基于深度学习的主题模型研究

黄佳佳[1)] 李鹏伟[1)] 彭 敏[2)] 谢倩倩[2)] 徐 超[1)]

[1)](南京审计大学信息工程学院 南京 211815)

[2)](武汉大学计算机学院 武汉 430072)

摘 要 主题模型作为一个发展二十余年的研究问题,一直是篇章级别文本语义理解的重要工具.主题模型善于从一组文档中抽取出若干组关键词来表达该文档集的核心思想,因而也为文本分类、信息检索、自动摘要、文本生成、情感分析等其他文本分析任务提供重要支撑.虽然基于三层贝叶斯网络的传统概率主题模型在过去十余年已被充分研究,但随着深度学习技术在自然语言处理领域的广泛应用,结合深度学习思想与方法的主题模型焕发出新的生机.研究如何整合深度学习的先进技术,构建更加准确高效的文本生成模型成为基于深度学习主题建模的主要任务.本文首先概述并对比了传统主题模型中四个经典的概率主题模型与两个稀疏约束的主题模型.接着对近几年基于深度学习的主题模型研究进展进行综述,分析其与传统模型的联系、区别与优势,并对其中的主要研究方向和进展进行归纳、分析与比较.此外,本文还介绍了主题模型常用公开数据集及评测指标.最后,总结了主题模型现有技术的特点,并分析与展望了基于深度学习的主题模型的未来发展趋势.

关键词 主题模型;深度学习;潜在主题;词向量;神经网络

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2020.00827

Review of Deep Learning-Based Topic Model

HUANG Jia-Jia[1)] LI Peng-Wei[1)] PENG Min[2)] XIE Qian-Qian[2)] XU Chao[1)]

[1)](*School of Information Engineering, Nanjing Audit University, Nanjing* 211815)

[2)](*School of Computer Science, Wuhan University, Wuhan* 430072)

Abstract As a research hotspot for more than twenty years, topic model plays an important role in semantic analysis of multi-documents. The topic model is adept in extracting groups of keywords from documents to represent their core idea, and thus provides crucial support for document classification, information retrieval, automatic summarization of multi-documents, sentiment analysis and so on. Conventional topic model based on three-layers Bayesian network has been well studied in the past ten years. However, combining with deep learning techniques makes topic model grow new lease of life in recent years due to the wide applications of deep learning in natural language processing, such as word embeddings training, text generation and knowledge graph building. In deep-learning-based topic models, it has become a major task of designing a more accurate and effective model by introducing advanced ideas and techniques from deep learning, such as word embeddings, neural network (e.g., recurrent neural network, RNN), variational auto-encoder (VAE) and knowledge graph. In this review, we first comparatively discuss four probabilistic topic models and two sparse additive topic models from model assumption, document generation process and parameter inference. There are latent Dirichlet allocation (LDA), Dirichlet multinomial mixture model (DMM), biterm topic model (BTM), sparse topical coding (STC) and sparse

收稿日期:2019-02-19;在线出版日期:2019-10-31.本课题得到国家自然科学基金(61802194,61902190)、江苏省高等学校自然科学研究项目(17KJB520015)资助.**黄佳佳**,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为自然语言处理、审计数据分析.E-mail:huangjj@nau.edu.cn.**李鹏伟**,博士,讲师,主要研究方向为软件安全与数据分析.**彭 敏**,博士,教授,主要研究领域为自然语言处理、信息检索.**谢倩倩**,博士研究生,主要研究方向为自然语言处理、信息检索.**徐 超**,博士,教授,主要研究领域为审计数据分析.

additive generative model (SAGM) respectively. The above six models are the typical representations of the conventional topic model and have various improvement versions and applications since they have been proposed. Then, we introduce the latest research progress of deep-learning-based topic models in detail, which can be summed up as three different types of models. The first type of the model is named word-embedding-based probabilistic topic model, which improves one of the conventional topics model (e.g., LDA, DMM or BTM) with auxiliary pre-trained word embeddings while still complying with the basic assumption of the original model. In these models, word embeddings that pre-trained from large volume of corpus like Wikipedia are introduced to evaluate the similarity between word pair. Based on the evaluation, similar words are more likely to be assigned to the same topic during topic sampling process, and thus the topic coherence and text classification accuracy are improved eventually. The second type of the model is named neural-network-based topic model, which employs neural network structure, such as Multilayer Perceptron (MLP) or RNN, to model the document generation process with introducing latent topic structure. In these models, bag-of-words of a text is feeded into neural topic model and transferred into embeddings, then topic distribution and topic-word distribution are inferred out by the neural network. To further improve the performance of the neural topic model, VAE is employed to transfer the text embeddings into latent space before topic inference process, and sparsity constraint of topic-word distribution is enforced into the model to generate more expressive topical words. The third type of the model is named jointly training model of topic and language, which can train a topic model and language model simultaneously. In these models, token sequence of a text is feeded into a neural network to generate text with the guidance of latent topics. Furthermore, we summarize the public datasets (e.g., 20NewsGroups) and evaluation metrics (e.g., Pointwise Mutual Information) used in above topic models. Finally, we end up with discussing some potential trends of topic model's future development.

Keywords topic model; deep learning; latent topic; word embeddings; neural network

1 引 言

主题模型一直是自然语言处理和信息检索等领域的一个基础性研究问题,其研究成果不仅广泛应用于文本聚类/分类[1]、查询检索[2]、话题检测与演化追踪[3]、多文档自动摘要[4]等任务,还在情感分析[5]、产品推荐[6]、本体生成[7]、词向量训练[8]等研究中扮演重要角色.中国知网的统计数据表明,自2014开始,有关"主题模型"的学术文献逐年递增,其引起的下载量和引用率也随之快速增长.与此同时,学者们也从若干角度对主题模型的研究进展进行综述.表1总结了近5年来有关主题模型的国内外中英文综述文献,这些文献主要讨论了主题模型的几个热点研究问题,包括模型参数估计[9],潜在主题可视化[10],主题模型在微博文本[11-12]、时序文本[13]、软件工程[14]中的应用,主题模型综合应用[15]等.值得关注的是,Sharma[16]不仅详细分析了经典主题模型的研究进展,包括基于LDA[17](Latent Dirichlet Allocation)和基于非LDA的模型,还总结了早期神经网络主题模型的基本思想,如NTM[18](Neural Topic Model)、SCNTM[19](Supervised Citation Network Topic Model)等.总的来说,这些综述文献主要关注概率主题模型及其变形形式的相关技术与方法,很少涉及结合深度学习的主题模型研究进展.

主题模型主要利用吉布斯采样、变分推断、非负矩阵分解等机器学习算法从高维稀疏的文本特征空间中推断出潜在主题信息.其中,概率主题模型是面向多文档语义分析的一种重要工具,旨在从大规模文档中抽取出表达这些文档的若干主题并以一组词汇及其概率的形式表达,并可基于主题对文档进行分类或聚类.如此,每个主题可表达成词汇的概率分布形式,而每个主题也有一个生成概率来表达其出

表 1　已有主题模型研究综述

作者	年份	刊物名	关键词
Boyd-Graber等人[15]	2017.11	Foundations and Trends in Information Retrieval	主题模型,应用
Sharma[16]	2017.07	International Journal of Modern Education and Computer Science	主题模型,LSA,LDA
杜慧等人[9]	2017.04	计算机科学	主题模型,LDA,参数估计
王燕鹏[28]	2017.05	科学观察	主题模型,LDA,文献计量
陈静等人[11]	2017.02	信息工程大学学报	概率主题模型,微博文本,社团发现
桂小庆等人[13]	2017.02	计算机科学	主题模型,时态主题模型
Sun 等人[14]	2016.12	IEEE International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing	主题模型,软件工程,数据分析
孙国超等人[10]	2015.12	情报工程	主题模型,LDA,可视化
Alghamdi 等人[12]	2015.03	International Journal of Advanced Computer Science & Applications	主题模型,LDA,文本挖掘

现的可能性.早期的概率主题模型主要以 pLSA[20](probabilistic Latent Semantic Analysis)为代表.随后,Blei 等人[17]在 2003 年提出的 LDA 模型使得人们对主题模型的研究进入热潮,并发展出以变分贝叶斯推断[17](Variation Bayesian Inference,VBI)和吉布斯采样[21](Gibbs Sampling,GS)为两种主要的参数推断方式.基于 LDA 模型的基本框架,大量研究从模型假设、主题数量、参数推断方式、监督式模型等角度提出各类改进方法.如基于文档-主题高斯先验假设的相关主题模型 CTM[22](Correlated Topic Model)、可自动推断主题数量的 GSDMM[23](collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture Model)、基于社团分析的主题模型[24]和考虑标签信息的监督式主题模型[25-27]等.

与此同时,社交网络上产生的海量用户自媒体内容(User-Generated-Content,UGC)也是研究人员关注的对象.这些文本具有规模大、更新速度快、语义信息不丰富、噪声信息高等特点,使得传统的 pLSA 和 LDA 等模型常遭遇模型泛化能力弱、主题词可解释性差和分类准确性低等挑战.为此,研究者们在 LDA 模型基础上不断提出各种改进方法,以构建更适用于短文本的主题模型.其中 DMM[23,29](Dirichlet Multinomial Mixture Model)和 BTM[30](Biterm Topic Model)模型可较为有效地缓解短文本特征空间高维稀疏给主题模型带来的挑战.然而,由于社交网络短文本中词汇共现信息不丰富,DMM、BTM 等仅能依靠语料本身提供的信息进行主题推断,其效果依然不够理想.

自 2006 年以来,深度学习逐渐成为机器学习的一个热点研究方向,并受到工业界和学术界的广泛关注.深度学习在计算机视觉、图像处理和自然语言处理等领域均有突破性进展.在自然语言处理方面,基于深度学习的词向量(word embedding)模型①率先取得成功突破.该模型基于 Harris 提出的分布假说,即上下文相似的词汇,其语义也相似[31].基于神经网络的词向量模型能更为有效地表达词汇的语义信息,因而在度量词汇之间的语义相似性以及发现概念之间的潜在关系方面显著优于传统的独热(one-hot)模型.伴随着词向量的发展,结合深度学习思想的主题模型开始崭露头角,在相关国际顶级会议(如 ACL、ICJAI、AAAI、SIGIR、KDD 等)上引起新的研究热潮.

深度学习技术已在中文分词[32]、实体关系抽取[33]、事件抽取[34]、机器翻译[35-36]、自动问答[37]等自然语言处理的各项任务中发挥显著优势.相对而言,主题建模由于在多文档层面进行全局文本语义分析,需综合考虑各文档之间的语义关系,目前主要解决方案还是概率主题模型.特别地,为应对文本中词汇共现信息不丰富或领域知识匮乏等挑战,目前结合先验知识(如领域知识、词向量等)的概率主题模型取得了较大进展.其中,词向量技术在不显著增加主题模型复杂度的条件下可显著提升主题语义一致性、分类准确率和主题的可解释性.

为能够系统性综述基于深度学习技术的主题模型研究成果,我们整理和分析近 5 年来自然语言处理、人工智能、机器学习和数据挖掘等相关领域的国际顶级会议和国内外知名学术期刊相关研究成果发现,目前基于深度学习的主题模型主要包括如下几类:

(1)词/文档向量辅助增强的概率主题模型.这类模型主要利用预训练的词向量来度量词汇之间的语义相似度,并在传统概率主题模型的文本生成过程中将语义相似的词汇或文档同时增强到同一主题下.这类方法的典型代表有 GLDA[38](Gaussian LDA)、WEI-FTM[39](Word Embedding Informed Focused Topic Model)、GPU-DMM[1](Generalized Pólya

① Efficient estimation of word representations in vector space. https://arxiv.org/pdf/1301.3781.pdf, 2013

Urn Dirichlet Multinomial Mixture)和 KGE-LDA[40](Knowledge Graph Embedding LDA)等.

(2)基于神经网络的主题模型.这类方法主要利用神经网络,如前馈神经网络、变分自编码网络等重构主题模型的文本生成过程,并在建模过程中添加主题-词汇的稀疏约束以生成更具表达能力的主题词.这类方法的典型代表有 NTM[18]、AVITM[41](Auto-encoded Variational Inference for Topic Model)、NSTC[42](Neural Sparse Topical Coding)和 SCHOLAR[43](Sparse Contextual Hidden and Observed Language Autoencoder)等.

(3)主题与语言模型联合训练模型.该类方法从主题模型角度刻画文档-主题分布,并从语言模型角度利用文档中词序列之间的语义依赖生成自然文本.这类方法不仅能够从文档中推断出潜在主题,还能够利用语言模型生成特定主题下的自然语句和词向量表示.因而联合模型既可作为主题模型用于文本分类,还可作为语言模型用于句子生成、词向量训练等任务,其典型代表有 SLRTM(Sentence Level Recurrent Topic Model)①、GMNTM[44](Gaussian Mixture Neural Topic Model)、TWE[45](Topical Word Embeddings)、TopicVec[46]和 STE[8](Skip-gram Topical word Embedding)等.

总的来说,过去十几年里传统主题模型得到较多关注并广泛应用于自然语言处理的各类任务中.与此同时,若干学者也从不同侧面综述了传统主题模型的研究进展,如表 1 所示.这些文献较少涉及深度学习技术在主题模型中的应用进展.然而,随着深度学习逐步深入应用于自然语言处理的各项任务,近 5 年不断有学者提出结合深度学习思想的主题模型.这些模型往往比传统方法表现出更优良的性能并具有其他功能,如生成自然文本和训练出词向量等.因此,本文以主题模型为出发点,围绕深度学习与主题模型的各种创新性结合,对现有相关方法进行总结和归纳,重点对比讨论各方法在模型假设、文本生成过程和主题推断等方面的异同点,并探讨了主题模型的未来研究趋势.本文的研究成果是从深度学习视角对现有主题模型研究进展的进一步补充和完善.

本文第 2 节给出主题模型的形式化定义和经典主题模型的基本框架;第 3 节至第 5 节对结合深度学习主题模型的各类技术方法分类论述;第 6 节对比总结学术界在主题模型研究中所采用的主要公开数据集和评估指标,以便研究者开展实验评估;第 7 节对未来值得关注的研究方向进行初步探讨;第 8 节总结全文.

2 传统主题模型简介

2.1 主题模型问题定义

以 LDA 为代表的概率主题模型一般为生成模型,即每篇文档的每个词都是通过"以一定概率选择某个主题,并从这个主题中以一定概率选择某个词汇"这一过程得到.模型基本结构包含两个分布:即文档为关于主题的多项式分布、主题为关于词汇的多项式分布.

为形式化描述主题模型的生成过程,我们首先引入相关符号,如表 2 所示.设 $D=\{d_1,d_2,\cdots,d_N\}$ 为包含 N 个文档的语料集合,$V=\{w_1,w_2,\cdots,w_M\}$ 为 D 中所有词汇集合.指定主题个数 K 后,主题建模的目标是:

(1)生成 K 个主题 $z_k(k=1,2,\cdots,K)$ 及每个主题的生成概率 θ_k;

(2)每个主题以词汇的概率分布形式表示 $\phi_k\in\mathbb{R}^M$;

(3)生成文档关于主题的条件概率 $p(z|d)$,以便对文档进行分类或聚类.

表 2 主题模型中符号及解释

符号	解释
D	文档集合
K	潜在主题个数
V	词汇表中单词集合
B	文档集中双词(biterm)集合
N	文档集中文档数量
M	词汇集合中词汇数量
N_d^w	词汇 w 在文档 d 中出现次数
z	潜在主题
n_k^w	词汇 w 分配给主题 k 的次数
n_k	文档(或 biterm)分配给主题 k 的次数
$\boldsymbol{\theta}$	主题分布向量(或矩阵)
$\boldsymbol{\phi}$	主题-词汇分布矩阵
$z_{d,w}$	第 d 个文档中第 w 个词汇的主题分配序列
α	文档-主题分布的 Dirichlet 先验参数
β	主题-词汇分布的 Dirichlet 先验参数

主题模型主要采用变分贝叶斯推断[17]和吉布斯采样[21]两种参数推断方法.变分贝叶斯推断是一种近似算法,该方法将模型参数的后验概率表达式用一个简单的变分分布来近似,并采用 EM 算法迭代最大化变分下界来估计参数.而吉布斯采样是一

① Sentence level recurrent topic model: Letting topics speak for themselves. https://arxiv.org/pdf/1604.02038.pdf, 2016

种从马尔科夫链中抽取样本的随机算法. LDA 和 DMM 等概率主题模型常采用坍缩吉布斯采样[47-48](Collapsed Gibbs Sampling,CGS)进行估计参数.

2.2　基于狄利克雷假设的概率主题模型

综合分析近几年主题模型的研究进展发现,LDA 模型常用作各类改进模型的基准方法. 为解决短文本聚类(或分类)问题,Yan 等人[30]提出的 BTM 模型和 Yin 等人[23]提出的 DMM 也受到广泛关注,并也常作为基准模型使用. 由于主题模型一般需指定主题数量 K,当 K 值较大时,主题之间往往存在一定的相关性. 为刻画这种相关性,2006 年 Blei 等人[22]提出相关主题模型 CTM,该模型也常作为其它相关主题模型的基准模型.

本文后面综述的神经网络主题模型常以 LDA、DMM、BTM 和 CTM 模型的基本思想和推断过程为基准. 因此,为更好理解和对比当前结合深度学习思想的主题模型研究方法,本文首先简要对比分析了上述四个基准模型在模型假设、文本生成过程、目标函数及参数推断等方面的异同点. 由于大部分针对基准模型的改进方法均使用坍缩吉布斯采样(CGS)实现参数估计,下文将着重介绍模型在 CGS 方式下的参数推断过程.

2.2.1　LDA

LDA 模型①为一个三层贝叶斯概率模型,其基本假设为:每个文档是关于主题的多项式分布,而每个主题是关于词汇的多项式分布. LDA 模型的总体目标是根据文档集合 D 及先验参数 α 和 β 推断每个文档中每个词汇的主题分配序列 $z_{d,w}$,并根据该序列得到文档-主题分布概率矩阵 $\boldsymbol{\theta}$ 和主题-词汇分布概率矩阵 $\boldsymbol{\phi}$. 该模型示意图如图 1 所示.

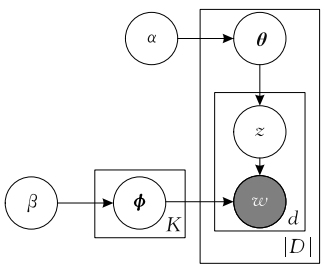


图 1　LDA 模型示意图

在该模型中,文档生成过程如下:

(1)确定需要生成的文档数目 N,以及每个文档中词汇数量 d_n;

(2)对于每个文档 d:

从参数为 α 的 Dirichlet 先验中生成文档-主题分布:$\boldsymbol{\theta}_d \sim Dir(\alpha)$;

(3)对于每个主题 $k(k=1,2,\cdots,K)$:

从参数为 β 的 Dirichlet 先验中生成主题-词汇分布:$\boldsymbol{\phi}_k \sim Dir(\beta)$;

(4)对于文档 d 中第 $i(i=1,2,\cdots,d_n)$个位置:

从 $\boldsymbol{\theta}_d$ 中生成一个主题分配:$z_{d,i} \sim Multi(\boldsymbol{\theta}_d)$;

根据主题分配 $z_{d,i}$ 生成一个词汇:$w_{d,i} \sim Multi(\boldsymbol{\phi}_{z_{d,i}})$.

