文章编号：XXXX-XXXX（2017）00-0000-00

基于双向 LSTM 语义强化的主题建模

彭敏 杨绍雄 朱佳晖
（武汉大学 计算机学院， 武汉 430072）

摘要：当前，双向 LSTM 神经网络等深度学习方法在文本语义特征表达方面取得了突破性的进展，为构建深层次的具有语义连贯性的主题模型提供了可能。但是，现有方法在文本的概率主题建模方面，提升的效果还比较有限。本文提出了一个基于双向 LSTM 语义强化的概率主题模型 DGPU-LDA（Double Generalized Polya Urn with LDA）。该模型一方面结合双向 LSTM 文档语义编码框架 DS-Bi-LSTM（Document Semantic Bi-directional LSTM）来实现文档宏观语义的嵌入表示，另一方面采用文档-主题和词汇-词汇双 GPU（Generalized Polya Urn）语义强化机制以及 LSTM 来刻画参数推断过程中的吉布斯采样过程。在搜狗新闻数据集以及 20 新闻组数据集上的实验结果表明， DGPU-LDA 模型在主题语义连贯性、文本分类准确率方面相对于一些比较前沿的主题模型具有一定的优势，同时也表明了该模型在文本语义特征表达方面的有效性。

关键词：双向 LSTM；语义强化；主题模型
中图分类号：TP391 **文献标识码：**A

Semantic Reinforcement of Topic Modeling with Bi-directional LSTM

Peng Min , Yang Shaoxiong, and Zhu Jiahui
(School of Computer, Wuhan University, Wuhan 430072, China)

Abstract：Nowadays, deep learning models, such as bi-directional LSTM neutral networks, have achieved breakthroughs in text semantic representation. This makes it possible for constructing a semantic coherent topic model with a deep architecture. But the research is not wide and profound yet. Based on the deep semantic reinforcement from bi-directional LSTM, we propose a probabilistic topic model DGPU-LDA (Double Generalized Polya Urn with LDA). In DGPU-LDA, we design a document-wise semantic encoder DS-Bi-LSTM (Document Semantic Bi-directional LSTM) to embed the semantics of each document. Then document-topic GPU semantic reinforcement, word-word GPU semantic reinforcement, together with LSTM iterative dependency modeling, can be exploited to capture the Gibbs sampling process in model inference. Experimental results on SogouCA dataset and 20 Newsgroup dataset demonstrate that the proposed DGPU-LDA outperform some of the state-of-the-art topic models in topic semantic coherence and text classification. Meanwhile, these remarkable improvements also indicate the effectiveness of our DGPU-LDA in text semantic feature representation.

Key words: bi-directional LSTM; semantic reinforcement; topic model

收稿日期：2017-XX-XX；**定稿日期**：2017-XX-XX
基金项目：《社会网络的主题演化分析与传播趋势预测研究》（基金号 61472291）

1 引言

当前，对于自然语言处理的诸多任务而言，深度神经网络虽然已经发挥出了一定的作用，但是对于文本的概率主题建模方面，提升的效果还比较有限。主题建模主要针对文档层面，旨在从文档集合中挖掘出表达语义主旨的相关词汇或短语。常用的概率主题模型一般采用潜在狄利克雷分布（Latent Dirichlet Allocation, LDA）[1]及其变种。概率主题模型通过主题作为中间层特征的表达方式，能够显式地抽取文本的语义信息，从而胜任文本分类、突发事件检测、主题演化分析、推荐系统等相关任务。

虽然主题模型的应用领域十分广泛，但是在特定环境下，其建模过程仍然存在一系列挑战。首先，当前的主题模型在结构上的可扩展性十分有限，往往通过在基本的生成模型中加入额外的随机变量来实现功能上的增强。其次，当前的主题建模没有较好地结合相关语义强化机制，使得产生的主题语义连贯性较差，很难为人所理解。同时，当前的主题模型在参数推断过程中没有显式地考虑上下文信息，使得对于主题的分配过程难以收敛到一个较好的状态。此外，对比深度学习的框架结构，传统的主题模型仍然是一种浅层的表示结构，存在特征表达能力不强等缺点。

结合上述问题与挑战，本文基于双向 LSTM 神经网络[2]构建文本主题模型，旨在增强模型整体的可扩展性，主题表达的语义连贯性，参数推断的上下文一致性，以及文档语义表征的准确性。

本文研究了传统概率主题模型的深度语义强化问题，并提出了具有文档-主题和词汇-词汇双方面语义强化的主题模型 DGPU-LDA。首先，针对文档级别的语义编码问题，本文提出了一个基于双向LSTM的文档语义编码框架DS-Bi-LSTM，能够从文档的实体关系标注词汇序列中聚合出文档的宏观语义编码。其次，本文结合广义玻利亚瓮模型（Generalized Polya Urn, GPU）[3]分别从文档-主题和词汇-词汇两方面对词汇的主题分配采样进行强化。最后，考虑双 GPU 模型的强化以及主题分配序列的前后依赖性，设计相关采样算法来完成 DGPU-LDA 模型的参数推断。

本文的贡献主要如下：

(1) 本文提出了一个文档级别的语义编码框架 DS-Bi-LSTM。DS-Bi-LSTM 编码框架能够有效地实现对于文档语义信息的嵌入表示，为文档语义信息的抽取和表达提供了一个深度学习的实现方式。

(2) 本文将双向 LSTM 网络融入到传统的概率主题建模过程中，提出了 DGPU-LDA 模型。DGPU-LDA 有效结合了文档-主题和词汇-词汇两个级别的 GPU 深度语义强化机制。

