文章编号：1003-0077(2021)02-0089-10

基于神经自回归分布估计的涉案新闻主题模型构建方法

毛存礼[1,2]，梁昊远[1,2]，余正涛[1,2]，郭军军[1,2]，黄于欣[1,2]，高盛祥[1,2]

（1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 云南省人工智能重点实验室，云南 昆明 650500）

摘 要：神经主题模型能有效获取文本的深层语义特征，但现有的神经主题模型忽略了外部知识对获取主题分布的帮助。因此，针对涉案主题分析任务，该文提出了一种基于神经自回归分布估计的涉案新闻主题模型构建方法。以案件要素作为外部知识对 iDocNADEe 模型进行了扩展，通过计算案件要素与主题词的相关度来构建注意力机制对 iDocNADEe 模型双向编码的隐状态进行加权，利用神经自回归算法计算加权后的主题词双向隐状态的自回归条件概率实现涉案新闻文本主题模型构建。实验结果表明，该文提出方法较基线模型困惑度降低了 0.66%、主题连贯性提高了 6.26%，并且在文档检索精确率方面也明显高于基线模型。

关键词：案件要素；iDocNADEe；注意力机制；神经自回归分布估计；涉案新闻；主题模型

中图分类号：TP391 **文献标识码**：A

Topic Model of Judicial News Based on Neural Autoregressive Distribution Estimator

MAO Cunli[1,2], LIANG Haoyuan[1,2], YU Zhengtao[1,2], GUO Junjun[1,2],
HUANG Yuxin[1,2], GAO Shengxiang[1,2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: The neural topic models can effectively obtain the deep semantic features of the text, but the existing topic models are defected in negligence of the contextual information and the external knowledge. This paper proposes a topic model of judicial news based on neural autoregressive distribution estimator. The iDocNADEe is expanded with case elements as external knowledge, the attention mechanism is constructed by calculating the correlation between case elements and topic-relevant words to adjust weights of the hidden states in iDocNADEe. Then, the neural autoregressive algorithm is applied to calculate the weighted autoregressive conditional probability of the bidirectional hidden state of topic-related words. Experimental results show that compared with the baseline model, the perplexity is reduced by 0.66%, and the topic coherence is improved by 6.26% with the proposed method, as well as a significant higher document retrieval accuracy.

Keywords: case elements; iDocNADEe; attention mechanism; neural autoregressive distribution estimator; news involved in the case; topic model

0 引言

涉案新闻是指与司法案件相关的新闻，准确抽取涉案新闻主题信息对进一步开展涉案新闻检索、涉案新闻事件分析等研究具有重要价值。传统主题模型主要考虑词频统计特征，而忽略了文档中的词语出现的次序及上下文信息[1-2]。例如，“窃取”一

收稿日期：2020-06-12 **定稿日期**：2020-07-21

基金项目：国家重点研发计划(2018YFC0830105, 2018YFC0830101, 2018YFC0830100)；云南省应用基础研究计划重点项目(2019FA023)；云南省中青年学术和技术带头人后备人才项目(2019HB006)；云南省高新技术产业专项(201606)；云南省重大科技专项计划项目(202002AD080001)

词,可以是窃取个人财产,也可以是窃取国家机密,但是案件性质完全不同,前者涉及盗窃罪,而后者则是触及了非法获取国家秘密罪。神经主题模型由于能够获得文本的深层语义信息,既可以捕获文中词汇之间的局部依赖关系,又可以利用潜在主题捕获全局语义信息,有效弥补传统主题模型的缺陷,近年来在文本检索、文本分类、文本摘要等自然语言处理任务中表现出较好的效果[3-11]。

案件要素是案件关键信息的体现,涉案新闻与普通新闻的根本区别在于是否出现案件要素。对涉案新闻主题分析的核心是对案件要素相关词汇的主题进行预测,故可以将案件要素作为涉案领域知识来捕获文本中涉案相关词语的主题分布和文本的主题表征。然而,现有的神经主题模型忽略了领域知识对特定领域主题分析任务的作用。为此,本文针对涉案新闻主题分析任务,提出一种基于神经自回归分布估计的涉案新闻主题模型构建方法,是对Gupta等人[3]提出的神经主题模型(document informed neural autoregressive topic models with distributional prior,iDocNADEe)的进一步扩展,把案件要素作为涉案领域知识,通过计算案件要素与涉案新闻文本中主题词的相关度构建注意力机制,对文本经过双向语言模型编码后的前后向隐状态进行加权,以此增强涉案新闻文本中主题词的涉案语义特征表示,最后利用神经自回归算法计算加权后双向隐状态的自回归条件概率实现涉案新闻文本主题抽取。在构建的涉案新闻数据集上进行了实验,结果表明,所提方法较iDocNADEe模型困惑度降低了0.66%、主题连贯性提高了6.26%;将构建的涉案新闻主题模型用于涉案新闻检索对比实验,准确率也明显高于基线模型。

本文第1节介绍主题模型的相关工作;第2节描述基于神经自回归分布估计的涉案新闻主题模型构建方法;第3节通过实验对比了所提方法在主题构建及文本检索方面的优势;第4节进行总结并提出未来的研究方向。