为估计 LDA 模型中的各参数,可最大化整个模型所有可见变量(即 w)与隐藏变量(即 $\boldsymbol{\theta},\boldsymbol{\phi},z$)的联合分布:

$$L(D)=\prod_{d\in D}\prod_{w\in d}\sum_{z}p(w,z,\boldsymbol{\theta},\boldsymbol{\phi}\mid\alpha,\beta)\propto \prod_{d\in D}\prod_{w\in d}\sum_{z}p(\boldsymbol{\theta},\boldsymbol{\phi},z\mid w,\alpha,\beta) \tag{1}$$

在坍缩吉布斯采样中,求解模型参数的联合概率 $p(\boldsymbol{\theta},\boldsymbol{\phi},z|w,\alpha,\beta)$可转化为求解 $p(z|w,\alpha,\beta)$,并最终转为求解一个主题分配 $z_{d,w}$在剩余主题下的条件概率 $p(z_{d,w}|z_{-(d,w)},w,\alpha,\beta)$. 其中,$z_{d,w}$表示当前词汇所分配的主题,$z_{-(d,w)}$表示除去当前词汇的其他所有词汇的主题分配序列. 更进一步,引入狄利克雷分布和 Δ 函数,上述条件概率可转化为如下求解过程:

$$p(z,w\mid\alpha,\beta)\propto p(z_{d,w}\mid z_{-(d,w)},w,\alpha,\beta) \propto(n_{d,k}^{-(d,w)}+\alpha_k)\frac{(n_{k,w}^{-(d,w)}+\beta_w)}{\sum_{v=1}^{V}(n_{k,v}^{-(d,v)}+\beta_v)} \tag{2}$$

其中,$n_{d,k}^{-(d,w)}$表示排除文档 d 中当前词汇 w 的情况下,文档 d 中的剩余词汇分配给主题 k 的次数;$n_{k,w}^{-(d,w)}$表示排除文档 d 中词汇 w 的情况下,词汇 w 分配给主题 k 的次数.

2.2.2　DMM

狄利克雷多项式混合模型 DMM 来源于 Nigam 等人[29]提出的混合语言模型(Mixture of Unigrams Model). 该模型是一个基于朴素贝叶斯假设的生成模型:即文档中每个词汇独立产生自满足狄利克雷先验的多项式分布 $\boldsymbol{\phi}_k \sim Dir(\beta)$,且每个文档只由一个主题构成,所有主题依旧产生于满足狄利克雷先验的多项式分布(即 $\boldsymbol{\theta} \sim Dir(\alpha)$). 2014 年,Yin 等人[23]提出基于坍缩吉布斯采样的狄利克雷多项式混合分布模型 GSDMM②,并第一次将 DMM 模型用于短文本聚类. 该模型示意图如图 2 所示.

① The implementation of GibbsLDA is available at https://github.com/jasperyang/GibbsLDApy

② The implementation of DMM is available at https://github.com/atefm/pDMM

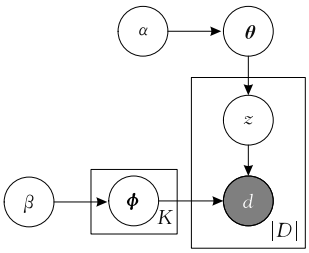


图 2 DMM 模型示意图

在 DMM 模型中,文档生成过程如下:

(1)从参数为 α 的 Dirichlet 先验中生成主题分布:$\boldsymbol{\theta}\sim Dir(\alpha)$;

(2)对于每个主题 $k(k=1,2,\cdots,K)$:

从参数为 β 的 Dirichlet 先验中生成主题-词汇分布:$\boldsymbol{\phi}_k\sim Dir(\beta)$;

(3)对于每个文档 d:

从 $\boldsymbol{\theta}$ 中生成一个的主题分配:$z_d\sim Multi(\boldsymbol{\theta})$;

对于文档 d 中第 $i(i=1,2,\cdots,d_n)$ 个位置:

根据主题分配 z_d 生成一个词汇:$w_{d,i}\sim Multi(\boldsymbol{\phi}_{z_d})$.

GSDMM 采用坍缩吉布斯采样实现模型参数推断,即通过最大化如下联合概率来求解模型参数:

$$\begin{aligned}L(D)&=\prod_{d\in D}\sum_{z}p(d,z,\boldsymbol{\theta},\boldsymbol{\phi}\mid\alpha,\beta)\\&=\prod_{d\in D}\sum_{z}p(\boldsymbol{\theta},\boldsymbol{\phi},z\mid d,\alpha,\beta)\end{aligned}\tag{3}$$

类似于 LDA,在坍缩吉布斯采样过程中,条件概率 $p(\boldsymbol{\theta},\boldsymbol{\phi},z|d,\alpha,\beta)$ 可转化为求解一个主题分配 z_d 在剩余主题下的条件概率:

$$p(z\mid z_{\neg d},d,\alpha,\beta)\propto\frac{n_{k,\neg d}+\alpha}{|D|-1+K\alpha}\times\frac{\prod_{w\in d}\prod_{j=1}^{N_d^w}(n_{k,\neg d}^w+\beta+j-1)}{\prod_{j=1}^{N_d^w}(n_{k,\neg d}^w+V\beta+j-1)}\tag{4}$$

其中,$n_{k,\neg d}$ 为排除当前文档 d 后剩余文档分配给主题 k 的计数,$n_{k,\neg d}^w$ 为排除文档 d 中的词汇 w,剩余词汇 w 分配给主题 k 的计数.

相比于 LDA,GSDMM 能更好地缓解短文本的高维稀疏问题. 相比于 DMM,GSDMM 采用电影分组过程(Movie Group Process,MGP)可自动推断出主题数量.

2.2.3 BTM

为解决短文本的特征稀疏问题,Yan 等人[30]提出 BTM① 模型. 该模型首先从文本集中挖掘所有双词(即 biterm),然后直接在 biterm 集 B 上进行主题推断. 与 LDA 假设不同,BTM 模型假设每个 biterm 中的两个词汇均是采样于一个主题 z,而每个主题是关于词汇的多项式分布. 该模型示意图如图 3 所示.

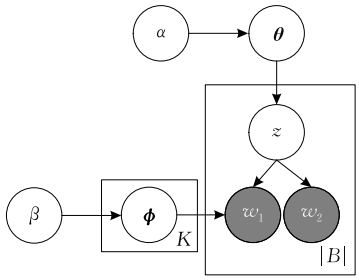


图 3 BTM 模型示意图

设 $b=\{w_1,w_2\}$ 是由词汇 w_1 和 w_2 构成的一个 biterm,所有 biterm 构成集合 $B=\{b\}$. 在 BTM 中,双词生成过程如下:

(1)从参数为 α 的 Dirichlet 先验中生成主题分布:$\boldsymbol{\theta}\sim Dir(\alpha)$;

(2)对于每个主题 $k(k=1,2,\cdots,K)$:

从参数为 β 的 Dirichlet 先验中生成主题-词汇分布:$\boldsymbol{\phi}_k\sim Dir(\beta)$;

(3)对于每个双词 b:

从主题分布 $\boldsymbol{\theta}$ 中生成双词 b 的主题分配:$z_b\sim Multi(\boldsymbol{\theta})$;

对于 b 中第 $i(i=1,2)$ 个位置,根据主题分配 z_b 生成一个词汇:$w_{b,i}\sim Multi(\boldsymbol{\phi}_{z_b})$.

该模型通过最大化如下联合概率来求解参数:

$$\begin{aligned}L(B)&=\prod_{b\in B}\sum_{z}p(b,z,\boldsymbol{\theta},\boldsymbol{\phi}\mid\alpha,\beta)\\&=\prod_{b\in B}\sum_{z}p(z,\boldsymbol{\theta},\boldsymbol{\phi}\mid b,\alpha,\beta)\end{aligned}\tag{5}$$

同样地,利用坍缩吉布斯采样,BTM 将条件概率 $p(\boldsymbol{\theta},\boldsymbol{\phi},z|b,\alpha,\beta)$ 转换为求解一个主题分配 z_b 在剩余主题下的条件概率 $p(z_b\mid\boldsymbol{z}_{-(b)},\boldsymbol{B},\alpha,\beta)$:

$$p(z_b\mid\boldsymbol{z}_{-(b)},\boldsymbol{B},\alpha,\beta)\propto(n_k+\alpha)\frac{(n_{w_i|k}+\beta)(n_{w_j|k}+\beta)}{\left(\sum_{w}n_{w|z}+V\beta\right)^2}\tag{6}$$

其中,n_k 为双词 b 分配给主题 z 的次数,$n_{w_i|k}$、$n_{w_j|k}$ 分别为词汇 w_1、w_2 分配给主题 k 的次数.

该模型不关注文档是否隶属一个主题或多个,而是直接从双词集合入手实施主题推断. 因而相比 LDA 和 DMM,BTM 模型的主题聚合性和文本分类准确性显著提高[1,30].

① The implementation of BTM is available at https://github.com/xiaohuiyan/BTM

2.2.4　CTM

在主题建模任务中，学者们不仅关注模型的主题词抽取能力和文本分类准确性，同时还希望模型能够刻画主题之间的相关性，即认为潜在主题之间并非相互独立. 在这种情景下，LDA 模型中的狄利克雷先验假设不能很好地建模主题之间相关性. 为此，2006 年 Blei[22] 提出相关主题模型 CTM. 该模型假设文档-主题满足超参数为$\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$的逻辑斯谛-正态分布，即先从正态分布中采样 $\boldsymbol{\eta}_d \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$，并使用 softmax 函数将其归一化 $\boldsymbol{\theta}_d = softmax(\boldsymbol{\eta}_d)$. 与 LDA 类似，每个文档中的词汇依旧根据多项式分布采样主题分配序列，即 $z_{d,i} \sim Multi(\boldsymbol{\theta}_d)$. 该模型示意图如图 4 所示. 在 CTM 模型中，文档生成过程如下：

(1) 对于每个主题 $k(k=1,2,\cdots,K)$：

从参数为 β 的 Dirichlet 先验中生成主题-词汇分布：$\boldsymbol{\phi}_k \sim Dir(\beta)$；

(2) 对于每个文档 d：

从参数为$\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$的高斯先验中生成文档-主题分布：$\boldsymbol{\eta}_d \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$；

(3) 对于文档 d 中第 $i(i=1,2,\cdots,d_n)$个位置：

从归一化的 $\boldsymbol{\eta}_d$(即 $\boldsymbol{\theta}_d$)中生成一个主题分配：$z_{d,i} \sim Multi(\boldsymbol{\theta}_d)$；

根据主题分配 $z_{d,i}$ 生成一个词汇：$w_{d,i} \sim Multi(\boldsymbol{\phi}_{z_{d,i}})$.

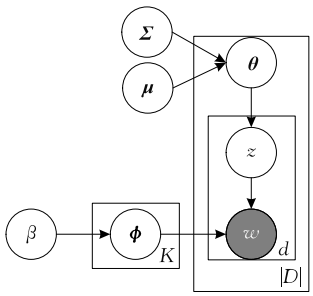


图 4　CTM 模型示意图

从上述过程可以看出，CTM 与 LDA 模型的文本生成过程类似，不同之处在于：在 CTM 中，主题从高斯先验而非狄利克雷先验中产生. 这样，协方差矩阵 $\boldsymbol{\Sigma}$ 即可表达主题之间的相关性.

由于高斯先验与多项式分布不是共轭分布，这给模型的求解增加了难度. Blei 等人[22] 使用变分推断方法求解参数，但这不能保证模型的稳定性. 随后，Chen 等人[49] 采用吉布斯采样求解 CTM 模型参数，使得参数推理难度得以降低. 由于分布的非共轭性，无法根据 θ_d^k 进行采样. 为此，在文献[49]中，模型假设超参数$\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$是满足共轭正态逆维沙特(Normal-Inverse-Wishart，NIW)先验的随机变量，即 $\boldsymbol{\Sigma} \sim IW^{-1}(\kappa,\boldsymbol{W}^{-1})$，$\boldsymbol{\mu} \sim N(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}/\rho)$. 这样，利用吉布斯采样，文档 d 中词汇 w 的主题分配条件概率为

$$p(z,w \mid \alpha,\beta) \propto p(z_{d,w} \mid z_{-(d,w)},w,\alpha,\beta) \propto \frac{e^{\eta_d^k}}{\sum_{j=1}^{K} e^{\eta_d^j}} \frac{(n_{k,w}^{-(d,w)}+\beta_w)}{\sum_{v=1}^{V}(n_{k,v}^{-(d,v)}+\beta_v)} \tag{7}$$

2.2.5　四个基准模型对比分析

最后，我们对比四种方法的复杂度、适用场景及其在线推断方法，如表 3 所示. 其中 DMM 与 LDA 模型的时间复杂度相接近，且均低于 BTM 和 CTM. CTM 时间复杂度最高，这主要是因为 CTM 需要额外的时间来从超分布中估计超参数$\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$. BTM 的复杂度次之，其复杂度主要取决于从文档集中挖掘出的双词规模. 当应用于短文本分类时，由于文档平均长度较短，挖掘出的双词集合规模相对也较小，因而 BTM 的时间复杂度与 LDA 差距较小.

表 3　四个模型对比

模型	复杂度		适用场景	在线推断方法
	时间复杂度	空间复杂度		
LDA	$O(K\|D\|l)$	$\|D\|K+VK+\|D\|l$	常规文本	Online-LDA[50]
DMM	$O(K\|D\|l)$	$\|D\|+VK+\|D\|l$	短文本	—
BTM	$O(K\|B\|)$	$K+VK+\|B\|$	短文本	Online-BTM[51]
CTM	$O(K\|D\|l+K^2+SK)$	$\|D\|K+VK+\|D\|l$	常规文本	—

注：l 为每个文档的平均长度，$|B| \approx |D|l(l-1)/2$[30].

此外，从应用场景方面，LDA 和 CTM 模型更适用于常规文本，如新闻、博客、维基百科等文本；而 DMM 和 BTM 由于其提出动机和基本假设，使其应用于短文本时优势更加显著. 最后，LDA 和 BTM 均有针对大规模文本流的在线推断方式.

2.3　基于稀疏约束的主题模型

在传统概率主题模型中，主题由整个词表的多项式分布表示. 然而在实际问题(如信息检索、关键词抽取等)中，每个主题只与若干个词汇相关，即主题-词汇分布应该是稀疏的. 为此，一部分学者提出基于主题分布稀疏假设的主题编码模型并在优化函数上直接约束主题稀疏度. 在编码模型中，最具代表性方法即为 Zhu 等人[52] 提出的非概率型稀疏主题编码模型 STC①(Sparse Topical Coding). 另一部分研究者通过在概率主题模型中引入辅助变量来控制

① The implementation of STC is available at http://ml.cs.tsinghua.edu.cn/~jun/stc.shtml

稀疏度[53-54].其中,具有代表性的方法是稀疏附加生成模型 SAGE[53]① (Sparse Additive Generative Model).为更好地理解稀疏约束主题模型的思想及其在神经主题模型中的应用,下面将简要介绍 STC 模型和 SAGE 模型.

2.3.1 STC

在 STC 模型中,假设每个文档 $d=\{w_1,\cdots,w_{N_d}\}$ 表示词汇空间 $\mathbb{R}^M$ 上的特征向量,其中 w_i 表示词汇 i 在文档 d 中的出现次数.相比于 LDA 模型,STC 引入如下几个变量:(1) 主题字典(类似 LDA 中的主题-词汇分布)$\boldsymbol{\phi}\in\mathbb{R}^{K\times M}$,其中每个 $\boldsymbol{\phi}_k$ 为词汇空间中的均匀分布;(2) 文档编码(类似 LDA 中的文档-主题分布)$\boldsymbol{\theta}\in\mathbb{R}^{N\times K}$;(3) 词编码(类似 LDA 中文档 d 中词汇 i 的主题分配序列)$\boldsymbol{s}_{di}\in\mathbb{R}^K$.STC 的目的是根据观测到的文档向量 $\boldsymbol{D}=\{\boldsymbol{d}\}_{d=1}^D$ 学习文档编码 $\boldsymbol{\theta}$、全局主题字典 $\boldsymbol{\phi}$ 以及每个文档 d 的词编码 $\boldsymbol{s}_d\in\mathbb{R}^{N_d\times K}$.与 LDA 所采用的狄利克雷-多项式分布不同,STC 模型从超参数为 λ 的拉普拉斯先验中采样文档编码:$\boldsymbol{\theta}_d\sim Laplace(\lambda^{-1})$,并根据文档编码 $\boldsymbol{\theta}_d$ 从超高斯分布中采样该文档的词编码:$\boldsymbol{s}_i\sim superGaussian(\boldsymbol{\theta}_d,\gamma^{-1},\rho^{-1})$.此外,STC 模型还添加如下约束:(1) 文档 d 中词汇 i 的出现次数 w_i 可由相应词编码和主题字典的线性组合 $\boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{\phi}_{\cdot i}$ 重构出;(2) 文档 d 中词汇 i 的词编码 $\boldsymbol{s}_{di}$ 近似于文档编码 $\boldsymbol{\theta}_d$,即文档中每个词汇的主题分布应与该文档的主题分布尽可能一致.稀疏编码模型结构如图 5 所示.

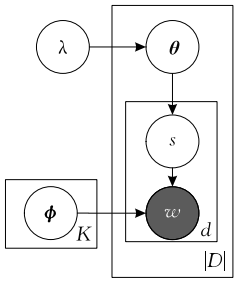


图 5 STC 模型示意图

在 STC 模型中,文档生成过程如下:

(1) 对于每个文档 d:

从 Laplace 先验概率中生成文档编码:$\boldsymbol{\theta}_d\sim\lambda\|\boldsymbol{\theta}\|_1$;

(2) 对于文档 d 中第 $i(i=1,2,\cdots,d_n)$ 个位置:

从 $\boldsymbol{\theta}_d$ 中生成一个词编码向量:$p(\boldsymbol{s}_{d,i}|\boldsymbol{\theta}_d)\sim superGaussian(\boldsymbol{\theta}_d,\gamma^{-1},\rho^{-1})$;

从均值为 $\boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{\phi}_{\cdot i}$ 的指数函数族中采样该词汇的词频:$p(w_{d,i}|s_{d,i}\times\boldsymbol{\phi}_{\cdot i})\sim Possion(s_{d,i}\times\boldsymbol{\phi}_{\cdot i})$.

为使得分布具有稀疏性,模型对优化函数添加 ℓ_1-正则化约束.设 $\Theta=\{\boldsymbol{\theta}_d,\boldsymbol{s}_d\}_{d=1}^D$ 为模型参数集,则模型的整体优化函数即为在稀疏约束条件下最小化文档集的整体重构误差.

$$\min_{\boldsymbol{\Theta},\boldsymbol{\phi}}\sum_{d,i}\ell(\boldsymbol{s}_{di},\boldsymbol{\phi})+\gamma\sum_d\|\boldsymbol{\theta}_d\|_1+\sum_{d,i}(\gamma\|\boldsymbol{s}_{di}-\boldsymbol{\theta}_d\|_2+\rho\|\boldsymbol{s}_{di}\|_1)\tag{8}$$

式(8)中的优化函数包括三项:第一项为文档集合的重构误差,其中误差函数为 $\ell(s_i,\boldsymbol{\phi})=-\log Possion(w_i;\boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{\phi}_{\cdot i})$;第二项为对文档编码的稀疏约束;第三项为对文档中的词编码与文档编码相似性约束,以及词编码的稀疏约束.针对优化函数中的三类非负变量$\{\boldsymbol{\theta},\boldsymbol{s},\boldsymbol{\phi}\}$,模型使用坐标下降法求解,即依次固定其中两个变量来优化剩余一个.

与 LDA 等概率主题模型相比,STC 在构建优化函数时,通过对文档编码和词汇编码直接添加稀疏正则化约束即可达到稀疏分布的效果.此外,STC 模型比 LDA 训练速度更快,文本分类准确性更高.

