

面向主题模型的主题自动语义标注研究 综述*

凌洪飞 欧石燕

(南京大学信息管理学院 南京 210023)

摘要:【目的】对面向主题模型的主题自动语义标注方法进行总结与评述,以促进主题模型的发展与应用。

【文献范围】在 Web of Science 和 CNKI 数据库中分别以“Topic Labeling OR Topic Labelling OR Topic Tagging OR Topic Indexing”和“主题模型 AND (标注 OR 标签)”等检索式进行检索,通过手工筛选获得代表性文献 57 篇。【方法】对相关论文进行深入阅读与分析,以主题标注过程中主题标签的生成来源为线索,对已有方法进行分 类与比较分析。【结果】面向主题模型的主题自动语义标注包括候选标签生成与排序两个主要步骤,根据候选标签的生成来源可分为依靠自身语料库和依靠外部语料库两类方法。【局限】目前该领域的研究还不是很丰富,分析与评述不够系统和全面。【结论】该领域的研究仍具有较大探索空间,面向社交媒体内容的主题语义标注是未来研究方向,可结合更丰富的知识库并采用深度学习技术进行改进提升。

关键词: 主题语义标注 概率主题模型 隐含狄利克雷分布

分类号: G350 TP39

DOI: 10.11925/infotech.2096-3467.2018.1127

1 引 言

主题模型(Topic Model)是一类生成式概率模型,利用该类模型能够从大规模文本数据中自动抽取出隐含、抽象的主题信息。自 Blei 于 2003 年提出隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)模型以来^[1],主题模型得到极大关注与快速发展,在信息

检索、文本挖掘、自然语言处理等领域得到广泛应用^[2]。

在主题模型中,主题的抽取结果是一系列词汇的概率分布。一个关于“物化视图”的主题示例如表 1 所示,该主题由若干个词汇及其出现概率表示,这些词汇在语料库中的共现概率较高,彼此之间具有一定关联性,在一定程度上能够反映出一个主题概念。

表 1 关于“物化视图”的“主题-词汇”概率分布

词汇	views	view	materialized	maintenance	warehouse	tables	summary	updates	...
出现概率	0.01	0.01	0.05	0.05	0.03	0.02	0.02	0.02	...

虽然主题模型在许多领域得到广泛应用,但在实际应用中发现其抽取得到的“主题-词汇”概率分布并不易于用户理解^[3]。如何以一种更直观的方式对主题抽取结果进行解释和描述是一个亟待解决的问题,因此主题语义标注(Topic Labelling)应运而生。主题语义

标注是指使用一系列词汇作为语义标签,对主题模型抽取出的“主题-词汇”概率分布的语义进行显性描述,其目的是增强主题模型抽取结果的可读性与可理解性^[4-6]。在已有研究与应用中,一些学者直接使用“主题-词汇”分布中概率最高的若干个词汇构成的集合作为

通讯作者: 欧石燕, ORCID: 0000-0001-8617-6987, E-mail: oushiyan@nju.edu.cn。

*本文系国家自然科学基金重点项目“基于关联数据的学术文献内容语义发布及其应用研究”(项目编号: 17ATQ001)的研究成果之一。

主题的语义标签^[1]，这类方法虽然简单直接，但标签往往难以理解(尤其是针对专业领域的主题)。譬如，对于表 1 所示的“主题-词汇”分布，选择出现概率最高的若干个词汇构成的集合为“materialized、maintenance、warehouse、tables”，该词汇标签对于数据库方面的专家来说可能是一个好的表示，但对于普通用户则可能很难理解。此外，有学者通过人工标注方式，人为地选择一个短语作为主题标签表示该主题的语义。但人工标注不仅需要耗费较多的时间精力，而且由于标注者的主观性，往往会造成标注结果不一致，如表 1 所示的主题有可能使用“物化视图(Materialized View)”或“数据仓库(Data Warehouse)”两种标签表示。


有鉴于此，一些学者开始探索如何采用文本挖掘和自然语言处理方法进行主题自动语义标注，为“主题-词汇”概率分布自动生成并选择一个合适的语义标

签，显性表示该主题的语义信息。这不仅避免了人工参与，减少主观性和人工劳动，而且通过机器进行标签选择和比较，也使语义标注结果更具客观性。本文对面向主题模型的主题自动语义标注相关研究进行总结和评述。

2 主题语义标注中语义标签的种类

给定一个采用“主题-词汇”概率分布表示的隐性主题，可以采用多种方式表示该主题的语义信息。在已有研究中，主流方式多采用词汇集合或短语作为语义标签来表示“主题-词汇”概率分布的语义，但近来也有学者尝试使用单个词汇^[7]、句子^[8-9]以及图片^[10-11]等作为标签来表示主题。采用不同种类的语义标签对表 1 所示“主题-词汇”分布进行语义标注，结果如表 2 所示。

表 2 采用不同类型语义标签对主题进行标注的结果示例

“主题-词汇”概率分布		标签类别	标签
views	0.10	词汇集合	views, view, materialized, maintenance, warehouse, tables
view	0.10	单个词汇	view, maintenance
materialized	0.05	短语	materialized view, data warehouse
maintenance	0.05	句子	materialized view selection and maintenance using multi-query optimization.
warehouse	0.03	图片	
tables	0.02		
summary	0.02		
updates	0.02		

不同类型的语义标签各有特点，其适用场景也有所差异。一些学者对不同主题语义标注方法在信息检索中的效果进行比较研究，结果发现，相较于词汇集合和图片，短语标签能够更好地概括主题的语义信息，更易于用户理解，使用户能够在更短时间内检索到更多的相关文档^[12-13]。也有研究表明，不同主题适合于不同类型的标签，有的适合于文本标签，而有的则更适合于图片表示^[14]。

