以此作为 7 种启发式规则的设计基础. 利用此类启发式规则, 实现高可靠性的核心句抽取方法, 并在实验环节检验了基于核心句的 CRF 属性抽取性能, 证明了核心句抽取对于属性判别的重要支撑性作用.

近期, Wei 等人(2020)[37]详细分析了现有神经属性抽取模型的输出错误, 发现错误样本中存在“边界异常”的个体并不在少数. 根据他们的界定标准, “边界异常”特指属性表述的前后端存在多字少词(或少字多词)的现象. 事实上, “边界异常”隐含地反映出: 模型能够一定程度上较为准确地鉴别属性的表述位置, 但在抽取环节对文字片段边缘的微妙差异并未进行有效矫正. 为此, Wei 等人提出在现有神经抽取模型之后耦合一套边界重定位(repositioning)的后处理模块. 然而, 重定位模型需要边界正确和边界错误的样本进行训练, 而后者却难以获取. 获取困难的根本原因并不是人工标注的时效性低, 而是不同抽取模型具有自身的“个性”(建模机理不

同, 输出的正例和负例分布不同, 且负例的表现形式也不同), 从而导致人工伪造的边界判定错误, 并不绝对是模型真实的输出错误, 其用于训练的价值具有极高的不确定性. 因此, Wei 等人提出了运行时回收策略. 其目的是在训练和开发阶段实时监控抽取模型产出的错误, 并借助边界的字符串比对算法, 认定并回收“边界异常”的真实样本(即特定模型输出的“边界异常”样本). 在此基础上, Wei 等人开发了基于指针网络的边界重定位模型 BBCR (BERT-BiSELF-CRF-repositioning), 并验证了 BBCR 可以耦合在不同神经抽取模型之后, 且普遍地对抽取性能产生优化.

- 数据扩展与增强

以监督学习为基础的神经网络抽取模型往往受训练数据稀疏和质量的影响, 并不能极大地发挥其强大的深度学习能力. 数据扩展和增强是弥补这一不足的重要手段.

Zhang 等人(2015)[74]和 Wang 等人(2015)[75]基于 WordNet 获取已知属性词项的同义词, 并通过同义词替换的方式制作“具有相同上下文和迥异属性表述”的句子级伪样本, 从而刚性地扩展了训练数据集. 进一步地, Kobayashi (2018)[76]和 Wu 等人(2019)[77]通过预训练语言模型实现同义词替换. 该类数据扩展方法更为“柔性”, 且利用了预训练语言模型强大的语义感知和语言生成能力, 从而对新型的属性观测样本依然具有替换能力, 使得数据扩展方法一定程度上能够满足时新性的实际应用需求. 但上述方法的基本思路都是依照“同义”这一特质进行属性表述的改写, 不可避免地面临“多义词”导致的替换误差问题. 显然, 替换偏差将形成错误的训练样本, 反而误导抽取模型的表示学习过程. 近期, Li 等人(2020)[72]提出并实现了一种有条件的句子级数据增强方法 MASS (masked sequence-to-sequence). 与上述基于预训练语言模型的方法类似, MASS 同样使用了遮蔽策略. 不同点在于: MASS 并不遮蔽可观测的属性词项, 反而遮蔽其上下文中的其他词项, 并利用 Seq2Seq (sequence-to-sequence, 即序列到序列)生成模型, 专门生成被遮蔽的非属性词项, 从而实现“保留属性词项”条件下的语境复述生成. 换言之, MASS 重塑了一句话, 但却保留了属性表述不变. Li 等人在实验环节对每个训练样本所在的语句重塑了 4 个变体语句, 不仅扩展了训练样本, 也同时增强了训练数据的语用多样性. 实验结果验证了 MASS 提供的增强数据, 切实地提高了属性抽取模型的性能.