随后,不断有学者基于 STC 提出改进方法,如用于大规模主题文本建模的在线稀疏主题编码模型 OSTC[55] (Online Sparse Topical Coding),基于主题-词汇、文档-主题双稀疏假设的对偶稀疏主题模型 DsparseTM[56] (Dual-Sparse Topic Model) 和组稀疏主题编码模型 STCSG[57] (Sparse Topical Coding with Sparse Groups).基于双稀疏假设的模型在短文本主题建模上表现出了更好的主题一致性,但是由于引入变量和参数过多,模型计算复杂度较高.

2.3.2 SAGE

在较大规模的文档集中,词汇表中所有词汇的出现频率一般满足对数分布,该分布即为词汇的背景分布.SAGE 模型的一个核心假设是每个潜在主题的主题词可从均值为 0 的高斯先验中生成 $\eta_{k,i}\sim N(0,\tau_{k,i})$,并结合该文档集的词汇背景分布 $\boldsymbol{m}$ 从而生成主题-词汇分布 $\boldsymbol{\phi}$.通过这种方式既可在高斯先验中增加主题稀疏性约束以避免过拟合,即为 $\boldsymbol{\eta}$ 再添加一个指数先验约束 $\tau_{k,i}\sim\varepsilon(\lambda)$;又可在主题-词汇估计时直接将文档中词汇的背景分布概率考虑进来.相比于直接从文档的词袋子中推断主题,从词汇的背景分布中采样主题可有效识别文档中不能表达主题信息的无意义高频词汇,从而提高模型的主题词的语义一致性.

① The implementation of SAGE is available at https://github.com/jacobeisenstein/SAGE

SAGE 模型无需采用狄利克雷-多项式分布来实现文档生成过程，而是首先生成全局的主题-词汇分布，并基于此直接生成当前文档的词汇分布 $w_{d,i}\sim\boldsymbol{\phi}_{z_d}$. 相比于 LDA 模型，SAGE 模型不需要在文档生成过程中重复地为每个文档中的每个词汇计算主题分配概率，而是将全局主题-词汇分布映射为对应的文档词汇. 在 SAGE 模型中，文档生成过程如下：

(1) 为整个词汇表生成一个对数先验分布 $\boldsymbol{m}$；

(2) 对于每个主题 k：

对于每个词汇 i：

从 Laplace 分布中采样参数：$\tau_{k,i}\sim\varepsilon(\lambda)$，采样主题-词汇的先验分布：$\eta_{k,i}\sim N(0,\tau_{k,i})$；

结合词汇背景分布 $\boldsymbol{m}$ 生成当前文档集的主题-词汇分布：$\boldsymbol{\phi}_k\sim softmax(\boldsymbol{\eta}_k+\boldsymbol{m})$

(3) 对于每个文档 d：

从均匀分布中采样一个主题分配 z_d；

对于文档 d 中第 $i(i=1,2,\cdots,d_n)$ 个位置：

采样一个词汇：$w_{d,i}\sim\boldsymbol{\phi}_{z_d}$.

从上述生成过程中可以看出，SAGE 模型的文档生成过程更加简洁，因而运算效率得以提升. 此外，SAGE 在文本分类方面不仅准确率高于 LDA，且当训练样本较少时，其分类准确性优势更加显著. SAGE 构建的主题模型虽然稀疏性更高但在困惑度指标方面往往并不比 LDA 更优，即模型泛化能力并未得到提升.

3　词向量辅助的概率主题模型

词向量辅助增强的概率主题模型旨在利用已训练好的词向量知识提升主题模型性能. 这类模型应用于短文本和领域文本时，往往使得产生的主题词具有更强的语义一致性.

在引入词向量技术之前，针对短文本稀疏问题，相关研究者假设每个文档只与若干个主题相关，并提出基于稀疏假设的主题模型，如焦点主题模型 FTM[58] (Focused Topic Model)、ICD[59] (IBP Compound Dirichlet Process) 等. 这两个模型分别是在 LDA 和 HDP[60] (Hierarchical Dirichlet Processes) 基础上通过印度自助餐过程[61] (Indian Buffet Process，IBP) 来生成文档-主题的稀疏表达，从而使得每个文档只包含若干个主题. 此外，针对文本简短问题，相关研究者还提出伪文档主题模型思想，即将若干短文本合并为一个长文本 (即伪文档)，再在伪文档上实施常规主题推断. 例如，Quan 等人[62] 提出的自聚合主题模型 SATM (Self-Aggregation based Topic Model) 和 Zuo 等人[63] 提出的伪文档主题模型 PTM (Pseudo-document-based Topic Model). 这两个模型均假设每个短文本采样自某个潜在的长文本，隶属于这个伪文档的所有短文本都包含同一个主题.

针对领域文本，相关研究者引入背景知识，如利用频繁共现的词汇 (即 must-link words) 或不能同时出现的词汇 (即 cannot-link words) 来约束文档生成过程. 这类模型假设语义相近的词汇处于同一个 must-link 集合中，而不太可能共现的词汇则处于 cannot-link 集合中. 这些背景知识可通过挖掘领域先验知识的方式获得，如频繁项集、同义词、多义词等. 基于此构建的主题模型有 MDK-LDA[64] (LDA with Multi-Domain Knowledge)、AMC[65] (Automatically generated Must-links and Cannot-links)、LML[66] (Lifelong Machine Learning) 等. 然而，这类模型需要事先在较大规模的领域文本中挖掘领域背景知识，因而对领域知识匮乏的应用来说，这类方法的应用范围受到限制.

自 Bengio 等人[67] 提出基于神经网络的词向量分布式表示以来，词向量技术成为词汇的分布式表示学习中最具代表性的方法[68]. 该方法的主要思想是将词汇表示成低维空间中的稠密实值向量，使得词汇之间的语义关系度量更加准确. 其中 Mikolov 等人[69] 提出的 Skip-gram 和 CBOW 模型是最具代表性的词向量模型. 随后，斯坦福大学[70] 提出一种语义表达能力更强的 GloVe① 词向量模型. 这两类模型均是利用大规模语料 (如维基百科、新闻文本) 训练浅层神经网络从而获得词汇的向量表示. 由于利用了词汇所在的上下文与目标词之间的共现关系建模，模型可以捕获复杂的上下文信息. 因而训练出的词向量不仅能更为准确地表达词汇的语义信息，还能够刻画词汇之间的潜在关系与概念层次，因而广泛应用于命名实体识别、信息检索及推荐系统等领域.

在词向量辅助的概率主题模型中，尽管各类模型均使用预训练的词向量来直接或间接度量词汇之间的语义相似性使得相似词汇增强到同一主题中，但词向量在模型中的使用方式不尽相同. 这主要包括：(1) 在词向量空间中直接推断主题，即基于高斯分布的词向量主题模型；(2) 遵循传统概率主题模

① The implementation of GloVe is available at https://nlp.stanford.edu/projects/glove/

型的文本生成过程,但在主题推断过程中利用词向量度量词汇之间相似度并增强到同一主题中,即基于词向量增强的主题模;(3)在词向量基础上利用领域知识向量辅助主题建模,即基于知识向量的主题模型.

3.1 基于高斯先验的词向量主题模型

LDA 等基于狄利克雷先验的主题模型均从文本的词袋子中采样主题,即从词表规模固定的离散空间中采样.随着词向量在语义相似度度量方面表现出显著优势,学者们开始尝试直接从连续的词向量空间中采样主题,以期提高主题词的语义一致性和解决未登录词(Out-Of-Vocabulary words,OOV words)问题.这时,主题-词汇的狄利克雷-多项式分布假设将不再适用,而多元高斯分布利用欧氏距离能够在连续空间中刻画词汇之间的语义相似度,使得语义相关的词汇能以更大概率聚合到同一主题下.

3.1.1 高斯 LDA 主题模型

最早提出从词向量空间中采样的主题模型是 GLDA[38] 模型.在该模型中,观测变量不再是离散的词汇,而是稠密的词向量,进而每个文档向量表示为相应词向量的拼接.此外,GLDA 假设每个主题-词汇采样自多元高斯分布.在 GLDA 模型,文档生成过程如下:

(1)对于每个主题 $k(k=1,2,\cdots,K)$:

生成其协方差矩阵和均值:$\boldsymbol{\Sigma}_k \sim IW^{-1}(\boldsymbol{\psi}_0,\boldsymbol{v}_0)$,$\boldsymbol{\mu}_k \sim N\left(\boldsymbol{\mu}_0,\frac{1}{\tau}\boldsymbol{\Sigma}_k\right)$;

(2)对于每个文档 d:

从参数为 α 的 Dirichlet 先验中生成文档-主题分布:$\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha})$;

(3)对于文档 d 中第 $i(i=1,2,\cdots,d_n)$ 个位置:

从 $\boldsymbol{\theta}_d$ 中生成一个主题分配:$z_{d,i} \sim Multi(\boldsymbol{\theta}_d)$;

根据主题分配 $z_{d,i}$ 生成一个词向量:$\boldsymbol{v}_{d,i} \sim N(\boldsymbol{\mu}_{z_{d,i}},\boldsymbol{\Sigma}_{z_{d,i}})$.

可以看到,GLDA 模型先从狄利克雷先验中生成文档-主题分布,再从多元高斯先验中生成主题-词汇分布.每个词汇向量 $\boldsymbol{v}_{d,i}$ 在主题 k 上的概率满足均值为 $\boldsymbol{\mu}_k$、方差为 $\boldsymbol{\Sigma}_k$ 的多元高斯分布.

此外,作者也提出坍缩吉布斯采样实现模型参数估计,即文档 d 中词汇 w 的主题分配条件概率为

$$p(z_{d,w}=k \mid z_{-(d,w)},\boldsymbol{V}_d,\varsigma,\boldsymbol{\alpha}) \propto (n_{d,k}+\alpha_k) \times t_{v_k-M+1}\left(\boldsymbol{v}_{d,w} \mid \boldsymbol{\mu}_k,\frac{\tau_k+1}{\tau_k}\boldsymbol{\Sigma}_k\right) \quad (9)$$

其中,$\boldsymbol{V}_d$ 为文档 d 的相应词向量拼接,$t_v(\boldsymbol{x}\mid\boldsymbol{\mu},\boldsymbol{\Sigma})$ 是自由度为 $\boldsymbol{v}$、参数为 $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的 t-分布.

GLDA 模型使用了大规模维基百科语料预先训练出大规模词汇的词向量,这样在推断出文档的主题-词汇向量分布后,可从预训练的词向量库中计算出与各个主题词向量最匹配的词汇,因而能有效应对新文档主题建模过程中出现的未登录词问题,进而提高主题模型的鲁棒性.此外,相比于 LDA 模型,GLDA 模型提高了主题词的语义一致性.

在 Gaussian 分布假设与 LDA 模型基础上,Hu 等人[71]在 2016 年的 ACL 会议上提出潜在概念主题模型 LCTM(Latent Concept Topic Model).LCTM 模型中引入一个新的变量:潜在概念(如概念"*publish*"包含词汇"*publications*","*magazine*","*print*"等),即由语义相似的词向量构成.这样,整个模型的基本假设分别是文档-主题满足狄利克雷-多项式分布、主题-概念满足狄利克雷-多项式分布、概念-词汇满足多元高斯分布.因而,在文档生成过程中,首先分别生成主题分布 $\boldsymbol{\phi}_k \sim Dir(\beta)$ 和概念分布 $\boldsymbol{\mu}_c \sim N(\boldsymbol{\mu},\sigma^2\boldsymbol{I})$;对于每个文档 d 中第 $i(i=1,2,\cdots,d_n)$ 个位置而言,生成主题分配序列 $z_{d,i} \sim Multi(\boldsymbol{\theta}_d)$ 后,依次生成该主题对应的概念 $c_{d,i} \sim Multi(\boldsymbol{\phi}_{z_{d,i}})$ 和该概念对应的词向量 $\boldsymbol{v}_{d,i} \sim N(\boldsymbol{\mu}_{c_{d,i}},\sigma^2\boldsymbol{I})$.相比于 LDA 和 GLDA 等模型,由于 LCTM 模型引入概念变量,使得每个相似词汇可以聚集在相同的概念下面.在采样过程中,主题先生成概念再生成词汇,因而该模型在解决文档中出现未登录词汇方面的分类准确率优于 LDA 和 GLDA.

此外,与 GLDA 模型思想类似还有 Li 等人[72]提出的 MvTM(mix-von Mises-Fisher Topic Model).该模型也是从词向量空间中采样主题,但与 GLDA 采用的多元高斯分布不同,MvTM 认为使用词向量之间的余弦相似度而非欧氏距离来度量词汇的语义相似度更为恰当,因而假设主题-词向量满足混合 von Mises-Fisher(vMF)分布[73].这样在文档生成过程中,每个文档的词向量满足 $\boldsymbol{v}_{d,i} \sim vMF(\Delta_{z_{d,i}})$,其中 $\Delta_k=\{\mu_k,\kappa_k\}$ 为 vMF 分布的参数.相比于 GLDA 模型,MvTM 能获得更高的主题词语义一致性和分类准确性.

3.1.2 高斯稀疏主题模型

在稀疏主题建模方面,Zhao 等人[39]在焦点主题模型 FTM[58] 基础上提出一种结合词向量的焦点主题模型 WEI-FTM 并应用于短文本聚类.在 FTM 模型中,主题的稀疏性体现在文档只与若干个主

题相关；而在 WEI-FTM 中，稀疏性体现在每个主题只与若干个词汇相关. 为实现主题-词汇的稀疏分布，WEI-FTM 在 LDA 模型假设的基础上，引入二值变量 $b_{k,w}\sim Bern(sigmoid(\pi_{k,w}))$ 来表示采样过程中主题 k 是否选择词汇 w，其中，变量 $\pi_{k,v}$ 由词向量 $\boldsymbol{v}_w$ 和主题向量 $\boldsymbol{u}_k$ 之间的相似性及主题偏置 c_k 来决定，即 $\pi_{k,w}=\boldsymbol{v}_w\boldsymbol{u}_k^{\mathrm{T}}+c_k$. 为采样主题向量 $\boldsymbol{u}_k$. WEI-FTM 假设主题-词汇分布满足超参数为 σ_0^2 高斯先验，即主题向量 $\boldsymbol{u}_k$ 与偏置向量 $\boldsymbol{c}$ 均采样自均值为 0、方差为 σ_0^2 的高斯分布：$\boldsymbol{u}_k,:,\boldsymbol{c}\sim N(0,\sigma_0^2\boldsymbol{I})$. 因此，在 WEI-FTM 模型的主题采样过程中，首先依据高斯先验生成主题向量 $\boldsymbol{u}_k$，再根据伯努利分布决定主题 k 是否选择词汇 w. 若主题 k 选择词汇 w，则依据狄利克雷先验生成主题-词汇分布. 由于在采样过程中根据词汇与主题之间的相似性来决定某个词汇是否分配给当前主题，使得模型中的主题只聚焦于若干个词汇，进而提升了模型的泛化能力.

3.1.3　高斯相关主题模型

为进一步提升相关主题模型的主题词抽取能力和分类准确性，He 等人[74] 和 Xun 等人[75] 在 2017 年的 KDD 和 IJCAI 会议上分别提出一种利用分布式表示学习的相关主题模型. 其中，He 等人[74] 引入主题向量 $\boldsymbol{u}_k$ 和文档向量 $\boldsymbol{a}_d$，$\boldsymbol{u}_k$ 采样自超参数为 α 高斯先验 $\boldsymbol{u}_k\sim N(0,\alpha^{-1}\boldsymbol{I})$，而文档采样自超参数为 ρ 的高斯先验分布 $\boldsymbol{a}_d\sim N(0,\rho^{-1}\boldsymbol{I})$. 与 CTM 不同，文档-主题分布将不再从超参数为 $\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ 的高斯先验中采样，而是依据主题向量 $\boldsymbol{U}$ 与文档向量 $\boldsymbol{a}_d$ 之间的相似性采样，即 $\boldsymbol{\eta}_d\sim N(\boldsymbol{U}\boldsymbol{a}_d,\tau^{-1}\boldsymbol{I})$. Xun 等人[75] 提出的相关高斯主题模型 CGTM[75]（Correlated Gaussian Topic Model）综合了 GLDA 和 CTM 模型的基本假设，即文档-主题和主题-词汇均采样自满足共轭 NIW 先验的高斯分布. 在模型推断方面，上述两个相关主题模型分别采用了随机变分和吉布斯采样. 相比于 CTM 和 LDA，两模型在文本分类准确率方面均有一定提升.

3.2　基于词向量增强的主题模型

基于词向量增强的主题模型一般遵从 LDA、DMM 等基准模型假设，而在文本生成过程中将语义相似的词汇以更大概率分配到同一主题中，以产生性能更好的主题模型.

3.2.1　主题向量与词向量相似增强

最先使用该思想的是 2015 年 ACL 会议上提出潜在特征主题模型 LFTM[76]①（Latent Feature Topic Model），该模型分别以 LDA 和 DMM 为基准模型，并在主题-词汇的狄利克雷多项式分布构件基础上增加一个潜在特征（latent feature）构件. 根据使用基准模型的不同形成 LF-LDA 和 LF-DMM 模型.

LF-LDA 模型首先引入词向量 $\boldsymbol{v}_w$ 和主题向量 $\boldsymbol{u}_k$，根据两个向量之间的相似度 $\mu_w=\boldsymbol{u}_k\boldsymbol{v}_w^{\mathrm{T}}$ 作为主题 k 选择词汇 w 的权重，并将潜在构件定义为 $CatE(w|\mu)\sim softmax(\mu_w)$. 在文档生成过程中，首先为文档 d 采样一个满足狄利克雷先验的主题序列 $z_d\sim Multi(\boldsymbol{\theta}_d)$；其次，利用一个满足伯努利分布的指示器 s_d 来决定文档 d 中每个词汇 w 产生自狄利克雷多项式分布还是潜在特征分布. 这样，词汇 w 的生成概率即为两个构件的混合：$w_d\sim(1-s_d)Multi(\boldsymbol{\phi}_{z_d})+s_dCatE(\boldsymbol{u}_{z_d}\boldsymbol{v}^{\mathrm{T}})$. LF-DMM 的采样过程类似，但使用 DMM 作为基准模型.

LFTM 模型将主题向量与词向量之间的相似度引入到文本生成过程中，由于额外引入了主题向量这一变量并使用最大后验估计求解该变量，使得模型复杂度比基准模型（LDA 和 DMM）有较大幅度的提高. 另一方面，引入潜在构件特征使得模型的主题词抽取和分类能力有较大程度的提升，特别是针对短文本而言，LFDMM 比 DMM 模型提升显著.

与 LFTM[76] 建模思想类似的还有 LFBTM[77]（Latent Feature Biterm Topic Model），不同点在于 LFBTM 使用 BTM 而非 LDA 或 DMM 作为基准模型. 该模型利用主题-词汇多项式构件和潜在特征构件的混合构件替代了 BTM 中的主题-词汇多项式分布假设. 在 LFBTM 中，每个双词 $b=\{w_1,w_2\}$ 中的两个词汇均采样自两个构件的混合分布.

3.2.2　词向量相似增强

另一种使用词向量增强思想的主题模型是 Li 等人[1] 在 2016 年 SIGIR 会议上提出的 GPU-DMM②. 该模型基于 DMM 结构，并利用广义波利亚翁（generalized Pólya urn，GPU）增强语义相似的词汇在同一主题上的分配概率. 与之前论述的其他词向量主题模型不同，该模型只使用词向量之间的语义相似度来将相似词汇增强到同一主题中，而不用引入主题向量或文本向量等其他变量.