(3) 本文将 LSTM 网络引入到 DGPU-LDA 模型的吉布斯采样过程中，保证了参数推断过程中的上下文一致性。

2 相关工作

2.1 循环神经网络

循环神经神经网络（Recurrent Neural Network, RNN）[4]是一种能够较好地胜任文本建模的深度神经网络。在 RNN 中，序列当前时刻下的输出不仅与当前时刻下的输入有关，还与上一时刻下的输出有关，即对历史信息具有记忆功能。然而，单纯的 RNN 无法有效地解决信息编码长距离依赖的问题，基于此，Hochreiter 等人提出了一种新的 RNN，即长短时记忆网络（Long Short-Term Memory, LSTM）[5]。LSTM 通过记忆单元保存当前的输入和之前的隐藏层状态，并通过三个门限控制输入、输出以及信息的遗忘，能够有效刻画语义的长距离依赖。为了同时捕获历史以及未来两个方向上的语义信息，双向 LSTM（Bi-directional LSTM）[2]应运而生。双向 LSTM 网络有前向和后向两个 LSTM 隐藏层，且每一对前向和后向隐藏层都连接着同一个输出单元，能够为输出层中每一个时刻提供更加完整的上下文信息，因此表达效果相对于单向的 LSTM 有一定的提升。

2.2 基于知识语义强化的主题建模

基于知识语义强化的主题建模旨在将领域知识进行特征化表达并融合到主题建模过程中，从而增强主题的语义连贯性，使抽取出的主题具有更好的可读性。基于知识语义强化的主题建模的基础是在词汇生成的过程中施加约束，比如使语义相近词汇的共现概率增大的 Must-Link 机制和使语义冲突词汇共现概率减小的 Cannot-Link 机制[6]。Hu 则等人提出了一个基于知识库实体分类体系的主题模型 LGSA[7]，每一个主题的词汇分布都通过实体层次结构上的随机游走过程来生成。然而，在某些复杂情景下，领域知识本身也需要在建模过程中不断获得或持续更新。LTM[8]是一个面向多领域生命周期学习的主题模型，能够通过频繁项集挖掘自动地从历史数据中获取知识模式来为当前的主题建模服务。AMC[9]在 LTM 的基础上引入了“像人一样学习”的思维方式，虽然

也使用频繁项集挖掘来获得知识，但是强调知识的集成与调整，因此不仅能够检测错误的领域知识，而且能够被应用于大数据场景中。

2.3 基于神经网络的主题建模

基于神经网络的主题建模旨在用神经网络结构替代有向概率图结构，既避免了引入复杂的分布和参数，同时又能实现文本主题特征的良好表达。传统的神经网络主题模型往往都是基于受限玻尔茨曼机[10]，例如基于神经自回归估计的 DocNADE[11]，基于深度受限玻尔茨曼机的 Over-Replicated Softmax[12]等等。然而，受限玻尔茨曼机模型训练相对比较困难，且不太能够适应文本的序列化建模方式。Cao 等人提出了一个基于前馈神经网络的主题模型 NTM[13]，将文档-主题和主题-词汇这两个生成过程用两个隐藏层来分别表示，并通过点积来得到最终的文档-词汇生成概率。Tian 等人则提出了一个基于 RNN 的句子级别的主题模型 SLRTM[14]以生成主题相关的句子。在 SLRTM 中，每一个词汇的生成不仅与该词汇所在的句子的主题相关，还与其之前的所有词汇相关。随着 Encoder-Decoder 框架[15]的出现，基于端对端方式和注意力机制[16]的主题建模近年来也逐渐兴起。Xing 等人将 Twitter-LDA[17]融入到 Encoder-Decoder 框架中，同时结合文本注意力和主题注意力来生成对话[18]。而 Li 等人则将注意力机制有机地融入到文本序列的生成过程中，提出了循环注意力主题模型[19]，并使用循环注意力贝叶斯过程来生成注意力权值。

3 DGPU-LDA 模型

3.1 文档语义编码

词嵌入（Word Embedding）[20, 21]虽然能够提供具有语义的词汇特征表达，但是无法直接提供更高粒度的语义信息，例如句子级别和文档级别的语义信息。概率主题模型大部分构建在文档集合之上，因而要实现其语义强化必须先要具备文档级别的语义信息。

传统的语义信息抽取往往需要对文本进行实体关系抽取，进而构造语义模板或语义约束来挖掘语义信息。然而，本文基于双向 LSTM 神经网络构建主题模型，底层词汇的特征表达均采用嵌入的形式。因此，为了更好地融入模型，需要对文档语义进行嵌入表示。

近年来，立足于深度神经网络，文档级别的语义嵌入表示方法往往依赖实体和关系的词汇特征[22]、实体和关系的文本描述[23]等等。此类方法采用词嵌入的组合或是深度神经网络的编码来实现整个文档的语义嵌入表达。

本文认为文档整体表达的语义与其所包含的主题密切相关，因此，本文基于文档语义编码来进行主题建模。为了能够同时捕获历史和未来两个方向上的上下文，本文使用双向的 LSTM 神经网络来对文档的宏观语义进行编码，进而提出了 DS-Bi-LSTM（Document Semantic Bi-directional LSTM）编码框架。

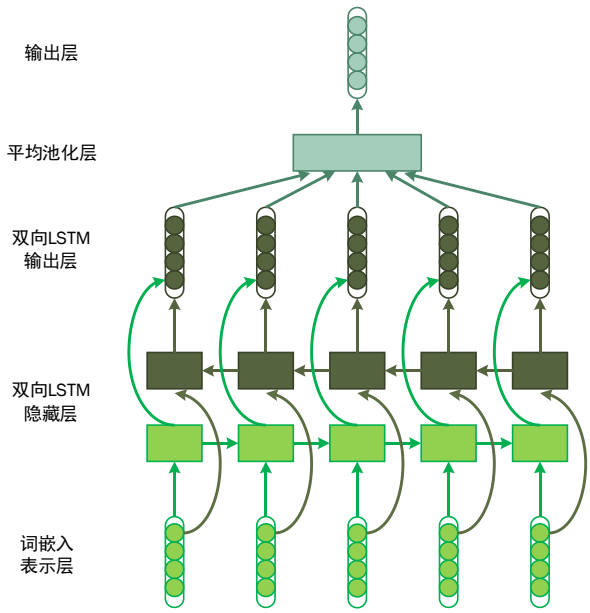


图 1 DS-Bi-LSTM 文档语义编码框架

本文提出的 DS-Bi-LSTM 编码框架如图 1 所示。该框架主要分为 5 层，以文档中所有关键词的词嵌入表达为输入，以该文档的整体语义嵌入表达为输出。下面分别详细介绍这 5 个层次。