比较有特色的人工数据增强工作来自 Fan 等人(2019)[29], 他们重启了数据的人工标注. 但与以往不同, Fan 等人并未提供新的属性表述和上下文样本(即并未扩展观测数据), 相反地, 其在原始观测数据上进行了精细加工, 不仅标注了属性词的相对位置信息, 同时标记了相关情感词的位置信息, 并将训练样本限定在属性词项和相关情感词同时出现的句子上. 相应地, Fan 等人开发了基于位置信息的分段编码模型 TOWE (target-oriented opinion words extraction), 能够对候选属性词项(即句子中的任意词项)实施同向的上下文依赖的 LSTM 编码(inward-LSTM 编码)和逆向的上下文依赖的 LSTM 编码(outward-LSTM), 形成以候选词项为中心的前文双向编码和后文双向编码模型. 在此基础上, 该模型结合了句子级的全局信息进行了联合编码. 实验结果表明: TOWE 不仅能够支持“前端语境”“属性词”和“后端语境”的三元分类, 也能支持基于 BIO 的 CRF 自动标记过程, 从而为属性抽取研究额外提供了一套策略级的处理模式. 更为重要的是, Fan 等人提供的标记数据能够支持属性和情感分析的多任务学习(如第 4.4 节).

此外, 跨领域应用完备训练的神经抽取模型往往遭遇目标领域标注数据不足的情况, 借助跨领域自监督学习技术, 可一定程度上缓解上述问题. 跨领域自监督学习的核心是进行信息的传递与借鉴, 即通过学习源领域标注数据的语言信息, 实现针对目标领域“同构”数据的学习模式, 呈现为一种“草船借箭”的学习模式. 事实上, 跨领域自监督与数据增强有着异曲同工的作用, 区别仅在于前者利用外部领域数据的隐含信息, 后者则较为直观地利用另一领域数据的信息. 在这一方面, Chen 等人(2021)[73]提出一种新颖的主动跨领域自适应方法, 目标是通过积极补充可传输的知识, 进行属性词信息的跨领域传递. 具体地, 该方法为所有单词构建句法桥和语义桥, 并将其视为信息传递的媒介. 句法桥旨在识别每个单词在不同领域中扮演的句法角色, 并依照角色进行信息传递. 在实现过程中, 句法角色相同的词, 即使词形不一致也可作为桥梁, 将源领域上下文编码信息传递到目标领域的语境编码过程中, 从而补充目标领域属性词的上下文信息. 语义桥是源领域属性词和目标领域属性词之间的另一“桥梁”. Chen 等人(2021)将目标领域的句子作为查询(即查询语句), 利用句法

结构匹配, 在源领域数据中获取语法结构近似的样本, 并将其视为外部信息, 融入目标领域中查询语句的句子级语义编码. 借助上述理论基础和方法学设计, Chen 等人(2021)能够通过门控操作融合语义桥和句法桥的信息, 形成更健壮的目标领域抽取模型.

4.4 结合情感词的属性抽取方法

情感分析和属性抽取任务是自然语言领域的“双子座”, 关系极为密切. 原因在于: 首先, 情感词与属性词都是评论文本中的高频语言单位(如第 2.2 节所示, 绝大部分语料集都来自评论数据); 其次, 两者往往在局部紧凑的语言单位(句子、小句或语块)中共同出现; 最后, 两者之间存在语义层面的关联性(比如“先进性能”中, 属性“性能”与情感词“先进”有着极为密切的语义关系). 也因此, 对同一语句的语义编码适用于两种任务中不同目标的判别, 且情感与属性可以互为推理线索. 从而, 针对属性抽取的相关研究已将情感分析技术作为重要的辅助手段予以运用, 并在多任务学习(multi-task learning)架构下实现了不同抽取模型. 表 9 对结合情感词的属性抽取方法的设计特色、优势和实验环境进行了罗列, 方法解析如下文所示.

