计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191000163

一种低频词词向量优化方法及其在短文本分类中的应用

程　婧[1,2]　刘娜娜[1,2]　闵可锐[3]　康　昱[4]　王　新[1,2]　周扬帆[1,2]
1 复旦大学计算机科学技术学院　上海 201203
2 上海市智能信息处理重点实验室　上海 201203
3 上海市秘塔网络科技有限公司　上海 200135
4 微软亚洲研究院　北京 100080
(jcheng17@fudan.edu.cn)

摘　要　众多自然语言处理(Natural Language Processing,NLP)任务受益于在大规模语料上训练的词向量。由于预训练的词向量具有大语料上的通用语义特征,因此将这些词向量应用到特定的下游任务时,往往需要通过微调进行一定的更新和调整,使其更适用于目标任务。但是,目标语料集中的低频词由于缺少训练样本,导致在微调过程中无法获得稳定的梯度信息,使得词向量无法得到有效更新。而在短文本分类任务中,这些低频词对分类结果同样有着重要的指示性。因此,在具体的短文本分类任务上获得一个更好的低频词词向量表示是有必要的。针对这个问题,文中提出了一种与下游任务模型无关的低频词词向量更新算法,通过基于K近邻的词向量偏移计算方法,利用通用词向量中与低频词相似的高频词所获得的任务特征信息,来指导低频词的信息更新,从而获得更准确的且适用于当前任务语境的低频词词向量表示;并以TextCNN作为基准模型,基于word2vec和GloVe得到的两个通用预训练词向量,在3个公开的短文本数据集上进行了优化算法的效果验证。实验结果表明,使用优化算法更新低频词词表示后,模型分类准确率能达到84.3%～94%,较更新前提升了0.4%～1.4%,体现了优化算法的有效性,也进一步证明了短文本分类任务中低频词对分类结果的影响,为短文本分类的研究工作提供了一定的借鉴。

关键词:词向量;低频词;微调;短文本分类
中图法分类号　TP391

Word Embedding Optimization for Low-frequency Words with Applications in Short-text Classification

CHENG Jing[1,2],LIU Na-na[1,2],MIN Ke-rui[3],KANG Yu[4],WANG Xin[1,2] and ZHOU Yang-fan[1,2]
1 School of Computer Science,Fudan University,Shanghai 201203,China
2 Shanghai Key Laboratory of Intelligent Information Processing,Shanghai 201203,China
3 META SOTA,Shanghai 200135,China
4 Microsoft Research,Beijing 100080,China

Abstract　Many Natural Language Processing (NLP) tasks have benefitted from the public availability of general-purpose vector representations of words trained with large-scale datasets. Since pre-trained word embeddings only have general semantic features from large corpus,it is often necessary to fine-tune these embeddings to make them more suitable for target tasks when it is applied to certain downstream tasks. But,the words with low occurrence frequencies can hardly receive stable gradient information when fine-tuning. However,low-frequency terms are likely to convey important class-specific information in tasks for short text classification. Therefore,it is necessary to obtain a better low-frequency word embedding on the specific task. To address the problem,this paper proposes a model-agnostic algorithm,which optimizes the vector representations of these words according to the task specifics. This approach leverages the update information from common words to guide the embedding updating on rare words. It helps achieve more effective embeddings for the low-frequency words. Our evaluation on three public short-text classification tasks shows that the proposed algorithm produces better task-specific embeddings for rarely occurring words,as a result,the model performance is improved from 0.4% to 1.4% on these tasks. It proves the positive influence of low frequency words on short-text classification tasks,which can shed light on short text classification tasks.