1 相关工作

随着狄利克雷多项式混合模型(dirichlet multinomial mixture, DMM)[1]、潜在狄利克雷分布(latent dirichlet allocation,LDA)[2]等概率主题模型的广泛应用,越来越多的研究聚焦于如何将主题模型应用于各类特定领域的自然语言处理任务。如张绍武等人[12]基于一种动态主题模型实现了新疆暴恐舆情分析;吴彦文等人[13]基于LDA模型与长短期记忆网络(long short-term memory,LSTM)模型实现了短文本情感分类;陈琪等人[14]基于支持向量机和LDA模型提出了一种评论分析方法。上述方法普遍基于早期的概率主题模型,而这些概率主题模型存在泛化能力弱、主题可解释性差等缺陷。基于神经网络的方法来构建主题模型能有效解决这些问题。Cao等人[4]提出了基于前馈神经网络的主题模型(neural topic model,NTM),该模型将传统主题模型的主题—词分布以及主题—文档分布转换为两个权重矩阵,并使用了后向传播(back propagation,BP)算法训练参数。Kingma等人[5]在2014年提出变分自编码器(variational auto-encoder, VAE),能够训练一个直接将文档映射到后验分布的神经网络。因此,Miao等人[6]使用VAE构建了一种神经变分文档模型(neural variational document model, NVDM),并在此基础上加入主题—词分布,进而形成了基于VAE的主题模型。基于Larochelle等人[7]提出的神经自回归分布估计器(neural autoregressive distribution estimator, NADE),Lauly等人[8]提出了一种生成式主题模型——文档的神经自回归分布估计(document neural autoregressive distribution estimator, DocNADE),通过词的序列学习主题,即对某个词 v_i 进行预测时,需要其前文作为输入。对比于概率主题模型,神经主题模型能够更好地利用词汇之间的语义相似度。随着循环神经网络的发展,文档的词序列作为输入能够更加充分利用词汇的上下文信息。Dieng等人[9]利用循环神经网络(recurrent neural network, RNN)捕获词之间的依赖,结合循环神经网络与主题模型提出了TopicRNN。Lau等人[10]利用卷积神经网络(convolutional neural network, CNN)和LSTM提出了主题-语言模型联合训练模型(topically driven language model,TDLM),利用CNN提取文本特征,并使用LSTM刻画词汇之间的语义,将文本的主题信息与LSTM的隐藏层结合。这两种模型可认为是多任务学习模型,由主题推断和文本生成两个子任务组成,由此模型生成的文本语义更加自然,但这些方法更加侧重于对语言模型的优化。

Gupta等人[11]利用LSTM语言模型,经过训练后能够根据给定词序列来预测后续单词的特性,提

出了基于词嵌入的语境化文档神经自回归分布估计器(contextualized document neural autoregressive distribution estimator with embeddings，ctx-DocNADEe)，但并没有考虑到文档的双向语义。而 Gupta 等人[3]受双向语言模型[15]和递归神经网络[16-17]的启发，提出了一种纳入完整上下文语义信息的主题模型(document informed neural autoregressive distribution estimator with embeddings，iDocNADEe)，该模型将上下文同时作为输入，并引入 Glove 词嵌入作为先验知识，将语言模型的预测方式应用到了主题模型。通过对以上工作的分析可以看出，无论是传统主题模型还是神经主题模型，都是基于通用领域，而在涉案领域暂无相关研究。因此，我们考虑如何将这些方法应用到涉案领域以获得更好的主题表示。

2　基于神经自回归分布估计的涉案新闻主题模型

2.1　涉案新闻案件要素库构建

案件要素是指案件的内在组成部分及各部分之间的相互关系和排列状况，如刑事案件由何事、何时、何地、何物、何情、何故、何人等 7 要素构成，对案件构成要素进行分析能够从根本上把握案件发生、发展的趋势和规律[18]。对于涉案新闻主题抽取任务，分析涉案文本与案件要素之间的关联关系有助于提高涉案主题分布的准确性。为构建案件要素库，我们从互联网中收集了有关重庆公交坠江案、丽江唐雪反杀案、昆明孙小果涉黑案等刑事案件的相关新闻，并基于韩鹏宇等人[19]的方法定义且抽取了涉案新闻中的案件要素，包括“案件名称、涉案人员、涉案地点、涉案触发词”，为涉案新闻主题建模提供了领域知识。以涉案新闻“还女司机清白！重庆万州公交坠江系乘客殴打公交车司机导致!”为例，其案件要素构成如表 1 所示。

表 1　案件要素实例

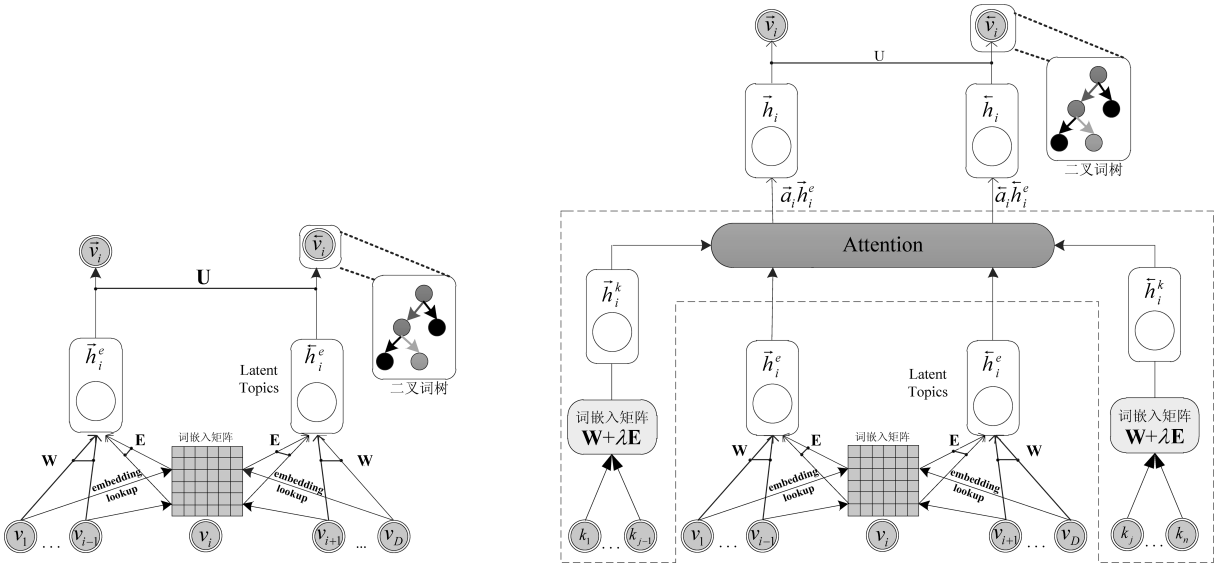
案件要素	实例
案件名称	重庆公交坠江案
涉案地点	重庆、万州
涉案人员	女司机、乘客、司机
涉案触发词	殴打、坠江

