DOI:10.19651/j.cnki.emt.1802050

结合字词向量的主题向量模型*

张 青[1] 韩立新[1] 刘合兵[2]
(1.河海大学 计算机与信息学院 南京 211100; 2.河南农业大学 信息与管理科学学院 郑州 450046)

摘 要: 为了将已有的英文主题向量模型更好地应用于中文的主题向量训练,并且解决主题个数事先确定的缺点。本文将原有模型中,文档向量和词向量线性相加的方式改为内积的方式,并结合文档向量、字向量和词向量三者一起训练主题向量。当得到主题向量后通过聚类方法将相似的主题聚集在一起,以此来确定主题个数。实验表明,该方法训练出的主题词的相关性较原有模型和传统模型有所提升,并且能够获得较为合理的主题个数,同时,还能够得到词向量,主题向量和文档向量。

关键词: 主题模型;字向量;主题向量;词向量;文档向量;字词嵌入
中图分类号: TP183;TN01 **文献标识码**: A **国家标准学科分类代码**: 520.2040

Mixing topic models and character word embeddings to make lda2vec

Zhang Qing[1] Han Lixin[1] Liu Hebing[2]
(1. College of Computer and Information, Hohai University, Nanjing 211100, China;
2. College of Information and Management Science, Henan Agricultural University, Zhengzhou 450046, China)

Abstract: In order to better apply the original English topic vector model to the training of Chinese topic model vector, and solve the shortcomings of setting the topic number. This paper changes the linear addition of the document vector and the word vector in the original topic vector model to the inner product. and combines with the document vector, character vector and word vector to train the topic vector. When the topic vector is obtained, the similar topics are gathered together by the clustering method. Meanwhile, it can determine the number of topics. Experiments show that the relevance of the topic words trained by this method is improved compared with original and traditional model, and the number of themes can be obtained reasonably. At the same time, word vector, topic vector and document representation can be obtained.

Keywords: topic model; character vector; topic vector; word vector; document vector; character word embedding

0 引 言

主题模型被广泛地运用于文档的主题提取,与降维等方法不同的是,主题模型提取出的主题信息是可以解释,并且容易被理解的。传统的方法有潜在狄利克雷分布[1](latent dirichlet allocation,LDA),该模型假设一篇文章由若干主题组成,每个主题又由若干词按一定概率组成。在此基础上延伸出的改进模型有从语料中估计出主题个数的层级 LDA[2]、假设主题随着时间的变化而变化的时态主题模型[3]、半监督的主题模型[4]、利用变分自动编码去加速主题训练的模型[5]和利用神经网络去训练主题的模型[6]等。尽管针对主题模型有各种改进,但是绝大多数的主题模型都是词袋模型,没有较好地利用文本的语义和上下文信息,且在训练模型时需要事先设置主题个数,而主题个数的设置对训练出的主题质量有一定地影响,如果设置的主题个数太大则会使得训练出的主题交叉性较大。设置的太小则不能很好的训练出主题。

为了解决词袋模型的缺点,本文将引入能够捕捉语言中的语义和句法规则的潜在向量。目前,已经有很多研究中融入了潜在向量。Le 等[7]提出了段落向量;Kiros 等[8]提出了句子向量;Ghosh 等[9]构造上下文长短期记忆网络(long short-term memory,LSTM)去提取句子特征向量。Moody[10]将 LDA 和词向量(word2vector)相结合,在运用负采样方式进行词向量训练的模型[11]中,融入文档向量,即主题向量模型(lda2vec)模型。该方法可以很好地运用于英文文本的主题训练。相较于英文文本,在中文文本

收稿日期:2018-09-13
* 基金项目:河南省科技攻关项目(162102110120)资助

中，分词技术的好坏对词向量的训练的影响较大，为了降低由于分词带来影响并同时提高主题向量和词向量的训练结果，本文将在 lda2vec 模型中加入字向量，同时为了更适用于应用场景，在计算上下文向量时，将文档向量和词向量线性相加的方式改为文档向量与字词向量的内积方式。简而言之，本文提出了结合字向量、词向量和文档向量的主题向量训练方式。在训练模型时可以将主题个数的初始值设置相对大些，当训练好主题向量后通过聚类方法把相似的主题聚集在一起，从而可以得到较好的主题个数。同时，与 lda2vec 已有的采用将 Chainer 框架[12]进行训练的方式不同，本文采用 tensorflow 和 GPU 相结合，也很好地提高了模型训练的速度，并且避免了大量包库的安装。