Keywords　Word embedding,Low-frequency word,Fine-tuning,Short text classification

到稿日期:2019-10-24　返修日期:2019-12-20　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61702107);赛尔网络下一代互联网技术创新项目(NGII20180611)
This work was supported by the National Natural Science Foundation of China (61702107) and CERNET Innovation Project (NGII20180611).
通信作者:康昱(kangyu159@gmail.com)

1 引言

在使用深度学习进行自然语言处理(NLP)的任务中,词语的向量化表示是基本步骤。它学习每个单词从原始文本数据到密集的低维向量空间的变换,将人类使用的自然语言符号转化成便于机器处理的连续高维向量形式。为了方便使用,目前业界已有一些用于预训练词向量的模型,如 Mikolov 等[1-2]提出的 word2vec 模型和 Pennington 等[3]提出的 GloVe 模型。利用上述两个模型在大规模语料上训练的具有上下文信息的词向量,被广泛应用在各种自然语言处理任务上[4-8],并取得了不错的效果。

由于将词语向量化是所有深度学习 NLP 任务的基础,其向量化的合理性和准确性将显著影响自然语言处理任务的效果。在一些特定的目标任务语料中,使用通用的预训练词向量固然可以取得较好的效果,然而部分词语的语义在目标场景的语料中会与预训练语料有所偏差,因此需要更适用于当前语料的词语向量化表示来提升任务的完成效果[9]。一种常用的方法是:用预训练的词向量初始化目标任务的嵌入层参数后,让其在模型训练过程中同样进行更新(微调),从而获得具有目标语料特征的词向量表示[10]。然而,这一方法仅可以有效地更新目标语料中高频次出现的单词,而对于其中低频次出现的词,由于训练数据量不足,难以对它们进行有效更新。

研究发现,在一些任务中,低频词同样扮演着重要的角色,短文本分类就是其中之一。该任务的需求是基于内容将短文本(如搜索查询、商品评价、微博等)分类到 n 个类别中的一个或多个,这一场景在如今的生活中十分常见。但这一类短文本由于自身单词数量少、数据稀疏的特点,能够提供的信息有限,即使某些单词出现频次低也会对文本理解起到帮助作用。因此,如果可以有效改善低频词的向量表示,丰富单词在目标语料上的语义信息,将会提升分类效果。

因此,为了在目标语料上得到更准确的低频词词向量表示从而提升分类效果,本文提出了一种利用相似高频词的任务特征信息来更新低频词词向量的算法(Learn and Update Task-specific Information for Low-frequency Words' Embedding Through High-frequency Words, LeUp)。该方法的设计理念借鉴了迁移学习的思想,并认为利用微调方法对高频词的向量表示进行更新是可靠的,因此可以利用这些更新信息来指导低频词的更新,使得原本没有得到有效微调的低频词,甚至是在模型训练中没有出现过的词语,也能获得具有目标任务特征信息的词向量表示,从而提升分类效果。

现有的在短文本分类场景下的词向量研究多数是通过加入一些额外信息来构造向量,以提升效果,例如特殊构造的主题信息[11-12]等。但这些方法需要一些领域文本知识的支持,缺乏在不同任务场景中的通用性;同时,它们也没有考虑低频词在这类任务中的影响。而本文方法可以视为一个通用的、一致的算法框架,可以应用于不同主题的短文本分类任务中。另一方面,我们充分挖掘了低频词在提升短文本分类任务效果中的潜力,通过设计优化算法,成功获得低频词更有效的向量表示,从而提升了分类模型的表现。

本文在多个短文本分类数据集上验证了优化算法的效果,实验结果显示,在采用本文的低频词词向量更新算法后,整体任务的分类准确率均有显著提升。因此,本文认为该方法是有效的、通用的。

综上,本文的主要贡献可以总结为:

1) 关注低频词在短文本分类中的重要作用,从改善低频词在目标任务中的向量表示的角度来提升短文本分类任务的效果;

2) 设计了简单高效的算法,利用高频词的向量更新指导低频词的向量更新,解决了微调过程中低频词词向量无法获得有效更新的问题,使得在不损失原有语义特征的基础上丰富了低频词在目标任务上的向量表示;

3) 以 word2vec 和 GloVe 两种模型得到的预训练词向量为基础,在 3 个短文本分类数据集上验证了所提算法的表现,均取得了效果的提升,体现了该算法的有效性。

2 相关工作

与长文本不同,短文本数据稀疏,且缺乏上下文背景,分类器能够获取的知识有限,于是研究人员尝试从增加额外特征信息和改善词向量表示两个方面来更好地理解短文本,提升短文本分类任务的效果。