具体来说，GPU-DMM 模型包含以下三个核心步骤：(1) 利用词向量构建词汇的语义相似度矩阵

① The implementation of LFTM is available at https://github.com/datquocnguyen/LFTM/

② The implementation of GPU-DMM is available at https://github.com/NobodyWHU/GPUDMM

$\boldsymbol{A}$;(2)在坍缩吉布斯采样过程中,对于文档 d 中的每个词汇 w,根据 w 在当前主题 z_d 下的分配概率构建指示变量 $s_{d,w}\sim Bern(\lambda_{w,z_d})$ 以决定该词汇是否采用主题增强策略;(3)若词汇 w 采用增强策略,则利用 w 的相似词汇集 $M_w=\{w_j\mid A_{w_j,w}>\varepsilon\}$ 增强该主题计数 $\tilde{n}_k\leftarrow\tilde{n}_k+N_d^w A_{w,w'}$,及其相似词汇 $w'\in M_w$ 在主题 k 上的计数 $\tilde{n}_k^{w'}\leftarrow\tilde{n}_k^{w'}+N_d^w A_{w,w'}$. 该模型的优化函数及推断过程与 DMM 类似,这里不再赘述.

GPU-DMM 在 DMM 模型的基本假设基础上,巧妙利用了词向量之间的相似度提升模型的主题词抽取和分类能力. 模型简洁高效易于推广,且在个评估指标上优于基准模型 DMM 和 LFDMM.

LDA 模型假设每个文本包含若干个主题,而 DMM 模型假设中每个文本只包含一个主题. 针对短文本分类问题,每个文本包含多少个主题是一个值得商榷的问题. 往往较短的短文本(如 30 字以内)包含一个主题,而较长的短文本(如 100~140 个字)可能包含多个主题. 为此,Li 等人[78] 再次提出 GPU-PDMM(Poisson-based Generalized Pólya Urn Dirichlet Multinomial Mixture)模型. 该模型在 GPU-DMM 基础上引入 Poisson 分布来约束每个文档所包含的主题数目,即希望每个文档只包含 1~3 个主题. 在文档生成过程中,对于每个文档 d,首先从泊松分布中采样主题数目 t_d 及相应的主题分配序列向量 $\boldsymbol{Z}_d$;然后对 d 中的每个词汇 w,从分配的主题序列 $\boldsymbol{Z}_d$ 中使用均匀分布采样一个主题分配序列 $z_{d,w}$;最后使用与 GPU-DMM 模型相同的词向量增强策略. 由于修正了文档的主题数目假设,模型的主题词语义表达一致性和分类准确率有所提升,但是时间消耗增加显著.

3.2.3 文档向量与词向量相似增强

与 GPU-DMM 模型类似的还有彭敏等人[79] 提出基于文档词汇双向增强主题模型 DGPU-LDA (double generalized Pólya Urn with LDA). DGPU-LDA 模型以 LDA 为基准同时引入文档向量以期利用文档与词汇以及词汇与词汇之间的相似度增强主题模型能力. 为此,模型首先利用双向长短时记忆网络(Long Short-Term Memory,LSTM)构建文档的语义向量;其次在主题采样过程中,基于 GPU 使得文档向量和词汇向量同时对主题进行增强,即当主题 z 采样某个文档 d 中的词汇 w 后,若该词汇的词向量与文档语义向量相似则增强主题 z 在文档 d 上的计数,同时增强词汇 w 的相似词汇在主题 z 上的计数.

DGPU-LDA 在 GPU-DMM 基础上,增加了相似文档对主题计数策略,并利用 LSTM 构建文档向量,因而虽然能一定程度提高主题质量,但需要消耗额外的时间来生成文档向量.

3.3 基于知识向量的主题模型

传统的领域文本主题建模常使用频繁共现的词汇或不能同时出现的词汇来约束文档生成过程[64-66],从而提升主题模型性能. 然而,领域知识不仅包括共现词汇,还包括文本所蕴含的事件信息和实体关系等,这些知识及其向量表示也可融入到主题建模过程中.

在利用事件信息建模方面,Sun 等人[80] 提出一种利用事件知识的主题模型 Event-BTM-GPU (Event Biterm Topic Model based on Generalized Pólya Urn),该模型使用结构化的事件信息而非词汇来表示主题. 为获取事件信息,利用斯坦福大学的依存句法分析工具 CoreNLP 从每条句子中抽取满足“名词性主语(nsubj)”与“直接宾语(dobj)”依存关系的事件二元组(Predicate, Subject)和(Predicate,Object)并最终表达成三元组(Subject, Predicate,Object)形式,并利用词汇向量构建事件向量 $\boldsymbol{Sub},\boldsymbol{Verb},\boldsymbol{Obj}=\boldsymbol{Verb}\cdot(\boldsymbol{Sub}\otimes\boldsymbol{Obj})$. 模型的主题采样过程与 GPU-DMM 类似,即对于每一个双事件 $b=\{e_i,e_j\}$,借助 GPU 增加当前事件的语义相关事件在该主题分配上的计数. 此外,由于两个事件可能表达不同的主题,该模型进一步引入伯努利指示变量 $l_b\sim Bern(\lambda_b)$ 以决定 biterm 中的两个事件是否分配同一个主题. 使用事件而非词汇来表示主题,一定程度上增强了主题表达能力和可解释性. 然而,从文档中抽取事件的准确性以及事件知识的稀疏性都一定程度上影响了模型的性能.

另一种利用知识的主题模型是 Yao 等人[40] 在 2017 年 AAAI 会议提出的 KGE-LDA,即将知识图谱中的实体及其向量表示引入到 LDA 模型中. 在 KGE-LDA 中,每个文档 d 不仅由 $N_{d,w}$ 个词汇构成,还由 $N_{d,e}$ 个实体构成,其中每个实体由 TranE[81] 模型训练出对应的向量表示. 那么在文档生成过程中,不仅需要为每个词汇 w 采样一个主题,还需要为每个实体 e 采样一个主题. 由于实体向量是球形空间中的稠密向量,因此,模型假设主题-实体满足 vMF[73] 分布. 与之前针对主题-词汇向量采样所使用的高斯分布假设不同,vMF 分布能有效建模有向数据,进而捕获实体之间的潜在关系. KGE-LDA 是知识图谱与主题模型结合的一次尝试,由于文档生成过程分别从词汇和实体两个角度进行,因而丰富

了文档的语义信息，进而一定程度上提升了模型的主题词抽取和长文本分类能力.

3.4　词向量概率主题模型对比

最后，表 4 对比分析了词向量辅助的概率主题模型在模型假设、分布表示学习的作用以及基准模型上的异同点.

总体来说，上述几种主题模型直接利用已训练好的词向量来丰富文本特征空间表达，以期使得语义相似的文本和词汇能够被分配到同一主题中，从而提高主题词语义一致性和文本分类准确率. 但是各模型基本假设和使用词向量方面具有较大差异. 具体来说：

(1) 为从连续空间中直接采样词向量(或文档向量)，基于高斯分布的主题模型均改进了其基准模型的假设. 这类模型均采用一个或多个高斯先验替换狄利克雷先验，而高斯先验与多项式分布不是共轭分布，因而需要将高斯先验参数作为随机变量，并引入 NIW 先验求解模型参数. 然而，由于引入更多分布，模型的复杂度均比 LDA 和 CTM 更高，但可获得比基准模型更好的主题词抽取和文本分类能力. 此外，除 He 等人[74] 提出的相关主题模型使用随机变分推断外，其余几种模型均采用坍缩吉布斯抽样推断参数. 这样文档中词汇的主题分配条件概率可由剩余主题及词向量来估计.

(2) 基于词向量增强的主题模型一般不更改基准模型的假设，而是直接利用词向量、文档向量及其之间的相似度将相似的文档/词汇聚合到相同主题中. 这类模型基本沿用基准模型的吉布斯采样方法，但比其基准模型在主题词聚合性和分类准确率等指标上更胜一筹，特别是 GPU-DMM.

(3) 目前有关知识向量与主题模型结合的研究成果相对较少，尚处于探索阶段. 此外，从实验结果来看，这类模型效果一般，这可能是由于知识抽取的准确性不高以及待建模文本中所含知识稀疏性较高导致的. 如何抽取可信度更高的知识以及探索其他结合方式也是未来可研究的问题之一.

表 4　词向量辅助的概率主题模型对比

模型	分布表示学习的作用	文档-主题分布	主题-词汇分布	基准模型
GLDA[38]	词向量拼接构成文档向量，从词向量空间中采样	$Dir(\boldsymbol{\alpha})\rightarrow Multi(\boldsymbol{\theta}_d)$	$(IW^{-1}(\boldsymbol{\psi}_0,\nu_0),N\left(\boldsymbol{\mu},\frac{1}{\tau}\boldsymbol{\Sigma}_k\right))\rightarrow N(\boldsymbol{\mu}_z,\boldsymbol{\Sigma}_z)$	LDA
LCTM[71]	相似词向量构成概念空间，词向量拼接成文档向量，从词向量空间中采样	$Dir(\boldsymbol{\alpha})\rightarrow Multi(\boldsymbol{\theta}_d)$	$Dir(\beta)\rightarrow Multi(\boldsymbol{\phi}_z)\rightarrow N(\boldsymbol{\mu}_c)\leftarrow N(\boldsymbol{\mu},\sigma^2\boldsymbol{I})$	LDA
MvTM[72]	词向量拼接构成文档向量，从词向量空间中采样	$Dir(\boldsymbol{\alpha})\rightarrow Multi(\boldsymbol{\theta}_d)$	$(\mu_k,\kappa_k)\rightarrow vMF(\Delta_z)$	LDA、GLDA
WEI-FTM[39]	主题向量与词向量之间相似度决定主题-词汇分布	$Dir(\boldsymbol{\alpha})\rightarrow Multi(\boldsymbol{\theta}_d)$	$N(0,(\sigma_0)^2\boldsymbol{I})\rightarrow Dir(\beta\boldsymbol{b}_k)\rightarrow Multi(\boldsymbol{\phi}_z)$	LDA
He 等人[74]	主题向量与文档向量之间相似决定文档-主题分布	$\left.\begin{matrix}N(0,\alpha^{-1}\boldsymbol{I})\\N(0,\rho^{-1}\boldsymbol{I})\end{matrix}\right\}\rightarrow N(\boldsymbol{U}\boldsymbol{a}_d)\rightarrow Multi(\boldsymbol{\theta}_d)$	$Dir(\beta)\rightarrow Multi(\boldsymbol{\phi}_{z_d})$	CTM
CGTM[75]	从词向量空间中采样	$\left(IW^{-1}(\boldsymbol{\psi},\nu),N\left(\boldsymbol{\mu},\frac{1}{\kappa}\boldsymbol{\Sigma}_k\right)\right)\rightarrow N(\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)\rightarrow Multi(\boldsymbol{\theta}_d)$	$\left(IW^{-1}\left(\boldsymbol{\psi}_0,\nu_0\right),N(\boldsymbol{\mu}_0,\frac{1}{\tau}\boldsymbol{\Sigma}_k\right)\right)\rightarrow N(\boldsymbol{\mu}_z,\boldsymbol{\Sigma}_z)$	CTM
LFTM[76]	主题向量与词向量之间相似度决定主题-词汇分布	$Dir(\boldsymbol{\alpha})\rightarrow Multi(\boldsymbol{\theta}_d)$	$Dir(\beta)\rightarrow(1-s_d)Multi(\boldsymbol{\phi}_z)+s_dCatE(\boldsymbol{u}_{z_d}\boldsymbol{v}^{\mathrm{T}})$	LDA、DMM
LFBTM[77]		$Dir(\alpha)\rightarrow Multi(\boldsymbol{\theta})$	$Dir(\beta)\rightarrow(1-s_b)Multi(\boldsymbol{\phi}_{z_b})+s_bCatE(\boldsymbol{u}_{z_b}\boldsymbol{v}^{\mathrm{T}})$	BTM
GPU-DMM[1]	在主题采样过程中，将待采样词汇的相似词汇增强到同一主题中	$Dir(\alpha)\rightarrow Multi(\boldsymbol{\theta})$	$Dir(\beta)\rightarrow Multi(\boldsymbol{\phi}_{z_d})$	DMM
GPU-PDMM[78]		$\begin{matrix}Dir(\alpha)\\Possion(\lambda)\end{matrix}\rightarrow Multi(\boldsymbol{\theta})$	$Dir(\beta)\rightarrow Multi(\boldsymbol{\phi}_{z_d})$	BTM
DGPU-LDA[79]	构建文档向量，并将待采样词汇的相似词汇和相思文档增强到同一主题中	$Dir(\boldsymbol{\alpha})\rightarrow Multi(\boldsymbol{\theta}_d)$	$Dir(\beta)\rightarrow Multi(\boldsymbol{\phi}_{z_d})$	LDA
Event-BTM-GPU[80]	利用词汇向量构建事件向量，并将相似事件增强到同一主题中	$Dir(\alpha)\rightarrow Multi(\boldsymbol{\theta})$	$Dir(\beta)\rightarrow Multi(\boldsymbol{\phi}_{z_d})$	BTM
KGE-LDA[40]	利用知识图谱中实体向量，并同时为文档的词汇和实体生成主题-词汇/实体分布	$Dir(\alpha)\rightarrow Multi(\boldsymbol{\theta})$	主题-词汇：$Dir(\beta)\rightarrow Multi(\boldsymbol{\phi}_{z_d})$，主题-实体：$\left.\begin{matrix}vMF(\mu_0,C_0)\\\log N(\mu,\sigma^2)\end{matrix}\right\}\rightarrow vMF(\mu_{z_e},\kappa_{z_e})$	LDA

4 基于神经网络结构的主题模型

神经网络主题模型旨在利用神经网络刻画包含潜在主题信息的文本生成过程.这类模型中一般以文档词袋子为输入,并增添对应的词向量层和其他网络层以产生文档.模型一般使用后向传播算法逐层更新参数.

早期的神经网络主题模型主要是直接利用前馈神经网络构建主题模型并以其中的权重矩阵表示文档-主题分布和主题-词汇分布.随后,包含潜在结构层的变分自编码器[82](Variational Auto-Encoder,VAE)被用于构建主题模型.但是上述神经主题模型并未考虑分布的稀疏性.基于此,稀疏约束逐步被引入到神经主题模型中.

4.1 基于前馈神经网络的主题模型

早期神经主题模型主要采用受限玻尔兹曼机[83]或深度信念网络[84-85]构建模型的输入特征表示.例如,Wan 等人[85]提出一个深度信念网络与层次主题模型相结合的混合模型.在该模型中,神经网络用于特征抽取与非线性变换,为主题模型提供文本的低维向量表示.然而受限玻尔兹曼机模型复杂度较高,训练难度较大,不能适应文本序列建模[86].随后,Cao 等人[18]在 2015 年 AAAI 会议上提出了基于前馈神经网络的主题模型 NTM①(Neural Topic Model),开始尝试从神经网络视角构建主题模型.

在 LDA、DMM 等经典的三层贝叶斯主题模型中,文档表示为主题的多项式分布 $\boldsymbol{\theta}$,主题表示为词汇的多项式分布 $\boldsymbol{\phi}$,那么文档 d 关于词汇 w 的分布概率即表示为 $p(w|d)=\boldsymbol{\phi}_{\cdot,w}\times\boldsymbol{\theta}_d^{\mathrm{T}}$.NTM 从前馈神经网络角度解释上述两个分布,并构建出如图 6 所示的网络结构.

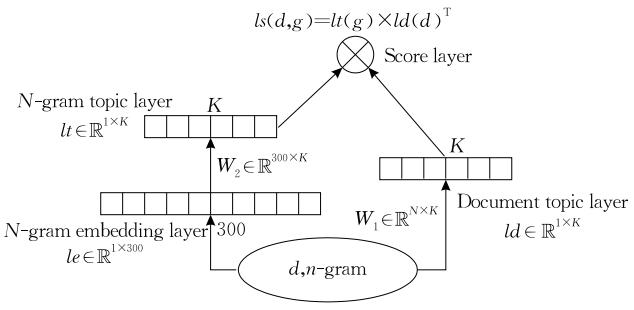


图 6 NTM 的网络结构示意图

在 NTM 模型中,每个词汇和每个文档在主题上的分布分别用两个隐藏层 n-gram-topic layer(lt)和 topic-document layer(ld)表示.那么输出的文档-词汇分布 $p(w/d)$即为两个隐藏层向量的点乘积.此外,与一般神经网络处理文本数据类似,NTM 模型的输入层即为每个文档的 n-gram 并添加词向量层.与传统概率主题模型不同的是,NTM 无需指定先验分布,而是分别使用神经网络常用的 sigmoid 和 softmax 函数从权重矩阵中生成隐藏层,即:

$$lt(g)=sigmoid(le(g)\times W_2) \tag{10}$$

$$ld(d)=softmax(W_1(d)) \tag{11}$$

NTM 模型使用后向传播(Back-Propagation,BP)算法即可训练出两个分布及对应的权重矩阵 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$.相比于 LDA 等概率主题模型,NTM 模型结构简洁且无需先验假设,但可获得质量更高的主题表示和分类准确率.

4.2 基于变分自编码器的主题模型

变分自编码器(VAE)是 Kingma 等人[82]在 2014 年提出的一种编码-解码网络,其模型结构如如图 7 所示.在该网络中,编码器将输入数据 d 压缩为潜在特征 z,而解码器根据数据在潜在空间中的分布重构出信号 $\hat{d}$.

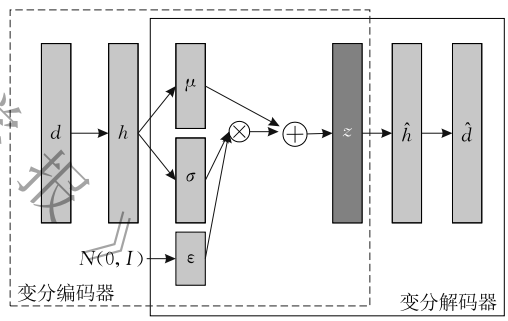


图 7 VAE 模型结构图

通常情况下,VAE 模型假设输入数据 d 在潜在特征 z 下的后验概率近似满足高斯分布,即 $q(z|d)\sim N(z;\boldsymbol{\mu},\boldsymbol{\sigma}^2\boldsymbol{I})$,其中 $\boldsymbol{\mu}$ 和 $\boldsymbol{\sigma}^2$ 均是数据 d 通过神经网络生成的,即 $\boldsymbol{\mu}_d=f_1(d)$,$\log\boldsymbol{\sigma}_d^2=f_2(d)$;另一方面,VAE 假设潜在特征 z 满足多元标准高斯先验,即 $p(z)\sim N(0,\boldsymbol{I})$.那么在解码阶段,通过采样一个 $z\sim q(z|d)$,即可从解码网络重构出 $\hat{d}=g(p(z_d),p(d|z_d))$.为使重构数据尽可能接近原始数据,VAE 最终的优化目标即为在最大化 d 的生成概率 $p(d)$的同时利用 KL 散度使得从数据中训练出的后验概率 $q(z|d)$ 尽可能逼近其理论变分概率 $p(z|d)$,这样优化函数的最终表达式如下:

① The implementation of NTM is available at https://github.com/elbamos/NeuralTopicModels

$$l=E_{z\sim Q}[\log(p(d|z)]-D_{KL}(q(z|d)\parallel p(z)) \quad (12)$$

此外，在解码阶段，模型使用了重参数化技巧(reparameterization trick)，即不直接从后验分布 $N(\boldsymbol{\mu},\boldsymbol{\sigma}^2)$ 采样 z_d，而是从标准高斯分布 $N(0,\boldsymbol{I})$ 中采样 $\boldsymbol{\varepsilon}$，并令 $\boldsymbol{z}_d=\boldsymbol{\mu}+\boldsymbol{\varepsilon}\boldsymbol{\sigma}^2$，从而使得模型可使用后向传播算法训练参数.