Mei 等指出，一个高质量的主题语义标签应该具有以下 4 个特性：

- (1) 可理解性，即标签对于用户是易于理解的；
- (2) 语义相关性，即标签能够真正反映“主题-词汇”分布的语义信息；
- (3) 全面覆盖性，即标签能够覆盖“主题-词汇”分布所包含的全部语义信息；
- (4) 区分性，即标签能够反映不同主题所表达语义的不同^[4]。

根据这 4 个标准，对不同类型语义标签的优缺点进行总结与比较，如表 3 所示。

表 3 主题语义标签的种类及其优缺点

主题语义标签		优点	缺点
词汇集合	简单、易实现	对于特定领域的主题，不易被用户理解表达的语义过于笼统，无法覆盖“主题-词汇”分布所表达的全部语义信息	与单个词汇表示法相反，该方法表达的语义过于具体
单个词汇	简单、易实现		
单个或多个句子	易于理解，主题区分度明显		
短语	介于单个词汇和句子这两种表示方式之间，是目前使用最多的一种表示主题语义的方式，易于理解		
图片	直观，用户能快速理解，且与语种无关		

3 主题自动语义标注方法

在主题模型诞生和广泛应用之前,一些学者就对文本聚类结果的语义标注进行了探索。基于主题模型的主题抽取与文本聚类具有一定的相似性,两者都属于无监督学习,都可以识别文档集合中蕴含的主题并将包含相同主题的文档进行聚集。不同之处在于多数文本聚类算法通常认为每个文档只属于一个主题(簇),每个主题(簇)是一组语义相似的文档,而主题抽取则认为一个文档可属于多个主题,每个主题是一组词汇的概率分布。文本聚类的语义标注是采用语义标签对同一簇中一组相关文档的含义进行概括和描述,针对的是信息更加丰富和具体的文档;而主题语义标注针对的则是更加抽象和细粒度的词汇概率分布。

针对文本聚类结果的语义标注,早在2000年就有学者采用关键词抽取法从文档簇中抽取出一组关键词,用关键词及其权重构成的集合表示簇的含义,这种“关键词-权重”表示有些类似于主题抽取得到的“主题-词汇”概率分布^[15]。在此基础上,一些学者进一步对文本聚类的词汇集合表示方法进行了完善,探索如何以单个词汇或短语的形式对词汇集合进行概括,代表性方法包括:使用与簇中心最相似的文档的标题表示整个文档簇^[16];结合统计与网络分析法从词汇集合中选择最重要的单个词汇表示该簇^[16-17];通过计算外部知识库中已有的知识单元与文档簇特征向量的相似性选择最相似的知识单元作为簇的标签等^[18-19]。由于文本聚类最初采用的“关键词-权重”表示法与主题模型得到的“主题-词汇”概率分布具有一定的相似性,因此其语义标注方法也为主题模型的语义标注提供了参考和借鉴,但面向主题模型的语义标注在标签的种类、生成方式和外部知识库的应用上有很大拓展、丰富和完善。

3.1 基本步骤

主题自动语义标注最早由Mei等^[4]开始研究,随后得到了众多学者的关注与进一步探索,如Aletas等^[10-11]、Lau等^[7,20]、Hulpus等^[21-22]。其基本步骤为:

- (1) 获取或生成能够表达给定主题语义的候选标签,可以是单个词汇、短语、句子或者图片等;
- (2) 设计一种度量方法,对候选标签集合中的标

签进行排序,据此选择排序最高的标签标注主题。

根据语义标签在生成过程中是否借助于外部语料库,现有方法可分为基于自身语料库和基于外部语料库两种类型。

3.2 基于自身语料库的方法

基于自身语料库生成主题语义标签的前提是原始文本语料内容足够丰富,这样才有可能从中解析出高质量的主题标签。此类方法的共同点是从进行主题抽取的原始语料库中提取出一系列候选标签(短语、句子等),然后对候选标签进行排序,通常转化为标签和“主题-词汇”分布的相似性度量问题。

早在2007年,Mei等首先提出基于自身语料库的主题自动语义标注方法,采用浅层句法分析和n元语法模型两种方法从原始语料中提取出短语作为候选标签,并使用互信息衡量候选标签和主题(即“主题-词汇”的概率分布)间的相似度^[4],为主题自动语义标注奠定了基础。之后,Kou等对该方法进行改进,在选择候选标签时,使用多种词向量方法^[23-24]在原始语料上训练词向量,采用词向量加权求和的方式分别表示候选标签和主题,最后根据向量之间的余弦相似度对候选标签进行排序^[25]。该方法在新闻、Twitter文本和学术论文语料上进行了测试,结果表明,采用词向量改进的方法在不同类型语料上均优于Mei等的方法。但是,该方法理论上需要依赖大规模训练语料生成优质词向量,而进行主题抽取和标注任务的语料往往规模较小,这在一定程度上间接影响了标注效果。为使生成的候选短语与主题具有更强的相关性,部分学者在候选短语生成时对原始语料进行筛选。譬如,Cui等仅从摘要中抽取候选短语标签^[26];Nolasco等则从共享同一个主题的文档集合中抽取该主题的候选短语标签^[27];Atapattu等在在线课程文本进行主题抽取和语义标注时,选择从教师的讲义材料以及学生对课程的实时评论中抽取候选短语标签^[28]。

从自身语料库中抽取短语作为主题标签,虽然简单易行,但是短语表达的语义比较概括,同时也可能存在一词多义等现象,不免给用户的理解造成困扰。因此也有学者选择一个或多个句子作为主题标签,使得对于主题的描述更加具体和明确。Basave等使用与Nolasco等相同的方法为每个主题生成相关文档集合,

