文章编号：2095-2163(2020)03-0058-06 中图分类号：TP391.1 文献标志码：A

基于 Multi-TWE 模型的短文本分类研究

王云云，张云华
（浙江理工大学 信息学院，杭州 310018）

摘 要：针对目前词向量无法解决短文本中一词多义的问题，提出融合词向量和 BTM 主题模型的 Multi-TWE 多维主题词向量模型。将 BTM 模型训练得到目标词与相应主题进行不同方式的连接，形成多维主题词向量来表示多义词词义，最后将 Multi-TWE 模型应用于短文本分类，提出基于 Multi-TWE 模型的短文本分类方法，与 SVM、BTM 和 Word2Vec 分类方法进行对比实验，实验结果表明本文提出的短文本分类方法在平均 F_1 值上比前三种方法分别提升了 3.54%、11.41%和 2.86%。

关键词：短文本分类；一词多义；BTM 主题模型；词向量

Research on short text classification based on multi-TWE model

WANG Yunyun, ZHANG Yunhua
(School of Informatics Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

【Abstract】 Aiming at the current problem that word vectors cannot solve the problem of polysemy in short texts, a Multi-TWE multi-dimensional topic word vector model combining word vectors and BTM topic models is proposed. The BTM model is trained to connect the target word with the corresponding topic in different ways to form a multi-dimensional topic word vector to represent the meaning of the polysemous word. Finally, the Multi-TWE model is applied to short text classification, and a short text classification method based on the Multi-TWE model is proposed. Compared with the SVM, BTM, and Word2Vec classification methods, the experimental results show that the short text classification method proposed in this paper improves the average F_1 value by 3.54%, 11.41%, and 2.86% compared with the previous three methods, respectively.

【Key words】 short text classification; polysemy; BTM topic model; word vector

0 引 言

短文本是人们生活信息口语化在互联网上的体现。顾名思义，短文本字数少、篇幅小，导致在分析短文本时，很难准确地分析出短文本的语义信息，并且有不少的词语具有多种词义和词性，会根据不同的使用场景表达出不同的语义[1]，这更加剧了短文本分析的难度。

2013 年，Mikolov 等人[2]提出的 word2vec 模型，利用上下文语义关系将词语映射到一个低维稠密的空间，使相似词语在空间中的距离相近，通过空间位置获得对应的词向量表示[3]。Zhu 等人[4]在使用词向量来表示文本向量的基础上，融合了改进的 TF-IDF 算法，并利用 SVM 分类器进行短文本分类。Yao 等人[5]使用词向量来表示新闻标题类短文本，并通过判断语义相似度来扩展文档表示。以上研究表明词向量模型应用在文本表示上的可行性。在使用词向量进行词义消歧方面，Niu 等人[6]融合了知网的义原信息和注意力机制，实现自动地根据上下文选取合适的词语词义的方法，在判断语义相似度和词义消歧方面取得了更好的效果。Liu 等人[7]提出将词向量与主题模型相结合，组成主题向量用于词语消歧。曾琦等人[8]提出了一种将多义词的不同语义用不同主题来表示，最后训练多义词词向量。深度学习方法中利用词向量训练的便捷性与主题模型能挖掘主题语义的这一能力相结合，既能保证准确率又能降低邻域依赖。本文利用这一复合方法进行一词多义的研究。

1 相关工作

1.1 BTM 主题模型

由于传统的主题模型是获取文档级别的词共现[9]，短文本的数据稀疏性导致传统主题模型效果不好。针对这一问题，Yan 等人[10]提出了 BTM 主题模型，来进行短文本建模。BTM 通过语料级别的词共现来为短文本建模。设有语料库 L，语料库 L 中有一个二元词组集合 $|B|$，表示语料中所有的词对，图模型如图 1 所示。

图 1 中，$b=(b_i,b_j)$ 表示其中的任一词对，b_i,b_j 分别表示词对中的词语，z 表示词语的主题，K 表示

作者简介：王云云(1992-)，女，硕士研究生，主要研究方向：软件工程技术；张云华(1965-)，男，教授，主要研究方向：软件工程。
通讯作者：王云云 Email：1528723134@qq.com