词嵌入表示层：该层为输入层。由于某些文档的词汇数目往往较多，且某些不重要的词汇对于文档主题的表达没有直接影响，因此本文只抽取每个文档中的实体关系标注词汇作为输入。实体关系标注词汇即为在 DBpedia 知识库[1]三元组中出现的相关词汇，可通过实体关系链接操作进行获取。获取实体关系标注词汇之后，通过查表即可得到这些词汇的词嵌入向量，所有词汇的词嵌入通过 Skip-Gram 模型[20, 21]得到。

双向 LSTM 隐藏层：包含前向和后向两个 LSTM 隐藏层，在同一时刻下，每个输入词嵌入均同时连接到前向和后向两个 LSTM 隐藏层单元，同时这两个隐藏层单元又连接到同一个输出。设当前 t 时刻下的输入词嵌入为 E_t，上一个时刻的

[1] http://wiki.dbpedia.org/.

前向 LSTM 隐藏层单元的输出为 h_{t-1}^{f}，后向 LSTM 隐藏层单元的输出为 h_{t+1}^{b}，则当前时刻下的前向和后向隐藏层单元的输出为：

$$\begin{aligned} h_t^f &= H\left(E_t, h_{t-1}^f, c_{t-1}, b_{t-1}\right) \\ h_t^b &= H\left(E_t, h_{t+1}^b, c_{t-1}, b_{t-1}\right) \end{aligned} \tag{1}$$

其中 $H(\cdot)$ 表示 LSTM 的隐藏层操作，c_{t-1} 表示上一个时刻下的 Cell 单元的状态值，b_{t-1} 泛指上个时刻下所有的偏置。

双向 LSTM 输出层：每个输出单元同时连接到该时刻下的前向和后向两个 LSTM 隐藏层单元，即：

$$g_t = \sigma\left(W_{hg}^{f} h_t^{f} + W_{hg}^{b} h_t^{b} + b_g \right) \tag{2}$$

其中 W_{hg}^{f} 和 W_{hg}^{b} 分别为前向隐藏层和后向隐藏层与双向 LSTM 输出层之间的连接权值，b_g 为偏置。该层的输出形式为向量，且每个向量的维数与输入向量一致。

平均池化层：池化操作可以将原始的特征值进行处理并构造出新的特征，进而实现对于原始有效特征的降维、强化以及对于噪声的过滤。常用的池化操作有最大池化和平均池化两种。由于文档的宏观语义与文档中的每一个实体关系标注词汇均密切相关，因此本文采用平均池化操作。平均池化即对某一范围内的所有神经元值取平均，能够将每一个方面的局部信息都考虑在内，避免信息的丢失。平均池化的操作如公式(3)所示：

$$pool(g) = \sum_{t=1}^{T} \frac{g_t}{T} \tag{3}$$

其中 T 为输入的实体关系标注词汇序列的长度。

语义编码输出层：将平均池化的结果通过激活函数即可得到最终的整个文档的语义编码向量，即：

$$s = \sigma\left(pool(g) \right) \tag{4}$$

该语义编码向量的维数与输入的实体关系标注词汇嵌入的维数一致，以便于进行相似度计算。通过 DS-Bi-LSTM，可由文档的实体关系标注词汇获取整个文档的宏观语义编码，实现了文档级别语义信息的抽取，为后文基于文档语义进行主题建模奠定了基础。

3.2 DGPU-LDA 主题建模

3.2.1 LDA 主题模型

本文基于由 Blei 等人提出的潜在狄利克雷分布 LDA 构建主题模型[1]，它是当前大部分概率主题模型的基础。LDA 是一个典型的三层贝叶斯概率生成模型，自上而下依次由文档、主题和词汇三个层次构成。主要包含 2 个狄利克雷-多项式共轭结构：文档-主题以及主题-词汇。文档-主题分布 θ 以及主题-词汇分布 φ 均服从多项式分布，并且 θ 的参数服从以 α 为先验参数的狄利克雷分布，而 φ 的参数服从以 β 为先验参数的狄利克雷分布。

LDA 的生成过程如下：

(1) 确定要生成的文档的数目 M，每个文档相应的词汇数目 N_d（即每个文档的长度）。

(2) 对于每一篇文档的每一个词汇：1) 从文档-主题分布 θ_m 采样当前词汇的主题分配 $z_{m,n}$：$Mult(\theta_m)$；2) 根据主题分配选定该主题相应的主题-词汇分布 $\varphi_{z_{m,n}}$；3) 根据该主题-词汇分布采样一个词汇 $w_{m,n}$：$Mult(\varphi_{z_{m,n}})$。

LDA 的生成过程为 LDA 的训练过程（即文档-主题分布参数、主题-词汇分布参数推断过程）提供了概率计算的基础。

3.2.2 基于 GPU 模型的语义强化

通过 DS-Bi-LSTM 得到的文档宏观语义编码能够反映出文档的语义元素，因此能够作为主题产生的约束。对于当前文档中的每一个词汇，若该词的词嵌入与该文档的语义编码极其相似，则该词汇为文档语义的代表词，应将其进行数量上的增强，以增加其被该文档的相应主题选择的概率。