其中，涉案地点可能是案件中涉及到的地名，可能是省份、城市或更加具体的场所。涉案人员包括案件中涉及的人员，如嫌疑人、受害人、目击者等。涉案触发词指某些与司法领域相关或描述案件关键的词，如表 1 中的“殴打”及“坠江”。

2.2　融合案件要素的涉案新闻主题建模

2.2.1　基于神经自回归分布估计的主题模型

作为一种无监督的生成式主题模型，iDocNADEe 的结构如图 1(a)所示，该模型从文档中抽取其潜在特征，并据此重新生成文本，以生成文本的对数似然函数为最终的优化目标。



(a) iDocNADEe　　(b) 本文模型

图 1　iDocNADEe 和本文模型架构，虚线框内代表本文所加入的案件要素注意力机制

首先，将一篇词数为 D 的文档表示为一个序列 $v=[v_1,v_2,\cdots,v_D]$，其中 $v_i\in\{1,\cdots,V\}$ 表示文档中第 i 个词在词表中的位置，V 表示语料库词表的大小。

基于 iDocNADEe 模型，文档的每个词汇 v_i 都有两个包含了上下文信息的隐状态，分别是前向隐状态 $\overrightarrow{h}_i^e$ 以及后向隐状态 $\overleftarrow{h}_i^e$，这两个隐状态由 v_i 的上下文信息 $v_{<i}=[v_1,\cdots,v_{i-1}]$ 与 $v_{>i}=[v_{i+1},\cdots,v_D]$ 以及引入预训练的词向量作为先验知识计算得到，即 $[\overrightarrow{h}_i^e,\overleftarrow{h}_i^e]$ 包含了 v_i 的完整上下文信息，如式(1)、式(2)所示。

$$\overrightarrow{h}_i^e(v_{<i})=g(\overrightarrow{c}+\sum_{j<i}\boldsymbol{W}_{:,v_j}+\lambda\sum_{j<i}\boldsymbol{E}_{:,v_j}) \quad (1)$$

$$\overleftarrow{h}_i^e(v_{>i})=g(\overleftarrow{c}+\sum_{j>i}\boldsymbol{W}_{:,v_j}+\lambda\sum_{j>i}\boldsymbol{E}_{:,v_j}) \quad (2)$$

其中，$g(\cdot)$ 代表任意非线性激活函数，$\overrightarrow{c}\in\mathbb{R}^H$，$\overleftarrow{c}\in\mathbb{R}^H$ 为偏置向量，H 表示隐藏层大小，即主题数量。$\boldsymbol{W}\in\mathbb{R}^{H\times V}$ 为参数矩阵，$\boldsymbol{E}\in\mathbb{R}^{H\times V}$ 是预训练的词向量矩阵，λ 是权重系数。$W_{:,v_i}$，$E_{:,v_i}$ 分别代表矩阵 $\boldsymbol{W}$，$\boldsymbol{E}$ 中的第 v_i 列。值得注意的是，矩阵 $\boldsymbol{W}$ 是一个可学习的参数矩阵，其代表了主题模型的主题—词分布，每一行 $W_{l,:}$ 编码了第 l 个潜在主题的主题信息，而每一列 $\boldsymbol{W}_{:,v_i}$ 则是词 v_i 的向量表示，如图 2 所示。

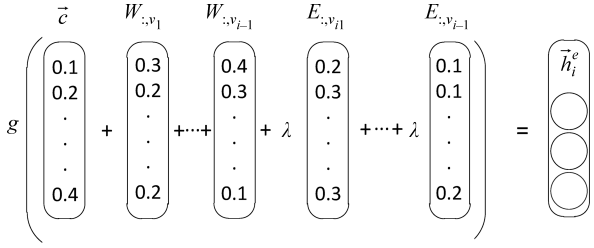


图 2 前向隐状态 $\overrightarrow{h}_i^e$ 的计算

其次，iDocNADEe 模型将文档中所有词的联合分布 $p(v)$ 分解为每个词 v_i 的条件分布的乘积，即 $p(v)=\prod_{i=1}^{D}p(v_i)$，并据此对文档建模。其中，每个词的前后向自回归条件 $p(v_i)$（如图 1(b) 上方的 $\overrightarrow{v}_i$，$\overleftarrow{v}_i$）分别由前向隐状态 $\overrightarrow{h}_i^e$ 和后向隐状态 $\overleftarrow{h}_i^e$ 通过神经网络计算得到，如式(3)、式(4) 所示。

$$p(v_i=w\mid v_{<i})=\frac{\exp(\overrightarrow{b}_w+U_{w,:}\overrightarrow{h}_i^e(v_{<i}))}{\sum_{w'}\exp(\overrightarrow{b}_{w'}+U_{w',:}\overrightarrow{h}_i^e(v_{<i}))} \quad (3)$$

$$p(v_i=w\mid v_{>i})=\frac{\exp(\overleftarrow{b}_w+U_{w,:}\overleftarrow{h}_i^e(v_{>i}))}{\sum_{w'}\exp(\overleftarrow{b}_{w'}+U_{w',:}\overleftarrow{h}_i^e(v_{>i}))} \quad (4)$$