表 9 结合情感词的属性抽取比对

年份	模型	设计特色	数据集	语言	优点	来源
2016	RNCRF	RNN+CRF	L-14, R-14	英文	提出了一种新的联合模型, 将递归神经网络和条件随机场集成到一个统一的框架中, 用于显式属性和情感词的共同提取	Ref.[78]
2017	CMLA	Attention+GRU+DP	L-14, R-14, R-15	英文	提出了一种耦合多层注意力模型, 其提供了端到端的解决方案, 并且不需要任何依存关系分析器或其他辅助语言资源进行预处理	Ref.[11]
2018	HAST	THA+STN	L-14, R-14, R-15, R-16	英文	利用情感摘要和属性检测历史两个线索进行属性抽取; 情感摘要是从整个输入句子中提炼出来的, 以每个当前令牌为条件进行属性预测; 属性检测历史是从以前的属性预测中提炼出来的, 以便利用坐标结构和标记模式约束升级属性预测	Ref.[39]
2020	SpanMlt	BiLSTM+BERT+Attention	L-14, R-14, R-15, R-16	英文	首次将属性词和情感词成对抽取, 提出一个端到端的模型, 将情感词作为属性抽取的线索或辅助提升属性抽取的性能	Ref.[36]
2020	LOTN	BiLSTM+Attention	L-14, R-14, R-15, R-16	英文	通过迁移学习的知识, 把预训练的情感分类中得到的情感词传递给 TOWE, 增强了性能, 与 IOG 相比降低了复杂度	Ref.[79]

Wang 等人(2016)[78]首次将属性与情感的关系纳入编码过程, 构建了融入依存特征的递归神经条件随机场模型(recursive neural conditional random fields, RNCRF). RNCRF 以依存关系树为桥梁, 在属性词和依存情感词之间进行了信息交互与融合, 形成了依存感知(dependency-aware)条件下的语义编码模式. 特别地, Wang 等人在属性抽取和情感分类两种任务中训练 RNCRF, 使得上下文特征和关系特征的表示学习经历了两种任务的不同考验, 提升了特征的表征能力.

考虑到 RNCRF 对依存分析的准确性有着极高的要求,Wang 等人(2017)[11]提出了一种替代方案, 即耦合多层注意力的 CMLA 模型(coupled multi-layer attention). 区别于 RNCRF, CMLA 并不利用现有的依存分析工具预先甄别属性和情感的关系, 而是借助交互注意力实现关联性的感知学习以及属性与关联情感的信息交互. 具体地, CMLA 具有多层的双通道注意力网络, 每层中相互耦合的注意网络分别关注属性和情感的注意力分布, 并进行特征信息的双向传递和交互加权. 实验结果表明, CMLA 能够在依存关系未知(dependency-unaware)的条件下, 依然较好地实现多任务表示学习, 并助力属性抽取的性能优化.

类似地, Li 等人(2018)[39]也利用属性和情感词的交互注意力构建多任务网络. 不同的是, Li 等人利用了历史注意力和选择性转换机制 HAST (history attention and selective transformation)对多任务模式进行升级. 具体地, HAST 蕴含两套 BiLSTM 循环神经网络, 分别为“节制性注意力加权通道”THA 和“选择性注意力转移通道”STN. 其中, THA 专门独立学习和记录历史属性的预测信息(历史信息是指当前时刻之前的编码信息), 进而仅仅利用(即有节制的利用)来自历史属性的注意力, 影响当前时刻的隐状态编码; STN 则聚焦在情感词的识

别和表示问题上, 在充分利用 THA 提供的历史属性和句子级全局信息的条件下, STN 将对潜在的情感词提供更为确切的表示, 并将这类表示逆推式地转移回 THA 通道, 参与其中对属性词的注意力加权. 事实上, HAST 体现了一种有节制且重点突出的思维模式, 仅仅在“属性-属性”和“情感-属性”之间的关联性上寻找突破口, 专注于紧密相关且使任务直接受益的上下文特征, 进行编码优化和表示学习.

值得指出的是, 单一语句中的情感词可能较为丰富且分布位置不同, 其与特定一个或多个属性词的关联性并不明确, 从而影响了上述多任务学习架构中属性与情感词的准确注意力交互. 针对这一问题, Wu 等人(2020)[79]采纳了 Fan 等人(2019)[29]增强后的数据样本(如第 4.3 节的第 2 部分), 并充分利用了其中明确标定的属性和相关情感词标记, 建立了一种基于启发式规则的严密交互注意力加权策略(严密交互注意力计算特指关联情感的注意力计算, 而非全局所有情感词的综合注意力加权), 且借此将 Fan 等人的 TOWE 模型升级为多任务架构下的观点转移网络(latent opinions transfer network, LOTN). 实验结果表明, LOTN 切实产生了明显的性能优势. 该工作为面向属性分析的多任务学习架构提供了最为可信的佐证, 即情感信息的传递和多任务框架下的参数共享, 对提升属性抽取有着直接的贡献. 近期, Zhao 等人(2020)[36]提出了基于 BERT 和 BiLSTM 的端到端多任务网络 SpanMlt (span-based multi-task), 并首次实现了属性词和情感词的成对抽取.