本文采用 GPU 模型[3]来模拟与文档语义相关的词汇的增强过程。GPU 模型已被广泛用于概率主题模型中的词汇采样强化[8, 9, 24]，能够显著改善吉布斯采样的效果。然而，以往的绝大部分工作对于词汇采样的强化仅仅停留在语义相近的词汇之间，没有从文档语义的角度进行加强。文本考虑了文档宏观语义对于主题建模的影响，因此主题语义的增强不仅体现在词嵌入之间的语义关联，还体现在词汇与所处文档之间的语义关联。因此，在主题建模的场景下，本文采用的 GPU 模型具有以下两层含义：

(1) 当某一词汇 w 被主题 z_w 采样到后，若该词汇与当前文档 d 的语义编码相近，则该主题 z_w 与文档 d 的共现次数会大大增加。

(2) 当某一词汇 w 被主题 z_w 采样到后，与该词汇语义相近的所有词汇 $w^* \in \mathbf{RW}_w$ 均会被加强，使得 w^* 被主题 z_w 采样到的概率大大增加。

首先，对于文档-主题部分的 GPU 增强，可通过计算当前词汇的词嵌入与当前文档的语义编

码向量的余弦相似度来判断。若两者之间的余弦相似度大于一定的阈值 κ，则认为存在文档-主题增强矩阵 **A**，对相应文档下的相应主题的共现次数增加 $a(0<a<1)$，否则不进行任何增强，即：

$$A_{d,z}=\begin{cases}0, & dist\ s_d(e_{w_z},<\kappa)\\ a, & dist\ s_d(e_{w_z},\geq\kappa)\end{cases} \tag{5}$$

其中 $dist(s_d,w_z)$ 表示词汇 w_z 的词嵌入 e_{w_z} 与文档 d 的语义编码 s_d 之间的余弦相似度。

其次，对于词汇-词汇部分的 GPU 增强，可通过计算词嵌入与词嵌入之间的余弦相似度来判断。对于当前被采样的词汇 w，所有与其余弦相似度大于阈值 ρ 的词汇构成了该词汇的相关词汇集合 $\mathbf{RW}_w$。设词汇-词汇增强矩阵为 **B**，则对于当前词汇 w，需要对所有属于 $\mathbf{RW}_w$ 中的词汇同步进行增强，将相应位置上的共现次数增加 $b(0<b<1)$；而对于当前词汇自身与自身之间的增强，将共现次数增加 1；其他情况下不进行任何增强。具体的增强方式如公式(6)所示：

$$B_{w,w^*}=\begin{cases}1, & w=w^*\\ b, & w^*\in\mathbf{RW}_w\text{且}w\neq w^*\\ 0, & \text{其他}\end{cases} \tag{6}$$

通过文档-主题和词汇-词汇两部分的 GPU 增强，不仅使得与文档语义相关的主题占比提高，而且使得语义相近的词汇在一个主题中共同出现的概率提高。

3.2.3 DGPU-LDA 模型结构

结合文档-主题 GPU 增强和词汇-词汇 GPU 增强，本文提出了双 GPU 语义增强主题模型 DGPU-LDA（Double Generalized Polya Urn with LDA）。DGPU-LDA 立足于基本的 LDA 结构，在词汇的主题分配采样过程中融合文档-主题 GPU 增强和词汇-词汇 GPU 增强这两个方面。

DGPU-LDA 的模型结构如图 2 所示，其中绿色部分表示文档-主题 GPU 强化和词汇-词汇 GPU 强化。文档-主题 GPU 强化同时依赖于 DS-Bi-LSTM 文档语义编码和词嵌入两部分，而词汇-词汇 GPU 强化只依赖于后者。DGPU-LDA 模型整体上与 LDA 相比并没有引入额外的随机变量，也没有引入额外的先验分布，因此保证了其参数推断过程的简便性。由于 DGPU-LDA 的词汇生成过程与 LDA 差别不大，只是在词汇采样的方式上有所差别，因此对于其生成过程不再赘述。

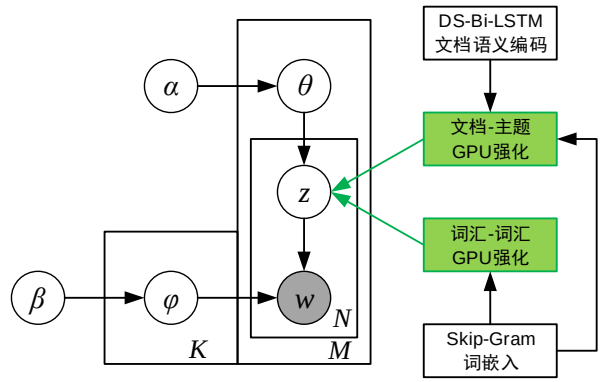


图 2 DGPU-LDA 模型结构图

3.3 模型参数推断

坍缩吉布斯采样[25]是主题模型参数推断过程中常用的一种随机算法。由于其基于马尔可夫假设，操作简便，且在小规模数据上具有较高的计算效率，因此被广为使用。本文亦基于坍缩吉布斯采样来进行模型参数的推断。

本文提出的 DGPU-LDA 采用文档-主题 GPU 强化和词汇-词汇 GPU 强化来约束词汇的主题分配，因此在进行坍缩吉布斯采样时需要考虑两方面的 GPU 模型的融合。此外，传统的坍缩吉布斯基于马尔可夫假设，即当前时刻下的主题分配仅与上一时刻下的主题分配相关，然而，不同时刻下的主题分配序列之间的真实依赖关系仍尚待研究。因此，本文对 DGPU-LDA 的参数推断过程也采用深度神经网络进行改进，以增强前后时间片之间主题的相关性。结合横向的双 GPU 强化和纵向的 LSTM 神经网络主题分配依赖建模，本文提出了针对 DGPU-LDA 模型的参数推断算法。