然后利用 4 种自动文摘技术从中抽取指定长度的文摘句作为对应主题的标签^[8]。实验结果表明,该方法可抽取可读性良好、语义表达完整的句子作为主题标签^[8]。但其不足之处在于主要使用抽取式文摘技术,只能直接从文档中抽取原始句子标注主题,未能使用生成式文摘技术自动生成概括性更好的摘要句作为主题标签。除此之外, Wan 等同样也从文档集合中选取句子作为主题标签,通过计算“主题-词汇”分布与文档集合中各句子的相对熵确定候选句子,从语义相关性、全面覆盖性、区分度三个角度对候选句进行评分,最后选择综合得分最优的若干个句子作为主题标签^[29]。相比 Basave 等的方法, Wan 等对句子的筛选更加综合全面,但是仍使用抽取式文摘技术生成候选标签,且在文摘的连贯性、可读性和易理解性上还有提升的空间^[29]。

除了对采用 LDA 主题模型抽取出的平行主题进行语义标注,也有学者着眼于如何对具有层次结构的主题抽取结果进行语义自动标注。譬如, Mao 等根据层次主题模型(Hierarchical LDA, hLDA)的抽取结果将文档与主题进行一一映射,按照主题的层次结构将原始文档排列成类似的层次结构,并采用与 Mei 等相同的方法从文档中生成候选标签,使得

生成的候选标签同样具有层次结构。最后,根据主题之间的包含、并列等层次结构关系,采用基于词汇加权和基于 JS 散度两种方法对候选标签进行排序^[30]。

对已有的基于自身语料库的主题自动语义标注方法进行汇总,结果如表 4 所示。总的来说,基于自身语料库的主题自动语义标注思路简单、清晰,生成和抽取的标签在主题语义相关性上更接近原始语料库。但采用此类方法生成语义标签的前提是:原始文本语料内容足够丰富,能从中析出高质量的主题标签。这对待处理语料提出较高要求,不仅要能够获得高质量的“主题-词汇”概率分布,还要能获得高质量的候选语义标签,因而此类方法尤其不太适用于内容短小且表述不规范的用户生成内容。此外,在候选标签生成上,随着语料规模的增加,此类方法生成的候选短语数量会急剧增加,文本解析也需要耗费更长时间^[20]。不仅如此,通过自然语言处理方式从原始语料中抽取的候选短语标签往往良莠不齐,一些标签语义表达不完整或可理解性较差。此外,此类方法难以捕捉多个候选短语标签之间存在的包含、并列等语义关系,对于一词多义、一义多词、词语之间蕴含关系等情况的处理也较为粗糙。

表 4 基于自身语料库的主题自动语义标注方法汇总

方法	标签类别	候选标签生成方式	候选标签排序方式
Mei 等 ^[4]	短语	通过浅层句法解析和 n 元语法模型从原始语料中抽取出名词性短语	计算“主题-词汇”分布与候选短语标签的互信息
Kou 等 ^[25]	短语	通过浅层句法解析和语块技术从原始语料中抽取出名词性短语	采用词向量分别表示“主题-文档”分布和候选短语标签,然后计算两者之间的余弦相似度
Cui 等 ^[26]	短语	通过依存句法分析从原始语料的摘要中抽取出名词性短语	计算“主题-文档”分布和候选“短语标签-文档”分布之间的相对熵(KL 距离)
Nolasco 等 ^[27]	短语	从特定主题所关联的原始文档中抽取短语	使用基于词频及其变体的方法为该主题的所有候选短语标签排序
Basave 等 ^[8]	句子	通过自动摘要方法从特定主题所关联的原始文档中抽取指定长度的文摘句	由自动摘要算法决定
Wan 等 ^[29]	句子	通过计算“主题-词汇”分布与原始语料库中每个句子的相对熵来生成该主题对应的候选句子集合	从语义相关性、全面覆盖性和区分度三个方面对候选句进行综合评分
Mao 等 ^[30]	短语	通过 n 元语法模型从原始语料中抽取出名词性短语	基于主题之间的层次结构,采用词汇加权和 JS 散度两种方法计算“主题-词汇”分布与候选短语标签的相似性

3.3 基于外部语料库的方法

除了利用原始文本语料产生主题标签,也有学者尝试直接将外部语料库(主要是知识库)中的信息作为主题的候选标签,实现主题自动语义标注。在已有研

究中,使用的外部语料库主要有分类目录、搜索引擎、维基百科、关联数据等。

(1) 基于分类目录的主题语义标注

分类目录是一种描述概念间上下位关系的分类体

系结构^[31], 如 Google Directory^①和 Apache Open Office English Thesaurus^②等都属于此类。2009 年, Magatti 等首次尝试将分类目录应用到主题自动语义标注中, 将 Google Directory 树结构中前 5 层节点的概念作为候选标签, 并通过 Jaccard 系数^③、谷本系数^④等 6 种指标衡量主题与候选标签间的相似度^[31-32]。虽然该方法为主题标签的来源提供了新思路, 能够通过分类目录的树形结构获取候选标签间的上下位关系, 但在计算主题与候选标签的相似度时, 使用的度量方法过于肤浅和表面化, 只有两个单词具有相同的拼写方式才可以被认为是“相似”, 完全忽略了单词的语义信息, 其效果并不理想。此外, 随着信息组织技术的发展, 诸如 Google Directory 等分类目录已停止服务, 因此该方法难以拓展和广泛应用。

(2) 基于搜索引擎/维基百科的主题语义标注

维基百科作为 Web 2.0 的产物, 包含现实世界中诸多概念与知识, 可以辅助于主题模型的语义标注。2011 年, Lau 等首次尝试使用维基百科词条生成主题候选标签, 将“主题-词汇”分布中概率最高的前 10 个词汇组成一个查询字符串, 分别使用维基百科和 Google 搜索引擎提供的 API 进行检索, 随后将返回的搜索结果作为初始标签, 并结合语块划分技术和 RACO 度量^⑤等方法对初始标签进行扩展和筛选, 最后采用基于统计的无监督排序和基于回归分析的有监督排序两种方法对候选标签排序^[20]。在博客、新闻和医学论文语料上的测试结果表明, 通过该方法得到的主题标签要优于 Mei 等提出的基于自身语料的主题语义标注方法^[20]。