5 总结与展望

总体上, 属性抽取问题已经得到自然语言领域学者的广泛关注, 近几年的研究热度持续走高, 相关的数据加工和方法学设计水平也相应地得到不断的提高. 借助基于神经网络的深度表示学习技术, 属性抽取研究已经进入了“神经”时代, 相应的技术也有效提升了抽取精度, 以及抽取模型的健壮性和泛化能力. 特别地, 以 Transformer 为基础的预训练语言模型以及这类模型蕴藏的自监督学习模式, 正广泛影响着相关学者的研究思路, 促进了突破性工作的不断产出. 尽管如此, 属性抽取研究尚存在如下短板, 其反映出的科学问题将可能主导或引领未来的研究趋势.

- 低资源属性抽取: 低资源问题是自然语言领域中不同任务面临的共性问题, 其特指领域切换情境下的标注数据短缺和监督学习无法实施的关键问题. 目前, 科研人员已在面向属性分析的数据扩展和数据增强(如第 4.3 节)方面进行了尝试, 并取得了较好的效果. 但是, 低资源问题并不能简单依赖扩展和增强技术而得到解决. 原因在于: 低资源场景只给出极为少量的种子标记样本, 无法覆盖相对全面的语言现象, 仅仅能够用于指导人的初步认知理解(任务定义的理解), 而无法为机器的认知学习提供足够的参考. 扩展与增强技术(如基于同义词替换、同质异构复述生成等)仅能对种子进行微小的变化, 产生局限性较高的变种样本, 难以提升语用和语义模式的覆盖率. 从而, 解决低资源问题需要从技术和数据两方面入手: 在技术层面, 开发利用远程监督学习、主动学习和跨领域迁移学习技术, 将对低资源场景下的属性抽取方法研究提供重要的支撑; 在数据层面, 人工数据建设是必不可少的一环, 设计辅助标注技术(比如利用知识库进行文本结构化分析和属性结构化数据的挖掘)和标记质量评估方法, 将有助于以最为直接的方式扩展属性数据库. 特别地, 利用变分自编码和生成对抗网络, 能够分别在监督学习的前期(预处理)和中期(运行时)生成迷惑性样本, 以“柔性”的方式实现数据增强.
- 跨语言属性抽取: 值得特别指出的是, 现有属性抽取领域中关于跨语言问题的工作极为稀缺. 事实上, 在多语种数据上围绕属性进行语义表示学习研究, 将产生具有极高应用价值的成果. 比如前文提到的低资源场景下, 稀疏的种子数据难以支撑神经模型的监督学习, 其根本原因是语言现象(语用模式和语义知识范畴)覆盖面小. 那么针对特定目标语言的低资源场景, 利用标注资源充沛的源语言抽取模型进行适当迁移, 将能够快速形成基准性(baseline)的目标语言属性抽取模型. 特别地, 当结合多任务学习架构, 分别建立双语翻译通道、单语种遮蔽语言模型通道和双语分治的属性抽取通道后, 借助参数共享和特征监听机制, 能够形成多套跨语种的语义表示学习和特征抽象模型. 此外, 不同语种属性的表述模式(语法结构)、语用特点(多义词)及其与情感和语境的关联模式(依存模式)必然存在语言层面的固有差异. 针对不同语种进行现有抽取模型的性能检验, 并借此探究适应多语种个性化语

言处理的通用模型, 将有望扩充属性抽取领域的研究内容, 并且切实解决跨语言属性抽取中的实际应用问题.