其中，$w\in\{1,\cdots,V\}$。$\overrightarrow{b}\in\mathbb{R}^V$，$\overleftarrow{b}\in\mathbb{R}^V$ 分别为前后向的偏置向量。$\boldsymbol{U}\in\mathbb{R}^{V\times H}$ 代表连接隐藏层与输出层的权重矩阵。

由此，任意文档的对数似然函数如式(5)所示。

$$\log p(v)=\frac{1}{2}\sum_{i=1}^{D}(\log p(v_i\mid v_{<i})+\log p(v_i\mid v_{>i})) \quad (5)$$

2.2.2 融合案件要素特征构建的注意力机制

涉案新闻主题模型中，与案件相关的主题特征词之间应具有较大的相关度。例如，“重庆公交坠江”“万州公交坠江”，这两个文本中“重庆”和“万州”的涉案相关度明显高于这两个词作为地名之间的相关度。可见，将案件要素融入文本特征表示中，可以增强文本中与案件相关的词汇之间的主题相关度。为此，本文对 iDocNADEe 模型进行了改进，模型架构如图 1(b)所示。从图中可以看出，通过计算案件要素双向编码隐状态向量 $\overrightarrow{h}_i^k$，$\overleftarrow{h}_i^k$ 与涉案文本的双向隐状态向量 $\overrightarrow{h}_i^e$，$\overleftarrow{h}_i^e$ 的相关度构建注意力机制，实现案件要素特征对主题特征词的加权，在此基础上利用神经自回归算法计算加权后的双向隐状态向量 $\overrightarrow{h}_i$，$\overleftarrow{h}_i$ 的自回归条件概率，由此得到涉案文本的主题分布，下面我们将详细介绍这一过程的实现。

首先，本文模型的输入不仅包括了新闻的文本序列 v，还有案件要素集合 $k=[k_1,\cdots,k_n]$。与文本隐状态计算类似，我们首先计算案件要素的前后向隐状态，如式(6)、式(7)所示。

$$\overrightarrow{h}_i^k(k_{<i})=g(\overrightarrow{c}+\sum_{j<i}W_{:,k_j}+\lambda\sum_{j<i}E_{:,k_j}) \quad (6)$$

$$\overleftarrow{h}_i^k(k_{>i})=g(\overleftarrow{c}+\sum_{j>i}W_{:,k_j}+\lambda\sum_{j>i}E_{:,k_j}) \quad (7)$$

其中，$k_{<i}\in v_{<i}$，$k_{>i}\in v_{>i}$，$W_{:,k_j}$，$E_{:,k_j}$ 分别代表文档 v 中第 j 个要素在参数矩阵 $\boldsymbol{W}$ 以及预训练词向量矩阵 $\boldsymbol{E}$ 中的向量表示。然后我们使用得到的案件要素隐状态计算出案件要素双向注意力向量 $[\overrightarrow{y},\overleftarrow{y}]$，如式(8)、式(9)所示。

$$\overrightarrow{y}=\tanh\left(\frac{1}{n}\sum_{i=1}^{n}\overrightarrow{h}_i^k\right) \quad (8)$$

$$\overleftarrow{y}=\tanh\left(\frac{1}{n}\sum_{i=1}^{n}\overleftarrow{h}_i^k\right) \quad (9)$$

其中，注意力向量 $[\overrightarrow{y},\overleftarrow{y}]$ 编码了案件要素所包含的信息，利用该向量对文本的隐状态进行加权可以得到带有案件信息的隐状态。我们使用文档中第 i 个词汇的隐状态 $[\overrightarrow{h}_i^e,\overleftarrow{h}_i^e]$ 与注意力向量 $[\overrightarrow{y},\overleftarrow{y}]$ 计算第 i 个词汇处的双向注意力权值 $[\overrightarrow{a}_i,\overleftarrow{a}_i]$，如式(10)、式(11)所示。

$$\overrightarrow{a}_i(k_{<i})=\frac{\exp(\mathrm{score}(\overrightarrow{h}_i^e(v_{<i}),\overrightarrow{y}))}{\sum_{i=1}^{D}\exp(\mathrm{score}(\overrightarrow{h}_i^e(v_{<i}),\overrightarrow{y}))} \tag{10}$$

$$\overleftarrow{a}_i(k_{>i})=\frac{\exp(\mathrm{score}(\overleftarrow{h}_i^e(v_{>i}),\overleftarrow{y}))}{\sum_{i=1}^{D}\exp(\mathrm{score}(\overleftarrow{h}_i^e(v_{>i}),\overleftarrow{y}))} \tag{11}$$

其中，$\mathrm{score}(h_i^e,y)=(h_i^e)^T\cdot y$ 为注意力机制的对齐函数。最终的隐状态$[\overrightarrow{h}_i,\overleftarrow{h}_i]$由文本隐状态$[\overrightarrow{h}_i^e,\overleftarrow{h}_i^e]$与注意力权值$[\overrightarrow{a}_i,\overleftarrow{a}_i]$计算得到，如式(12)、式(13)所示。

$$\overrightarrow{h}_i(v_{<i},k_{<i})=\overrightarrow{h}_i^e(v_{<i})*\overrightarrow{a}_i(k_{<i}) \tag{12}$$

$$\overleftarrow{h}_i(v_{>i},k_{>i})=\overleftarrow{h}_i^e(v_{>i})*\overleftarrow{a}_i(k_{>i}) \tag{13}$$