Lau 等结合搜索引擎与维基百科的主题语义标注方法为该领域开辟了一个新的研究视角, 在其基础上, 不断有学者对此类方法进行进一步探索。2014 年, Aletras 等采用与 Lau 等类似的方式生成候选短语标签, 但在标签排序上使用了图论方法, 将搜索引擎返回的元数据中出现的单词按照其在原始语料中的共现情况构造一张单词之间的无向图, 并通过计算候选标签中

各单词在该无向图中的 PageRank 值对候选标签进行排序, 实验结果表明, 该方法要优于 Lau 等提出的方法^[33]。在 Aletras 等的研究基础上, Mirzagitova 等对无向图中单词之间的权重计算方法进行拓展, 并将该方法应用在俄语文本的主题语义标注中, 取得了良好效果^[34]。2016 年, 鉴于 Lau 等利用的外部搜索 API 已无法使用, Bhatia 等直接使用词向量^[23]和文档向量^[35]方法表示候选标签(即百科词条), 同时采用构成主题的若干个高概率词汇词向量的线性组合表示主题, 最后通过计算主题向量与候选词条向量间的相似性选择最佳标签^[36]。该方法扩展了 Lau 等方法的鲁棒性, 实验结果表明能够较大幅度提升在图书和医学论文数据集的主题标注效果^[36]。

此外, Aletras 等还探索了如何利用搜索引擎和维基百科自动选择一张合适的图片作为主题的语义标签, 他们使用图片检索功能, 根据搜索引擎返回的图片元数据信息构造图片之间的加权无向图, 并根据 PageRank 算法选择最优图片^[10]。同时, Aletras 等又结合主题词的词向量、图片标题的词向量、图片自身的特征信息等, 提出一种运用深度神经网络自动学习和选择最佳图片标注主题的方法^[11]。这些方法为主题模型的可视化研究与应用提供了参考。

与采用检索方式直接从维基百科中选择候选标签不同, Lauscher 等另辟蹊径, 于 2016 年提出一种间接使用维基百科词条进行主题标注的方法^[37]。从每篇文档中抽取命名实体, 并通过实体链接技术将其与维基百科中的对应词条相关联, 这样每篇文档就拥有了若干个由维基百科词条组成的标签。随后使用有监督的标签主题建模法(Labeled LDA)^[38]进行主题抽取和主题标注。该模型是 LDA 模型的一个变体, 接收带有标签的一组文档, 从中抽取出“主题-词汇”分布并将其自动关联到相应的文档标签。Lauscher 等的方法不依赖外部搜索 API, 且主题建模和语义标注能够同时完成, 但存在的问题是并非所有类型的文档都能抽取和链接到特定的维基百科词条实体, 且在处理过程中也忽视

①Google Directory 是由 Google 创建的网络目录, 用于网络信息资源分类, 已于 2011 年 7 月 20 日停止使用。

②<https://www.openoffice.org/>。

③Jaccard 系数是计算两个集合之间相似程度的值(取值在 0-1 之间), 该值越大, 表示两个集合越相似。

④谷本系数又称 Tanimoto 系数或广义 Jaccard 系数, 是对 Jaccard 系数的扩展。

⑤RACO (Related Article Conceptual Overlap): 一种通过维基百科的词条类别和词条之间的链接结构识别词条之间相关性的度量方法。

了百科词条之间的关联性^[37]。

总的来说,利用搜索引擎对主题进行语义标注,不仅可以使使用短语进行主题语义标注,还可以使用搜索到的相关图片进行标注,使得语义标签的种类更为丰富。同时,将维基百科的词条作为候选主题标签,充分利用现成的知识库信息,且维基百科词条具有多语种版本,这在非英文语料的主题抽取与语义标注上具有一定的参考价值和借鉴意义。但是,此类方法的不足之处在于:对于一些“主题-词汇”分布,可能存在搜索结果为空的情况,或者通过搜索引擎、实体识别与链接技术均难以获取与之最相关的维基百科词条,从而难以借助现有的外部语料库直接获取候选标签。在该问题上,已有研究通常只是简单地将“主题-词汇”分布中概率最高的若干个单词也添加到候选标签集合中^[20],还未提出更好的解决方法。另外,由于互联网上网页的出现和消亡比较频繁,使用搜索引擎进行搜索可能会产生搜索结果不稳定的现象。在进行搜索和对大规模维基百科语料进行加工处理时,也需要花费较长时间。

(3) 基于概念知识库的主题语义标注

除分类目录和维基百科这两种外部语料库外,还有学者利用本体、关联数据等外部概念知识库进行主题自动语义标注。2014 年, Hulpus 等利用维基百科的语义网版 DBpedia^①中的概念生成主题候选标签^[21]。使用词义消歧方法^[22],将“主题-词汇”分布中的每个单词映射到 DBpedia 中的一个概念,每个概念及其上下位概念构成一个词义图;将一个“主题-词汇”分布对应的所有概念的词义图连接起来,即得到该主题的主题图,图中的节点就是该主题的候选标签;最后采用网络分析法中的节点中心性度量选择中心度最高的节点作为主题标签。Aker 等将此方法应用在新闻评论的主题语义标注中,也取得了良好效果^[39]。此外, Allahyari 等也利用 DBpedia 对主题进行语义标注,但与 Hulpus 等的应用方式不同,他们对 LDA 主题模型进行修改,在主题和词汇之间加入一个概念层,将 DBpedia 中的概念嵌入模型中,形成具有“主题-概念-词汇”三个层级的主题模型,从而直接实现 DBpedia 中的概念到主