收稿日期：2019-12-28

主题数目，$z \in [1,K]$，θ 表示每篇文档的主题分布，φ 表示不同主题下的词分布，两者皆服从狄利克雷分布。α 和 β 分别是两者的先验参数。

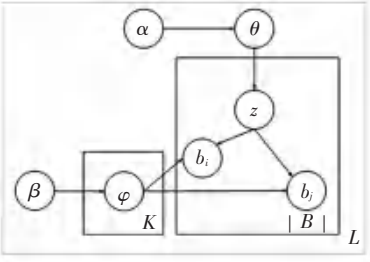

图 1　BTM 主题模型的图模型结构

Fig. 1　Graph model structure of BTM topic model

1.2　SE-WRL 词向量模型

SE-WRL 模型是由 Niu 等人[6]提出，该模型是在词汇表示学习模型中融入词义的义原知识，是一种基于 Skip-gram 模型的改进模型。SE-WRL 模型充分考虑词语的义原信息，使用义原信息帮助模型更好地理解词语语义。具体做法是，根据上下文词语来对中心词做词义消歧，使用注意力机制（attention）计算上下文对该词语各个词义的权重，然后使用语义向量的加权平均值表示词向量。SE-WRL 模型从 3 个角度融入义原信息，提出了 3 种不同的策略模型：

（1）简单义原聚集模型（SSA）。对于每一个词，SSA 模型把表示该词语义的所有义原都考虑进来。原本是用目标词的上下文向量来表示当前词，该模型是用目标词的所有上下文的义原向量的平均值来表示当前词，即词向量形式如下：

$$w = \frac{1}{m} \sum_{s_i(w) \in S(w)} \sum_{x_j(s_i) \in X_i(w)} x_j(s_i), \tag{1}$$

其中，m 表示所有属于词 w 的义原数量。

（2）基于上下文的义原注意力模型（SAC）。该模型结构如图 2 所示。

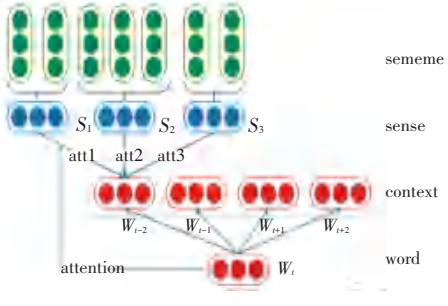

图 2　SAC 模型结构

Fig. 2　SAC model structure

由图 2 可以看出，SAC 模型中目标词 w 使用了原始的词向量，上下文使用义原向量来表示上下文词向量 w_c。模型使用目标词的词向量作为一个注意力，去选取更合适的语义来生成上下文词向量。形式化上下文词向量 w_c。此时需用到如下数学公式：

$$w_c = \sum_{j=1}^{|S(w_c)|} att(s_j(w_c)) \cdot s_j(w_c), \tag{2}$$

其中，$s_j(w_c)$ 表示 w_c 的第 j 个语义向量，$att(s_j(w_c))$ 表示关于目标词 w 的第 j 个语义的注意力得分。

（3）基于目标词的义原注意力模型（SAT）。该模型结构如图 3 所示。SAT 为上下文词语学习原始的词向量，但是为目标词学习义原向量。模型在目标词 w 的多个语义上，应用上下文词语作为注意力，来建立 w 的向量，公式如下：