现有的很多研究都在尝试挖掘内部主题信息或使用外部知识的方式来捕获额外的特征信息。有一些研究工作提出[11-12]为文本语料库中的每个单词分配主题,通过引入潜在的主题信息来增强词向量的表示能力。还有些研究工作[13-14]尝试从大型网络语料库中构建一个概念库,并为每个短文本分配一组相关概念,利用概念信息来学习短文本的概念向量表示,以增强短文本的语义表示能力。文献[15]引入大规模语义网络提供的词汇语义知识来加深短文本的语义表达。然而,这些方法通常需要依赖于大量高质量的外部数据,而这些数据可能无法适用于某些特定领域或语言。

预训练的词向量通常用于神经网络模型嵌入层的权重初始化,在训练过程中可以选择使其保持不变或者参与参数更新。文献[10,16-17]都尝试在模型训练过程中微调词向量,并取得了效果的提升。文献[18]在多个小规模短文本分类数据集上进行了实验,证明了无论是使用 Word2vec 还是 GloVe,预训练词向量经过微调后都能一定程度上解决词义偏差问题,提升了分类效果。但是,此种方法忽略了微调过程中低频词无法准确更新的问题。而在短文本分类任务中,低频词同样对文本理解有指向性意义。文献[19]同样注意到了低频词的作用,其尝试用相似词汇在文本中的信息来丰富低频词的词袋表示。但作者利用的是单词在文本中的词频统计信息,不同单词之间的词频统计信息并没有严格的语义关联,因此利用这种方法来丰富低频词的词袋表示有一定的局限性。针对上述问题,本文提出了一种基于分布式词表示的、利用高频词的任务特征信息更新低频词词向量的算法。

3 相关背景知识及实例说明

3.1 相关背景知识

3.1.1 词向量(word embedding)

Word embedding 原本指一类将词的语义映射到低维向量空间中的自然语言处理技术。随着它的广泛应用,现多直接表示为词向量,即用来表示词的向量,也可被认为是词的特

征向量或表征。

3.1.2　微调(fine-tune)

微调是迁移学习中的一种手段，常用来形容迁移学习的后期微调，一般指在目标任务数据集上对预训练模型的部分或全部网络结构进行重新训练，使其提取出目标任务特征的过程。而在本文中，微调指对权重初始化的预训练词向量及后续分类网络参数的更新。

3.1.3　低频词与高频词

低频词指在任务数据集中出现频次较少的单词，词频阈值视具体的数据集而定，Word2vec 模型中默认的低频词阈值为 5。本文根据数据集的词表大小及不同词频的占比来确定低频词。除去低频词，词表中剩下的单词被定义为高频词。

低频词尽管在文本中出现的频次很低，但仍有助于文本理解，下一小节将会通过一些具体实例来进行说明。

3.2　实例说明

为了便于理解低频词对短文本分类的作用，本文从实验数据集中选出几个包含低频词的样本来进行说明，如表 1 所列，其中字体加粗的部分是语料中的低频词。可以看出，训练样本的文本长度较短，具有有效分类信息的单词数量有限，其中标粗的单词与分类标签间有较强的关联性。对于上下文信息有限的短文本来说，出现频次低的单词也可能传达重要的分类类别信息，因此保证低频词同高频词一样，获得特定于当前任务的特征信息是有必要的。

表 1　文本分类数据集中的示例

Table 1　Example sentences from text classification dataset

The example sentences	Label
What is considered the costliest **disaster** the insurance **industry** has ever faced?	ENTY:event
What is the highest **waterfall** in the United States?	LOC:other
What is the best **hospital** for orthopedics in the country?	LOC:other
What are some good **exercises** for kids to do?	ENTY:sport
What's the largest U. S. **agricultural crop** by weight?	ENTY:food
What is a research expedition in mountain **climbing** ?	ENTY:sport
What are the **snakes** of New England?	ENTY:animal
What is the world's best selling **cookie**?	ENTY:food