- 多模态属性抽取: 现有研究往往忽视的一项重要前提是属性框架体系的完备性. 无论是人工标注或是机器自动分类属性, 都需要专家预先对特定领域和特定目标建立系统且严密的属性体系(包括属性类标记、定义和可区分性的界定标准). 当领域和目标进行切换时, 新领域的属性框架体系往往是未知的, 使得现有数据标注和机器建模技术都无法快速施展. 针对这一问题, 利用多模态信息处理技术实现属性体系的自动认知, 将是一项极富挑战且具趣味性的工作. 特别地, 利用图像模态的匹配技术, 将有助于属性体系在同类目标中进行共享(比如, 汽车的属性体系能够大范围地覆盖航空器的属性体系). 此外, 利用跨模态的特征表示转换, 将能够极大地助力文本层面的属性语义感知效果(比如, 结合图注生成模型 captioning 和属性抽取模型的多任务学习).
- 精细化研究: 除了上述低资源、跨语种和多模态属性抽取问题, 在现有针对高资源、单语种和独立模态的属性抽取研究中(主干研究内容), 技术方法的攻关对象仍存在一些空白, 包括: (1) 隐式属性的抽取问题, 即有指代词指向的属性和零指代属性的抽取问题, 针对这一问题的初步解法应集中于指代消解和生成模型的应用; (2) 多任务架构下的运行时推理(test-time inference)和强化学习可组成新型的神经属性抽取模型, 其可利用测试用例上非属性抽取类任务的执行和产出(比如预测情感), 形成运行时的奖励机制, 同步促进属性抽取模型的性能优化, 这类模型的设计工作在属性抽取领域尚未出现, 但近期已在其他自然语言理解任务中得以实践; (3) 基于知识图谱的属性抽取研究相对薄弱, 目前, 利用诸如 FrameNet(框架语义网络)、ConceptNet(知识概念网络)及其框架的其他抽取技术已经得到较为广泛的研究, 其丰富的语义和知识资源、关系架构和样例, 能够辅助基于图和外部知识的抽取方法设计.

References:

[1] Liu B. Sentiment analysis and opinion mining. In: Proc. of the Synthesis Lectures on Human Language Technologies. 2012. 1−167.

[2] Wagner J, Arora P, Cortes S, Barman U, Dasha B, Jennifer F, Lamia T. DCU: Aspect-based polarity classification for semeval task 4. In: Proc. of the 8th Int'l Wokshop on Semantic Evaluation (SemEval 2014). 2014. 223−229.

[3] Dai HL, Song YQ. Neural aspect and opinion term extraction with mined rules as weak supervision. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. 5268−5277.

[4] Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S. SemEval-2014 task 4: Aspect based sentiment analysis. In: Proc. of the 8th Int'l Workshop on Semantic Evaluation (SemEval 2014). Dublin: Association for Computational Linguistics, 2014. 27−35.

[5] Pontiki M, Galanis D, Papageorgiou H, Manandhar S, Androutsopoulos I. Semeval-2015 task 12: Aspect based sentiment analysis. In: Proc. of the 9th Int'l Workshop on Semantic Evaluation (SemEval 2015). Denver: Association for Computational Linguistics, 2015. 486−495.

[6] Pontiki M, Galanis D, Papageorgiou H, Androutsopoulos I, Manandhar S, Al-Smadi M, Al-Ayyoub M, Zhao YY, Qin B, De CO. SemEval-2016 task 5: Aspect based sentiment analysis. In: Proc. of the 10th Int'l Workshop on Semantic Evaluation (SemEval 2016). San Diego: Association for Computational Linguistics, 2016. 19−30.

[7] Liu Q, Gao ZQ, Liu B, Zhang YL. Automated rule selection for aspect extraction in opinion mining. In: Proc. of the 24th Int'l Joint Conf. on Artificial Intelligence. 2015. 1291−1297.

[8] Mukherjee A, Liu B. Aspect extraction through semi-supervised modeling. In: Proc. of the 50th Annual Meeting of the Association for Computational Linguistics. 2012. 339−348.

[9] Chen ZY, Mukherjee A, Liu B. Aspect extraction with automated prior knowledge learning. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014. 347−358.

[10] Jiang SY, Guo LD, Wang LX, Fu SH. Survey on opinion target extraction. Acta Automatica Sinica, 2018, 44(7): 1165−1182 (in Chinese with English abstract).