首先，根据坍缩吉布斯采样的公式，融合文档-主题 GPU 强化和词汇-词汇 GPU 强化的关于当前词汇的主题分配的采样概率为：

$$p\left(z_{d,n}=ct\,|\,\mathbf{z}_{-(d,n)},\mathbf{w},\alpha,\beta\right)$$
$$\propto\frac{C_{d,ct}^{-(d,n)}A_{d,ct}+\alpha}{\sum_{k=1}^{K}\left(C_{d,k}^{-(d,n)}A_{d,k}+\alpha\right)}\frac{\sum_{u=1}^{V}C_{ct,u}^{-(d,n)}B_{u,n}+\beta}{\sum_{v=1}^{V}\left(\sum_{u=1}^{V}C_{ct,u}^{-(d,n)}B_{u,v}+\beta\right)}$$

其中 ct 表示当前词汇的主题分配编号，$C_{d,ct}^{-(d,n)}$ 表示排除当前词汇的情况下，第 d 个文档中的词汇被分配给第 ct 个主题（也就是当前主题）的次数，$C_{ct,u}^{-(d,n)}$ 表示排除当前词汇的情况下，第 u 个词汇被分配给第 ct 个主题的次数。

其次，本文额外考虑了当前词汇的历史主题分配对于最终主题分配序列的影响，构建了一个基于 LSTM 的主题分配依赖网络。该网络由一层

前向 LSTM 神经网络构成，其原理类似于预测句子的下一个词汇。网络的输入层为相应时刻下的主题分配序列，隐藏层为 LSTM 隐藏层，输出层为 Softmax 函数，对应于下一个时刻下的属于每一个主题分配的概率。所有的主题分配均需采用 one-hot 模型表示成向量的形式。

由于主题建模过程往往是无监督的，且进行主题抽取的文本往往不带任何标注，因此，为了训练主题分配依赖网络的权值，需要对整个文档集合的每一个词汇进行主题标注。本文采用 LDA 模型来无监督地标注文档集合中的每一个词汇上的主题，即将坍缩吉布斯采样收敛后的主题分配序列视为该文档集合的主题标注序列。

最终，结合基于双 GPU 强化的坍缩吉布斯采样和主题分配依赖网络，可给出当前词汇下的相应主题分配的采样概率：

$$p\left(z_{d,n}=ct\right)=\pi\cdot p\left(z_{d,n}=ct\mid \mathbf{z}_{-(d,n)},\mathbf{w},\alpha,\beta\right) +\left(1-\pi\right)\cdot LSTM\left(z_{d,n}\right) \tag{8}$$

其中 $LSTM\left(z_{d,n}\right)$ 为主题分配依赖网络的输出，π 为两部分的平衡因子。

根据该采样算法可有效地从历史主题分配中获得有助于当前主题分配采样的信息，增强了参数推断的上下文一致性。

4 实验结果及分析

4.1 实验数据集

本文主要采用一中一英两个数据集来验证本文提出的基于深度学习的主题模型的有效性。中文数据集采用搜狗新闻数据集（SogouCA）2012 版[2]。英文数据集采用 20 新闻组数据集（20 Newsgroups）[3]。关于这两个数据集的描述详见表 1。

对于搜狗新闻数据集，本文从 20 个主题中选择新闻数目较多的 12 个主题，分别是“汽车”、“财经”、“健康”、“体育”、“教育”、“文化”、“军事”、“政治”、“国际”、“娱乐”、“女性”、“科技”。本文在每个主题下随机选择 4000 条新闻（3000 条用于训练，1000 条用于测试）构成用于主题建模的文档集合，总计 48000 条（训练集 36000 条，测试集 12000 条）。而对于 20 新闻组数据集，本文对相关噪声文本进行剔除，最终得到其训练集的大小为 11328，测试集大小为 7511。

[2] http://www.sogou.com/labs/resource/ca.php.

[3] http://kdd.ics.uci.edu/databases/20newsgroups/20-newsgroups.html.

表 1 数据集描述

	搜狗新闻数据集	20 新闻组数据集
文本数	1294233	约 20000
来源	搜狐、雅虎、网易等十个国内媒体网站	UseNet
主题分类数	20	20
主题内容	汽车、IT、健康、财经、体育、旅游、教育、文化、军事、政治、社会、国内、国际、房产、娱乐、时尚、传媒、公益、女性、科技	无神论、计算机图形学、Windows 操作系统、IBM 硬件、Mac 硬件、Windows.x、二手市场、汽车、摩托车、棒球、曲棍球、加密、电子、医学、航空、基督徒、枪支、中东、政治杂谈、宗教杂谈

构建好训练集和测试集后，本文针对文本内容进行分词，去停用词和词性标注等预处理工作。中文分词工具采用 Jieba 分词[4]，并在词性标注的基础上对新闻文本中的命名实体进行识别，得到每一个文档的候选实体关系标注词汇。针对英文文本，本文采用 NLTK[5]来进行分词、去停用词、词性标注和命名实体识别。

4.2 实验设置

本文分别采用 LDA[1]、GPU-DMM[24]、NTM[13]作为对比主题模型。所有模型的介绍及相关参数设置如下：

LDA：经典的概率主题模型。本文利用传统的向量空间模型来进行 LDA 建模，并使用 Python genism 主题建模工具包[6]来实现 LDA 模型。

GPU-DMM：一个基于狄利克雷多项式混合模型以及词嵌入的语义强化主题模型[7]。在该模型中，词汇-词汇 GPU 强化部分语义相关性数值设置为 0.5，强化值为 0.2。

NTM：一个基于神经网络的主题模型[8]。NTM 中文本表示的基本单位并非是单个词汇，而是 N-gram，因此在本实验中将使用 Bi-gram（二元词组）作为其文档基本表示。

DGPU-LDA：本文提出的基于双 GPU 深度语义强化的概率主题模型。本文根据相似度分布来

[4] https://github.com/fxsjy/jieba/.