1　lda2vec

lda2vec 模型将 word2vector 和 LDA 模型相结合。在训练词向量时，不直接使用词向量来预测上下文向量。而是使用词向量和文档向量一起来预测上下文向量，模型框架如图 1 所示[1]。

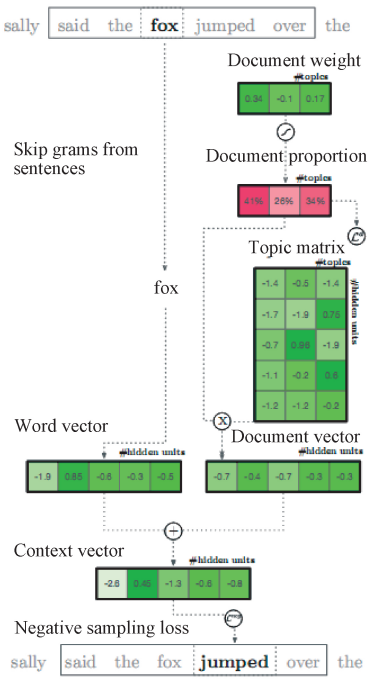

图 1　lda2vec 模型

其中，文档主题权重是文档有关于主题的 Dirichlet 分布，通过 softmax 函数转换成文档主题概率。再与主题向量相乘得到文档向量。而文档向量 $\vec{\boldsymbol{d}}_j$ 和单词向量 $\vec{\boldsymbol{w}}_j$ 相加，为文档中的每个单词生成上下文向量 $\vec{\boldsymbol{c}}_j$，如式(1)所示。

$$\vec{\boldsymbol{c}}_j = \vec{\boldsymbol{w}}_j + \vec{\boldsymbol{d}}_j \tag{1}$$

该方法的优势在于，当去寻找与词语“北京”最相似的词时，一般的 word2vec 模型很大概率会预测出“天津”、“南京”、“中国”等词。但如果该篇文章主题是与天气相关的，运用 lda2vec 模型，预测出的相似词还会包含“雾霾”、“沙尘暴”等，该方法能够很好的将词语与语境相结合。

lda2vec 不仅能学习单词的词向量，同时还能学习主题向量和文档向量。在 lda2vec 模型中，文档向量，主题向量和词向量三者的隐层维度是相同的。所以在得到主题向量后，只需要计算与该主题最相近的词向量即可得到该主题下的词。

lda2vec 的损失函数由两部分构成，一部分是负采样损失(skipgram negative sampling loss，SGNS)[11]，另一部分是有关于文档主题的 Dirichlet 分布。如式(2)所示。

$$L = \sum_{ij} L_{ij}^{neg} + L^d \tag{2}$$

$$L_{ij}^{neg} = \log_2\sigma(\vec{\boldsymbol{c}}_j \cdot \vec{\boldsymbol{w}}_i) + \sum_{l=0}^{n} \log_2\sigma(-\vec{\boldsymbol{c}}_j \cdot \vec{\boldsymbol{w}}_l) \tag{3}$$

$$L^d = \lambda \sum_{jk} (\alpha - 1)\log_2 p_{jk} \tag{4}$$

式中：$\sum_{ij} L_{ij}^{neg}$ 是 SGNS 损失，与利用负采样方式训练 word2vec 的损失函数是一致的，具体形式如式(3)所示；$\vec{\boldsymbol{c}}_j$ 是上下文向量；$\vec{\boldsymbol{w}}_j$ 是中心词向量；$\vec{\boldsymbol{w}}_i$ 是目标词向量；$\vec{\boldsymbol{w}}_l$ 是负采样的词向量。有关文档主题分布的损失如式(4)所示，p_{jk} 表示文章 j 属于主题 k 的概率。参数 λ 是调节负采样损失和文档主题分布损失的比重。当希望文档的主题分布是稀疏的，即文章很大概率的属于某些主题时，则将参数 α 设置为 0～1。而 $\alpha > 1$ 代表，文章等概率的选择主题。当 $\alpha = 1$ 时，则式(4)的损失函数为 0，即在训练词向量时，只考虑负采样的损失，等价于不利用文档信息的负采样词向量训练。