[11] Wang WY, Pan SJ, Dahlmeier D, Xiao XK. Coupled multilayer attentions for co-extraction of aspect and opinion terms. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. 2017. 3316−3322.

[12] Le R, Roth D. Design challenges and misconceptions in named entity recognition. In: Proc. of the 13th Conf. on Computational Natural Language. 2009. 147−155.

[13] Fan Z, Wu Z, Dai XY, Huang S, Chen J. Target-oriented opinion words extraction with target-fused neural sequence labeling. In: Proc. of the 2019 Association for Computational Linguistics. 2019. 2509−2519.

[14] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification usingmachine learning techniques. In: Proc. of the 2002 Conf. on Empirical Methods in Natural Language Processing. 2002. 79−86.

[15] Hu MQ, Liu B. Mining and summarizing customer reviews. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2004. 168−177.

[16] Farra N, Mckeown K, Habash N. Annotating targets of opinions in Arabic using crowdsourcing. In: Proc. of the ACL-2015 Workshop on Arabic Natural Language. 2015. 89−98.

[17] Xu R, Xia Y, Wong KF, Li WJ. Opinion annotation in online Chinese product reviews. In: Proc. of the 6th Int'l Conf. on Language Resources and Evaluation. 2008. 1625−1632.

[18] Akhtar MS, Ekbal A, Bhattacharyya P. Aspect based sentiment analysis in Hindi: Resource creation and evaluation. In: Proc. of the 10th Int'l Conf. on Language Resources and Evaluation. 2016. 2703−2709.

[19] Bhattacharya A, Debnath A, Shrivastava M. Enhancing aspect extraction in Hindi. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics. 2021. 140−149.

[20] Yang H, Zeng B, Yang JH, Song YW, Xu RY. A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction. Neurocomputing, 2021, 419: 344−356.

[21] Cruz I, Gelbukh AF, Sidorov G. Implicit aspect indicator extraction for aspect based opinion mining. Journal of Computational Linguistics and Application, 2014, 5(2): 135−152.

[22] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the 18th Int'l Conf. on Machine Learning. 2001. 282−289.

[23] Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 1962, 160(1): 106−154.

[24] Elman JL. Finding structure in time. Cognitive Science, 1990, 14(2): 179−211.

[25] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735−1780.

[26] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing. 2014. 1724−1734.

[27] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Advances in Neural Information Processing Systems. 2017, 30: 6000−6010.

[28] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies. 2018. 4171−4186.

[29] Fan ZF, Wu Z, Dai XY, Huang SJ, Chen JJ. Target-oriented opinion words extraction with target-fused neural sequence labeling. In: Proc. of the Annual Conf. of the North American Chapter of the Association for Computational Linguistics. 2019. 2509−2518.

[30] Cho K, Courville A, Bengio Y. Describing multimedia content using attention-based encoder-decoder networks. IEEE Trans. on Multimedia, 2015, 17(11): 1875−1886.

[31] Xu S, Li H, Yuan P, Wu YZ, He XD, Zhou BW. Self-Attention guided copy mechanism for abstractive summarization. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. 1355−1362.

[32] Wang YX, Guo Y, Zhu SQ. Slot attention with value normalization for multi-domain dialogue state tracking. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. 2020. 3019−3028.

[33] Dai ZH, Yang ZL, Yang YM, Carbonell JM, Le QV, Salakhutdinov R. Transformer-xl: Attentive language models beyond a fixed-length context. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. 2978−2988.

[34] Child R, Gray S, Radford A, Sutskever I. Generating long sequences with sparse transformers. arXiv:1904.10509, 2019.

[35] Dehghani M, Gouws S, Vinyals O, Uszkoreit J, Kaiser L. Universal transformers. arXiv:1807.03819, 2018.

[36] Zhao H, Huang LT, Zhang R, Lu Q, Xue H. Spanmlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. 3239−3248.

[37] Wei ZK, Hong Y, Zou BW, Cheng M, Yao JM. Don't eclipse your arts due to small discrepancies: Boundary repositioning with a pointer network for aspect extraction. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. 3678−3684.

[38] Li X, Lam W. Deep multi-task learning for aspect term extraction with memory interaction. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. 2017. 2886−2892.