[5] http://www.nltk.org/.

[6] http://radimrehurek.com/gensim/.

[7] https://github.com/NobodyWHU/GPUDMM.

[8] https://github.com/elbamos/NeuralTopicModels.

设置相似度阈值，即截取前 10%左右。文档-主题 GPU 强化部分的相似度阈值 κ 为 0.2，强化值 a 为 0.2。词汇-词汇 GPU 强化部分的相似度阈值 ρ 为 0.5，强化值 b 为 0.1。主题分配依赖网络中平衡因子 π 的值设定为 0.8。

此外，所有模型中选取的词汇列表长度均为 5000。LDA、GPU-DMM、DGPU-LDA 模型中的文档-主题分布先验参数 α 均为 $50/K$，主题-词汇分布先验参数 β 均为 0.05，模型参数推断的迭代次数均为 50。NTM 中文档-主题隐藏层、主题-词汇（N-gram）隐藏层的神经元数均为 256，训练的迭代次数为 50。

4.3 主题语义连贯性评价

主题语义连贯性最常用的自动评价指标是点对互信息（Point-wise Mutual Information, PMI）[26]。相关研究表明，PMI 自动评价的结果往往与人工评价高度一致[27]，PMI 值越高，则该主题的语义连贯性也越强，因此本文采用 PMI 作为主题语义连贯性自动评价的标准。对于某个主题-词汇分布 φ_k，其 PMI 值计算如下：

$$PMI(\varphi_k)=\frac{2}{V(V-1)}\sum_{1\le i<j\le V}\log\frac{p(w_i,w_j)}{p(w_i)p(w_j)} \qquad (9)$$

其中 $p(w_i)$ 是词汇 w_i 在测试文档集中的出现概率，$p(w_i,w_j)$ 表示词汇 w_i 和词汇 w_j 在测试文档集中的联合概率，V 为词汇列表的维数。由于词汇列表的维数较长，因此在本文中，对于每一个主题-词汇分布，只选取概率值最高的前 10 个词汇进行 PMI 值计算。

图 3 和图 4 展示了 LDA、GPU-DMM、NTM、DGPU-LDA 这 4 个主题模型在不同主题数 K 下的 PMI 值。

由图 3 中可知，在搜狗新闻数据集上，本文提出的 DGPU-LDA 的 PMI 值总体较高，表明其抽取出的主题具有较强的语义连贯性。传统的 LDA 模型由于没有考虑到文档以及词汇的语义强化，因此 PMI 值最低。NTM 是 LDA 模型的神经网络重构，但是并未考虑到词汇既语义的强化，因此 PMI 值相对于 GPU-DMM 以及 DGPU-LDA 均较低。GPU-DMM 融合了基于词嵌入的语义相似度强化，因此产生的主题具有一定的语义连贯性，但是相比于 DGPU-LDA 模型，并没有同时进行文档-主题层面的语义强化。

结合图 4 可知，上述结论在 20 新闻组数据集上仍成立。然而在该数据集上，各个模型的 PMI 值的差异并不如在搜狗新闻数据集上明显。而且，随着主题数目的增加，PMI 值也基本没有变化。这可能是因为中文主题建模与英文主题建模的差别，英文主题建模情景下往往需要一个更长的词汇列表。

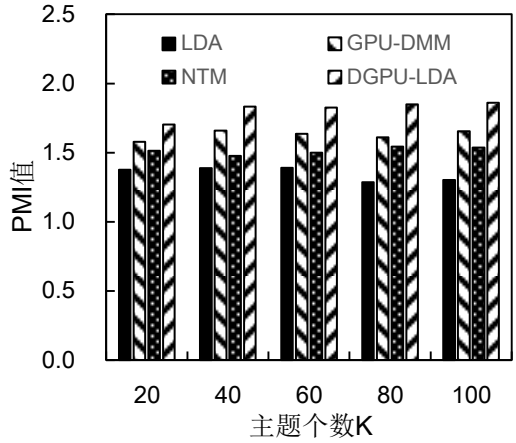


图 3 搜狗新闻数据集主题 PMI 值

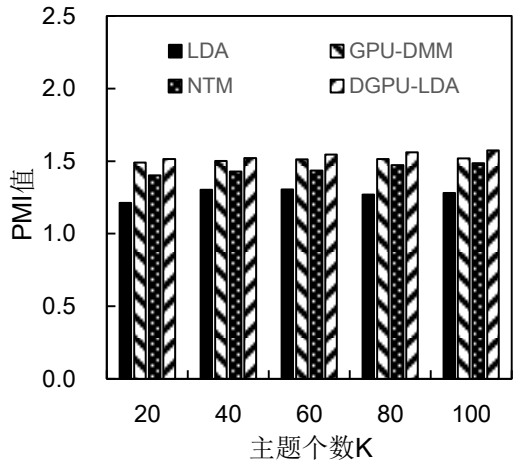


图 4 20 新闻组数据集主题 PMI 值

4.4 主题词效果展示

为了更直观地展现主题的语义连贯性，本文需要对各个主题模型中抽取出的主题的代表词汇进行分析。对于每个主题，本文只选取主题-词汇分布中概率值最高的前 5 个词进行展示。对于 NTM，本文采用二元词组进行展示。由于篇幅所限，本文只展示各个主题模型在搜狗新闻数据集上的“汽车”、“财经”和“体育”这 3 个类别下的主题代表词汇以及其中的与主题无关的冲突词（用加粗黑体标识），详细结果如表 2 所示，这里所有主题模型的主题个数 K 设置为 40。