题的映射,利用概念对主题进行语义表示^[5, 40-43]。

除了将主题中的词汇直接映射到本体知识库中的概念外,还有学者直接计算本体中的概念与“主题-词汇”分布的相似性,选择一个最相似的概念作为最优主题标签,例如 Adhitama、Davoudi、Hindle 等人分别使用不同的本体在新闻报道和软件维护需求文档中进行主题抽取与自动语义标注实验,并通过人工评分或与人工标注结果相比较的方式,证明了此类方法能取得良好效果^[44-47]。

针对从原始语料中抽取的候选标签可能会出现多个标签映射到同一概念的情况, Mehdad 等提出一种优化方法。训练分类器识别候选标签之间是否存在语义蕴含关系,将语义相同的标签删除;在此基础上借助 WordNet 中词汇间的上下位关系选择上一层级概念对在语义上存在并列关系的候选标签进行泛化,使得到的主题标签更具有普遍性^[48]。

与基于维基百科的主题语义标注相比,基于概念知识库的主题语义标注能够利用的知识更为丰富。本体、关联数据等概念知识库预先定义了概念与概念、概念与属性等关系,使得在进行主题自动语义标注时,能够更加充分地利用概念之间的关系处理候选标签之间存在的语义蕴含关系,但候选标签生成和排序的计算方法相对来说更加复杂。此外,与基于搜索引擎和维基百科的自动语义标注方法类似,此类方法也可能存在无法将“主题-词汇”分布映射到概念知识库中知识单元的情况。

(4) 基于其他外部语料库的主题语义标注

除了使用上述较为成熟和完整的语料库和知识库进行主题标注外,还有学者借鉴迁移学习的思想,将在其他外部语料库训练得到的主题标签应用到待处理的主题语义标注任务中。例如, Herzog 等通过对“Comparative Agendas Project(CAP)”文本语料^②进行迁移,实现了对英国下议院演讲文本主题抽取结果的自动语义标注^[49]。他们从带有主题标签的 CAP 语料中训练得到 19 个带标签主题,分别计算这 19 个主题与从目标语料(英国下议院演讲文本)中抽取到的待标注主题间的相似度,最后将这 19 个主题中相似度最高的主

①<http://wiki.dbpedia.org/>。

②CAP 是一个集成和编码了世界各国政策信息的项目,其中英国政策被分为 19 个主题类别,并附有相关文字介绍。

题的标签迁移到待标注主题。类似地, Mao 等采用标签迁移方法构建了一个快速主题自动语义标注框架^[50]。采用 Labeled LDA 模型训练带有主题标签的外部文本语料, 形成预训练好的“主题-词汇”分布及其对应的主题标签数据库, 当从新的语料中获得新的“主题-词汇”分布时, 将训练得到的最相似的主题标签迁移到当前待标注的主题上。

此类方法充分利用互联网上带有标签的监督学习数据进行主题标签迁移, 但是, 对于外部语料库的选择有所限制, 即使用的外部语料库需与待处理语料库具有一定相似性, 这样两个语料库的主题才具有共通

点, 从而实现标签迁移。此外, 要找到针对特定主题的有标签的外部语料库也存在一定难度, 这在一定程度上限制了其应用的广度。

对基于外部语料库的主题自动语义标注方法进行汇总, 结果如表 5 所示。与基于自身语料库的方法相比, 使用外部知识库进行标注可以充分利用知识库中的先验知识。由于外部语料库中包含的知识较为丰富, 基于此产生的主题候选标签也更加多样。在候选标签的排序上, 此类方法也更为丰富, 除传统的基于相似度的方法, 还可采用网络分析、机器学习等方式对候选标签进行排序。

表 5 基于外部语料库的主题自动语义标注方法汇总

外部语料库类别	标签类别	候选标签生成方式	候选标签排序方式
分类目录	单个词汇或短语 ^[31-32]	将分类目录中已有的主题概念作为候选标签 ^[31-32]	计算“主题-词汇”分布与候选标签的余弦相似度、Jaccard 系数等 ^[31-32]
搜索引擎、维基百科	单个词汇或短语 ^[20,33-34,36-37] 、图片 ^[10-11]	对搜索引擎的搜索结果进行解析, 生成候选标签 ^[10-11,20,33-34] ; 或将维基百科的词条作为候选标签 ^[20,36-37]	计算“主题-词汇”分布与候选标签的相似度 ^[20,36] ; 采用机器学习方法进行排序 ^[11,20] ; 采用网络分析中节点中心性度量方法进行选择 ^[10,33-34] ; 采用有监督主题模型同时进行主题抽取与标注 ^[37]
概念知识库	单个词汇或短语 ^[5,21,39-48]	将概念知识库中的概念及其属性作为候选标签 ^[5,21,39-48]	根据网络分析中的节点中心性度量方法进行选择 ^[21,39] ; 或对主题模型进行扩展 ^[5,40-43] ; 或计算知识库中的概念与“主题-词汇”分布的相似度 ^[44-48]
其他外部语料库	单个词汇或短语 ^[49-50]	使用具有主题相关性的外部语料库自带的主题标签作为候选标签 ^[49-50]	计算外部语料库主题抽取得到的“主题-词汇”分布与待解决问题的“主题-词汇”分布的相似度, 进行主题标签迁移 ^[49-50]

4 结 语

主题模型在文本挖掘中具有非常广泛的应用, 但其抽取结果是一组词汇的概率分布, 无法直观地揭示主题含义, 用户难以理解, 这是主题模型最大的弊端, 也是主题模型在各领域应用中的难点。目前, 一些学者已经关注到主题模型的这一缺陷, 尝试利用自然语言处理、机器学习、网络分析和知识库技术予以解决。本文对面向主题模型的主题自动语义标注相关研究进行总结和梳理, 为主题模型的使用者在对主题抽取结果进行解释与描述时提供参考。