4　面向任务的低频词词向量更新算法

4.1　算法概览

图 1 给出了 LeUp 算法的基本框架。可以看到，该更新算法分为两个步骤。1)通过常规的微调操作，对预训练的词向量在目标短文本语料集上更新出任务特定的词向量表示。这一过程将分别得到高频词和低频词的更新后的词向量，我们认为那些高频词的词向量得到的更新方向是近似正确的，因此该更新方法可用于指导低频词词向量进行更有效的更新。2)本文通过一个更新策略，找到与低频词最相似的几个高频词，并利用这些高频词的更新量来辅助调整低频词在任务数据集上的词向量，最终获得所有单词面向目标任务的词向量表示，以用于下游任务的模型训练。

其中，对低频词向量更新的指导调整可以视为一个对向量的偏移操作。即对于一个低频词 w_L 及其预训练词向量 $\boldsymbol{E}[w_L]$，通过该 LeUp 算法得到该词的更新量 $\Delta[w_L]$，再由式(1)得到该低频词在目标数据集上的预期词向量 $\boldsymbol{E}'[w]$：

$$\boldsymbol{E}'[w_L]=\boldsymbol{E}[w_L]+\Delta[w_L] \tag{1}$$

通过对整体算法框架的介绍可以看出，对低频词向量更新的核心(即得到偏移量 $\Delta[w_L]$)，来源于第二个步骤中利用相似的高频词对它的更新量进行指导的更新策略。

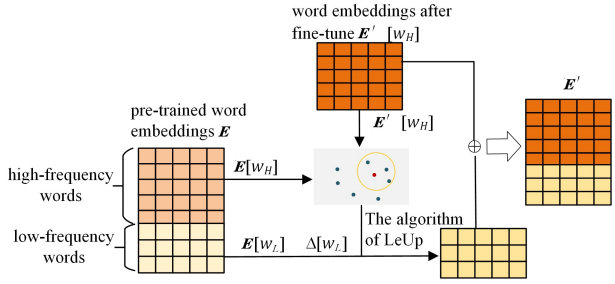


图 1　LeUp 算法的基本框架

Fig. 1　Framework of LeUp

4.2　基于 K 近邻的低频词更新策略

前文提到，需要利用相似的高频词来指导对低频词的向量更新，即可以利用这些高频词的更新信息，来获得更好的低频词更新量 $\Delta[w_L]$。因此，我们设计了一种基于 K 近邻算法的策略，通过快速检索到与低频词相近的多个高频词 w_H，进而利用这些高频词在微调过程中的更新量 $\Delta[w_H]$来计算低频词词向量更合理的更新量 $\Delta[w_L]$。

K 近邻算法[20]本身是一种被广泛用于机器学习分类任务的算法，它的思想是对于一个输入样本点，可以找到在训练集中与它最靠近的前 k 个样本点，并根据这 k 个样本点的标签决定输入样本点的分类。我们摈弃了 K 近邻算法在分类上的功能，而是借鉴了它检索 k 个最相似样本的功能。具体而言，经 word2vec 或 Glove 训练后，每一个单词都会被表达成一个向量表示(即 $\boldsymbol{E}[w]$)，这些向量可以视作向量空间中的点，并且 word2vec 和 GloVe 模型可以捕获局部上下文的相似性，这意味着若两个单词具有相似的上下文，则这两个单词在向量空间中相互靠近。对于一个低频词 w_L，本文通过 K 近邻检索来找到预训练词向量空间中与它最相似的前 k 个高频词的向量样本点 $\boldsymbol{E}[w_H^1],\cdots,\boldsymbol{E}[w_H^K]$。其中，向量样本点的相似性度量可以通过余弦相似性(见(式 2))获得：

$$sim(\boldsymbol{E}[w^a],\boldsymbol{E}[w^b])=\frac{\sum\boldsymbol{E}[w^a]_i\cdot\boldsymbol{E}[w^b]_i}{\sqrt{\sum(\boldsymbol{E}[w^a]_i)^2}\cdot\sqrt{\sum(\boldsymbol{E}[w^a]_i)^2}} \tag{2}$$