表 2 主题代表词汇展示

	LDA	GPU-DMM	NTM	DGPU-LDA
汽车	产销 **标准** **酸性** 高档 **利率**	耐用 **参考** 合格 速度 广汽	**折价 减持** **阴雨 方法** 速腾 制动 **南方 讲述** 消费 群体	单车 车辆 消费 速度 搭载
冲突词数	3	1	3	0
财经	回暖 成本 **特例** 煤炭 **款**	地产 **合格** 股权 社会 利率	范围 下浮 **版本 现象** 有限 公司 行内 酒店 都市报 记者	下跌 冲击 股权 市场 **合格**
冲突词数	2	1	1	1
体育	**成本** 球队 **标准** 互动 收入	欧洲 宝贝 **合格** 体育 赛季	雅虎 体育 **欧元 上涨** 球员 巡回 **技术 层面** 宝贝 参与	体育 欧洲 记者 决赛 **合格**
冲突词数	2	1	2	1

由表 2 可知，在这 4 个模型中，LDA 所产生的主题可读性最差，每个主题下均包含大部分与主题无关的冲突词，例如“汽车”主题下的“标准”、“酸性”和“利率”均为冲突词。NTM 虽然能够通过二元词组增强主题的可读性，但是其没有显式地考虑词汇语义强化，冲突词组数目仍然偏多。而 GPU-DMM 以及本文提出的 DGPU-LDA 均表现良好，基本所有的主题代表词汇均与该主题所要表达的内容相关。相比于 GPU-DMM，本文提出的 DGPU-LDA 出现冲突词的情况更少，但是优势并不明显。造成该现象的原因可能是 DGPU-LDA 在文档-主题语义强化上的效果并不能直接体现在主题代表词汇的语义连贯性上，需要从文档层面进行评价才能得出进一步的结论。

4.5 文本分类效果评价

本节将 DGPU-LDA 模型用于文本分类任务中，以验证模型整体有效性。本文分两部分进行文本分类评价：一是对于本文提出的文档语义编码框架 DS-Bi-LSTM 在文档表征有效性方面的评价，二是对于主题模型整体在文本分类方面的准确性评价。采用的分类模型为支持向量机。

(1) DS-Bi-LSTM 文档编码有效性评价

DS-Bi-LSTM 文档语义编码框架是 DGPU-LDA 的基础，能够实现对于文档的语义嵌入表示。本文将训练集中的所有文档通过 DS-Bi-LSTM 进行嵌入表示，得到特征矩阵，进而将其用于文本分类中，得出其在测试集上的分类准确率。由于该编码框架是底层特征表示方法，因此本文选择传统的向量空间模型、M-Skip-Gram 作为基准方法。向量空间模型选取 5000 个词汇作为特征词汇，将文本表示成特征词汇上的词频向量。M-Skip-Gram 是将每个文档的所有词汇的词嵌入求平均得到整个文档的特征向量，词嵌入通过 Skip-Gram 模型构建。这三个特征表示方法下的文本分类准确率（每种特征表示方法下的分类实验重复 10 次，取均值和标准差）如表 3 所示。

表 3 不同特征表示方法下的文本分类效果

	向量空间模型	M-Skip-Gram	DS-Bi-LSTM
搜狗新闻数据集	0.545±0.012	0.662±0.015	**0.690±0.016**
20 新闻组数据集	0.577±0.023	0.634±0.014	**0.663±0.017**

根据表 3 的准确率结果可知，本文提出的 DS-Bi-LSTM 文档语义编码框架在文档特征表示方面相对于传统的向量空间模型和平均的词嵌入模型具有一定的优越性。DS-Bi-LSTM 文档编码框架的有效性为 DGPU-LDA 在文本分类上的优异表现打下了关键基础。

(2) 主题模型整体有效性评价

文本分类任务是主题模型整体评价的有效手段之一，文本分类准确率越高，则代表主题模型抽取出的主题的特征表达能力越强。图 5 和图 6 展示了 LDA、GPU-DMM、NTM、DGPU-LDA 这 4 个主题模型在不同主题数 K 下的文本分类准

确率。

由图 5 和图 6 可知，在两个数据集上，本文提出的 DGPU-LDA 模型的文本分类准确率均为最高：在搜狗新闻数据集上，主题数 $K=100$ 的情况下分别达到了 0.767 和 0.788；在 20 新闻组数据集上，主题数 $K=100$ 的情况下分别达到了 0.804 和 0.813。这说明基于双向 LSTM 语义强化的主题模型在文档特征表示方面具有较强的刻画能力。