[39] Li X, Bing LD, Li P, Lam W, Yang ZM. Aspect term extraction with history attention and selective transformation. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. 2018. 4194−4200.

[40] George AM. Wordnet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39−41.

[41] Li Z, Feng J, Zhu XY. Movie review mining and summarization. In: Proc. of the ACM 15th Conf. on Information and Knowledge Management. ACM, 2006. 43−50.

[42] Qiu G, Liu B, Bu J, Chen CL. Opinion word expansion and target extraction through double propagation. Computational Linguistics, 2011, 37(1): 9−27.

[43] Jakob N, Darmstadt TU, Gurevych I. Extraction opinion targets in a single and cross-domain setting with conditional random fields. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. 2010. 1035−1045.

[44] Xu B, Zhao TJ, Wang SY, Zheng DQ. Extraction of opinion targets based on shallow parsing features. Acta Automatica Sinica, 2011, 37(10): 1241−1247 (in Chinese with English abstract).

[45] Chernyshevich M. IHS R&D Belarus: Cross-domain extraction of product features using conditional random fields. In: Proc. of the 8th Int'l Workshop on Semantic Evaluation (SemEval 2014). 2014. 309−313.

[46] Liu PF, Joty S, Meng H. Fine-grained opinion mining with recurrent neural networks and word embedding. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. 2015. 1433−1443.

[47] Toh Z, Su J. NLANGP at SemEval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In: Proc. of the 10th Int'l Workshop on Semantic Evaluation (SemEval-2016). 2016. 282−288.

[48] Xu H, Liu B, Shu L, Philip SY. Double embeddings and CNN-based sequence labeling for aspect extraction. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. 2018. 592−598.

[49] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011, 12(1): 2493−2537.

[50] Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. 2015. 2539−2544.

[51] Ma DL, Li SJ, Wu FZ, Xie X, Wang HF. Exploring sequence-to-sequence learning in aspect term extraction. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. 3538−3547.

[52] Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora. In: Proc. of the 42th Annual Meeting of the Association for Computational Linguistics. 2004. 415−422.

[53] Chen JX, Ji DH, Tan CL, Niu ZY. Unsupervised feature selection for relation extraction. In: Proc. of the Conf. on Including Posters/Demos and Tutorial Abstracts. 2005. 262−267.

[54] Popescu AM, Etzioni O. Extracting product features and opinions from reviews. In: Proc. of the Conf. on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2007. 9−28.

[55] García-Pablos A, Cuadros M, Rigau G. V3: Unsupervised aspect based sentiment analysis for semeval2015 task 12. In: Proc. of the 9th Int'l Workshop on Semantic Evaluation (SemEval 2015). 2015. 714−718.

[56] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 1998, 30(1−7): 107−117.

[57] Liu Q, Gao Z, Liu B, Zhang YL. Automated rule selection for opinion target extraction. Knowledge-based Systems, 2016, 104: 74−88.

[58] Jin W, Ho HH. A novel lexicalized HMM-based learning framework for web opinion mining. In: Proc. of the 26th Annual Int'l Conf. on Machine Learning. New York: ACM, 2009. 465−472.

[59] Poria S, Cambria E, Gelbukh A. Aspect extraction for opinion mining with a deep convolutional neural network. Knowledge-based Systems, 2016, 108: 42−49.

[60] Shu L, Xu H, Liu B. Lifelong learning CRF for supervised aspect extraction. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. 148−154.

[61] Chen Z, Liu B. Lifelong machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2018, 12(3): 1−207.

[62] Wei ZK, Cheng M, Zhou XB, Li ZF, Zou BW, Hong Y, Yao JM. Convolutional interactive attention mechanism for aspect term. Journal of Computer Research and Development, 2020, 57(11): 208−218. (in Chinese with English abstract)

[63] Venugopalan M, Gupta D, Bhatia V. A supervised approach to aspect term extraction using minimal robust features for sentiment analysis. In: Proc. of the Progress in Advanced Computing and Intelligent Engineering. Singapore: Springer, 2021. 237−251.

[64] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321−357.

[65] Ho TK. Random decision forests. In: Proc. of the 3th Int'l Conf. on Document Analysis and Recognition, Vol.1. IEEE, 1995. 278−282.