$$w = \sum_{j=1}^{|S(w)|} att(s_j(w)) \cdot s_j(w), \tag{3}$$

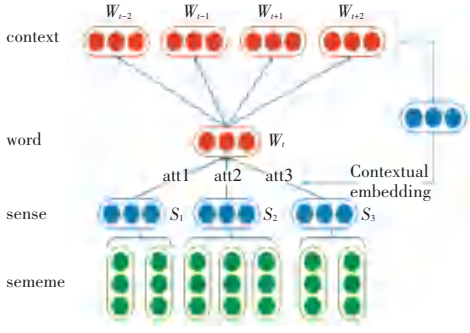

图 3　SAT 模型结构

Fig. 3　SAT model structure

这时，w'_c 是由 $C(w_i)$ 中受限窗口内的词向量组成的上下文向量，公式如下：

$$w'_c = \frac{1}{2K'} \sum_{k=i-K'}^{k=i+K'} w_k, k \neq i. \tag{4}$$

2　基于 Multi-TWE 模型的短文本分类研究

传统词向量模型无法很好地处理汉语中存在的一词多义问题，主要是因为词向量模型对于多义词中各种语义信息的处理不敏感，训练出的单一词向量容易混淆多义词的含义。2015 年，Liu 等人[7]提出了主题词向量的概念，即将主题融入到基本的词向量表示中，并允许由此产生的主题词向量在不同语境下获得一个词的不同含义。

根据上述思想，本文将主题词向量的概念应用到短文本语义挖掘中，本文的算法使用基于义原信息和注意力机制的 SE-WRL 词向量模型来训练词

向量,该模型使用注意力机制在一定程度上能够消除多义词的影响,但由于短文本本身具有的上下文特征稀疏性,SE-WRL 词向量模型在短文本上的应用效果有限。因为 BTM 主题模型能够有效地解决短文本的特征稀疏的问题,因而,本文引入了 BTM 主题模型来进行短文本的语义挖掘,提出了一种 Multi-TWE 多维主题词向量算法。

首先对于处理好的短文本语料进行 BTM 主题模型初始化,通过吉布斯采样过程获取词和主题,利用 SE-WRL 词向量模型分别进行向量的训练,得到不同的主题词向量,达到词义消歧的效果,实现文本分类。该算法框架包含 MuTWE-1 和 MuTWE-2 这两种主题词向量模型,接下来将具体分析 MuTWE-1 和 MuTWE-2 这两个模型算法。

2.1 MuTWE-1 主题词向量模型算法

MuTWE-1 模型算法具体的参数推理分为 2 步,分别是:BTM 模型参数推理和 MuTWE-1 主题词向量训练,具体算法流程如图 4 所示。首先对 BTM 主题模型进行参数推理,这里通过使用吉布斯采样方法抽取词对 b 和每个词对相对应的主题 z,然后将词对 b 和主题词 z 组合成伪词 (b,z),融入 SE-WRL 训练模型中,最后得到主题词向量 $W(b,z)$。

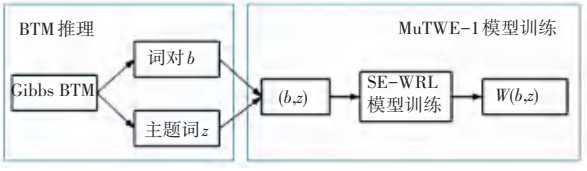

图 4 MuTWE-1 模型算法流程图

Fig. 4 MuTWE-1 model algorithm flowchart

2.2 MuTWE-2 主题词向量模型算法

MuTWE-2 模型算法的参数推理同样分为 2 步:分别是:BTM 模型参数推理和 MuTWE-2 主题词向量训练,与 MuTWE-1 模型算法的主要区别在于主题词向量的模型训练部分,MuTWE-1 是将伪词作为模型的输入,得到的是“伪词”向量,而 MuTWE-2 将词语和主题分开进行训练,分别得到词语向量和主题向量,再将两者进行连接得到最终的多维主题词向量,这里的词语向量与主题向量的长度不需要相同。MuTWE-2 模型算法框架如图 5 所示。