总的来说, 主题自动语义标注的重点和难点都集中于候选标签的生成与选择。现有研究主要依靠句法分析、统计、自动文摘、实体链接等技术从原始语料库或外部知识库中抽取出候选标签, 然后主要采用相似性度量, 或者利用网络分析法对候选标签间的关联网络进行分析, 选择最佳的候选标签对主题进行语义标注。

通过对已有研究的梳理与分析, 笔者认为, 面向主题模型主题自动语义标注的研究未来需聚焦以下几个方面:

(1) 面向社交媒体的主题语义标注: 现有研究在进行主题自动语义标注时主要以科学文献和新闻文本为语料, 虽然也有少数研究以 Twitter 等社交媒体中的用户生成内容为语料, 但总体上研究还不够深入。与科学文献等用词规范的语料相比, 用户生成内容由于数据稀疏、表述不规范等原因, 主题抽取效果往往不是很理想, 从而进一步影响了主题语义标注的效果。鉴于此, 一些学者提出在进行语义标注之前, 首先对“主题-词汇”概率分布进行修正, 以此提高语义标注效果。譬如, 将“主题-词汇”分布中高频但无意义的非停用词剔除^[51]、根据 WordNet 词汇间的关系提升语义相似词汇在“主题-词汇”分布中的概率权重^[52]等。还有学者提出使用社交媒体中全部类型的数据(包括文本、图片、视频)进行主题建模, 从多角度识别用户的主题倾向^[53]。

(2) 面向中文文本的主题语义标注: 已有研究主要以英文语料为主, 国内仅有少数学者如 Tang、周亦鹏等对中文语料的主题语义标注进行了探索, 但取得的效果均不能令人满意^[54-55]。当然, 这也与中文文本信息处理难度较大以及中文知识库不够成熟密切相关。对于中文语料主题抽取和主题自动语义标注的研究, 需得到更多关注。

(3) 候选标签生成方法的适用条件探索: 在候选标签生成上, 现有方法或采用自然语言处理技术从原始语料中抽取短语和句子作为候选标签, 或直接使用外部语料库和知识库中的信息作为候选标签。前者难以捕捉候选标签之间的语义关系, 且抽取得到的候选标签质量参差不齐; 后者则存在并非所有“主题-词汇”分布均能对应到知识库中知识单元的情况。因此, 有必要探索这两类候选标签生成方法在不同场景下的应用效果。譬如, 对于社交媒体语料, 其口语化文字较多、内容通常关联到最新的生活热点, 难以关联到知识库中的信息, 直接使用原始语料的 Hashtag 标签作为候选标签或许可以取得较好的效果; 而对于科技文献等语言表述规范、领域术语丰富的语料, 采用领域知识库中的概念作为候选标签则可能效果更佳。

(4) 混合排序方法的适用性探索: 虽然许多学者从不同角度对主题自动语义标注中候选标签的生成与排序方法进行探索并取得了一些成果, 但在实际应用中, 使用单一的主题自动语义标注方法只能对部分“主题-词汇”分布赋予质量较好的标签, 还有一些则并不理想^[6]。虽然这与主题抽取得到的“主题-词汇”概率分布的质量有关, 但是单一标注方法往往基于这样一个假设: 候选标签的生成与排序方式(如相似度计算)对任何类型的主题均适用, 这显然与实际并不完全相符。针对不同类型的“主题-词汇”分布, 如表示具体概念与抽象概念的主题、表示顶层概念与细粒度概念的主题, 在语义标注中使用不同的排序策略, 或者将不同的排序方法进行集成, 是否能取得更好的效果, 有待进一步探索。

(5) 深度学习方法的进一步应用: 近年, 深度学习在许多领域取得了突破性进展。在主题自动语义标注中, 有学者利用词向量技术进行文本表示, 从而能更精确地反映候选标签与主题的相似度; 有学者借鉴迁移学习的思想, 将使用外部带有主题标签的语料库

训练得到主题标签迁移到相似的待标注主题, 实现标签共享。但总体上, 深度学习在主题自动语义标注中的应用还不够广泛, 有着很大的拓展空间。譬如, 在对候选标签进行排序时, 通常需要将候选标签和“主题-词汇”分布转化为相同维度的向量表示, 然后计算两者的相似度。传统做法通常是对构成主题的单个词汇的向量进行加权平均得到“主题”的向量, 但近年, 有学者对如何基于词向量对句子或词汇集合进行向量表示进行了大量探索, 提出许多效果更优的表示方法^[56-57], 可尝试将此方法应用到“主题”的向量表示中, 以提升主题语义标注的效果。

随着自然语言处理、深度学习等技术的发展以及知识库的不断丰富, 探索更高效、适用范围更广的主题模型自动语义标注方法势必成为可能, 从而更大程度地促进以 LDA 为代表的主题模型在学术界和工业界的广泛应用。

参考文献:

- [1] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. *Journal of Machine Learning Research*, 2003, 3: 993-1022.
- [2] 徐戈, 王厚峰. 自然语言处理中主题模型的发展[J]. *计算机学报*, 2011, 34(8): 1423-1436. (Xu Ge, Wang Houfeng. The Development of Topic Models in Natural Language Processing[J]. *Chinese Journal of Computers*, 2011, 34(8): 1423-1436.)
- [3] Chang J, Gerrish S, Wang C, et al. Reading Tea Leaves: How Humans Interpret Topic Models[C]//*Proceedings of the 2009 International Conference on Neural Information Processing Systems*. 2009: 288-296.
- [4] Mei Q, Shen X, Zhai C X. Automatic Labeling of Multinomial Topic Models[C]//*Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2007: 490-499.
- [5] Allahyari M, Kochut K. Automatic Topic Labeling Using Ontology-Based Topic Models[C]//*Proceedings of the 14th International Conference on Machine Learning and Applications*, Miami, Florida, USA. IEEE, 2015: 259-264.
- [6] Gourru A, Velcin J, Roche M, et al. United We Stand: Using Multiple Strategies for Topic Labeling[C]//*Proceedings of the 23rd International Conference on Applications of Natural Language to Information Systems*, Paris, France. Springer, 2018: 352-363.
- [7] Lau J H, Newman D, Karimi S, et al. Best Topic Word