本文取相似性最高的 k 个样本点用于指导低频词的词向量更新，同时这些点分别对应着 k 个高频词在预训练词向量集合中的向量表示。图 2 直观地展现了这一过程。

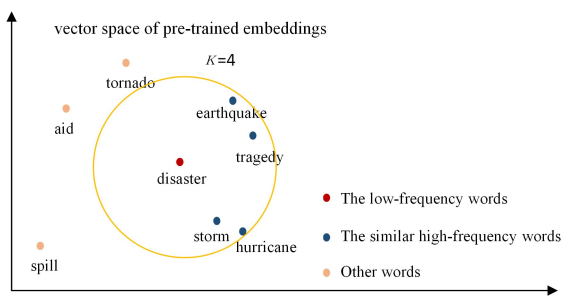


图 2　KNN 算法的直观解释

Fig. 2　Description of K-th nearest neighbor algorithm

通过 K 近邻检索的方法，可以在预训练词向量空间中找到目标任务数据集中某个低频词的 k 个相似高频词及其词向量。经过模型训练，当微调后的效果得到提升时，可以认为高频词词向量的更新方向近似准确，即认为 $\Delta[w_H]$ 是可靠的。因此，在这个过程中高频词的词向量更新可以借鉴到与它相近的低频词词向量的更新上。也就是说，低频词 w_L 在目标任务语料集上的更新量 $\Delta[w_L]$ 可以近似地由 $\Delta[w_H]$ 得到。

对于一个低频词向量 $\boldsymbol{E}[w_L]$，由前文的 K 近邻方法得到了与该向量最相近的 k 个高频词向量 $\boldsymbol{E}[w_H^1],\cdots,\boldsymbol{E}[w_H^K]$，而这些高频词在经过微调后，可以分别得到更新后的向量表示 $\boldsymbol{E}'[w_H^1],\cdots,\boldsymbol{E}'[w_H^K]$，进而，通过式(3)可以得到这些高频词在目标语料集上获得的向量表示更新：

$$\Delta[w_H^k]=\boldsymbol{E}'[w_H^k]-\boldsymbol{E}[w_H^k],k\in[1,K] \tag{3}$$

进而，综合这些高频词语的更新量，通过式(4)获得低频词在目标任务语料上的向量更新：

$$\Delta[w_L]=\frac{\sum_{k=1}^{K}\Delta[w_H^k]}{K} \tag{4}$$

这一整合方式可以视为对高频词的向量更新进行一次算术平均。之所以选择算术平均而没有考虑相似性因素进行加权平均是出于两点考虑：1)从下节的实验可以看出，使用算术平均可以取得不错的效果，简单有效；2)预训练词向量算法本身具有局限性。Word2vec 或 GloVe 都是基于单词上下文来进行训练的，当把单词映射到一个向量空间时，上下文相似的单词也会相应的比较靠近。但上下文相似的单词在语义上不一定相近甚至完全相反，因此若引入相似性权重也可能会带来误差，经过最终权衡，本文选择了算术平均来整合低频词的向量更新量。

综上，本文介绍了 LeUp 算法的核心词向量更新策略，该策略通过整合相似高频词的词向量更新量，获得了低频词在目标语料集上的更新量 $\Delta[w_L]$，进而将其代入式(1)，即可获得低频词在目标语料上的更准确的向量表达 $\boldsymbol{E}'[w_L]$。

5 实验设置与结果分析

前文详细介绍了获得任务相关的词向量过程中低频词词向量的更新算法，为了验证使用 LeUp 获得的任务相关的低频词词向量的有效性，本节在 3 个公开的短文本数据集上分别使用更新前后的词向量进行实验，并对实验结果进行了实验分析。

5.1 数据集

为了验证算法的效果，本文在 3 个数据集上进行了实验，数据集的相关统计数据如表 2 所列。其中 Train 和 Test 分别为训练集和测试集的大小，训练时会从训练集随机选择 10% 作为验证集；Classes 为分类标签；Avg. L 为平均句子长度；Vocab 为字典大小；Fre. 3 为字典中词频小于 3 的单词数；Fre. 2 为字典中词频小于 2 的单词数；Avg. frq 为所有单词的平均词频。