从上述生成过程中可以看出，VAE 是一种无监督模型，即不需要对数据 d 进行标注而仅通过对 d 进行重构即可构造出优化函数，进而训练出模型. VAE 模型作为一个无监督生成模型，目前主要用于图像分类[82]、文档生成[87-88]、自动摘要生成[89]和主题建模[90]等数据生成任务.

在主题建模方面，Miao 等人[87]首先尝试使用 VAE 构建神经变分文档模型(Neural Variational Document Model，NVDM)，并在此基础上考虑主题-词汇分布，进而形成基于神经自编码器结构的主题模型[90-92].

NVDM 是一个无监督的文档生成模型，旨在从文档的词向量空间中抽取潜在特征(即主题)，并据此生成文档. NVDM 遵从 VAE 模型的基本结构，即假设主题 z 满足标准高斯先验，这样每个文档 d 即可由多层感知机编码为主题上的分布 $q(z|d)=relu(\boldsymbol{\mu}_d+\boldsymbol{\varepsilon}\boldsymbol{\sigma}_d)$，其中 $\boldsymbol{\mu}_d$ 与 $\boldsymbol{\sigma}_d^2$ 由多层感知机网络产生；那么在解码阶段，通过神经网络采样 $z\sim q(z|d)$ 即可生成文本. 由于 NVDM 仅使用神经网络权重矩阵 $\boldsymbol{W}$ 来表示主题-词汇分布 $\boldsymbol{\phi}$，因而产生的模型在主题词语义一致性方面不及 LDA[92].

基于此，Ding 等人[91]提出一种关注主题语义表达一致性的神经主题模型，该模型使用预训练的词向量来度量词对之间的语义相似度，并将其作为 NVDM 优化函数的一部分. 相比于 NVDM，上述方法提高了主题词语义一致性.

在 NVDM 基础上，Miao 等人[90]在 2017 年 ICML 会议上提出了一系列具有不同形式的神经主题模型，各模型均是基于 VAE 结构，并着重考虑了潜在主题的表示方式. 在这一系列模型中，对于输入文档 d，首先通过多层感知机将其编码为在潜在特征下的概率分布 $\boldsymbol{x}\sim G(N(\boldsymbol{\mu}_d,\boldsymbol{\sigma}_d^2))$. 与 NVDM 不同的是，这里不直接使用 $\boldsymbol{x}$ 作为解码器的输入，而是在此基础上使用不同神经网络进一步生成文档-主题分布 $\boldsymbol{\theta}=q(z|d)$，分别是：(1) GSM(Gaussian Softmax Construction)：直接利用一个 softmax 函数构建 $\boldsymbol{\theta}=softmax(\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{x})$；(2) GSB(Gaussian Stick Breaking Construction)：利用 sigmoid 函数和折棍子模型构建一个稀疏分布 $\boldsymbol{\theta}=f_{\mathrm{SB}}(sigmoid(\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{x}))$；(3) RSB(Recurrent Stick Breaking Construction)：利用循环神经网络(Recurrent Neural Network，RNN)和折棍子模型构建 $\boldsymbol{\theta}=f_{\mathrm{SB}}(f_{\mathrm{RNN}}(\boldsymbol{W}_3^{\mathrm{T}}\boldsymbol{x}))$. 在主题-词汇推断方面，模型引入主题向量 $\boldsymbol{z}\in\mathbb{R}^{K\times H}$ 和词向量 $\boldsymbol{w}\in\mathbb{R}^{M\times H}$，那么每个主题在词汇上的分布即可使用两个向量的语义相似度来表示，即 $\boldsymbol{\phi}_k=softmax(\boldsymbol{w}\cdot\boldsymbol{z}_k^{\mathrm{T}})$. 进而，根据主题 $\boldsymbol{\theta}_d$ 即可表达文档 d 中词汇 w 的生成概率 $p(w|d)=\boldsymbol{\theta}_d\boldsymbol{\phi}_{\cdot w}^{\mathrm{T}}$.

可以看出，上述三个模型结构与 NVDM 类似，即基于主题的高斯先验，并由神经网络刻画出文档-主题分布 $p(\theta|d)$；不同点是主题-词汇不再是简单的神经网络权重矩阵，而是由主题向量与词向量的乘积产生，进而推断出文档-词汇分布. 此外，高斯神经网络主题模型 GSB 利用折棍子模型可刻画出文档-主题的稀疏分布，从而提升了神经主题模型的性能.

变分推断和坍缩吉布斯采样均无法灵活应用于新主题模型，即当模型假设产生微小变动时均需重新训练模型. 为此，Srivastava 等人[41]① 在 2017 年 ICLR 会议上提出另一种使用 VAE 推断思想的主题模型推断方法 AVITM. 该方法假设主题模型的文档-主题满足逻辑斯谛-正态先验而非多元高斯先验或狄利克雷先验. 这样模型既可使用 VAE 的重参数化技巧提升训练速度，也可避免自动变分推断[92]存在的主题同质性问题. 值得注意地是，该假设与相关主题模型 CTM[22]的逻辑斯谛-正态先验略有不同，AVITM 先从拉普拉斯先验中采样参数 $(\mu,\Sigma)\sim Laplace(\alpha)$，再从逻辑斯谛-正态分布中采样文档-主题分布 $\boldsymbol{\theta}\sim softmax(N(u,\Sigma))$. 此外，模型在处理主题-词汇分布时与 NVDM[87]类似，即在 VAE 生成网络的权重矩阵 $\boldsymbol{W}$ 上增加一个 softmax 层作为主题-词汇分布 $\boldsymbol{\phi}$. 由于 AVITM 是一个可应用于其他主题模型的黑盒推断方法，作者将其应用于一个 LDA 变形形式的主题模型 ProdLDA 上以应验证其在主题模型假设发生微小改变时的推断效果和效率. AVITM 由于使用了 VAE 结构的参数重采样技术和逻辑斯谛-正态先验假设，使其比变分推断方法下的 LDA 和 NVDM 均具有更高的主题词一致性水平和训练速度；而 ProdLDA 也比 LDA 具有更高的主题词一致性水平.

① The implementation of AVITM is available at https://github.com/akashgit/autoencoding_vi_for_topic_models

4.3 基于稀疏约束的神经网络主题模型

基于神经网络的稀疏主题模型也是近年来较为关注的研究方向.随着深度学习技术的不断发展,学者们在构建神经网络主题模型的同时向网络中添加主题-词汇的稀疏约束以提升模型的主题抽取能力.其中有代表性的方法是 Peng 等人[42]提出的神经稀疏主题编码模型 NSTC、Lin 等人[93]提出的神经稀疏最大化主题模型(Neural SparseMax Topic Models,NSMTM)和 Card 等人[43]提出的神经网络稀疏主题模型框架 SCHOLAR①,这些模型均基于主题-词汇稀疏分布假设.

NSTC 模型是在 STC[52]基础上使用词向量提升主题聚合性,并利用基于后向传播算法的神经网络简化参数推断过程.与 STC 相比,NSTC:(1)采用神经网络结构代替词编码的超高斯先验和主题字典均匀分布假设;(2)文档编码不再从先验分布中采样,而是依据文档中所有词编码 $s_{d,w}$ 在所有主题字典上的分布 $\boldsymbol{\phi}_{kw}$ 来采样:即 $\boldsymbol{\theta}_d \propto \sum_{d=1}^{D}\sum_{n=1}^{N_d} \boldsymbol{s}_{d,w}\boldsymbol{\phi}_{k,w}$;(3)引入词向量,即将文档 d 中每个词汇表达成词向量形式.而与 STC 假设相同的是文档 d 中每个词汇 $w_{d,n}$ 依旧利用泊松分布采样词编码.从神经网络视角分析,NSTC 模型结构与 NTM 类似,即包含文档词列表输入层、词向量层(le)、词汇编码层 s_d(ld)、主题字典层 $\boldsymbol{\phi}$(lt)以及文档-词汇生成输出层(ls).

NSTC 模型使用词向量序列作为输入并从神经网络视角重构了传统稀疏编码模型 STC,不仅简化了模型复杂度,还显著提高了主题模型的泛化能力和文本分类准确性.

基于变分自编码器的主题模型(如 GSM[90]、AVITM[41]等)一般在推断出的隐藏层上添加 softmax 函数转化为文档-主题分布 $\boldsymbol{\theta}_d$ 和主题-词汇分布 $\boldsymbol{\phi}_k$.由于 softmax 函数产生的结果仍然是稠密向量,使得模型不满足稀疏假设.为此,Lin 等人[93]在 GSM 模型框架基础上提出一个主题稀疏约束的神经主题模型 NSMTM.与 GSM 类似,首先利用多层感知机从文档 d 中推断出隐藏层 $\boldsymbol{x}\sim G(N(\boldsymbol{\mu}_d,\boldsymbol{\sigma}_d^2))$,但这里不采用 softmax 转化为主题 $\boldsymbol{\theta}_d$,而是使用 sparsemax[94]产生具有稀疏表示的文档-主题分布 $\boldsymbol{\theta}_d\sim sparsemax(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x})$.同样地,使用 sparsemax 函数采样主题-词汇的稀疏分布 $\boldsymbol{\phi}_k=\mathrm{sparsemax}(\boldsymbol{w}\cdot\boldsymbol{z}_k^{\mathrm{T}})$.此外,与 GSM 不同的是,在参数推断方面,NSMTM 未使用 VAE 模型中的 KL 散度而是采用 Wasserstein 散度②来度量变分分布之间的距离,以避免可能存在的训练过程不稳定问题.相比于 NVDM 和 AVITM,NSMTM 应用于短文本时具有更好的模型泛化能力和更高的主题词语义一致性.

本文所讨论的绝大多数主题模型均以文档的词汇为输入以推断主题.但实际上文档可能还包含可以辅助主题推断的其他元数据(meta-data),如作者、日期、文档来源等信息.基于此,Card 等人[43]提出一种以元数据为标签辅助建模的神经网络主题模型框架 SCHOLAR.该框架结合了监督式 LDA[25](Supervised LDA,SLDA)与 SAGE 模型的优点:(1)类似于 SLDA,可以使用各种元数据作为标签信息解决多标签分类问题;(2)利用 SAGE 模型的指数先验控制主题分布的稀疏性.

具体来说,其文本生成过程如下:(1)类似于 AVITM[41],SCHOLAR 从超参数为 α 的逻辑斯谛-正态先验中采样文档-主题分布 $\boldsymbol{\theta}_d$;(2)结合文档的元数据 $\boldsymbol{c}_d$ 并通过网络层转化为 $f_g(\boldsymbol{\theta}_d,\boldsymbol{c}_d)$,并据此生成文档中的词汇 $w_{d,i}\sim softmax(f_g(\boldsymbol{\theta}_d,\boldsymbol{c}_d))$;(3)由于对每个文档引入多个标签,在生成文档后利用一个多层感知机 f_y 计算所有标签 $\boldsymbol{y}_d$ 在当前主题下的概率分布 $\boldsymbol{y}_d\sim p(y\mid f_y(\boldsymbol{\theta}_d,\boldsymbol{c}_d))$.此外,为实现主题-词汇的稀疏约束,SCHOLAR 模型假设主题-词汇先验分布 $\boldsymbol{\eta}$ 满足超参数为 λ 的指数-正态分布.而在网络结构上,SCHOLAR 使用 VAE 结构实现模型参数推断,即在编码阶段可使用多层感知机先从输入数据$\{d,\boldsymbol{c}_d,y_d\}$中采样 $\boldsymbol{\mu}_d$ 和 $\boldsymbol{\sigma}_d$,再利用重参数化技巧采样潜在主题分布 $\boldsymbol{\theta}_d=softmax(\boldsymbol{\mu}_d+\boldsymbol{\varepsilon}\boldsymbol{\sigma}_d)$.

SCHOLAR 可以更加便利地融合元数据进而扩展为其他模型,如监督式主题模型.这使得该模型不仅可以用于纯文本数据的分类/聚类和主题建模任务,还可以应用在其他多元数据上,如情感分析、时序数据分析.

4.4 神经网络主题模型对比

总体来说,基于神经网络结构的主题模型与传统概率主题模型在分布假设与模型结构方法具有较大差异.表 5 对比分析了各神经网络主题模型在模型输入、网络结构和参数推断方法上的异同点.

一方面,神经主题模型基本摒弃了概率主题模型关于分布的狄利克雷先验假设与吉布斯采样方式,直接将分布转换为神经网络中的结点或权重矩

① The implementation of SCHOLAR is available at github.com/dallascard/scholar

② Relaxed Wasserstein with applications to GANs. ArXiv Preprint: 1705.07164, 2017

阵，并使用神经网络常用的后向传播算法或随机梯度算法训练模型参数.此外，由于构建出了神经网络结构，那么在输入层上即可叠加词向量，从而更好地利用词汇之间的语义相似度.此外，NTM 与后面讨论的其他主题模型结构也略有不同.

(1) NTM 模型是较为简单直接的神经网络主题模型，它从前馈神经网络角度直接重构两个分布矩阵而不添加其他约束.

(2) 基于 VAE 推断的若干模型(如 GSM、GSB、RSB、AVITM 等)在遵从 VAE 模型的文档-主题推断过程及神经变分参数估计方法的基础上，进一步优化主题-词汇的分布假设以符合主题建模需求.

(3) NSTC 和 SCHOLAR 分别从不同角度添加主题-词汇的稀疏约束.其中，NSTC 将该约束添加到模型的优化函数中，而 SCHOLAR 使用先验分布实现稀疏约束.

表 5　基于神经网络结构的主题模型对比

方法	输入层	文档-主题层	主题-词汇层	文档-词汇层
NTM[18]	文档的 n-gram 向量	$\boldsymbol{\theta}_d=softmax(\boldsymbol{W}_1(d))$	$\boldsymbol{\phi}_w=sigmoid(le(g)\times\boldsymbol{W}_2)$	$p(w\|d)=\boldsymbol{\phi}_w\boldsymbol{\theta}_d^{\mathrm{T}}$
NVDM[87]	文档词袋子	$\boldsymbol{\theta}_d=relu(N(\mu_d,\sigma_d^2))$	$\boldsymbol{\phi}=\boldsymbol{W}$	$p(w\|d)=\boldsymbol{\theta}_d\boldsymbol{\phi}_{\cdot w}^{\mathrm{T}}$
GSM[90]		$\boldsymbol{\theta}_d=softmax(N(\mu_d,\sigma_d^2))$		
GSB[90]	文档的词向量	$\boldsymbol{\theta}_d=f_{\mathrm{SB}}(sigmoid(N(\mu_d,\sigma_d^2)))$	$\boldsymbol{\phi}_k=softmax(\boldsymbol{w}\cdot\boldsymbol{z}_k^{\mathrm{T}})$	$p(w\|d)=\boldsymbol{\theta}_d\boldsymbol{\phi}_{\cdot w}^{\mathrm{T}}$
RSB[90]		$\boldsymbol{\theta}_d=f_{\mathrm{SB}}(f_{\mathrm{RNN}}(N(\mu_d,\sigma_d^2)))$	$\boldsymbol{\phi}_k=softmax(\boldsymbol{w}\cdot\mathrm{RNN}(\boldsymbol{z}_k^{\mathrm{T}}))$	
AVITM[41]	文档词袋子	$\boldsymbol{\theta}_d=softmax(N(Laplace(\alpha)))$	$\boldsymbol{\phi}_w=softmax(\boldsymbol{W})$	$p(w\|d)=\boldsymbol{\theta}_d\boldsymbol{\phi}_{\cdot w}^{\mathrm{T}}$
NSTC[42]	文档的词向量	$\boldsymbol{s}_{d,w}=relu(\boldsymbol{W}_1(w,:))$	$\boldsymbol{\phi}_w=relu(le(w)\times\boldsymbol{W}_2)$	$p(w\|d)=\boldsymbol{s}_{d,w}\boldsymbol{\phi}_{\cdot w}^{\mathrm{T}}$
NSMTM[93]	文档的词向量	$\boldsymbol{\theta}_d=sparsemax(N(\mu_d,\sigma_d^2))$	$\boldsymbol{\phi}_k=sparsemax(\boldsymbol{w}\cdot\boldsymbol{z}_k^{\mathrm{T}})$	$p(w\|d)=\boldsymbol{\theta}_d\boldsymbol{\phi}_{\cdot w}^{\mathrm{T}}$
SCHOLAR[43]	文档词袋子	$\boldsymbol{\theta}_d=softmax(N(\alpha))$	$\boldsymbol{\phi}_k\sim softmax(\boldsymbol{\eta}_k+\boldsymbol{m})$	$p(w\|d)\sim softmax(\boldsymbol{\theta}_d\boldsymbol{\phi}_{\cdot w}^{\mathrm{T}}+\boldsymbol{m})$

5　联合训练主题模型

传统概率主题模型和基于神经网络结构的主题模型都是一种文档生成模型，即可从数据中训练出文本生成过程所需参数，并产生潜在主题的词汇表达.这使得主题模型可用于文本分类/聚类、主题词抽取(或称为潜在面特征抽取)、情感分类、商品评论分析、多文档自动抽取摘要等任务.然而，在上面综述的主题模型中，往往以文档的词袋子形式作为模型输入，并产生主题-词汇的分布.随着循环神经网络(RNN)在序列数据处理中逐步显现出优势[32-33,35,95]，学者开始考虑以文档的词序列作为输入，使用 RNN 网络转换为隐藏层向量，并根据不同任务输出相应结果.相比于使用词袋子输入，使用词序列可更加充分地利用词汇的上下文信息，使得相关(分词、文本分类、实体关系抽取、事件抽取、自动摘要等)任务的性能得以显著提升.

另一方面，对于某些文本分析任务，如文本摘要、机器翻译、人机对话等，所需要的输出也是自然文本形式.这时，往往需要借助语言模型才能生成自然语句.然而，语言模型一般只刻画句子级别的词序列，对于较长文档甚至多文档无法有效处理.而传统主题模型能够捕获文档全局语义结构，但基于词袋子模型假设而忽略了词汇顺序[95].因而，近几年学者提出多种结合主题模型与语言模型的新型神经网络联合训练模型.这类模型融合了上述两种方法的优势，即既可以从输入词序列中捕获词汇之间的依赖关系，也利用了潜在主题结构捕获多文档的全局语义信息，因而既可作为主题模型实现文本分类和主题推断，也可作为语言模型生成自然文本.

当前考虑潜在主题结构的语言模型研究成果相对较少，并主要聚焦于两种方式：(1) 基于浅层神经网络结构的主题与词向量联合训练模型；(2) 基于 RNN 结构的主题与文本生成联合训练模型.两者一般均能产生潜在主题和词向量，但侧重点有所不同，前者侧重于生成表达不同主题(或概念)的词向量，而后者侧重于生成特定主题下的句子而非主题词.

5.1　神经语言模型简介

在自然语言处理中，语言模型是计算一个句子分配概率的模型.传统的统计语言模型一般基于 k-阶马尔科夫假设，即计算每个词汇在前 k 个词汇序列出现情况下的发生概率，进而完成整个句子的词序列分配，如式(13)所示：

$$p(w_{i+1}\mid w_1,\cdots,w_i)\propto p(w_{i+1}\mid w_{i-k+1},\cdots,w_i)\tag{13}$$

随后，Bengio 等人[67]提出的神经概率语言模型以 k-阶词汇的词向量为输入，预测下一个词的向量表示.这时，输入的词向量将被传入到一个神经网络层，并通过非线性变换预测句子输出：

$$\begin{cases}p(w_{i+1}\mid w_{i-k+1},\cdots,w_i)=p(w_{i+1}\mid \boldsymbol{h}_{i+1})\\ \boldsymbol{h}_{i+1}=f(\boldsymbol{h}_{i-1},w_i)\end{cases}\tag{14}$$

其中，$f(w)$ 函数一般是神经网络单元，如多层感知

机或循环神经网络；$\boldsymbol{h}_i$ 为词汇 w 对应的隐藏层向量.