2　结合字向量的主题向量训练

在词向量训练时，大多数都是以词为最小单元进行训练，忽略了词语的内部组成。尤其对于中文，一个词语由多个字组成，每个词都含有丰富的内在信息。词的语义和其组成的字有很大的关联。而 lda2vec 最初提出来是运用于英文的主题向量训练，并不是很适用与中文的主题向量训练。Chen 等[13]在训练中文词向量时融入了字向量，并做了大量对比实验证明字词联合的词向量训练可以得到较好的结果。所以本文在主题向量训练的过程中也融入了字向量。即本文采用结合字向量 $\vec{\boldsymbol{ch}}$，词向量 $\vec{\boldsymbol{w}}$ 和文档向量 $\vec{\boldsymbol{c}}$ 的主题向量训练方式。模型如图 2 所示。

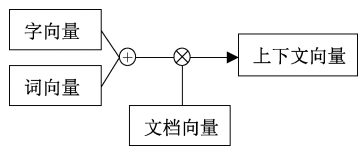

图 2　融合字向量的主题向量模型

在 lda2vec 模型中词向量和文档向量是通过相加得到上下文向量的。更有理由认为，文档向量是该文档在主题

的每个隐层的权重,与词向量的每个隐层进行内积更具有说服性。实验结果也证明了这一点。

在训练词向量时,把组成词语的汉字单独取出来,与词向量和文档向量一起进行训练。这样就使得那些共享汉字的词语之间产生了联系。而针对类似于沙发、巧克力等非合成词,导入了文献[13]中的词典库,对于这些词语不去进行单个字的拆分处理。

2.1　上下文向量

在用负采样方式训练上下文向量 $\vec{c}_j$ 时,它由词语向量 $\vec{w}_j$,单个汉字向量 $\overrightarrow{ch}$ 和文档向量 $\vec{d}_j$ 相结合的得到,如式(5)所示。

$$\vec{c}_j = (\vec{w}_j + \overrightarrow{ch}) \cdot \vec{d}_j \tag{5}$$

以上这 4 个向量隐层维度都是一样的。将词向量和字向量相加得到的向量称为字词向量。字词向量 $\vec{x}_j$ 的计算方法如式(6)所示。

$$\vec{x}_j = \gamma \cdot \vec{w}_j + (1-\gamma)\frac{1}{N_j}\sum_{k=1}^{N_j}\overrightarrow{ch}_k \tag{6}$$

式中:字词向量是由字向量和词向量按照一定的比列相加得到的。N_j 是词 w_j 中包含字的个数;$\sum_{k=1}^{N_j}\overrightarrow{ch}_k$ 是词语中的汉字向量等权相加。由于同一个汉字,在不同的词语中可能会有完全不同的语义,如果使用一个向量来表征一个字,那么可能无法标志出这些差异性,所以本文采用多个向量来表征同一个汉字。使用基于位置的字向量嵌入,即同一个汉字根据其在词语中出现的位置不同,对应不同位置的向量表现形式。因为考虑到一般情况下词语包含两个汉字,为了节省空间,本文将位置分为了非第一个和非最后一个词两个位置,实验结果表明,与将字的位置分为前中后 3 种的模型,可以取得一样好的效果。

2.2　文档向量

文档向量 $\vec{d}_j$ 是由潜在主题向量 $\vec{t}_0,\vec{t}_1,\cdots,\vec{t}_k$ 和文档主题的 Dirichlet 计算得到,如式(7)所示。

$$\vec{d}_j = p_{j0}\vec{t}_0 + p_{j1}\vec{t}_1 + \cdots + p_{jk}\vec{t}_k + \cdots + p_{jn}\vec{t}_n \tag{7}$$

式中:$\vec{p}_{jk}$ 是在文档 j 中主题 k 所占比重,由主题的 Dirichlet 分布通过 softmax 函数后计算得到,得到的 $0 < \vec{p}_{jk} < 1$,并且 $\sum_k \vec{p}_{jk} = 1$。当训练得到主题向量时,可以通过获取与主题向量最相关的前 N 个词,作为该主题下的词。当有任意两个主题中包含的词重复性较大或者主题向量之间的距离较小时,可以合并这两个主题。因此只要将主题个数设置的大些,并且通过合并比较相近的主题,就能够在无监督的语料训练中得到较好的主题个数。避免了现在大多数主题模型中主题个数的选择。