1)TREC6[20]。TREC6 是一个问题分类数据集，它将问题分为 6 个粗粒度标签，分别为 ABBREV，ENTITY，DESCRIPTION，HUMAN，LOCATION 和 NUMERIC。

2)TREC50[21]。它包含与 TREC6 相同的句子，但是将问题分为 50 个细粒度标签，如 NUMERIC：date，ENTITY：animal，LOCATION：city 等。

3)TagMyNews[22]。此数据集来自英语新闻，一共有 7 个标签，包括 sci-tech，sport 等。本文使用新闻标题作为分类样本。

表 2 数据集的统计信息

Table 2 Summary statistics for datasets after tokenization

Datasets	Train	Test	Classes	Avg. L	Vocab	Fre. 3	Fre. 2	Avg. fre
TREC6	5452	500	6	10	8766	6443	5105	6
TREC50	5452	500	50	10	8766	6443	5105	6
TagMyNews	26084	6520	7	8	23598	13547	10238	11

5.2 实验设置及度量指标

5.2.1 实验设置

为了验证算法的通用性，本文使用公开的 word2vec 和 GloVe 两种预训练词向量进行实验。其中，word2vec 使用 CBOW 模型在 Google 新闻上进行训练，词汇量有 300 万，词向量维度为 300；GloVe 在 Common Crawl 语料上进行训练，词汇量为 220 万，词向量维度为 300 维。另外，根据表 2 中的统计结果，考虑到优化算法是利用相似高频词词向量的变化来更新低频词的词向量，为了尽量在更多的高频词中找到合适的高频词，结合不同数据集上的单词数量及不同词频的占比，在 TREC6 及 TREC50 数据集上，本文将词频小于 2 的单词定义为低频词(即 LeUp 算法中需要更新的单词)；在 TagMyNews 中，本文将词频小于 3 的单词定义为低频词。对于分类模型，本文使用 Kim 等[10]提出的 TextCNN 模型，通过不同大小的卷积核提取文本中的 n-gram 字符特征。该模型很好地捕获了文本的局部特征，在多个不同场景下的短文本分类任务中均取得了很好的效果，并且该论文也提出用微调来优化模型效果。于是，本文使用 TextCNN 模型作为 baseline 以及实验分类器，并且模型的参数也遵循原论文中的设置。具体参数设置如表 3 所列。

表 3 模型参数配置

Table 3 Model configuration

Description	Values
Pre-trained word embedding	word2vec or GloVe
Filter region size	(3,4,5)
Feature maps	100
Pooling	1-max pooling
Dropout rate	0.5
L2 norm constraint	3
Gradient descent method	Adadelta

5.2.2 度量指标

实验采用准确率 accuracy 对分类效果进评价。由于本文中的实验都是多分类问题，此处以 3 分类为例，介绍 accuracy 的计算。

表 4　多分类问题的混淆矩阵

Table 4　Confusion matrix of multi-classification

True Label	Predict Label		
	$w1$	$w2$	$w3$
$w1$	$c11$	$c12$	$c13$
$w2$	$c21$	$c22$	$c23$
$w3$	$c31$	$c32$	$c33$

$$accuracy=\frac{\sum c_{ii}}{\sum\sum c_{ij}}*100\% \tag{5}$$

5.3　实验结果和分析

本文将 TextCNN 作为分类模型，分别通过以原始的预训练词向量初始化嵌入层权重并且在训练过程中保持不更新(no fine-tune)，预训练词向量初始化权重后在模型训练过程中进行微调(fine-tune)以及微调后使用 LeUp 来更新低频词向量 3 种方式进行训练，模型效果的对比如表 5 所列。