相同词的词向量一致,主要是表示不同语义的主题向量不同。对于模型训练得到的词向量和主题向量通过公式(5)进行连接:

$$b^z = b \oplus z. \tag{5}$$

其中,$\oplus$ 是连接运算,b^z 的长度是词向量 b 和主题向量 z 的长度之和。这种连接的方式可以降低向量的维度,缓解数据的维度灾难。

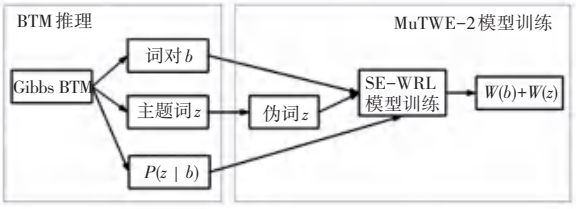

图 5 MuTWE-2 模型算法框架图

Fig. 5 MuTWE-2 model algorithm framework

2.3 基于 Multi-TWE 算法的短文本分类方法处理

应用以上改进模型的具体短文本分类方法的构建流程如图 6 所示。

Multi-TWE 算法模型是一种基于词粒度的语义理解模型,即最后的训练结果是主题词向量,若想将其应用于短文本分类,还需要转化为文本粒度的向量表示,因为短文本中的词语对于最终生成的短文本向量的贡献度各不相同,所以本文采用加权和取平均的方式计算文本向量。这里借助 TF-IDF 权重算法计算每个词语的贡献值。权重 ω_i 的定义可表示为:

$$\omega_i = \frac{tf_{i,j} \times idf_i}{\sqrt{\sum_{t_i \in d_j} [tf_{i,j} \times idf_i]^2}}, \tag{6}$$

其中,$tf_{i,j}$ 表示特征 t_i 在文本 d_j 中的出现频率,idf_i 表示特征 t_i 的逆向文件频率 IDF,具体公式可写为:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{7}$$

$$idf_i = \log \frac{|D|}{|\{j:t_i \in d_i\}|}, \tag{8}$$

综上,短文本向量的求解公式如下:

$$v_d = \frac{1}{N} \sum_{b^z \in d} \omega \cdot b^z. \tag{9}$$

其中,N 表示短文本训练出的主题词向量的个数,ω 表示权重因子。

得到所有短文本特征向量后,将其输入到分类器中,从而对分类器进行训练与测试。

3 实验及结果分析

3.1 实验数据

采用搜狗实验室提供的中文新闻标题文本分类语料来进行短文本分类研究。

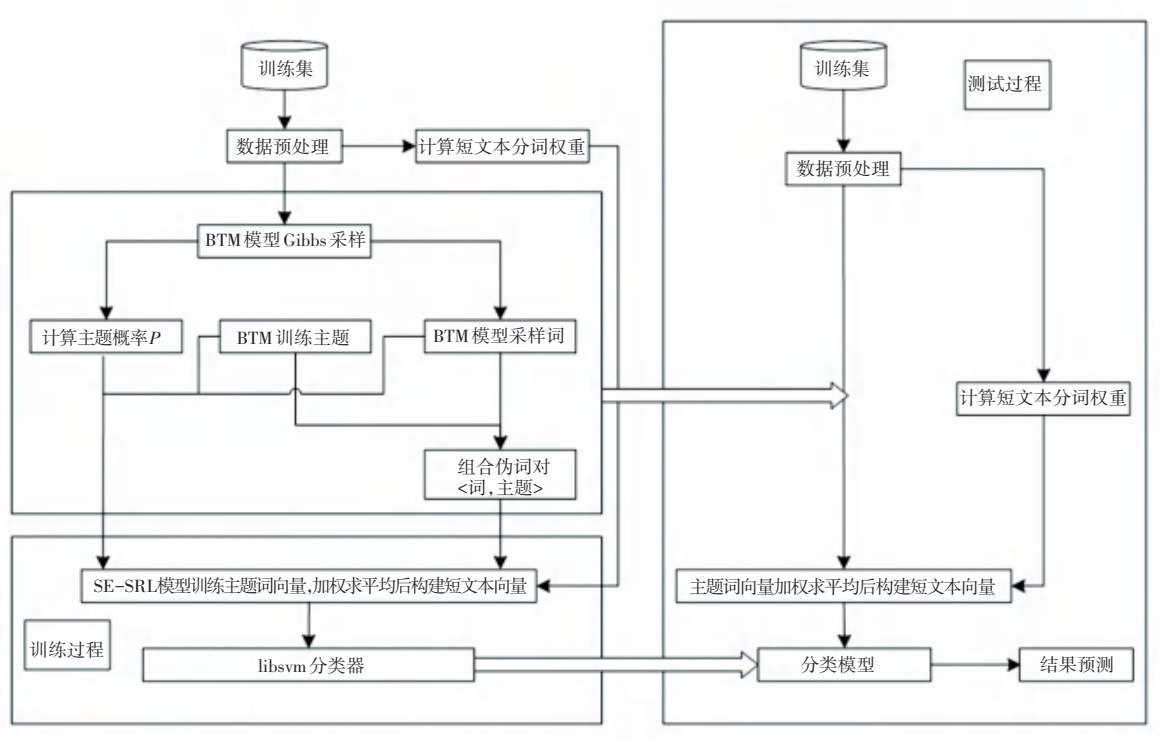

图 6　短文本分类方法的构建流程图

Fig. 6　Construction flowchart of short text classification method

该数据集中的新闻标题的长度主要集中于 10~30 字之间，所以该数据集中的新闻标题很适合作为短文本分类的研究对象。数据集共涉及了 9 个邻域，每篇文本内容都包括网址、标题、内容等，只抽取其中的新闻标题部分，最终获得的新闻文本数据分布见表 1。其中，每个类别的文档按照 80%作为训练数据，20%作为测试数据。

表 1　新闻文本数据分布

Tab. 1　News text data distribution

编号	领域	文档数据/个
1	娱乐	2 142
2	体育	2 065
3	财经	1 432
4	IT	2 268
5	军事	926
6	汽车	2 038
7	房产	2 003
8	健康	2 260
9	教育	1 904
总计		17 038