语言模型可在原始文档上进行训练，即对于一个 k-阶语言模型而言，只需从连续的文本中提取 $(k+1)$个元组，然后将第$(k+1)$个词汇看做监督信号. 这样几乎可以创造无限训练数据. 其中，作为语言模型的副产品-词向量，已广泛应用于自然语言处理的各项任务中.

基于多层感知机结构的浅层语言模型虽然能够生产句子，但是由于没有深入考虑输入词汇的上下文关系，因而产生的文本可读性不强. 随后，Mikolov 等人[96]将循环神经网络引入到语言建模中，使得语言模型的性能得到较大提升. 此外，基于循环神经网络的改进版本，如长短期记忆网络[97]以及门限循环单元(Gated Recurrent Unit，GRU)网络[98]，也相继用于进一步改善语言建模的性能. 这样的语言模型在人机对话[99]、机器翻译[97]、自动文摘等[89]等文本生成任务中起到关键作用.

5.2 基于 Skip-gram 的主题-词向量模型

词向量与主题联合训练模型旨在产生更具表达能力的词向量，特别是针对“一词多义”问题，可考虑不同词汇在不同主题下的语义差异，以训练出相应的词向量. 一般来说，这类模型往往也会产生文档-主题和主题-词汇分布这一副产品，因而也可以作为主题模型使用.

5.2.1 Skip-gram 模型

广泛使用的词向量模型 Word2Vec 是由 Mikolov 等人[69]在 2013 年提出的，其中包含两种上下文表示方法各异的模型：CBOW 和 Sikp-gram. 此外，还存在其他词向量表示模型，如 GloVe[70]等. 关于词向量模型的研究成果比较丰富①，但从主题模型角度而言，Sikp-gram 模型常作为基准模型来训练表达不同主题的词向量.

Skip-gram 模型利用目标词的上下文来预测目标词，即给定词汇序列 $D=\{w_1,\cdots,w_M\}$和其上下文 c，从窗口大小为 k 的上下文中选择一个词汇 w_{i+c}作为目标词 w_i的上下文表示，并根据上下文预测 w_i. Skip-gram 的目标是对整个词序列最大化：

$$L(D)=\frac{1}{M}\sum_{i=1}^{M}\sum_{-k\le c\le k,\ c\neq 0}\log p(w_{i+c}\mid w_i) \tag{15}$$

其中，$p(w_{i+c}\mid w_i)=softmax(\boldsymbol{w}_{i+c}\cdot\boldsymbol{w}_i)$，$\boldsymbol{w}_i$、$\boldsymbol{w}_{i+c}$为目标词 w_i与上下文词 w_{i+c}的词向量.

Skip-gram 模型使用浅层神经网络结构，即对输入的独热编码词汇，映射到一个不包含激活函数的隐藏层，并通过 softmax 函数转换为输出概率. 那么，通过最大化上述目标函数即可训练出隐藏层权重，该权重即为词汇的向量表示.

5.2.2 主题-词向量联合训练模型

由于 Skip-gram 等经典词向量模型对每个词汇只训练一个向量表达，无法解决“一词多义”问题. 另一方面，主题模型能够将文档映射到低维主题空间中，每个主题空间中的词汇往往表达同一主题(或概念)，因而结合主题模型思想而设计的词向量训练模型应运而生.

最早利用主题模型训练词向量的方法是由 Liu 等人提出的主题词向量模型 TWE[45]②，该模型假设每个词汇在不同主题下有不同的向量表示. 为此，首先引入 LDA 模型获取主题-词汇分布，并基于 Skip-gram 设计 3 种变形形式以刻画不同主题下的词向量：

(1) TWE-1. 假设每个主题 z 作为该主题下所有相关词汇的一个伪词汇，模型同时学习词汇 w_i和主题 z_k的向量表示，并将这两个向量拼接为主题词的向量 $\boldsymbol{w}^z=\boldsymbol{w}\oplus\boldsymbol{z}$. 这时，需最大化如下目标函数：

$$L(D)=\frac{1}{M}\sum_{i=1}^{M}\sum_{-k\le c\le k,\ c\neq 0}\log p(w_{i+c}\mid w_i)+\log p(w_{i+c}\mid z_i) \tag{16}$$

(2) TWE-2. 假设每个词汇-主题对(w_i,z_k)为一个伪词汇，并直接学习主题词的向量表示 $\boldsymbol{w}^z$. 这时，需最大化如下目标函数：

$$L(D)=\frac{1}{M}\sum_{i=1}^{M}\sum_{-k\le c\le k,\ c\neq 0}\log p(\langle w_{i+c}\mid z_{i+c}\rangle\mid\langle w_i,z_i\rangle) \tag{17}$$

(3) TWE-3. 保留并链接词汇和主题的向量表示，进而拼接出主题词向量 $\boldsymbol{w}^z=|\boldsymbol{w}|\oplus|\boldsymbol{z}|$. 这时，词向量和主题向量维度不一定相同.

图 8[45]对比了这三种方法与 Skip-gram 之间的差异，其中，深色圆圈代表词向量，浅色圆圈代表主题向量. TWE 是 LDA 与词向量模型的简单有效结合，可从同一语料中同时训练出词汇和主题的向量表示，并以两者的拼接作为主题词的向量表示. 最后，文档-主题分布即为该文档中所有主题词向量的平均值. 由于 TWE 直接使用 LDA 生成主题词，因

① 有关词向量模型的分析对比，可参阅：https://www.researchgate.net/publication/301779119_A_Survey_of_Word_Embedding_Literature_Context_Representations_and_the_Challenge_of_Ambiguity

② The implementation of TWE available at https://github.com/largelymfs/topical_word_embeddings

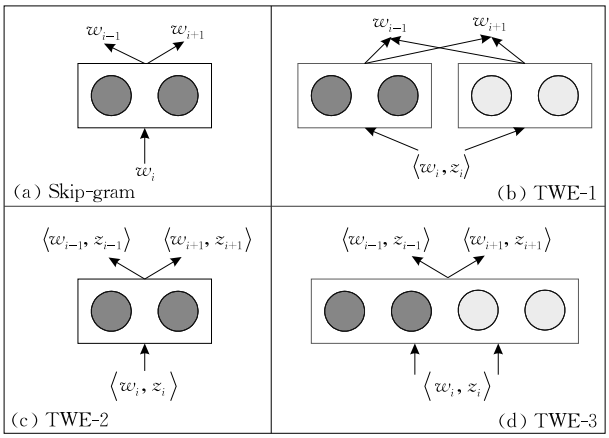


图 8　Skip-gram 与 TWE 对比

而该方法获得的主题词的语义一致性方面有待商榷，且模型训练往往需要大规模语料的支撑.

可以看出，TWE 本质上还是一个单纯的词向量模型，它使用 LDA 模型推断出的主题-词汇分布作为词向量模型的输入. 随后学者们开始尝试将主题与词向量联合训练，即以文档的词序列为输入，同时训练出主题-词汇分布和词汇在不同主题下的向量表示.

最早的主题-词向量联合训练模型是 Li 等人[46]在 2016 年 ACL 会议上提出的一种生成式主题向量模型 TopicVec①，该模型建立在 PSDVec[100]词向量模型基础上并结合了主题信息，即每个文档是由主题的多项式混合分布构成，每个主题由向量表示，而主题中每个词向量由所在上下文信息及主题共同决定. 该模型的文本生成过程遵循主题模型的一般步骤：

(1) 对于每个文档 d：

从参数为 α 的 Dirichlet 先验中生成文档-主题分布：$\boldsymbol{\theta}_d \sim Dir(\alpha)$；

(2) 对于每个主题 $k(k=1,2,\cdots,K)$：

从参数为 γ 的超球体分布中生成主题向量：$\boldsymbol{\phi}_k \sim Unif(B_\gamma)$；

(3) 对于文档 d 中第 $i(i=1,2,\cdots,d_n)$ 个位置：

从 $\boldsymbol{\theta}_d$ 中生成一个主题分配：$z_{d,i} \sim Multi(\boldsymbol{\theta}_d)$；

根据主题分配 $z_{d,i}$ 和词汇所在上下文 c 生成一个词汇：$w_{d,i} \sim P(w_{d,i} \mid w_{d,i-c}:w_{d,i-1}, z_{d,i}, d)$.

从上述生成过程中可以看出，TopicVec 与 LDA 等经典主题模型的不同之处在于：(1) TopicVec 中的每个主题向量从超球体分布中采样而来；(2) 文档 d 中每个词汇 w_c 的产生概率不仅由其所采样的主题 z_c 决定，还由其前面 c 个上文词汇 $(w_0:w_{c-1})$ 决定，即：

$$P(w_c \mid w_0:w_{c-1}, z_c, d) \approx P(w_c)\exp\left\{\boldsymbol{w}_c^{\mathrm{T}}\left(\sum_{l=0}^{c-1}\boldsymbol{w}_l+\boldsymbol{\phi}_{z_c}\right)+\sum_{l=0}^{c-1}a_l a_c+r_{z_c}\right\} \quad (18)$$

其中，$\boldsymbol{w}_c$ 为词向量，$a_l a_c$ 为无法由词汇语义相似度 $\boldsymbol{w}_c^{\mathrm{T}}\boldsymbol{w}_l$ 捕获的残差，r_{z_c} 为主题残差常量.

与 LDA 模型的文本生成过程类似的还有 Jiang 等人[101]提出的潜在主题向量模型(Latent Topic Embedding，LTE). 不同之处在于，在 LTE 中，文档中每个词汇由二项指示器来决定是从狄利克雷-多项式分布中产生还是从上下文中产生. 当从多项式分布中产生时，文本生成过程即 LDA 过程；而当依据上下文产生时，词汇生成概率由其词向量与上下文词向量及其主题的链接函数决定.

TopicVec 和 LTE 模型均使用词向量模型中的链接函数辅助生成文档中的词汇，这使得词汇之间的语义相似度可以参与到主题采样过程中，进而提高了主题词语义一致性.

2017 年 SIGIR 会议上还提出一种新的词向量与主题联合训练模 STE[8]以同时产生文档-主题分布和主题-词汇分布及相应的词向量表示. 该模型假设文档 d 中每个词汇 w_j 的生成概率直接由给定的目标词汇 w_i 及相应主题 z 决定，即 $p(w_j \mid w_i, d) = \sum_z p(w_j \mid w_i, z)p(z \mid d)$. 此外，根据 $p(w_j \mid w_i, d)$ 推断方式的不同，STE 产生两种变形形式：(1) STE-Same. 即 Skip-gram 中的每个词对 (w_i, w_j) 来自于同一主题；(2) STE-Diff. 即 Skip-gram 中的每个词对 (w_i, w_j) 中两个词汇的主题分配过程相互独立.

上述 4 种模型不仅可以产生表达不同主题的词向量，还可以产生文档-主题分布，因而模型也可用于文本分类等任务. 此外，这类模型与 LDA 假设类似，即主题在文档上的分布是也是稠密的. Moody 等人②提出一种新的结合 LDA 与词向量的主题模型，即 lda2vec，该模型依旧采用 Skip-gram 中的欠采样损失函数，但与上述 4 种模型不同之处在于模型它还可以产生稀疏可解释的文档向量.

5.3　基于 RNN 结构的主题-文本生成模型

RNN 可以处理任意长度的序列数据，并生成有效的特征，因而在语言模型[96]和机器翻译[35]等自然语言处理任务中取得一定的应用突破. 在基于 RNN 结构的主题-文本生成联合训练模型中，输入文本不

① The implementation of TopicVec is available at https://github.com/askerlee/topicvec

② Mixing Dirichlet topic models and word embeddings to make lda2vec. https://arxiv.org/pdf/1605.02019.pdf, 2016

再是词袋子形式，而是文本的词序列. 模型通过RNN网络生成隐藏层单元，并基于该单元生成特定主题下的自然文本. 从某种角度看，该类模型类似生成式多文档摘要模型，可为若干个文档按主题生成每个主题下的摘要句子.

Tian等人首次提出一个基于RNN的句子级别的主题模型SLRTM②，该模型假设一个句子中所有词汇采样自一个主题，且句子中每个词汇的生成依赖于该句子的主题及句子中该词汇前面的所有词汇. 由于该模型与传统概率主题模型有较大差异，图9展示了模型结构.

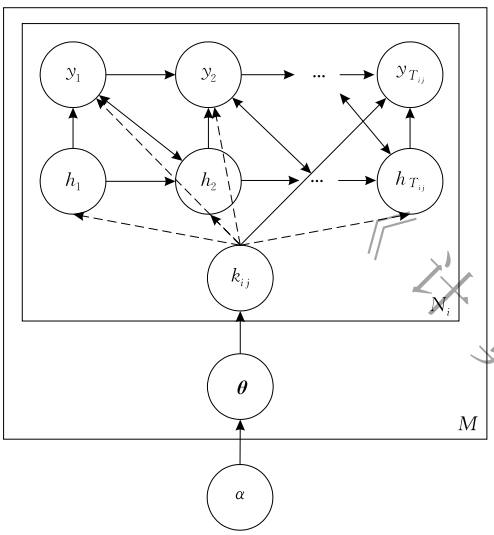


图9 SLRTM模型示意图

在SLRTM模型中，由N_d个句子组成的文档d的生成过程如下：

(1) 从参数为α的Dirichlet先验分布中生成文档-主题分布：$\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha})$；

(2) 对于文档d中第j个句子$s_{d,j}$：

从$\boldsymbol{\theta}_d$中生成一个主题分配：$z_{d,j} \sim Multi(\boldsymbol{\theta}_d)$；

对于句子s_j中的每个词汇$y_t \in \{y_1, y_2, \cdots, y_{T_j}\}$：

利用LSTM计算词汇y_t的隐藏层状态$\boldsymbol{h}_t = f(\boldsymbol{h}_{t-1}; \boldsymbol{y}_{t-1}; \boldsymbol{z}_{dj})$，其中$\boldsymbol{y}_{t-1}$和$\boldsymbol{z}_{dj}$分别为相应的词向量和主题向量；

根据主题分配$z_{d,j}$及隐藏层生成一个词汇：$p(w|y_{t-1}, \cdots, y_1; z_{dj}) \propto g(y_t; \boldsymbol{h}_t; \boldsymbol{y}_{t-1}; \boldsymbol{z}_{dj})$.

从上述生成过程可以看到，SLRTM使用LDA中文档-主题分布假设，但未采用主题-词汇分布假设. 为充分利用词汇的上下文语义信息，SLRTM采用神经语言模型的基本思想，即利用LSTM将句子中的词向量映射到隐藏层；然后结合主题向量生成在特定主题下的词序列(即句子).

与SLRTM模型的文档生成过程相似还有Dieng等人[95]提出的TopicRNN①，该模型也利用RNN捕获词汇之间的局部依赖并使用潜在主题捕获文档的全局语义信息. 然而，TopicRNN与SLRTM在模型假设方面有所不同：(1) TopicRNN直接在文档级别而非句子推断主题；(2) TopicRNN从高斯先验中为文档采样一个主题向量$\boldsymbol{\theta}_d$；(3) 在文档d的词汇生成过程中，TopicRNN引入一个满足伯努利分布的停用词指示器$l_t \sim Bern(sigmoid(\boldsymbol{\Gamma}^{\mathrm{T}} \boldsymbol{h}_t))$来决定待采样词汇是否为停用词. 这里，停用词指示器l_t用来控制主题向量$\boldsymbol{\theta}_d$是否影响词汇生成的条件概率估计，即当y_t为停用词，则$\boldsymbol{\theta}_d$对最终输出没有影响；反之，利用$\boldsymbol{\theta}_d$与潜在词向量$\boldsymbol{b}_i$之间的点乘积增强词汇到该主题的分配概率. 因此，该模型能够自动处理停用词问题，进而也可用作文档的特征抽取器.

另一个基于高斯先验的联合训练模型是高斯混合神经主题模型GMNTM[44]，该模型是一个主题与词向量、句子向量及文档向量的联合训练模型. GMNTM假设整个文档集是关于主题的多元高斯混合分布，并可从该分布中采样文档、句子及词汇的向量表示. 与此同时，在文档d中，GMNTM利用简单语言模型，根据词汇w的前k个词汇以及词汇所在的文档向量和句子向量共同生成词汇w.

从文档生成过程可以看到，SLRTM、TopicRNN和GMNTM均使用单神经网络结构，即文档-主题依旧从先验分布中生成，而主题-词汇由神经网络中产生. 在单神经网络基础上，Lau等人[102]和Wang等人[103]先后提出双神经网络的主题-语言联合训练模型TDLM②(Topically Driven Language Model)和TCNLM(Topic Composition Neural Language Model). 这时，整个模型可认为是一个多任务学习模型，两个子任务主题推断和文本生成均由神经网络产生.

TDLM模型以文档的词向量拼接为输入并使用卷积神经网络转换为文档向量$\boldsymbol{d}$. 为从文档中推断出潜在主题，模型使用带有注意力机制的单层网络将文档$\boldsymbol{d}$表达为主题$\boldsymbol{\theta}$的多项式分布形式，即$\boldsymbol{\theta} = \boldsymbol{B}^{\mathrm{T}} softmax(\boldsymbol{A}\boldsymbol{d})$，其中$\boldsymbol{A}$和$\boldsymbol{B}$为主题向量的不同表达形式. 在文本推断任务中，模型利用LSTM刻画句子中上下文词汇之间的语义关系，并将文档

① The implementation of TopicRNN is available at https://github.com/narratives-of-war/topic-rnn

② The implementation of TDLM is available at https://github.com/jhlau/topically-driven-language-model

的主题分布信息 $\boldsymbol{\theta}$ 融入到 LSTM 的隐藏层 $\boldsymbol{h}_t$ 中.

TCNLM 模型以文档的词汇特征空间为输入，并假设主题 $\boldsymbol{\theta}$ 是关于文档的高斯分布，使用 softmax 函数将输入层转换为主题向量 $\boldsymbol{z}=softmax(\boldsymbol{W\theta}+\boldsymbol{b})$. 这样，文档 d 中词汇 w 的生成概率表达式与 LDA 模型类似，即 $p(w|d;\boldsymbol{\phi},t)=\boldsymbol{\phi}_{\cdot,w}\times\boldsymbol{z}_d^{\mathrm{T}}$，其中 $\boldsymbol{\phi}$ 为主题-词汇分布. 此外，该模型引入混合专家(Mixture-of-Experts，MoE)语言模型，其中每个“专家”即是由一个 RNN 单元构建的潜在主题. MoE 在神经语言模型[67]基础上考虑了不同词汇在主题分布上的差异来生成文本. 这时，每个词汇的生成概率即为

$$p(w_{i+1})=softmax(\boldsymbol{V}\boldsymbol{h}_{i+1}) \tag{19}$$

$$\boldsymbol{h}_{i+1}=\sigma(\boldsymbol{W}(\boldsymbol{z})\boldsymbol{w}_i+\boldsymbol{U}(\boldsymbol{z})\boldsymbol{h}_i) \tag{20}$$

其中，$\boldsymbol{W}(\boldsymbol{z})$ 和 $\boldsymbol{U}(\boldsymbol{z})$ 为 RNN 单元的权重矩阵.

上述模型均由两个神经网络构成，因此整体目标函数也由神经主题模型和神经语言模型的优化函数组合而成. 实验结果表明，双神经网络联合训练模型优于只使用 RNN 结构的其他语言模型. 此外，包含神经语言模型结构的主题模型以一般句子的词序列而不是词袋子为输入，因此不仅可有效提升主题词抽取和文本分类能力，还可以作为文本生成模型，即给定一个主题，利用束搜索[104]等句子生成方法能够生成自然文本. 这使得该类模型在自动文本生成任务(如人机对话、自动摘要等)中展现出很大潜力.