表 5　不同的词向量处理方式在 3 个数据集上的效果对比

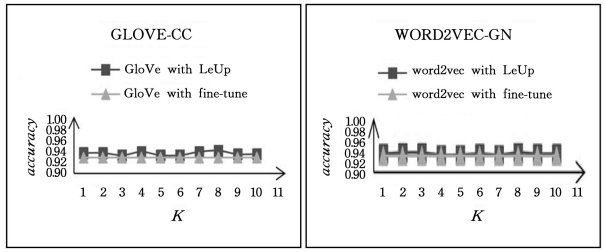
Table 5　Comparisons of accuracy on three benchmark datasets with different pre-trained embedding

(单位：%)

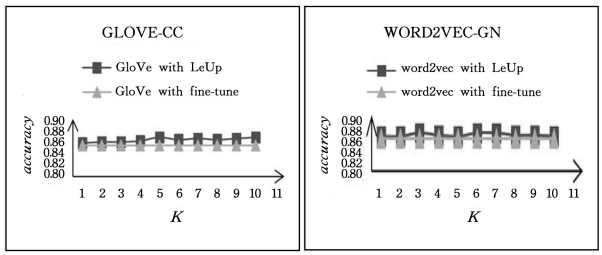
Embedding	Methods	TREC6	TREC50	TagMyNews
GloVe	no fine-tune	91.8	84.0	82.9
	fine-tune	92.8	85.4	83.9
	LeUp	94.0	86.8	84.3
word2vec	no fine-tune	92.2	84.8	81.3
	fine-tune	93.0	85.8	82.6
	LeUp	93.6	86.8	83.0

从表 5 可以看出，使用 GloVe 作为预训练词向量，微调后模型的 *accuracy* 提升了 1%～1.4%，说明微调操作使得高频词的词向量得到了可靠更新，通用的预训练词向量具有了目标任务的特征。在微调的基础上，使用 LeUp 算法对低频词的词向量进行更新后，模型的 *accuracy* 进一步提升了 0.4%～1.4%，分类误差较微调降低了 2.5%～16.7%。同样地，当使用 word2vec 时也有类似的效果。综上说明，无论是使用 word2vec 还是 GloVe 作为预训练词向量，微调在某些场景下都能一定程度地提升模型效果，并且在使用本文提出的 LeUp 方法调整低频词在任务数据上的词向量后，模型效果会得到进一步的提升。

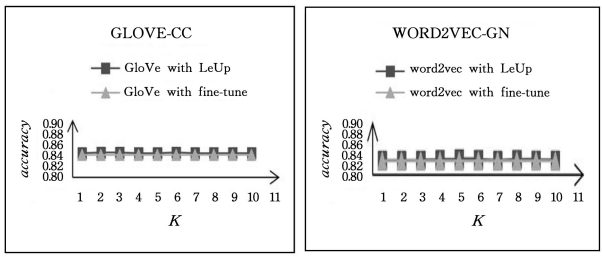
为了进一步探究 K 近邻算法中 K 对 LeUp 算法效果的影响，本文在实验过程中分别使用了不同的 k 进行实验，效果对比情况如图 3 所示。当 k 取较大值时，通过 K 近邻算法返回的 k 个词中会出现不相关的词汇，因此本组实验将 k 的取值范围定为[1,10]，以探究 LeUp 算法在不同 k 值时的效果变化。从图 3 可以看出，只要通过 LeUp 算法对低频词词向量进行更新，就可以取得不同程度的效果提升，说明 LeUp 能有效地赋予低频词词向量有关任务数据集的特征信息，能解决微调过程中低频词词向量未有效更新的问题，同时也从侧面说明了低频词在短文本分类任务中的作用。当 $k<4$ 时，随着 k 的增大，模型效果整体呈现上升趋势，但当 k 继续增大时，模型的准确率就出现了波动，这是因为在词向量空间中计算的相似词是上下文相似而不是严格的语义相似，因此随着 k 的增大，可能会增加语义误差。



(a) LeUp of different K on TREC6



(b) LeUp of different K on TREC50



(c) LeUp of different K on TagMyNews

图 3　不同 k 值的 LeUp 算法效果比较

Fig. 3　Performance comparison with LeUp of different k on three benchmark datasets