通过注意力机制，我们在隐状态$[\overrightarrow{h}_i,\overleftarrow{h}_i]$中融合了案件要素所包含的信息，因此，每个词最终的前后向自回归条件，即模型预测文档每个位置词汇 v_i 的概率不仅使用了上下文信息，而且融合了相关的案件信息，在这些案件信息的指导下，我们的模型能够更好地体现其作为生成模型的特点，获得质量更高的涉案新闻主题，如式(14)、式(15)所示。

$$p(v_i=w\mid v_{<i},k_{<i})=\frac{\exp(\overrightarrow{b}_w+U_{w,:}\overrightarrow{h}_i(v_{<i},k_{<i}))}{\sum_w{}'\exp(\overrightarrow{b}_w{}'+U_{w',:}\overrightarrow{h}_i(v_{<i},k_{<i}))} \tag{14}$$

$$p(v_i=w\mid v_{>i},k_{>i})=\frac{\exp(\overleftarrow{b}_w+U_{w,:}\overleftarrow{h}_i(v_{>i},k_{>i}))}{\sum_w{}'\exp(\overleftarrow{b}_w{}'+U_{w',:}\overleftarrow{h}_i(v_{>i},k_{>i}))} \tag{15}$$

最终我们通过式(5)、式(14)、式(15)计算融入了案件信息的生成文档的对数似然函数。

2.3　模型训练

在模型的训练过程中，直接由式(14)、式(15)进行计算会导致计算成本过高，因此我们遵从 Gupta 等人[3]的实验设计，使用二叉树进行计算（见算法1）。在二叉树中，从根到叶子的每个路径都对应一个词汇[20-21]。树中每个节点向左（或右）的概率由一组二元逻辑回归模型建模，然后通过这些概率来计算给定词的概率。

算法 1　使用二叉树计算涉案新闻文档的对数似然函数 $p(v)$ 算法伪代码

	输入：训练文档序列 v、文档中提取的要素集合 k 以及预训练的词向量矩阵 $\boldsymbol{E}$
	参数：$\{\overrightarrow{b},\overleftarrow{b},\overrightarrow{c},\overleftarrow{c},\boldsymbol{W},\boldsymbol{U}\}$
	输出：$p(v)$
1	$\overrightarrow{a}\rightarrow\overrightarrow{c}+\sum_{i>1}W_{:,v_i}+\lambda\sum_{i>1}E_{:,v_i}$
2	$\overrightarrow{k}\rightarrow\overrightarrow{c}+\sum_{j>1}W_{:,k_j}+\lambda\sum_{j>1}E_{:,k_j}$
3	$q(v)=1$
4	for $i=1$ to D **do**：
5	$\overrightarrow{h}_i^e\leftarrow g(\overrightarrow{a});\overleftarrow{h}_i^e\leftarrow g(\overleftarrow{a})$
6	$\overrightarrow{h}_i^k\leftarrow g(\overrightarrow{k});\overleftarrow{h}_i^k\leftarrow g(\overleftarrow{k})$
7	$\overrightarrow{y}\leftarrow\tanh\left(\frac{1}{n}\sum_{i<n}\overrightarrow{h}_i^k\right);\overleftarrow{y}\leftarrow\tanh\left(\frac{1}{n}\sum_{i>n}\overleftarrow{h}_i^k\right)$
8	$\overrightarrow{a}_i(k_{<i})\leftarrow\frac{\exp(\mathrm{score}(\overrightarrow{h}_i^e,\overrightarrow{y}))}{\sum_{i<D}\exp(\mathrm{score}(\overrightarrow{h}_i^e,\overrightarrow{y}))};\overleftarrow{a}_i(k_{>i})\leftarrow\frac{\exp(\mathrm{score}(\overleftarrow{h}_i^e,\overleftarrow{y}))}{\sum_{i>D}\exp(\mathrm{score}(\overleftarrow{h}_i^e,\overleftarrow{y}))}$
9	$\overrightarrow{h}_i(v_{<i},k_{<i})\leftarrow\overrightarrow{h}_i^e*\overrightarrow{a}_i;\overleftarrow{h}_i(v_{>i},k_{>i})\leftarrow\overleftarrow{h}_i^e*\overleftarrow{a}_i$
10	$p(v_i\mid v_{<i},k_{<i})=1;p(v_i\mid v_{>i},k_{>i})=1$
11	for $m=1$ to $\lvert\pi(v_i)\rvert$ **do**：
12	$p(\pi(v_i)_m\mid v_{<i},k_{<i})=g(\overrightarrow{b}_{l(vi)_m}+U_{l(vi)_m,:}\overrightarrow{h}(v_{<i},k_{<i}))$
13	$p(\pi(v_i)_m\mid v_{>i},k_{>i})=g(\overleftarrow{b}_{l(vi)_m}+U_{l(vi)_m,:}\overleftarrow{h}(v_{>i},k_{>i}))$
14	$p(v_i\mid v_{<i},k_{<i})\leftarrow p(v_i\mid v_{<i},k_{<i})p(\pi(v_i)_m\mid v_{<i},k_{<i})$

续表

15	$p(v_i \mid v_{>i}, k_{>i}) \leftarrow p(v_i \mid v_{>i}, k_{>i}) p(\pi(v_i)_m \mid v_{>i}, k_{>i})$
16	$q(v) \leftarrow q(v) p(v_i \mid v_{<i}, k_{<i}) p(v_i \mid v_{>i}, k_{>i})$
17	$\overrightarrow{a} \leftarrow \overrightarrow{a} + W_{:,v_i} + \lambda E_{:,v_i}$; $\overleftarrow{a} \leftarrow \overleftarrow{a} - W_{:,v_i} - \lambda E_{:,v_i}$
18	$\overrightarrow{k} \leftarrow \overrightarrow{k} + W_{:,k_j} + \lambda E_{:,k_j}$; $\overleftarrow{k} \leftarrow \overleftarrow{k} - W_{:,k_j} - \lambda E_{:,k_j}$
19	$\log p(v) \leftarrow \frac{1}{2} \log q(v)$