- Selection for Topic Labelling[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Beijing, China. 2010: 605-613.
- [8] Basave A E C, He Y, Xu R. Automatic Labelling of Topic Models Learned from Twitter by Summarisation[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014: 618-624.
- [9] Wan X, Wang T. Automatic Labeling of Topic Models Using Text Summaries[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 2297-2305.
- [10] Aletras N, Stevenson M. Representing Topics Using Images[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013: 158-167.
- [11] Aletras N, Mittal A. Labeling Topics with Images Using a Neural Network[C]//Proceedings of the 39th European Conference on Information Retrieval. Springer, 2017: 500-505.
- [12] Aletras N, Baldwin T, Lau J H, et al. Representing Topics Labels for Exploring Digital Libraries[C]// Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries. IEEE, 2014: 239-248.
- [13] Aletras N, Baldwin T, Lau J H, et al. Evaluating Topic Representations for Exploring Document Collections[J]. Journal of the Association for Information Science and Technology, 2017, 68(1): 154-167.
- [14] Sorodoc I, Lau J H, Aletras N, et al. Multimodal Topic Labelling[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017: 701-706.
- [15] Popescul A, Ungar L H. Automatic Labeling of Document Clusters[OL]. [2019-01-10]. <https://www.cis.upenn.edu/~ungar/Datamining/Publications/labels.pdf>.
- [16] Manning C D, Raghavan P, Schütze H. Introduction to Information Retrieval[M]. Cambridge: Cambridge University Press, 2008: 396-398.
- [17] Role F, Nadif M. Beyond Cluster Labeling: Semantic Interpretation of Clusters' Contents Using a Graph Representation[J]. Knowledge-Based Systems, 2014, 56: 141-155.
- [18] Carmel D, Roitman H, Zwerdling N. Enhancing Cluster Labeling Using Wikipedia[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2009: 139-146.
- [19] Tseng Y H. Generic Title Labeling for Clustered Documents[J]. Expert Systems with Applications, 2010, 37(3): 2247-2254.
- [20] Lau J H, Grieser K, Newman D, et al. Automatic Labelling of Topic Models[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011:1536-1545.
- [21] Hulpus I, Hayes C, Karnstedt M, et al. Unsupervised Graph-Based Topic Labelling Using DBpedia[C]//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. ACM, 2013:465-474.
- [22] Hulpus I, Hayes C, Karnstedt M, et al. An Eigenvalue-Based Measure for Word-Sense Disambiguation[C]// Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, Marco Island, Florida, USA. 2012.
- [23] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality[C]//Proceedings of the 2013 International Conference on Neural Information Processing Systems. 2013: 3111-3119.
- [24] Huang P S, He X, Gao J, et al. Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data[C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, 2013: 2333-2338.
- [25] Kou W, Li F, Baldwin T. Automatic Labelling of Topic Models Using Word Vectors and Letter Trigram Vectors[C]//Proceedings of the 11th Asia Information Retrieval Societies Conference. Springer, 2015: 253-264.
- [26] Cui L, Zhang X, Kimpton A, et al. Automatic Labelling of Topics via Analysis of User Summaries[C]// Proceedings of the 27th Australasian Database Conference. Springer, 2016: 295-307.
- [27] Nolasco D, Oliveira J. Detecting Knowledge Innovation Through Automatic Topic Labeling on Scholar Data[C]// Proceedings of the 49th Hawaii International Conference on System Sciences. IEEE, 2016: 358-367.
- [28] Atapattu T, Falkner K. A Framework for Topic Generation and Labeling from MOOC Discussions[C]// Proceedings of the 3rd ACM Conference on Learning @ Scale. ACM, 2016: 201-204.
- [29] Wan X, Wang T. Automatic Labeling of Topic Models Using Text Summaries[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 2297-2305.