3.2　文本分类评估标准

文本分类任务中常用的评价指标有准确率 $P(precision)$、召回率 $R(recall)$ 以及调和平均值 F_1，其数学定义如下所示：

$$P = \frac{A}{A + B}, \tag{10}$$

$$R = \frac{A}{A + C}, \tag{11}$$

$$F_1 = \frac{2PR}{P + R}. \tag{12}$$

其中，各参数含义解析详见表 2。

表 2　分类评价标准参数含义表

Tab. 2　Meanings of parameters for classification evaluation criteria

	分类为 T 类	分类为非 T 类
实际为 T 类	A	C
实际为非 T 类	B	D

3.3　主要相关参数分析

研究中拟分析的参数包含：BTM 的主题数目、向量维度以及向量窗口大小等。用 F_1 作为评价标准。实验结果如图 7~图 9 所示。

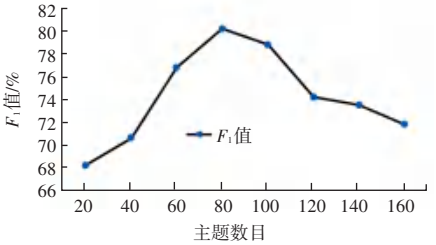

图 7　主题数目参数估计

Fig. 7　Parameter estimation of the number of topics

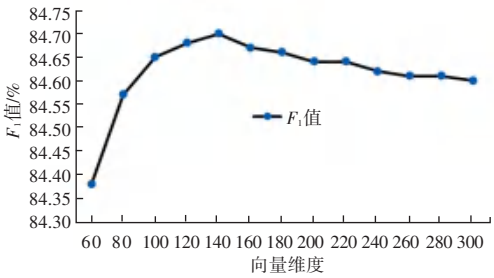

图 8 向量维度参数估计

Fig. 8 Vector dimension parameter estimation

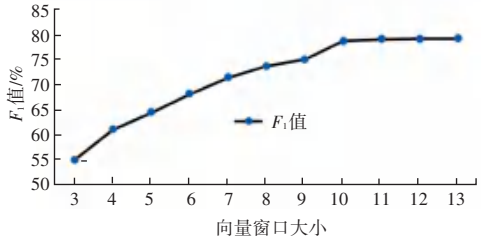

图 9 向量窗口大小参数估计

Fig. 9 Vector window size parameter estimation

分析图 7~图 9 后得出：

(1)由图 7 得出当主题数在 80 时 F_1 值取最大值。

(2)由图 8 得出词向量维度在 100~160 左右最有利，为了平衡各模型，最终将向量维度设置为 150，主题词向量的维度设置为 300。

(3)由图 9 看出，当窗口大小为大于 10 时，窗口长度的增长对于 F_1 值的增长并没有什么帮助，所以将实验的最佳向量窗口大小设置为 10。

3.4 分类对比实验

为了验证本文提出的基于 Multi-TWE 算法模型的短文本分类方法的有效性，分别选取 VSM 模型、BTM 主题模型和 TF-IDF 加权 word2vec 模型作为对比实验。所有分类方法选用 libsvm 作为分类器。实验采用五折交叉验证来评估各模型分类效果，测试结果见表 3。

表 3 对比实验测试结果

Tab. 3 Comparative experimental test results %

	VSM			BTM			Word2Vec			Multi-TWE		
	P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
娱乐	87.59	82.72	85.09	77.92	82.01	79.91	88.36	86.19	87.26	91.04	86.17	88.54
体育	78.92	70.06	74.23	81.39	71.30	76.01	84.73	82.44	83.57	89.98	86.71	88.31
财经	81.35	79.22	80.27	80.99	64.29	71.68	83.21	82.42	82.81	89.04	83.00	85.91
IT	82.74	88.37	85.46	90.57	80.58	85.28	90.55	84.37	87.35	90.79	87.11	88.91
军事	83.55	87.09	85.28	86.71	56.78	66.62	91.34	88.40	89.85	94.09	81.84	87.54
汽车	84.21	86.00	85.10	72.00	61.01	66.05	87.61	82.66	85.06	87.25	80.33	83.65
房产	87.03	76.33	81.33	59.74	87.93	71.14	80.85	78.49	79.65	80.36	85.75	82.97
健康	78.69	86.27	82.31	74.31	76.54	75.41	73.30	71.26	72.27	80.93	84.44	82.65
教育	85.32	74.57	79.58	70.87	81.28	75.72	79.24	74.77	76.94	83.51	80.58	82.02
avg	83.27	81.18	82.07	77.17	73.52	74.20	84.35	81.22	82.75	87.44	84.00	85.61

F_1 值对比图显示了 4 种分类算法在新闻标题短文本数据集上短文本分类评估结果对比，4 种算法在各类别上的 F_1 值如图 10 所示。由图 10 中可以看出：

(1)VSM 模型在文本分类上的表现一直很稳定，BTM 主题模型算法的分类效果表现最差。

(2)基于 Word2Vec 的文本分类方法在短文本分类上的表现优于 VSM 模型和 BTM 模型，证明了 Word2Vec 模型也能够很好地应用于短文本分类。

(3)本文的分类方法的分类效果明显优于前三种，在平均 F_1 值上分别提升了 3.54%、11.41% 和 2.86%。并且在“娱乐”、“体育”、“IT”、“教育”等多类新闻标题中取得了最高 F_1 值，充分证明了本算法模型具有很不错的短文本分类能力，这主要得益于多维主题词向量能够识别多义词在特定主题下的含义，增大了短文本向量的语义区分能力，从而提高了短文本分类的效果。

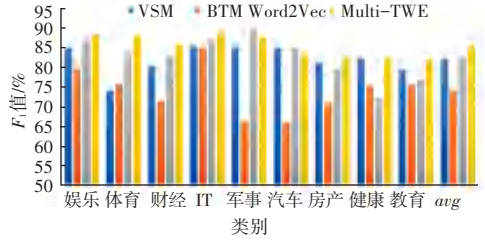

图 10 各类别 F_1 值对比图

Fig. 10 Comparison of F_1 values of each category

4 结束语

针对一词多义问题，本文提出了融合词向量和

(下转第 68 页)