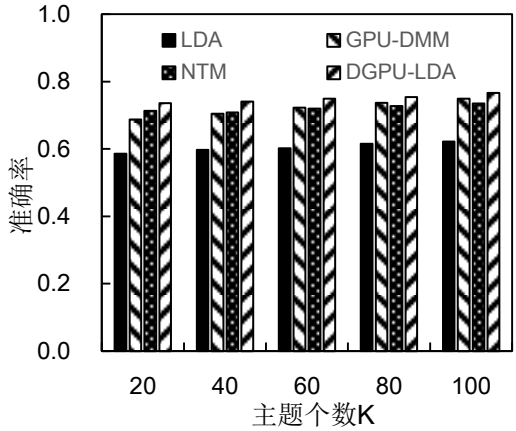


图 5 搜狗新闻数据集文本分类准确率

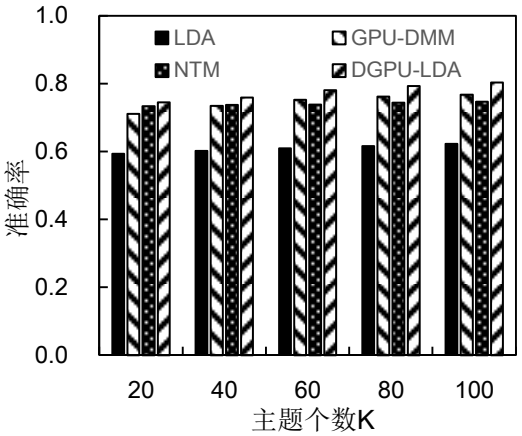


图 6 20 新闻组数据集文本分类准确率

5 总结

本文提出了一个基于双向 LSTM 语义强化的概率主题模型 DGPU-LDA。本文首先强调了文档宏观语义的重要性，并设计了一个基于双向 LSTM 的文档语义编码框架 DS-Bi-LSTM 来实现文档的语义嵌入表示。其次，将文档语义嵌入以及词嵌入分别用于吉布斯采样过程中的文档-主题 GPU 强化和词汇-词汇 GPU 强化。最后，将整个吉布斯采样的迭代过程用 LSTM 网络来刻画，从而推断出主题模型的参数。实验表明，DGPU-LDA 在主题语义连贯性和文本分类准确率方面均表现优异，能够利用好文档层面的语义信息。今后将改进 DS-Bi-LSTM 框架，通过知识库来扩展语义，使其适应于短文本建模场景。此外，还将研究 DGPU-LDA 的分布式训练算法，以适应于大规模训练集下的主题建模。

参考文献

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003 (3): 993-1022.

[2] Alex Graves, Jurgen Schmidhuber. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.

[3] Samuel Kotz, Hosam M. Mahmoud, Philippe Robert. On Generalized Polya Urn Models[J]. Statistics Probability Letters, 2000, 49(2): 163-173.

[4] Jeffrey L. Elman. Finding Structure in Time[J]. Cognitive Science, 1990, 14(2): 179-211.

[5] Sepp Hochreiter, Jürgen Schmidhuber. Long Short-term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[6] David Andrzejewski, Xiaojin Zhu, Mark Craven. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors[C]. In ICML'09, 2009: 25-32.

[7] Zhiting Hu, Gang Luo, Mrinmaya Sachan, Eric Xing, Zaiqing Nie. Grounding Topic Models with Knowledge Bases[C]. In IJCAI'16, 2016: 1578-1584.

[8] Zhiyuan Chen, Bing Liu. Topic Modeling Using Topics from Many Domains, Lifelong Learning and Big Data[C]. In ICML'14, 2014: 703-711.

[9] Zhiyuan Chen, Bing Liu. Mining Topics in Documents: Standing on the Shoulders of Big Data[C]. In KDD'14, 2014: 1116-1125.

[10] Geoffrey E. Hinton, Simon Osindero, Yee-Whye The. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[11] Hugo Larochelle, Stanislas Lauly. A Neural Autoregressive Topic Model[C]. In NIPS'12, 2012: 2708-2716.

[12] Nitish Srivastava, Ruslan Salakhutdinov, Geoffrey Hinton. Modeling Documents with a Deep Boltzmann Machine[C]. In UAI'13, 2013.

[13] Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, Heng Ji. A Novel Neural Topic Model and Its Supervised Extension[C]. In AAAI'15, 2015: 2210-2216.

[14] Fei Tian, Bin Gao, Di He, Tie-Yan Liu. Sentence Level Recurrent Topic Model: Letting Topics Speak for Themselves[C]. arXiv:1604.02038v2, 2016.

[15] Ilya Sutskever, Oriol Vinyals, Quoc V. Le. Sequence to Sequence Learning with Neural Networks[C]. In NIPS'14, 2014: 3104-3112.

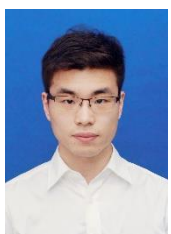
[16] Dzmitry Bahdanau, KyunghyunCho, Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate[C]. In ICLR'15, 2015.

[17] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, Xiaoming Li. Comparing Twitter and Traditional Media Using Topic Models[C]. In

ECIR’11, 2011: 338-349.
[18] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, Wei-Ying Ma. Topic Aware Neural Response Generation[C]. In AAAI’17, 2017: 3351-3357.
[19] Shuangyin Li, Yu Zhang, Rong Pan, Mingzhi Mao, Yang Yang. Recurrent Attentional Topic Model[C]. In AAAI’17, 2017: 3223-3229.
[20] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space[C]. CoRR, abs/1301.3781, 2013.
[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, Jeffrey Dean. Distributed Representations of Words and Phrases and Their Compositionality[C]. In NIPS’13, 2013: 3111-3119.
[22] Teng Long, Ryan Lowe, Jackie Chi Kit Cheung, Doina Precup. Leveraging Lexical Resources for Learning Entity Embeddings in Multi-relational Data[C]. In ACL’16, 2016: 112-117.
[23] Han Xiao, Minlie Huang, Xiaoyan Zhu. SSP: Semantic Space Projection for Knowledge Graph Embedding with Text Descriptions[C]. In AAAI’17, 2017: 3104-3110.
[24] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, Zongyang Ma. Topic Modeling for Short Texts with Auxiliary Word Embeddings[C]. In SIGIR’16, 2016: 165-174.
[25] Thomas L. Griffiths, Mark Steyvers. Finding Scientific Topics[C]. In PNAS’04, 2004: 5228-5235.
[26] David Newman, Jey Han Lau, Karl Grieser, Timothy Baldwin. Automatic Evaluation of Topic Coherence[C]. In HLT’10, 2010: 100-108.
[27] Jey Han Lau, David Newman, Timothy Baldwin. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality[C]. In EACL’14, 2014: 530-539.

彭敏（1973—），博士，教授，主要研究领域为自然语言处理。
E-mail：pengm@whu.edu.cn

杨绍雄（1993—），硕士研究生，主要研究领域为自然语言处理。
E-mail：yangshaoxxxxx@163.com



朱佳晖（1991—），硕士研究生，主要研究领域为自然语言处理。
E-mail：zhujiahui@whu.edu.cn