算法 1 展示了我们的模型如何在案件要素的指导下计算每篇涉案新闻的对数似然函数。其中，第 6～9 行展示了我们如何结合案件要素和注意力机制对新闻隐状态进行加权。而第 12～15 行表示了如何使用二叉树来降低模型的计算成本，$l(v_i)$表示从根到词 v_i 的路径上的树节点的序列，而 $\pi(v_i)$表示这些节点中的每个节点的左(或右)选择的序列[例如 $l(v_i)_1$ 将始终是树的根，如果词 v_i 的叶子节点在其左子树中，则 $\pi(v_i)_1$ 为 0，否则为 1]。因此，现在每个词的自回归条件的计算如式(16)～式(19)所示。

$$p(v_i = w \mid v_{<i}, k_{<i}) = \prod_{m=1}^{|\pi(v_i)|} p(\pi(v_i) \mid v_{<i}, k_{<i}) \tag{16}$$

$$p(v_i = w \mid v_{>i}, k_{>i}) = \prod_{m=1}^{|\pi(v_i)|} p(\pi(v_i) \mid v_{>i}, k_{>i}) \tag{17}$$

$$p(\pi(v_i)_m \mid v_{<i}, k_{<i}) = g(\overrightarrow{b}_{l(v_i)_m} + U_{l(v_i)_m,:} \overrightarrow{h}(v_{<i}, k_{<i})) \tag{18}$$

$$p(\pi(v_i)_m \mid v_{>i}, k_{>i}) = g(\overleftarrow{b}_{l(v_i)_m} + U_{l(v_i)_m,:} \overleftarrow{h}(v_{>i}, k_{>i})) \tag{19}$$

最终我们通过最大化对数似然函数 $\log p(v)$以优化参数$\{\overrightarrow{b}, \overleftarrow{b}, \overrightarrow{c}, \overleftarrow{c}, W, U\}$。

2.4 涉案新闻主题信息抽取

利用 2.3 节所述训练方法构建出涉案新闻主题模型后，通过共享训练好的模型参数$\{\overrightarrow{b}, \overleftarrow{b}, \overrightarrow{c}, \overleftarrow{c}, \boldsymbol{W}, \boldsymbol{U}\}$实现对新文档的主题预测。具体实现过程如下：

对于一篇未在训练集中出现过的涉案新闻 v^*，其词汇大小为 D^*，案件要素 k^* 的大小为 n^*。我们首先通过式(1)、式(2)计算新闻的双向隐状态 $\overrightarrow{h}(v^*)$和 $\overleftarrow{h}(v^*)$，其中参数$\{\overrightarrow{c}, \overleftarrow{c}, \boldsymbol{W}\}$由训练集训练得到，再由式(6)～式(11)计算案件要素双向注意力权值 $\overrightarrow{a}(k^*)$和 $\overleftarrow{a}(k^*)$，最终由式(12)、式(13)计算带有案件要素加权的隐状态 $\overrightarrow{h}(v^*)$和 $\overleftarrow{h}(v^*)$。最终得到涉案新闻经由案件要素加权的主题信息 $\overleftrightarrow{h}$，如式(20)所示。

$$\overleftrightarrow{h} = \overrightarrow{h}(v^*) + \overleftarrow{h}(v^*) \tag{20}$$

3 实验与结果分析

3.1 数据集

针对涉案新闻文本主题模型构建任务，由于目前还没有可用的公开数据集，本文使用的涉案新闻数据通过网络爬虫技术从新闻网站、微博以及微信公众号爬取了近年来部分热点案件的相关新闻，如重庆公交坠江案、丽江唐雪反杀案、孙小果涉黑案等。经过分析发现与案件相关的新闻正文的长度不均衡，而且文本中包含了大量的噪声，但新闻标题基本上都包含了跟案件相关的一些信息，如案件名称、涉案人员等重要信息。为此，本文仅选择了涉案文本的标题信息来构建涉案文本数据集，我们使用 HanLP[①] 对其进行分词，并按照 7∶3 的比例划分训练集与测试集。数据集具体信息如表 2 所示。

表 2 本文数据集的属性

属性	属性值
案件数	65
涉案新闻文档个数	44 000
案件要素总数	3 657
文本最大长度	26
文本平均长度	11.5

3.2 实验参数设置

实验涉及参数如表 3 所示。

① https://github.com/hankcs/HanLP

表 3　本文实验中各参数的设置

参数名	参数值	参数描述
V	31 483	词典大小
H	50/200	主题数量
λ	1.0	权重系数
l_r	0.003	学习率
$g(\cdot)$	sigmoid	激活函数

3.3　预训练词向量

在模型中，词向量作为对主题信息的补充，因此其维度需要与主题数一致，分别为 50/200 维，考虑到目前中文并没有基于大规模语料训练的开源 50/200 维词向量，我们利用开源库 gensim 中的 Word2Vec 工具包，联合了从中国裁判文书网爬取的裁判文书和本文中使用的语料（数据共计 17GB）以及开源中文新闻语料（news2016zh）①训练词向量，词向量的维度为 50/200 维。

3.4　评价指标

（1）困惑度（perplexity）

困惑度（PPL）用于检验主题模型的泛化能力，困惑度越低，则代表模型具备的泛化能力越好。我们通过计算测试集中涉案新闻的困惑度来评估主题模型作为生成模型的文档生成能力。困惑度的计算如式（21）所示。

$$\mathrm{PPL}=\exp\left(-\frac{1}{N}\sum_{t=1}^{N}\frac{1}{|v^t|}\log p(v^t)\right) \quad (21)$$

其中，N 是新闻数量，$|v^t|$ 则代表每篇新闻 $t\in\{1,\cdots,N\}$ 中的词汇数量。$\log p(v^t)$ 由式（5）得到。

（2）主题连贯性（topic coherence）

我们使用了 Röder 等人[22]提出的自动度量指标 C_V 来验证模型产生的主题的连贯性，并使用开源工具 gensim② 来完成这一项指标的计算。

该指标使用参考语料库上的滑动窗口来确定每个主题词的上下文特征。该指标越高，即代表主题的连贯性越好，主题模型效果越好。遵从 Gupta 等人[3]的实验设计，上下文滑动窗口的大小被设置为 110。