5.4　联合训练主题模型对比

最后，表 6 对比分析了主题联合训练模型在文档-主题生成、文本生成和词模型特点上的异同点.

表 6　联合训练主题模型对比

方法	文档-主题生成	文本生成	模型的特点
TWE[45]	狄利克雷-多项式分布	基于简单神经网络语言模型	在基于 LDA 产生的主题-词汇分布基础上，基于 Skip-gram 模型训练主题向量及不同主题下的词向量.
TopicVec[46]			以文档为单位推断主题，并根据主题、词向量及上下文词向量生成词汇.
LTE[101]	狄利克雷-多项式分布	狄利克雷-多项式分布与基于简单神经网络语言模型的混合	以句子为单位推断主题，并根据狄利克雷-多项式分布或上下文词向量及其主题生成词汇.
STE[8]	依据词汇在主题下的后验概率生成	基于简单神经网络语言模型	以文档为单位推断主题，并根据文档及目标词生成词汇向量.
SLRTM*	狄利克雷-多项式分布	基于 LSTM 的语言模型	以句子为单位推断潜在主题，并根据主题及上下文词汇生成词汇.
TopicRNN[95]	高斯分布	基于 RNN 的语言模型	以文档为单位推断主题，并根据主题及上下文词汇生成词汇，且可判别生成的词汇是否是停用词.
GMNTM[44]		基于简单神经网络语言模型	以文档为单位推断主题，使用上下文词汇及文档、句子生成词汇.
TDLM[102]	多层神经网络	基于 LSTM 的语言模型	以文档为单位推断主题，使用上下文词汇及文档的主题信息生成词汇.
TCNLM[103]	基于高斯先验的神经网络	基于 RNN 的语言模型	以文档为单位推断主题，使用上下文词汇及其主题生成词汇.

(1) 联合训练模型一般以句子/文档的完整词序列为输入，而非文档的词袋子. 因此，模型往往不再关注如何推断出主题-词汇分布，而是生成语义完整的自然句子. 此外，在神经语言模型基础上考虑了文档的主题分布差异，因而能够生成不同主题下的自然文本. 这使得这类模型不仅可用于文本分类，还可用于词向量、文档向量生成以及文本生成. 结合语言模型的主题模型由于其较高的使用价值，也将是未来的研究热点方向之一.

(2) 以训练词向量为目的的联合训练模型(如 TWE、TopicVec、LTE 和 STE)更加关注“一词多义”情况下的词向量表达问题，因而一般以 Skip-gram 模型为基准，并考虑词汇所表达的潜在主题的差异来训练词向量. 此外，由于不再关注自然文本生成问题，模型一般使用简单的神经网络结构来训练词向量，并产生文本的主题分布.

(3) 以语言模型为训练目的的联合训练模型(如 SLRTM、Topic RNN 等)更加关注自然文本生成. 为此，模型往往采用 RNN 等处理序列生成能力更强的深度神经网络来刻画文档序列中上下文之间的语义关系. 此外，由于考虑了文本中潜在主题的存在，模型一般采用类似 LDA 模型的文本生成基本流程，即文本中的每个词汇是在考虑所在主题信息下生成的.

6　主题模型的评测语料与指标

6.1　主题模型常用评测语料

学者们为了更加客观地对比主题模型的优劣，产生了若干个公开评测的数据集，如 20NewsGroups、

Web-Snippet 等.公开评测数据集使得不同模型能够在同一语料下进行横向对比.表 7 列出了主题建模中常用的公开数据集、使用该数据集的主题模型、以及这些数据集的下载资源.

表 7 主题模型评测常用公开语料对比

语料名	文档规模	使用该语料的主题模型
20NewsGroups①	20000	GLDA、LCTM、MvTM、He et al.、CGTM、LFTM、LFBTM、DGPU-LDA、KGE-LDA、NTM、Miao et al.、AVITM、NSTC、NSMTM、SCHOLAR、GMNTM、TDLM、SLRTM、TWE、TopicVec、STE
NIPS②	1740	GLDA、MvTM、KGE-LDA
Web-Snippet③	12000+	WEI-FTM、GPUDMM、GPUPDMM、NSTC
IMDB④	25000	SCHOLAR、TopicRNN、TDLM、TCNLM
Reuters-v2 Corpus⑤	804000	Miao et al.、GMNTM
Reuters-21578⑥	11367	WEI-FTM、CGTM
Twitter⑦	11109	WEI-FTM、LFTM
Wiki10+⑧	17000+	NTM、SLRTM

从表 7 中可以看出,20NewsGroups 是各个主题模型使用最多的评测数据集.该数据集是用于文本分类、挖掘和检索研究的国际标准数据集之一,包含 20 个不同类型、共约 20000 份的新闻文档.语料用语规范,各类别文本数量相当,其中某些类别的主题特别相似(如 comp. sys. ibm. pc. hardware vs. comp. sys. mac. hardware),还有些则完全不相关(如 misc. forsale vs. soc. religion. christian).此外,Reuters Corpus 和 Wiki10+ 也是使用较为广泛的测试语料,而 Web-Snippet 往往广泛应用于短文本主题建模.

6.2 主题模型常用评价指标

如何判断一个主题模型的好坏一直以来也是研究者们关注的问题.学术界认为主题模型的性能一般可从模型复杂度(包括时间复杂度和空间复杂度)、泛化能力、抽取的主题词质量与可理解性以及文本分类准确性等角度分别评估.

(1) 模型泛化能力.一般采用困惑度[17](Perplexity)或留存数据似然概率[22](Held-out Likelihood)指标,这两者虽表达形式不同,但原理基本一致.其中,困惑度是更为常用的指标,困惑度越低,通常表示模型的泛化能力越好.对于包含 N 个测试文本的语料库,其中 N_d 为文本 d 中词汇的数量,则主题模型的困惑度计算公式如下:

$$Perplexity = -\frac{1}{N}\sum_{d=1}^{N}\frac{1}{N_d}\log P(d) \tag{21}$$

(2) 主题词语义一致性(Topic Coherence).在主题模型中,能否产生语义一致、易于理解的主题词一直是主题模型关注的焦点问题.评价主题词语义一致性最常用的方法是点对互信息[105](Pointwise Mutual Information,PMI)或平均点对互信息[106](Normalized Pointwise Mutual Information,NPMI).给定 k 个主题,每个主题由 T 个最相关词汇组成,其中 $p(w_i)$ 为词汇 $w_i(i=1,\cdots,T)$ 在文档中出现概率,$p(w_i,w_j)$ 为词对 (w_i,w_j) 共同出现的概率,则 PMI 和 $NPMI$ 的计算公式分别如下:

$$PMI = \frac{1}{K}\sum_{k}\frac{2}{T(T-1)}\sum_{1\le i<j\le T}\log\frac{-p(w_i,w_j)}{p(w_i)p(w_j)} \tag{22}$$

$$NPMI = \frac{1}{K}\sum_{k}\frac{2}{T(T-1)}\sum_{1\le i<j\le T}\frac{\log\dfrac{p(w_i,w_j)}{p(w_i)p(w_j)}}{-\log p(w_i,w_j)} \tag{23}$$

在实验结果评估中,一般使用大规模语料库,如包含数百万个文档的维基百科或百度百科语料作为计算词汇出现概率和词对出现概率的依据.此外,由于百科文档较长,文档不同位置所包含的语义信息不尽相同,因而一般选择包含连续 10～20 个词汇的滑动窗口来计算上述概率[76].PMI 或 $NPMI$ 值越大,说明模型产生的主题词语义一致性越高.

另一种常用的主题词语义一致性评价指标是 Mimno 等人[107]提出的主题凝聚度[75,80](Coherence Score),该指标的基本思想是表达相同概念的词汇往往出现在同一文本中,因此可直接计算 top-N 个主题词及词对在当前文档集中的共现次数,而无需依赖外部数据.该指标比较合适度量高频主题词,而对于低频词汇往往无效.

近年来,随着词向量技术在词汇语义相似度度量方面体现出的显著优势,Fang 等人[108]提出一种基于词向量的主题语义表达一致性指标 WESim,该指标通过计算主题词中 top-N 个词对之间语义相

① The dataset is available at http://qwone.com/~jason/20Newsgroups/
② The dataset is available at http://www.cs.nyu.edu/~roweis/data.html
③ The dataset is available at http://jwebpro.sourceforge.net/data-web-snippets.tar.gz
④ The dataset is available at https://github.com/jhlau/topically-driven-language-model
⑤ The dataset is available at http://trec.nist.gov/data/reuters/reuters.html
⑥ The dataset is available at http://www.daviddlewis.com/resources/testcollections/reuters21578//
⑦ The dataset is available at http://trec.nist.gov/data/microblog.html
⑧ The dataset is available at http://www.zubiaga.org/datasets/wiki10+/

似度的平均值来衡量主题词语义一致性水平.大规模人工评价结果表明,WESim 指标在短文本数据集上能有效捕获主题词语义一致性,该指标也比 *PMI* 等效率更高、鲁棒性更好.

(3)文本分类能力(Text classification).主题模型作为监督式模型使用时,可也用于文本分类和信息检索,因而也常从分类的准确率、召回率和 *F*-score 角度评价模型的文本表达能力.此外,某些研究也使用准确率-召回率曲线(如 GMNTM)和 Micro-*F*1 score[109](如在 NTM、SLRTM 模型中)等方式评估模型的文本分类能力.更进一步,若干主题模型如 GLDA[38]和 LCTM[71]具有处理新文档中出现未登录词的能力.因此,这两个模型也从该角度验证模型文本表示性能,即比较当测试文档中未登录比例不断增大时,模型的文本分类准确率变化情况.

(4)人工评价方法.是一种比较可靠的评价主题是否有意义的主观方法.该方法比较适用于来源复杂且没有标注的文本语料.其中最常用的方法是直接对比展示不同模型产生的 top-5 或 top-10 个主题词,研究者和阅读者通过直观理解主题词语义来比较模型优劣.此外,文献[109]首次提出了人工评价主题的语义理解性,并将主题词中的冲突词数(Topic Intrusion Case)作为评价指标.冲突词是指一个主题中的另类词,即与该主题主要内容明显不相符的词汇.一般来说,冲突词数越少越好.

与冲突词检测指标有异曲同工之妙的是异常主题自动检测[110]①,即给定一个文本及其 top-*K* 相关主题,检测其中是否存在不能表达该文本的异常主题.与人工检测冲突词不同的是,该方法通过卷积神经网络来度量文本向量和主题向量之间的归一化 sigmoid 值(Normalised Sigmoid Score,NSS)以自动检测异常主题.

6.3　评价指标探讨

第 6.2 节讨论的若干评价指标从不同角度评估主题模型性能.总体来说,在主题建模相关文献中,一般使用困惑度评价模型在测试文本上的泛化能力,使用 *NPMI*/*PMI* 评估主题词语义一致性水平,使用分类准确率评估主题模型文本分类能力,最后通过展示主题词(或句子)、可视化主题词在空间中分布等方式评价模型所抽取的主题词或自然语句的可理解性和语义一致性.

表 8 对比展示了目前主题模型使用的主流评价指标及本文讨论的各方法所采用的评价指标.从表 8 中可以看出,一般情况下,主题建模相关文献均使用 2～3 种指标分别度量模型的泛化能力、文本分类能力和主题词抽取能力.其中,互信息(或 Coherence 指标等)和主题词展示是分别从客观计算和主观判断两个角度评价模型的主题词抽取能力的方式.这两类指标相辅相成,一般共同用于主题词语义一致性评价,使得研究者既可以从客观数值上实现模型间的横向比较,也可通过主观认知对比各模型产生的主题词优劣.

表 8　各主题模型使用评价指标对比

评价指标	主题模型
困惑度	WEI-FTM、He et al.、NSTC、Miao et al.、NSMTM、LTE、SLRTM、TopicRNN、GMNTM、TDLM、TCNLM
互信息	GLDA、LCTM、MvTM、WEI-FTM、LFTM、GPU-DMM、GPU-PDMM、DGPU-LDA、KGE-LDA、Miao et al.、AVITM、NSTC、NSMTM、SCHOLAR、SLRTM、TDLM、TCNLM、STE
分类准确率	LFTM、MvTM、GPU-DMM、GPU-PDMM、DGPU-LDA、Event-BTM-GPU、KGE-LDA、NTM、SCHOLAR、NSTC、SLRTM、GMNTM、TDLM、TWE、TopicVec、STE
主题词直观展示	GLDA、MvTM、He et al.、DGPU-LDA、Event-BTM-GPU、KGE-LDA、NTM、AVITM、NSTC、SLRTM、GMNTM、TopicRNN、TDLM、TCNLM、TWE

除第 6.2 节介绍的主流评价方式外,还有其它评价方式,如从聚类角度评价模型的聚类纯度[76]、从稀疏性角度评价模型的主题-词汇分布稀疏度[42-43,52,93]、从相似词展示角度评价模型的词向量刻画能力[8,45]等.总之,一般情况下,研究者在现有评价指标基础上根据主题模型主要任务的不同,选择最能体现模型优势的评价指标实施评价度量.

7　基于深度学习的主题模型研究趋势展望

主题模型自提出以来尽管已发展二十余年,但随着深度学习在自然语言处理领域的普及应用,利用深度学习思想和方法建立更加高效精确的主题模型仍然是一个非常有吸引力的研究方向.本文首先介绍了主题建模的基本框架,并对比介绍了传统主题模型中几个经典模型在基本假设、目标函数和模型参数求解等方面的异同点.随后,本文重点讨论了近几年利用深度学习,特别是词向量技术的主题模型的发展现状.最后,本文整理归纳了当前基于深度学习的主题模型所采用的公开评测语料与评价指

① The source code and dataset is available at https://github.com/sb1992/Topic-Intrusion-for-Automatic-Topic-Model-Evaluation

标,并对比各模型所使用语料及评测指标.

经过前几章的技术梳理发现,基于深度学习的主题模型得到蓬勃发展,现有模型正向更加智能、精确和通用的方向发展,这体现在:(1)将词向量、文档向量等引入主题模型中度量词汇(或文档)之间的语义相似性有助于产生更具语义一致性的主题;(2)采用神经网络对句子(或文档)序列编码而非直接使用词袋子模型作为主题模型输入,使得模型在文档生成方面能够产生更加符合自然语言特点的句子和文本;(3)结合深度学习的主题模型不仅可用于主题建模、文本分类和信息检索等领域,主题模型一定程度也与语言模型通用,使得主题模型具有更加广泛的应用价值.

综上所述,各类主题模型的成功主要得益于传统主题模型完备的理论基础与深度神经网络在处理复杂、非线性任务上的显著优势.本质上,现有主题模型还不具备产生符合人类期望的自然文本或特点主题下的自然文本,未来主题模型可能在如下几个方向有进一步发展:

(1)在开放、非规范文本中应用.微博、Twitter、问答对话等产生大量口语化、弱规范、高噪声的短文本,而这些具有实时性、规模大的非规范文本具有研究和应用价值,被广泛应用于情感分析、社交网络分析、事件监测、自动问答等任务.目前主题模型技术主要在 20NewsGroups 等规范语料上有不错的性能,但直接应用于开放非规范短文本将不可避免地导致低性能问题.因此,如何针对开放地非规范文本也能产生令人满意的结果还值得进一步探索.

(2)融合高质量知识的主题模型.目前,词向量与传统概率主题模型及神经网络主题模型有机结合,可产生更优越的主题质量和分类准确率.然而,词向量往往只能表示词汇之间的语义相似度或潜在概念之间的距离,缺乏对词汇/概念之间关系的表示与推理.而基于表示学习的知识图谱能够表达实体之间更丰富的语义关系进而实现实体链接预测与推理,并在信息检索[111]、对话系统[112]和推荐系统[113]等任务中表现出优越的性能.此外,已有模型[40]开始采用知识图谱中的实体信息提升主题模型能力.因此,能否借鉴知识图谱已有研究成果,在主题建模过程中融合高质量的先验知识、丰富文档语义信息,使得模型的语义理解能力进一步加强,将是未来值得探讨的一个研究问题.

(3)融合句子/文本序列建模.主题建模是建立在文档级别的语义分析,当前主流主题模型的输入依然是文档/句子的词袋子,即模型生成文档的词汇分布,而非完整的句子.然而,以文档序列而非词袋子作为主题模型输入更符合人类对篇章语义的理解.目前,NTM[18]、NSTC[42]、SLRTM* 等结合神经网络的主题模型已显示出一定的优越性,但总体来说,生成的文本尚不能达到人工撰写的水平,特别是对于篇幅较长的文档.另一方面,Seq2Seq[35]模型可将输入句子转换为输出句子,因而在机器翻译任务[35]、文本生成式摘要[89]、会话建模[99]等句子生成任务中体现出较大优势,但是该模型一般无法捕获主题信息,且针对较长文本效果不理想.因此,如何将主题模型与 Seq2Seq 等序列生成模型有机融合,生成语义完整、句法正确的句子或文档将会在未来成为一个研究热点.

(4)结合生成对抗网络[114](Generative Adversarial Networks,GAN)等实现自动作文、人机自然语言对话和多文档自动摘要等任务.目前,在主题-语言联合训练模型中,神经语言模型一般直接使用深度神经网络训练模型参数而不关心数据分布,而传统概率主题模型所使用的先验分布虽能较好刻画全局主题信息,但无法生成自然语句.GAN 作为一个生成模型与判别模型的联合训练模型,既可判别数据类别,也可推断数据分布.GAN 在图像生成领域获得广泛关注,并可根据文本描述生成指定图像[115].近年来,GAN 也开始尝试应用于单文本摘要[116]、人机对话[117]等句子级别的文本生成任务中.但针对包含特定主题的文本生成任务,如多文档摘要或长句子对话,目前尚无相关研究.因此,借鉴 VAE 在神经主题模型中的应用思想,能否使用生成对抗网络来从深度神经网络中推断文本的潜在分布,从而使得模型能够在全局主题信息捕获和自然文本生成方法都表现优良,这也是一个值得探讨的问题.

8 结束语

总的来说,主题模型虽然还有若干问题有待解决,但是这不影响它在自然语言处理领域的进一步发展与应用,它在未来很长时间内仍然会是一个研究热点.新的理论、技术和方法的纳入,特别是自然语言处理的新成果(如词向量模型、篇章分布式表示等)和深度学习的新模型(如 Seq2Seq、VAE、GAN)不断涌现也能使其得到进一步发展.