2.3　损失函数

本文模型的损失函数同 lda2vec 一样,由两部分构成,一部分是负采样的损失函数,一部分是如式(2)所示有关文档主题分布的损失函数,不同的是在计算公式中上下文向量 $\vec{c}_j$ 时融入了字向量。在式(3)中有关负采样词个数 n 的选择,也会对训练出来的主题有影响。一般会选择训练样本中文本词总个数 u 的 β 指数倍,即 u^{β}。通过实验,最终选择的 $\beta = 0.7$。模型的训练采用的是 Adam 方法进行优化。

3　实　　验

3.1　实验数据

实验采用清华的中文语料。在一般的主题模型中语料的清洗的好坏对于实验结果很重要。如果清洗的不好,就会造成有些高频词几乎出现在每一个主题中。但是在本文的主题向量模型中,不进行清洗也可以训练出好的主题。清华语料中共有 14 大类,共有 768 415 篇文章,利用结巴进行分词后得到 1 264 675 个词,由于文本太大,本文在每个类别中选择 50 000 篇进行训练,针对没有 50 000 篇的类别对该类别下的文章进行重复随机采样以达到 50 000 篇。最终得到 70 万篇文章,867 398 个词。为了验证本文在清洗语料后效果的好坏,对全部的清华语料进行清洗,去掉停用词和出现频率低于 10 的词,最终得到 216 483 个不同的词。

3.2　评估方法

采用的是归一化点互信息(normalized pointwise mutual information, NPMI)[14-15],其计算方法如式(8)所示。

$$NPMI(t) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{\log_2\dfrac{p(w_i,w_j)}{p(w_i)p(w_j)}}{-\log_2 p(w_i,w_j)} \tag{8}$$

在 NPMI 评价方法中,$NPMI(t)$ 表示主题 t 的相关性,取主题 t 的 top-N 主题词集 $w_1,w_2,\cdots,w_N$ 进行计算,最后取 K 个主题的平均主题相关性作为评估指标,平均主题相关性越大,表示模型越优。

3.3　结果分析

1)主题相关性对比实验

为了验证提出模型的准确度,本文选取了以下几个流行模型进行对比实验。

LDA 是 Blei 等[1]提出的挖掘文本语料隐含主题的方法,使用 Gibbs 采样法进行后验推断。

NVDM 是 Miao 等[16]提出的用神经网络推断后验一种方式。其中隐层结点为 100,学习率为 0.05,迭代次数为 100。

ProLDA 是 Srivastava 等[5]提出的基于变分贝叶斯自动编码器主题模型,同样采用变分自动编码器进行推断学习,设置了 2 层网络,隐层结点一共有 200 个,学习率为 0.005。

lda2vec 是结合词向量和文档向量的主题向量训练模型。

本模型中超参数 β 设置为 0.7,初始化主题个数设置为 100,当两个主题向量的相似度大于 0.9 时,将这两个主

题进行合并。最终得到 47 个主题，而其他模型中主题个数均设置为 100 个。实验结果用主题相关性进行分析。同时，将本文模型记为 C-lda2vec，结果如表 1 所示。从表 1 可以看出，本文的模型优于传统的 LDA 模型和本文参考的原始 lda2vec 模型。

表 1　平均主题相关性

C-lda2vec	lda2vec	LDA	NVDM	ProLDA
0.614	0.608	0.156	0.091	0.225

2)β 对主题相关性的影响

进行负采样时，超参数 β 的选择会对实验结果有所影响。β 的设置对主题相关性的影响如表 2 所示，当 β 选择 0.7 时，能够得到较好的主题相关性。

表 2　β 对主题相关性的影响

β	0.5	0.6	0.7	0.8	0.9
主题相关性	0.594	0.597	0.614	0.611	0.604

3)主题词质量查看

为了查看主题词的质量，在实验中选取 5 个主题进行显示，每个主题选择与该主题向量最相似的前 10 个单词进行显示，如表 3 所示。第 1 行是主题标签，每一列对应着每个主题下的词。通过观察，可以得知每一个主题代表着一个类别，且主题之间的交叉性较小。在清华语料中，福彩和体彩都属于彩票这一类，而本文的模型可以将其区分开来，并且“中奖”一词同时属于这两个类别，这一结果也是合理的。同样的，在体彩与体育，体育与教育也有类似的情况。同时，每个主题下的词语相关性比较大。由此可见，本文采用的方法获得了一定的效果。