3.5　基线模型

本文选择了在 ICLR、AAAI 等会议发表的几个具有代表性的神经主题模型作为基准模型。

（1）**DocNADE**[8]：由 Lauly 等人提出的一种神经主题模型，作为 NADE 和 RSM 的扩展模型，该模型使用神经自回归估计对文本进行主题建模。

（2）**TDLM**[10]：由 Lau 等人在 ACL2017 提出，该模型是一种基于卷积神经网络、注意力机制以及 LSTM 网络的双神经网络模型，是一种多任务学习模型，由主题推断与文本生成两个子任务组成。

（3）**ctx-DocNADEe**[11]：由 Gupta 等人在 ICLR2019 提出，该模型在 DocNADE 的基础上引入了 LSTM 语言模型和 Glove 词向量，其文本的隐藏状态由三者共同计算得到。

（4）**iDocNADEe**[1]：同样是 DocNADE 的扩展版，由 Gupta 等人在 AAAI2019 提出，详情见 2.2 节。

3.6　实验结果与分析

第一组实验是本文提出模型与 4 个基准模型在涉案新闻数据集上，主题数 H 设置为 50 时的困惑度（PPL）对比，实验结果如表 4 所示。

表 4　本文方法和基准模型困惑度对比

模型	PPL
TDLM	1 501
DocNADE	1 004
ctx-DocNADEe	926
iDocNADEe	904
本文方法	898

根据表 4 可以看出，TDLM 模型的困惑度最高，即该模型的泛化能力最差，我们认为这主要因为该模型是一种双任务模型，并且主要目标在于优化语言模型，因此其主题模型的效果并不明显。而本文提出方法的困惑度最低，较基线模型降低了 0.66%，说明了案件要素通过注意力机制融入主题模型中的确可以提升生成文本的质量，并且可以提升模型的泛化能力。虽然提升的效果有限，但主题模型的生成能力仅代表了主题模型的一项能力，我们更加注重主题的质量，即主题连贯性和基于主题的文档检索效果。

主题连贯性能够评估模型所发现的主题的意

① https://github.com/brightmart/nlp_chinese_corpus

② (radimrehurek.com/gensim/models/coherencemodel.html, coherence type=c_v)

义。本文第二组实验对比了本文提出方法与4个基准模型在涉案新闻数据集上，主题数 H 设置为50时的主题连贯性，实验结果如表5所示。其中T10和T20分别代表每个主题取前10个以及前20个主题词计算出的主题连贯性。

表5 本文方法和基准模型主题连贯性对比

模型	C_V	
	T10	T20
DocNADE	0.403	0.536
TDLM	0.283	0.331
ctx-DocNADEe	0.449	0.568
iDocNADEe	0.463	0.581
本文方法	**0.492**	**0.632**

根据表5的实验结果可以看出，TDLM模型所得到的主题连贯性分数最低，即该模型得到的主题词的语义连贯性较差，因为其主要目的是通过主题模型来优化语言模型，而DocNADE只考虑了文本的前向序列，并没有考虑反向序列，因此其效果较拓展类模型较差。而其他两种方法都考虑了文章的上下文信息，所以效果有明显提高。而本文方法取得的主题连贯性最高，10个主题词时，效果较基线模型提升了6.26%，20个主题词时，效果提升了8.78%，这也表明基于案件要素的注意力机制能够帮助模型找到连贯性更好的主题。

为了进行词汇向量表示的测试，本文使用构建的涉案新闻数据集对所提出的模型进行了训练，并使用 $W_{:,v_i}$ 作为每个词汇的向量表示(200维)。我们选取了三个词汇以及与其相似度最高的5个词汇进行展示，此处的相似度由余弦相似度计算得到。实验结果如表6所示，其中 s_y，s_w 分别代表使用本文提出方法计算得到的词的向量表示与使用Word2Vec训练得到的词的向量表示所计算出余弦相似度。

表6 词汇向量表示对比(%)

弒母			坠江			车主		
相邻词	s_y	s_w	相邻词	s_y	s_w	相邻词	s_y	s_w
弒	67	32	万州	79	33	哭诉	74	32
杀母	70	66	公交	73	32	西安	71	36
天生	59	14	出水	61	31	服务费	66	35
少年	58	41	互殴	56	27	维权	64	40
管教	57	24	大巴	55	37	撒泼	63	33

根据表6的实验结果可以看出，通过训练，我们提出的方法抽取到的主题词跟案件要素具有更大的语义相关性。

表7 新闻检索系统中的混淆矩阵

	相关新闻	无关新闻
检索到的新闻	TP	FP
未检索到的新闻	FN	TN

主题模型的一个重要用途就是得到文档的主题信息。我们通过执行一个涉案新闻检索任务以评估本文所提出方法以及对比方法所得到的新闻主题信息的质量。我们使用式(20)来抽取每篇新闻的主题信息，并将训练集中的新闻用作检索，而测试集中的新闻用作查询。

本文设置了多组不同的检索分数(fraction of retrieved documents)以进行对比。我们将用作查询的新闻的主题信息与所有检索集中的新闻的主题信息做相似度计算，返回相似度最高的前 N_p 条新闻。N_p 的计算如式(22)所示。

$$N_p = N_r * \text{检索分数} \tag{22}$$

其中，N_r 是检索集的新闻数量。最终我们通过查询新闻的标签和返回的 N_p 条新闻的标签计算检索精确率。新闻检索系统的精确率表示在检索到的文档中，相关文档所占比例。已知混淆矩阵如表7所示，则精确率计算如式(23)所示。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{23}$$