参 考 文 献

[1] Li C, Wang H, Zhang Z, et al. Topic modeling for short texts with auxiliary word embeddings//Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval. Pisa, Italy, 2016: 165-174

[2] He Xu-Feng, Chen Ling, Chen Gen-Cai, et al. A LDA topic model based collection selection method for distributed information retrieval. Journal of Chinese Information Processing, 2017, 31(3): 125-133(in Chinese)
(何旭峰，陈岭，陈根才等. 基于 LDA 主题模型的分布式信息检索集合选择方法. 中文信息学报，2017，31(3)：125-133)

[3] Huang J, Peng M, Wang H, et al. A probabilistic method for emerging topic tracking in microblog stream. World Wide Web-Internet & Web Information Systems, 2017, 20(2): 325-350

[4] Yang G. A Novel contextual topic model for query-focused multi-document summarization//Proceedings of the International Conference on Tools with Artificial Intelligence. Limassol, Cyprus, 2014: 576-583

[5] Dermouche M, Velcin J, Khouas L, et al. A joint model for topic-sentiment evolution over time//Proceedings of the IEEE International Conference on Data Mining. Atlantic City, USA, 2015: 773-778

[6] Jiang S, Qian X, Shen J, et al. Author topic model-based collaborative filtering for personalized POI recommendations. IEEE Transactions on Multimedia, 2015, 17(6): 907-918

[7] Rani M, Dhar A K, Vyas O P. Semi-automatic terminology ontology learning based on topic modeling. Engineering Applications of Artificial Intelligence, 2017, 63: 108-125

[8] Shi B, Lam W, Jameel S, et al. Jointly learning word embeddings and latent topics//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. Shinjuku, Tokyo, Japan, 2017: 375-384

[9] Du Hui, Chen Yun-Fang, Zhang Wei, et al. Survey for method of parameter estimation in topic models. Computer Science, 2017, 44(6a): 29-32(in Chinese)
(杜慧，陈云芳，张伟等. 主题模型中的参数估计方法综述. 计算机科学，2017，44(6a)：29-32)

[10] Sun Guo-Chao, Xu Shuo, Qiao Xiao-Dong. Review on visualization of topic models. Technology Intelligence Engineering, 2015, 1(6): 51-61(in Chinese)
(孙国超，徐硕，乔晓东. 主题模型可视化研究综述. 情报工程，2015，1(6)：51-61)

[11] Chen Jing, Liu Yan, Wang Xu-Zhong. Research on application of probability topic model in microblog topic mining. Journal of Information Engineering University, 2017, 18(1): 103-110(in Chinese)
(陈静，刘琰，王煦中. 主题概率模型在微博主题挖掘方面的研究综述. 信息工程大学学报，2017，18(1)：103-110)

[12] Alghamdi R, Alfalqi K. A survey of topic modeling in text mining. International Journal of Advanced Computer Science & Applications, 2015, 6(1): 147-153

[13] Gui Xiao-Qing, Zhang Jun, Zhang Xiao-Min, et al. Survey on temporal topic model methods and application. Computer Science, 2017, 44(2): 46-55(in Chinese)
(桂小庆，张俊，张晓民等. 时态主题模型方法及应用研究综述. 计算机科学，2017，44(2)：46-55)

[14] Sun X, Liu X, Li B, et al. Exploring topic models in software engineering data analysis: A survey//Proceedings of the IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. Shanghai, China, 2016: 357-362

[15] Boyd-Graber J, Hu Y, Mimno D. Applications of topic models. Foundations and Trends in Information Retrieval, 2017, 11(2/3): 143-296

[16] Sharma D. A survey on journey of topic modeling techniques from SVD to deep learning. International Journal of Modern Education and Computer Science, 2017, 9(7): 50-62

[17] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. Journal of Machine Learning Research, 2012, 3: 993-1022

[18] Cao Z, Li S, Liu Y, et al. A novel neural topic model and its supervised extension//Proceedings of the National Conference on Artificial Intelligence. Austin, USA, 2015: 2210-2216

[19] Lim K W, Buntine W. Bibliographic analysis on research publications using authors, categorical labels and the citation network. Machine Learning, 2016, 103(2): 185-213

[20] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 2001, 42(1-2): 177-196

[21] Carter C K, Kohn R. On Gibbs sampling for state space models. Biometrika, 1994, 81(3): 541-553

[22] Blei D M, Lafferty J D. Correlated topic models//Proceedings of the International Conference on Neural Information Processing Systems. Vancouver, Canada, 2005: 147-154

[23] Yin J, Wang J. A Dirichlet multinomial mixture model-based approach for short text clustering//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA, 2014: 233-242

[24] Gerlach M, Peixoto T P, Altmann E G, et al. A network approach to topic models. Science Advances, 2018, 4(7): 1-11

[25] Blei D M, Mcauliffe J. Supervised topic models//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2007: 121-128

[26] Huang M, Rao Y, Liu Y, et al. Siamese network-based supervised topic modeling//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 4652-4662

[27] Ren Y, Wang Y, Zhu J, et al. Spectral learning for supervised topic models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(3): 726-739

[28] Wang Yan-Peng. The research progress of topic model based on bibliometrics. Science Focus, 2017, 12(5): 9-20 (in Chinese)
(王燕鹏. 基于文献计量的主题模型研究进展分析. 科学观察, 2017, 12(5): 9-20)

[29] Nigam K, Mccallum A K, Thrun S, et al. Text classification from labeled and unlabeled documents using EM. Machine Learning, 2000, 39(2/3): 103-134

[30] Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts//Proceedings of the International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 1445-1456

[31] Harris Z S. Distributional structure. Word, 1981, 10(2/3): 146-162

[32] Yao Y, Huang Z. Bi-directional LSTM recurrent neural network for Chinese word segmentation//Proceedings of the International Conference on Neural Information Processing. Kyoto, Japan, 2016: 345-353

[33] Miwa M, Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 1105-1116

[34] Lin C Y, Xue N, Zhao D, et al. A convolution BiLSTM neural network model for Chinese event extraction//Proceedings of the National CCF Conference on Natural Language Processing and Chinese Computing. Kunming, China, 2016: 275-287

[35] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Computer Science, 2014: 1724-1734

[36] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 6000-6010

[37] Zhang P, Niu J, Su Z, et al. End-to-end quantum-like language models with application to question answering//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 5666-5673

[38] Das R, Zaheer M, Dyer C. Gaussian LDA for topic models with word embeddings//Proceedings of the Annual Meeting of the Association for Computational Linguistics and the Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing, China, 2015: 795-804

[39] Zhao H, Du L, Buntine W. A word embeddings informed focused topic model//Proceedings of the Asian Conference on Machine Learning Seoul. Korea, 2017: 423-438

[40] Yao L, Zhang Y, Wei B, et al. Incorporating knowledge graph embeddings into topic modeling//Proceedings of the National Conference on Artificial Intelligence. San Francisco, USA, 2017: 3119-3126

[41] Srivastava A, Sutton C A. Autoencoding variational Inference for topic models//Proceedings of the International Conference on Learning Representations. Toulon, France, 2017: 1-12

[42] Peng M, Xie Q, Zhang Y, et al. Neural sparse topical coding//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 2332-2340

[43] Card D, Tan C, Smith N A, et al. Neural models for documents with metadata//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Iryna Gurevych, Yusuke Miyao, 2018: 2031-2040

[44] Yang M, Cui T, Tu W, et al. Ordering-sensitive and semantic-aware topic modeling//Proceedings of the National Conference on Artificial Intelligence. Austin, USA, 2015: 2353-2360

[45] Liu Y, Liu Z, Chua T, et al. Topical word embeddings//Proceedings of the National Conference on Artificial Intelligence. Austin, USA, 2015: 2418-2424

[46] Li S, Chua T, Zhu J, et al. Generative topic embedding: A continuous representation of documents//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 666-675

[47] Griffiths T L, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences, 2004, 101(Suppl 1): 5228-5235

[48] Porteous I, Newman D, Ihler A T, et al. Fast collapsed Gibbs sampling for latent Dirichlet allocation//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 569-577

[49] Chen J, Zhu J, Wang Z, et al. Scalable inference for logistic-normal topic models//Proceedings of the International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 2445-2453

[50] Banerjee A, Basu S. Topic models over text streams: A study of batch and online unsupervised learning//Proceedings of the SIAM Conference on Data Mining. Minneapolis, USA, 2007, 7: 437-442

[51] Cheng X, Yan X, Lan Y, et al. BTM: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12): 2928-2941

[52] Zhu J, Xing E P. Sparse topical coding//Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence. Barcelona, Spain, 2011: 831-838

[53] Eisenstein J, Ahmed A, Xing E P, et al. Sparse additive generative models of text//Proceedings of the International Conference on Machine Learning. Bellevue, USA, 2011: 1041-1048

[54] Wang C, Blei D M. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process//Proceedings of the Conference on Neural Information Processing Systems. Vancouver, Canada, 2009: 1982-1989

[55] Zhang A, Zhu J, Zhang B. Sparse online topic models//Proceedings of the International World Wide Web Conference. Rio de Janeiro, Brazil, 2013: 1489-1500

[56] Lin T, Tian W, Mei Q, et al. The dual-sparse topic model: Mining focused topics and focused terms in short text//Proceedings of the International World Wide Web Conference. Seoul, Korea, 2014: 539-550

[57] Peng M, Xie Q, Huang J, et al. Sparse topical coding with sparse groups//Proceedings of the International Conference on Web Age Information Management. Nanchang, China, 2016: 415-426

[58] Williamson S, Wang C, Heller K, et al. Focused topic models//Proceedings of the NIPS Workshop on Applications for Topic Models: Text and Beyond. Vancouver, Canada, 2009: 1-4

[59] Williamson S, Wang C, Heller K A, et al. The IBP compound Dirichlet process and its application to focused topic modeling//Proceedings of the International Conference on Machine Learning. Haifa, Israel, 2010: 1151-1158

[60] Teh Y W, Jordan M I, Beal M J, et al. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 2006, 101(476): 1566-1581

[61] Ghahramani Z, Griffiths T L. Infinite latent feature models and the Indian buffet process//Proceedings of the International Conference on Neural Information Processing Systems. British, Canada, 2005: 475-482

[62] Quan X, Kit C, Ge Y, et al. Short and sparse text topic modeling via self-aggregation//Proceedings of the International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 2270-2276

[63] Zuo Y, Wu J, Zhang H, et al. Topic modeling of short texts: A pseudo-document view//Proceedings of the International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 2105-2114

[64] Chen Z, Mukherjee A, Liu B, et al. Leveraging multi-domain prior knowledge in topic models//Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 2071-2077

[65] Chen Z, Liu B. Mining topics in documents: Standing on the shoulders of big data//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 1116-1125

[66] Silver D L, Yang Q, Li L, et al. Lifelong machine learning systems: Beyond learning algorithms//Proceedings of the Association for the Advancement of Artificial Intelligence. Palo Alto, USA, 2013: 49-55

[67] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3(6): 1137-1155

[68] Yu Ke-Ren, Fu Yun-Bin, Dong Qi-Wen. Survey on distributed word embeddings based on neural network language models. Journal of East China Normal University (Natural Science), 2017, (5): 52-65(in Chinese)
(郁可人，傅云斌，董启文. 基于神经网络语言模型的分布式词向量研究进展. 华东师范大学学报(自然科学版)，2017，(5)：52-65)

[69] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. Advanced in Neural Information Processing Systems, 2013, 26: 3111-3119

[70] Pennington J, Socher R, Manning C D, et al. GloVe: Global vectors for word representation//Proceedings of the Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1532-1543

[71] Hu W, Tsujii J. A latent concept topic model for robust topic inference using word embeddings//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 380-386

[72] Li X, Chi J, Li C, et al. Integrating topic modeling with word embeddings by mixtures of vMFs//Proceedings of the International Conference on Computational Linguistics. Osaka, Japan, 2016: 151-160

[73] Gopal S, Yang Y. Von mises-fisher clustering models//Proceedings of the International Conference on Machine Learning. Beijing, China, 2014: 154-162

[74] He J, Hu Z, Bergkirkpatrick T, et al. Efficient correlated topic modeling with topic embedding//Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 225-233

[75] Xun G, Li Y, Zhao W X, et al. A correlated topic model using word embeddings//Proceedings of the International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 4207-4213

[76] Nguyen D Q, Billingsley R, Du L, et al. Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics, 2015: 299-313

[77] Liu Liang-Xuan, Huang Meng-Xing. Biterm topic model with word vector features. Application Research of Computers, 2017, 34(7): 2055-2058(in Chinese)
(刘良选，黄梦醒. 融合词向量特征的双词主题模型. 计算机应用研究，2017，34(7)：2055-2058)

[78] Li C, Duan Y, Wang H, et al. Enhancing topic modeling for short texts with auxiliary word embeddings. ACM Transactions on Information Systems, 2017, 36(2): 1-30

[79] Peng Min, Yang Shao-Xiong, Zhu Jia-Hui. Semantic enhanced topic modeling by bi-directional LSTM. Journal of Chinese Information Processing, 2018, 32(4): 40-49 (in Chinese)
(彭敏，杨绍雄，朱佳晖. 基于双向 LSTM 语义强化的主题建模. 中文信息学报，2018，32(4)：40-49)

[80] Sun Rui, Guo Sheng, Ji Dong-Hong. Topic representation integrated with event knowledge. Chinese Journal of Computers, 2017, 40(4): 791-804(in Chinese)
(孙锐，郭晟，姬东鸿. 融入事件知识的主题表示方法. 计算机学报，2017，40(4)：791-804)

[81] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data. Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2013: 2787-2795

[82] Kingma D P, Welling M. Auto-encoding variational Bayes// Proceedings of the International Conference on Learning Representations. Banff, Canada, 2014: 1-14

[83] Zheng Y, Zhang Y, Larochelle H, et al. A deep and autoregressive approach for topic modeling of multimodal data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(6): 1056-1069

[84] Hinton G E, Osindero S, Teh Y W, et al. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527-1554

[85] Wan L, Zhu L, Fergus R, et al. A hybrid neural network-latent topic model. Journal of Machine Learning Research, 2012: 1287-1294

[86] Larochelle H, Lauly S. A neural autoregressive topic model// Proceedings of the International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2012: 2708-2716

[87] Miao Y, Yu L, Blunsom P. Neural variational inference for text processing//Proceedings of the International Conference on Machine Learning. New York, USA, 2016: 1727-1736

[88] Mnih A, Gregor K. Neural variational inference and learning in belief networks//Proceedings of the International Conference on Machine Learning. Beijing, China, 2014: 1791-1799

[89] Li P, Lam W, Bing L, et al. Deep recurrent generative decoder for abstractive text summarization//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 2091-2100

[90] Miao Y, Grefenstette E, Blunsom P, et al. Discovering discrete latent topics with neural variational inference//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 2410-2419

[91] Ding R, Nallapati R, Xiang B, et al. Coherence-aware neural topic modeling//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 830-836

[92] Kucukelbir A, Tran D, Ranganath R, et al. Automatic differentiation variational inference. Journal of Machine Learning Research, 2017, 18(1): 430-474

[93] Lin T, Hu Z, Guo X, et al. Sparsemax and relaxed Wasserstein for topic sparsity//Proceedings of the ACM International Conference on Web Search and Data Mining. Melbourne, Australia, 2019: 141-149

[94] Martins A F, Astudillo R F. From softmax to sparsemax: A sparse model of attention and multi-label classification// Proceedings of the International Conference on Machine Learning. New York, USA, 2016: 1614-1623

[95] Dieng A B, Wang C, Gao J, et al. TopicRNN: A recurrent neural network with long-range semantic dependency// Proceedings of the International Conference on Learning Representations. Toulon, France, 2017: 1-13

[96] Mikolov T, Karafiat M, Burget L, et al. Recurrent neural network based language model//Proceedings of the Conference of the International Speech Communication Association. Makuhari, Japan, 2010: 1045-1048

[97] Sundermeyer M, Schluter R, Ney H, et al. LSTM neural networks for language modeling//Proceedings of the Conference of the International Speech Communication Association. Portland, USA, 2012: 194-197

[98] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation//Proceedings of the Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1724-1734

[99] Hou Y, Liu Y, Che W, et al. Sequence-to-sequence data augmentation for dialogue language understanding//Proceedings of the International Conference on Computational Linguistics. Santa Fe, USA, 2018: 1234-1245

[100] Li S, Zhu J, Miao C, et al. A generative word embedding model and its low rank positive semidefinite solution. Empirical Methods in Natural Language Processing, 2015: 1599-1609

[101] Jiang D, Shi L, Lian R, et al. Latent topic embedding// Proceedings of the International Conference on Computational Linguistics. Osaka, Japan, 2016: 2689-2698

[102] Lau J H, Baldwin T, Cohn T. Topically driven neural language model//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 355-365

[103] Wang W, Gan Z, Wang W, et al. Topic compositional neural language model//Proceedings of the International Conference on Artificial Intelligence and Statistics. Lanzarote, Spain, 2018: 356-365

[104] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3156-3164

[105] Newman D, Lau J H, Grieser K, et al. Automatic evaluation of topic coherence//Proceedings of the North American Chapter of the Association for Computational Linguistics. Los Angeles, USA, 2010: 100-108

[106] Bouma G. Normalized (pointwise) mutual information in collocation extraction//Proceedings of the Biennial GSCL Conference. Tubingen, Germany, 2009: 31-40

[107] Mimno D M, Wallach H M, Talley E M, et al. Optimizing semantic coherence in topic models//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011: 262-272

[108] Fang A, Macdonald C, Ounis I, et al. Using word embedding to evaluate the coherence of topics from Twitter data// Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy, 2016: 1057-1060

[109] Lewis D D, Yang Y, Rose T G, Li F. RCV1: A new benchmark collection for text categorization research. The Journal of Machine Learning Research, 2004, 5: 361-397

[110] Bhatia S, Lau J H, Baldwin T, et al. Topic intrusion for automatic topic model evaluation//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 844-849

[111] Zheng W, Zou L, Peng W, et al. Semantic SPARQL similarity search over RDF knowledge graphs. Proceedings of the VLDB Endowment, 2016, 9(11): 840-851

[112] Chen Y N, Wang W Y, Gershman A, et al. Matrix factorization with knowledge graph propagation for unsupervised spoken language understanding//Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing. Beijing, China, 2015: 483-494

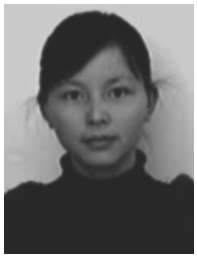
[113] Yang D, He J, Qin H, et al. A graph-based recommendation across heterogeneous domains//Proceedings of the Conference on Information and Knowledge Management. Melbourne, Australia, 2015: 463-472

[114] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Advances in Neural Information Processing Systems, 2014, 3: 2672-2680

[115] Zhang H, Xu T, Li H, et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 5908-5916

[116] Liu L, Lu Y, Yang M, et al. Generative adversarial network for abstractive text summarization//Proceedings of the National Conference on Artificial Intelligence. San Francisco, USA, 2017: 8109-8110

[117] Li J, Monroe W, Shi T, et al. Adversarial learning for neural dialogue generation//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 2157-2169



HUANG Jia-Jia, Ph. D., lecturer. Her main research interests focus on natural language processing and audit data analysis.

LI Peng-Wei, Ph. D., lecturer. His main research interests focus on software security and data analysis.

PENG Min, Ph. D., professor. Her main research interests focus on natural language processing and information retrieval.

XIE Qian-Qian, Ph. D. candidate. Her main research interests focus on natural language processing and information retrieval.

XU Chao, Ph. D., professor. His main research interests focus on audit data analysis.

Background

As a basic research issue in natural language processing and information retrieval, topic model plays an important role in various semantic analysis tasks of multi-documents, such as text classification, sentiment analysis, multi-document summarization, event detection and so on. Generally, topic model aims at extracting groups of keywords from document to express its main opinions. Since Blei et al. proposed LDA model, probabilistic topic model based on three-layers Bayesian network has gained wide attention in the past ten years. These researches focus on model assumption, parameter inference, topic sparsity and various applications

With the rapid development of deep learning, designing novel topic model based on neural network grows new lease of life in recent years. First of all, novel topic models that combining word embeddings into conventional topic model obtain significant improvement compared with their baseline models. Then, designing neural topic model equipped with neural network structure has shown new perspective of topic inference process. Last but not least, jointly training topic and language model is feasible and promising because topic model is also a document generative model. In this review, we will summarize and discuss the research progress of topic models and give an overview of the state-of-the-art methods from the above three aspects in detail.

Thanks to the support of Project Event Extraction and Evolution based on Representation Learning (Grant No. 61802194). This project aims at extracting events from user-generated-contents and detecting their evolution process automatically from granularities of meta-event element, element relation and topic. As an important part of this project, this work tries to investigate the progress of topic model based on representation learning and neural network. Furthermore, this work is also supported in part by College Natural Science Project of Jiangsu Province (Grant No. 17KJB520015).

Our research group mainly paid attention to topic model and event extraction and detection from user-generated-content in past years. More details can be found on the author's publications.