结束语　本文提出了一种面向任务的低频词词向量更新算法，该算法解决了微调方式调整预训练词向量过程中低频词无法获得有效更新的问题，使得低频词也能获得任务数据集上的特征信息。本文分别基于 word2vec 和 GloVe 两种预训练词向量，在 3 个公开数据集上进行了实验，验证了该算法的有效性和通用性。

未来将进一步探究 K 近邻的选择标准以及低频词词向量的更新方式，以期面向任务的低频词词向量得到更准确的更新，从而进一步提升分类任务的效果。

参 考 文 献

[1] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in Neural Information Processing Systems. New York: MIT Press, 2013: 3111-3119.

[2] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [J]. arXiv: 1301.3781, 2013.

[3] PENNINGTON J, SOCHER R, MANNING C. Glove: Global vectors for word representation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1532-1543.

[4] ZHANG Y, WALLACE B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence

classification[J]. arXiv:1510.03820,2015.
[5] CAMACHO-COLLADOS J,PILEHVAR M T,NAVIGLI R. Nasari:Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities[J]. Artificia Intelligence,2016,240:36-64.
[6] CALISKAN A,BRYSON J J,NARAYANAN A. Semantics derived automatically from language corpora contain human-like biases[J]. Science,2017,356(6334):183-186.
[7] WANG Y,HUANG M,ZHAO L. Attention-based LSTM for aspect-level sentiment classification[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg:ACL,2016:606-615.
[8] LIU Y,LIU B,SHAN L,et al. Modelling context with neural networks for recommending idioms in essay writing[J]. Neurocomputing,2018,275:2287-2293.
[9] REZAEINIA S M,GHODSI A,RAHMANI R. Improving the accuracy of pre-trained word embeddings for sentiment analysis[J]. arXiv:1711.08609,2017.
[10] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv:1408.5882,2014.
[11] LIU Y,LIU Z,CHUA T S,et al. Topical word embeddings[C]// Twenty-Ninth AAAI Conference on Artificial Intelligence. Menlo Park:AAAI,2015.
[12] ZENG J,LI J,SONG Y,et al. Topic memory networks for short text classification[J]. arXiv:1809.03664,2018.
[13] HUANG H,WANG Y,FENG C,et al. Leveraging Conceptualization for Short-Text Embedding[J]. IEEE Transactions on Knowledge and Data Engineering,2017,30(7):1282-1295.
[14] WANG J,WANG Z,ZHANG D,et al. Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification[C]// International Joint Conference on Artificial Intelligence. San Francisco:Morgan Kaufmann,2017:2915-2921.
[15] HUA W,WANG Z,WANG H,et al. Short text understanding through lexical-semantic analysis[C]// 2015 IEEE 31st International Conference on Data Engineering. Piscataway:IEEE,2015:495-506.
[16] MESNIL G,HE X,DEND L,et al. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding[C]// 14th Annual Conference of the International Speech Communication Association. Lous Tourils:ISCA,2013:3771-3775.
[17] YANG X,MAO K. Supervised fine tuning for word embedding with integrated knowledge[J]. arXiv:1505.07931,2015.
[18] ZHANG Y,WALLACE B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. arXiv:1510.03820,2015.
[19] HEAP B,BAIN M,WOBCKE W,et al. Word vector enrichment of low frequency words in the bag-of-words model for short text multi-class classification problems[J]. arXiv:1709.05778,2017.
[20] PETERSON L E. K-nearest neighbor[J]. Scholarpedia,2009,4(2):1883.
[21] LI X,ROTH D. Learning question classifiers[C]// Proceedings of the 19th International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics,New York:ACM 2002:1-7.
[22] VITALE D,FERRAGINA P,SCAIELLA U. Classification of short texts by deploying topical annotations[C]// European Conference on Information Retrieval. Heidelberg:Springer,2012:376-387.

CHENG Jing, born in 1993, postgraduate. Her main research interests include text classification and so on.

KANG Yu, born in 1988, Ph. D, is a member of China Computer Federation. His main research interests include data driven service intelligence, improving cloud computing service performance based on data analysis methods.