结果如图3所示，纵轴代表各模型取得的精确率，横轴代表检索分数。可以看到，检索分数与精确率成反比，因为检索分数越高，代表返回的新闻数量越多，而检索到无关新闻的数量也就越多，直接导致精确率的降低。当检索分数为1%时，检索系统所

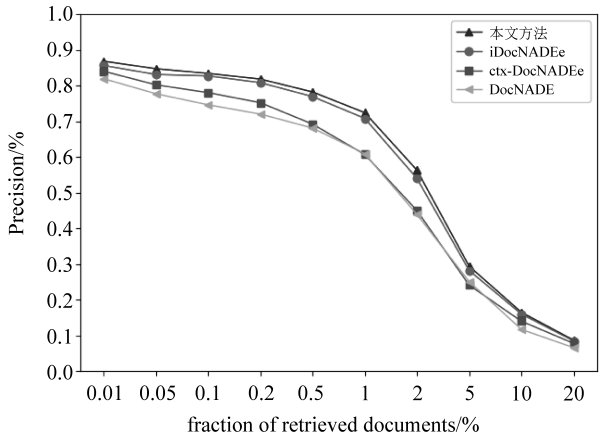


图3 各模型的文档检索精确率对比

返回新闻的数量恰好与检索集中一个类别新闻的平均数量相近，当检索分数继续升高时，返回的新闻几乎都是无关新闻，因此精确率大幅度降低。但无论检索分数的高低，利用我们提出模型所抽取的主题信息获得的检索精确率始终是最高的。这是因为我们将案件要素融入模型，因此模型所抽取的主题信息包含了案件信息，质量也就越高。这也证明了本文使用案件要素信息对模型进行注意力加权指导是有效的。

4　结束语

由于现有的主题模型忽略了上下文信息及外部知识对词语主题分布的帮助，本文对 iDocNADEe 模型做了进一步扩展，提出了一种基于神经自回归分布估计的涉案新闻主题模型构建方法。该方法通过融入案件要素作为外部知识，能较好地解决神经主题模型在涉案新闻领域效果不佳的问题，并能获得更低的困惑度以及更好的主题连贯性，在涉案新闻检索实验中也获得了更佳的性能。我们将在下一步工作中，研究如何利用除案件要素外的涉案领域知识，如裁判文书和法律条文等对涉案新闻主题模型的帮助。

参考文献

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

[2] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[3] Gupta P, Buettner F, Schuetze H, et al. Document informed neural autoregressive topic models with distributional prior[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019: 6505-6512.

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

[5] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv: 1312.6114, 2013.

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

[7] Larochelle H, Murray I. The neural autoregressive distribution estimator[C]//Proceedings of the 14th International Conference on Artificial Intelligence and Statis-tics, JMLR Workshop and Conference, 2011: 29-37.

[8] Lauly S, Zheng Y, Allauzen A, et al. Document neural autoregressive distribution estimation[J].Journal of Machine Learning Research, 2016, 18(113): 1-24.

[9] Dieng A B, Wang C, Gao J, et al. Topicrnn: A recurrent neural network with long-range semantic dependency[J]. arXiv preprint arXiv: 1611.01702, 2016.

[10] Lau J H, Baldwin T, Cohn T. Topically driven neural language model[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 355-365.

[11] Gupta P, Chaudhary Y, Buettner F, et al. Texttovec: Deep contextualized neural autoregressive topic models of language with distributed compositional prior[J]. arXiv preprint arXiv: 1810.03947, 2018.

[12] 张绍武，邵华，林鸿飞等.基于主题模型的新疆暴恐舆情分析[J].中文信息学报，2018，32(05)：105-113.

[13] 吴彦文，黄凯，王馨悦等.一种融合主题模型的短文本情感分类方法[J].小型微型计算机系统，2019，40(10)：2082-2086.

[14] 陈琪，张莉，蒋竞等.一种基于支持向量机和主题模型的评论分析方法[J].软件学报，2019，30(05)：1547-1560.

[15] Mousa A, Schuller B. Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Lin-guistics, 2017: 1023-1032.

[16] Vu N T, Adel H, Gupta P, et al. Combining recurrent and convolutional neural networks for relation classification [C]//Proceedings of NAACL-HLT 2016, 2016: 534-539.

[17] Vu N T, Gupta P, Adel H, et al. Bidirectional recurrent neural network with ranking loss for spoken language understanding[C]//Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, 2016: 6060-6064.

[18] 苗林林.刑事案件构成要素相关性分析应用研究[J].政法学刊，2017，34(05)：37-44.

[19] 韩鹏宇，高盛祥，余正涛等.基于案件要素指导的涉案舆情新闻文本摘要方法[J].中文信息学报，2020，34(05)：56-63，73.

[20] Morin F, Bengio Y. Hierarchical probabilistic neural network language model[C]//Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS 2005), 2005: 246-252.

[21] Mnih A, Hinton G E. A Scalable hierarchical distributed language model[C]//Proceedings of the 21st International Conference on Neural Information Processing Systems, 2008: 1081-1088.

[22] Röder M, Both A, Hinneburg A, et al. Exploring the space of topic coherence measures[C]//Proceedings of the 8th ACM International Conference on Web Search and Data Mining, 2015: 399-408.

毛存礼(1977—),博士,副教授,主要研究领域为自然语言处理、信息检索、机器翻译。
E-mail: maocunli@163.com

余正涛(1970—),通信作者,教授,博士生导师,主要研究领域为自然语言处理、信息检索、机器翻译。
E-mail: ztyu@hotmail.com

梁昊远(1995—),硕士研究生,主要研究领域为自然语言处理、信息检索。
E-mail: lianghaoyuan2749@foxmail.com