[66] Su FL, Xie QH, Huang QA, Qiu JY, Yue ZJ. Semi-supervised method for attribute extraction based on transductive learning. Journal of Shandong University (Natural Science), 2016, 51(3): 111−115 (in Chinese with English abstract).

[67] Qu SW, Wu CY, Wang XR. Aspects extraction based on semi-supervised self-training. CAAI Trans. on Intelligent Systems, 2019, 14(04): 635−641 (in Chinese with English abstract).

[68] Ansari G, Saxena C, Ahmad T, Doja MN. Aspect term extraction using graph-based semi-supervised learning. Procedia Computer Science, 2020, 167: 2080−2090.

[69] Kingma DP, Welling M. Autoencoding variational bayes. arXiv:1312.6114.2013, 2013.

[70] Liao M, Li J, Zhang H, Wang LZ, Wu XX, Wong KF. Coupling global and local context for unsupervised aspect extraction. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language and the 9th Int'l Joint Conf. on Natural Language Processing. 2019. 4579−4589.

[71] Zhang L, Qian LF, Xu X. Comment target extraction based on nuclear sentences and syntactic relations. Journal of Chinese Information Processing, 2011, 25(3): 23−30 (in Chinese with English abstract).

[72] Li K, Chen CB, Quan XJ, Ling Q, Song Y. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. 7056−7066.

[73] Chen Z, Qian TY. Bridge-based active domain adaptation for aspect term extraction. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics. 2021. 317−327.

[74] Zhang X, Zhao JB, Cun YL. Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems. 2015. 649−657.

[75] Wang WY, Yang DY. That's so annoying!!! A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve Tweets. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. 2015. 2557−2563.

[76] Sosuke K. Contextual augmentation: Data augmentation by words with paradigmatic relations. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Short Papers), Vol.2. 2018. 452−457.

[77] Wu X, Lv SW, Zang LJ, Han JZ, Hu SL. Conditional BERT contextual augmentation. In: Proc. of the Computational Science. 2019. 84−95.

[78] Wang WY, Pan SJ, Dahlmeier D, Xiao XK. Recursive neural conditional random fields for aspect-based sentiment analysis. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. 2016. 616−626.

[79] Wu Z, Zhao F, Dai XY, Huang SJ, Chen JJ. Latent opinions transfer network for target-oriented opinion words extraction. In: Proc. of the 34th Association for the Advancement of Artificial Intelligence. 2020, 34(5): 9298−9305.

附中文参考文献:

[10] 蒋盛益, 郭林东, 王连喜, 符斯慧. 评价对象抽取研究综述. 自动化学报, 2018, 44(7): 1165−1182.

[44] 徐冰, 赵铁军, 王山雨, 郑德权. 基于浅层句法特征的评价对象抽取研究. 自动化学报, 2011, 37(10): 1241−1247.

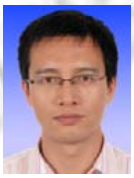
[62] 尉桢楷, 程梦, 周夏冰, 李志峰, 邹博伟, 洪宇, 姚建民. 基于类卷积交互式注意力机制的属性抽取研究. 计算机研究与发展, 2020, 57(11): 208−218.

[66] 苏丰龙, 谢庆华, 黄清泉, 邱继远, 岳振军. 基于直推式学习的半监督属性抽取. 山东大学学报(理学版), 2016, 51(3): 111−115.

[67] 曲昭伟, 吴春叶, 王晓茹. 半监督自训练的方面提取. 智能系统学报, 2019, 14(4): 635−641.

[71] 张莉, 钱玲飞, 许鑫. 基于核心句及句法关系的评价对象抽取. 中文信息学报, 2011, 25(3): 23−30.

徐庆婷(1994－), 女, 硕士, CCF 学生会员, 主要研究领域为属性抽取.



姚建民(1971－), 男, 博士, 主任研究员, 主要研究领域为机器翻译, 信息抽取.

洪宇(1978－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为信息抽取, 智能问答, 低资源机器翻译, 语篇分析.

周国栋(1967－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为自然语言处理.

潘雨晨(1996－), 男, 硕士生, 主要研究领域为属性抽取.