- [30] Mao X L, Ming Z Y, Zha Z J, et al. Automatic Labeling Hierarchical Topics[C]//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM, 2012: 2383-2386.
- [31] Magatti D, Calegari S, Ciucci D, et al. Automatic Labeling of Topics[C]//Proceedings of the 9th International Conference on Intelligent Systems Design and Applications. IEEE, 2009: 1227-1232.
- [32] Magatti D, Stella F. Probabilistic Topic Discovery and Automatic Document Tagging[A]// Brena R F, Guzman-Arenas A. Quantitative Semantics and Soft Computing Methods for the Web: Perspectives and Applications[M]. IGI Global, 2012: 25-49.
- [33] Aletras N, Stevenson M. Labelling Topics Using Unsupervised Graph-based Methods[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014: 631-636.
- [34] Mirzagitova A, Mitrofanova O. Automatic Assignment of Labels in Topic Modelling for Russian Corpora[C]// Proceedings of the 7th Tutorial and Research Workshop on Experimental Linguistics. 2016: 115-118.
- [35] Le Q, Mikolov T. Distributed Representations of Sentences and Documents[C]//Proceedings of the 31st International Conference on Machine Learning, Beijing, China. 2014: 1188-1196.
- [36] Bhatia S, Lau J H, Baldwin T. Automatic Labelling of Topics with Neural Embeddings[OL]. arXiv Preprint, arXiv: 1612.05340.
- [37] Lauscher A, Nanni F, Ruiz Fabo P, et al. Entities as Topic Labels: Combining Entity Linking and Labeled LDA to Improve Topic Interpretability and Evaluability[J]. Italian Journal of Computational Linguistics, 2016, 2(2):67-88.
- [38] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009: 248-256.
- [39] Aker A, Kurtic E, Balamurali A R, et al. A Graph-Based Approach to Topic Clustering for Online Comments to News[C]//Proceedings of the 38th European Conference on Information Retrieval. Springer, 2016: 15-29.
- [40] Allahyari M, Pouriye S, Kochut K, et al. A Knowledge-Based Topic Modeling Approach for Automatic Topic Labeling[J]. International Journal of Advanced Computer Science & Applications, 2017, 8(9): 335-349.
- [41] Allahyari M, Kochut K. Using Semantically-Extended LDA Topic Model for Semantic Tagging[J]. International Journal of Semantic Computing, 2016, 10(4): 503-525.
- [42] Allahyari M, Kochut K. Semantic Tagging Using Topic Models Exploiting Wikipedia Category Network[C]// Proceedings of the 10th International Conference on Semantic Computing, Laguna Hills, California, USA. IEEE, 2016: 63-70.
- [43] Allahyari M, Kochut K. OntoLDA: An Ontology-Based Topic Model for Automatic Topic Labeling[OL]. [2018-11-18]. <https://datasciencehub.net/system/files/ds-paper-492.pdf>.
- [44] Adhitama R, Kusumaningrum R, Gernowo R. Topic Labeling Towards News Document Collection Based on Latent Dirichlet Allocation and Ontology[C]//Proceedings of the 1st International Conference on Informatics and Computational Sciences. IEEE, 2017: 247-252.
- [45] Davoudi H, An A. Ontology-Based Topic Labeling and Quality Prediction[C]//Proceedings of the 21st International Symposium on Methodologies for Intelligent Systems. Springer, 2015: 171-179.
- [46] Hindle A, Ernst N A, Godfrey M W, et al. Automated Topic Naming to Support Analysis of Software Maintenance Activities[C]//Proceedings of the 33rd International Conference on Software Engineering. 2011.
- [47] Hindle A, Ernst N A, Godfrey M W, et al. Automated Topic Naming[J]. Empirical Software Engineering, 2013, 18(6): 1125-1155.
- [48] Mehdad Y, Carenini G, Ng R T, et al. Towards Topic Labeling with Phrase Entailment and Aggregation[C]// Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013: 179-189.
- [49] Herzog A, John P, Mikhaylov S J. Transfer Topic Labeling with Domain-Specific Knowledge Base: An Analysis of UK House of Commons Speeches 1935-2014[OL]. arXiv Preprint, arXiv: 1806.00793.
- [50] Mao X L, Hao Y J, Zhou Q, et al. A Novel Fast Framework for Topic Labeling Based on Similarity-Preserved Hashing[C]//Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. 2016: 3339-3348.
- [51] Chi J, Ouyang J, Li C, et al. Topic Representation: Finding More Representative Words in Topic Models[OL]. arXiv Preprint, arXiv:1810.10307.
- [52] Alkhodair S A, Fung B C M, Rahman O, et al. Improving

Interpretations of Topic Modeling in Microblogs[J]. Journal of the Association for Information Science and Technology, 2018, 69(4): 528-540.

[53] Chang S, Dai P, Chen J, et al. Got Many Labels?: Deriving Topic Labels from Multiple Sources for Social Media Posts Using Crowdsourcing and Ensemble Learning[C]// Proceedings of the 24th International Conference on World Wide Web. ACM, 2015: 397-406.

[54] Tang W, Wu X, Li Y, et al. A Topic Label Extraction Method for the University BBS[C]//Proceedings of the 1st International Conference on Data Science in Cyberspace, Changsha, China. IEEE, 2016: 678-682.

[55] 周亦鹏, 杜军平. 基于关联词的主题模型语义标注[J]. 智能系统学报, 2012, 7(4): 327-332. (Zhou Yipeng, Du Junping. Semantic Tagging of a Topic Model Based on Associated Words[J]. CAAI Transactions on Intelligent Systems, 2012, 7(4): 327-332.)

[56] Arora S, Liang Y, Ma T. A Simple but Tough-to-Beat Baseline for Sentence Embeddings[C]// Proceedings of the 5th

International Conference on Learning Representations. 2017.

[57] Yang Z, Zhu C, Chen W. Zero-Training Sentence Embedding via Orthogonal Basis[OL]. arXiv Preprint, arXiv: 1810.00438.

作者贡献声明:

凌洪飞: 确定研究思路, 文献调研, 论文起草和修改;
欧石燕: 提出研究问题, 论文修改、完善和最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: oushiyan@nju.edu.cn。

[1] 凌洪飞, 欧石燕. topic_labeling_papers.rar. 主题自动语义标注相关论文集。

收稿日期: 2018-10-16

收修改稿日期: 2019-02-27

Review of Automatic Labeling for Topic Models

Ling Hongfei Ou Shiyan

(School of Information Management, Nanjing University, Nanjing 210023, China)

Abstract: [Objective] This paper reviews methods of automatic topic labeling, aiming to promote the development of topic modelling. [Coverage] We used “Topic Labeling OR Topic Labeling OR Topic Tagging OR Topic Indexing” as search term for the Web of Science and CNKI databases. A total of 57 representative literatures on topic labeling were retrieved. [Methods] We categorized the existing methods and then conducted a comparative analysis for them. [Results] Automatic topic labeling usually had two steps: generating candidate labels from a corpus and then ranking them. These methods can be divided into two categories: label generation based on internal or external corpus. [Limitations] We might not be able to cover everything in this field. [Conclusions] More research could be done in automatic labeling, i.e. those for user-generated contents from social media using deep learning technologies.

Keywords: Topic Labeling Probabilistic Topic Models Latent Dirichlet Allocation (LDA)