表 3　主题词

福彩	体彩	体育	教育	时政
开奖	叫停	球员	文化	议论
中奖	足彩	球队	德育	政治
日期	私彩	比赛	教师	时事
七乐彩	胜负彩	球迷	学前教育	言行
双色球	大乐透	林书豪	义务教育	中国梦
奖池	竞彩	中超	体育	时弊
号码	足球	皇马	教学	理论
走势	中奖	奥运	学习	创新
金额	半全彩	欧冠	高等教育	社会
刮刮乐	进球彩	尼克斯	国民教育	精神

从以上实验可以得知，在中文语料中训练主题向量时融入字向量，得到的结果从主题的相关性和主题词的质量上都有所改善。并且不管语料事先是否清洗，都能得到较好的主题词。同时，实验得到的文档向量还可以用于文本分类。

4　结　　论

本文在 lda2vec 模型中加入字向量，将其更加适用于中文的主题向量训练。并通过相似主题向量的合并，得到合理的主题个数，避免了大多数主题模型需要认为设定主题个数的缺点。实验表明，本文提出的模型优于 lda2vec 模型，在主题相关性上得到了较理想的结果。未来的工作将考虑在该模型中加入时间序列和在由字向量组成词向量时，针对在同一个词中不同的字的权重可以加入 attention 机制。

参考文献

[1] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2012(3):993-1022.

[2] HUANG T, LI L, ZHANG Y. Multilingual multi-document summarization with enhanced hLDA features [C]. China National Conference on Chinese Computational Linguistics. Springer International Publishing, 2016:299-312.

[3] ZHANG H, ZHANG X, TIAN Z, et al. Incorporating temporal dynamics into LDA for one-class collaborative filtering[J]. Knowledge-Based Systems, 2018,150(6): 49-56.

[4] KANG D, PARK Y, CHARI S N. Hetero-labeled LDA: A partially supervised topic model with heterogeneous labels[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, Heidelberg, 2014.

[5] SRIVASTAVA A, SUTTON C. Autoencoding variational inference for topic models[C]. ICLR 2017, 2017:1-12.

[6] NGUYEN D Q, BILLINGSLEY R, DU L, et al. Improving topic models with latent feature word representations[J]. Transactions of the Association for Computational Linguistics, 2015(3):299-313.

[7] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]. International Conference on Machine Learning, JMLR.org, 2014: 1188.

[8] KIROS R, ZHU Y, SALAKHUTDINOV R, et al. Skip-thought vectors[C]. Computation and Language, 2015.

[9] GHOSH S, VINYALS O, STROPE B, et al. Contextual LSTM (CLSTM) models for large scale NLP tasks[C]. Computation and Language, 2016.

[10] MOODY C E. Mixing dirichlet topic models and word embeddings to make lda2vec[J]. CoNLL 2016.

[11] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]. Computation and Language, 2013:1-9.

[12] SEIYA T,KENTA O,SHOHEI H. Chainer: A next-generation open source framework for deep learning[J/OL].learningsys.org.

[13] CHEN X, XU L, LIU Z, et al. Joint learning of character and word embeddings [C]. International Conference on Artificial Intelligence, AAAI Press, 2015:1236-1242.

[14] MIKOLOV T, SUTSKEVER I, KAI C, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.

[15] LAU J H, NEWMAN D, BALDWIN T. Machine reading tea leaves: automatically evaluating topic coherence and topic model quality[C]. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014: 530-539.

[16] MIAO Y, YU L, BLUNSOM P. Neural Variational Inference for Text Processing[J]. Computer Science, 2015:1791-1799.

作者简介

张青,1993年出生,硕士研究生,主要研究方向为信息检索、自然语言处理。

E-mail:1547293301@qq.com

韩立新,博士、教授,主要研究方向为信息检索、自然语言处理、数据挖掘等。

刘合兵,博士研究生、副教授,主要研究方向为信息检索、数据挖掘。