基于 XLNet 和多粒度对比学习的新闻主题文本分类方法

陈　敏，王雷春，徐　瑞，史含笑，徐　渺

(湖北大学 计算机学院　湖北 武汉 430062)

摘要：新闻主题文本内容简短却含义丰富,传统方法通常只考虑词粒度或句粒度向量中的一种进行研究,未能充分利用新闻主题文本不同粒度向量之间的关联信息。为深入挖掘文本的词向量和句向量间的依赖关系,提出一种基于 XLNet 和多粒度特征对比学习的新闻主题分类方法。首先,利用 XLNet 对新闻主题文本进行特征提取获得文本中词、句粒度的特征表示和潜在空间关系;然后,通过对比学习 R-Drop 策略生成不同粒度特征的正负样本对,以一定权重对文本的词向量-词向量、词向量-句向量和句向量-句向量进行特征相似度学习,使模型深入挖掘出字符属性和语句属性之间的关联信息,提升模型的表达能力。在 THUCNews、Toutiao 和 SHNews 数据集上进行实验,实验结果表明,与基准模型相比,所提方法在准确率和 $F1$ 值上都有更好的表现,在三个数据集上的 $F1$ 值分别达到了 93.88%、90.08%、87.35%,验证了方法的有效性和合理性。

关键词：自然语言处理；文本分类；新闻主题；XLNet；对比学习

中图分类号：TP391.1　　**文献标志码**：A　　**文章编号**：1671-6841(2025)02-0016-08

DOI：10.13705/j.issn.1671-6841.2023164

A News Topic Text Classification Method Based on XLNet and Multi-granularity Contrastive Learning

CHEN Min, WANG Leichun, XU Rui, SHI Hanxiao, XU Miao

(*College of Computer Science*, *Hubei University*, *Wuhan* 430062, *China*)

Abstract: News topic text was typically concise but rich in meaning. However, traditional methods in most studies often only considered one type of granularity vector, either word or sentence-level, and failed to fully utilize the correlated information among different granularity vectors of news topic text. To address this issue and explore the dependence relationship between word vectors and sentence vectors in texts, a news topic classification method based on XLNet and multi-granularity feature contrastive learning was proposed. Firstly, features were extracted from the news topic text using XLNet to obtain the feature representations and potential spatial relationships of words and sentences in the text. Then, positive and negative sample pairs of different granularity features were generated using the R-Drop strategy in contrastive learning. Feature similarity learning was conducted on the word-word embedding, word-sentence embedding, and sentence-sentence embedding with certain weights, allowing the model to more deeply explore the related information between character attributes and sentence attributes, thereby enhancing the model's expression ability. Experiments were conducted on THUCNews, Toutiao, and SHNews datasets, the results showed that the proposed method outperformed other methods in terms of accuracy and $F1$ value, with $F1$ values reached 93.88%, 90.08%, and 87.35% respectively, thus verifying the effectiveness and rationality of the proposed method.

Key words: natural language processing; text classification; news topic; XLNet; contrastive learning

收稿日期:2023-06-30
基金项目:国家自然科学基金项目(62106069)。
第一作者:陈敏(1998—),女,硕士研究生,主要从事自然语言处理、文本分类研究,E-mail:1473828049@qq.com。
通信作者:王雷春(1974—),男,副教授,主要从事深度学习、大数据分析研究,E-mail:2430179820@qq.com。

0　引言

新闻主题文本分类通常是对新闻文稿所蕴含的主题类型进行总结和分类。但由于新兴媒介的不断涌现,某些自媒体撰写新闻内容缺乏专业性和规范性,导致新闻主题文本出现用词偏离实际、语义模糊等问题,给新闻主题文本分类研究带来极大挑战。

新闻主题文本分类的首要任务是如何捕捉新闻文本蕴含的语义信息并进行向量化表示[1]。传统的机器学习模型利用 One-Hot、TF-IDF 等方法记录词语在文本中出现的频率以便计算特征权重,但无法处理词与词之间的关系。近年来深度学习的快速发展极大地促进了自然语言处理领域的研究,词向量嵌入技术 Word2Vec[2] 和 GloVe[3] 被相继提出,可以将词语映射为高维空间向量并学习文本的上下文信息以提取特征,但由于上述两种词向量方式使用静态编码,导致词语在不同语境中出现相同词向量的问题[4]。预训练模型 BERT(bidirectional encoder representation from transformers)[5] 的出现使文本向量化工作更加高效和准确,通过对大规模语料库进行无监督和动态学习上下文信息,使生成的词向量具有更多的先验知识,XLNet 作为 BERT 模型的改进版,结合了自回归和自编码模型的优势,利用更多的语料信息实现双向预测,在新闻主题文本分类任务中表现更为优异。

由于新闻主题文本往往是对新闻文稿进行关键信息提炼,由一些高度概括内容的词汇组成,新闻主题文本的词、句等粒度都蕴含着关键的语义信息,然而当前新闻主题分类模型通常只考虑其中一个粒度向量开展研究工作,未能高效地利用词向量和句向量的依赖关系。如何通过模型去理解新闻主题的语义、实体信息,深入挖掘新闻主题文本的潜在信息,是解决新闻文本分类的关键。事实上,对比学习作为一种自监督学习方法,通过相似度学习出文本自身所蕴含的语义信息,可以有效挖掘出新闻主题文本的语义信息。因此,本文引入对比学习机制,提出在词向量-词向量、词向量-句向量和句向量-句向量的粒度上进行特征相似度对比,充分学习词、句之间的依赖,相比于相同粒度对比学习方法,能够更好地理解上下文信息,提升模型的分类能力。本文的主要贡献如下。

1) 提出一种基于 XLNet 和多粒度对比学习的新闻主题分类方法,通过学习不同粒度向量之间的依赖关系,挖掘新闻主题数据潜在的文本信息,以此提升模型的分类效果。

2) 本文提出的多粒度对比学习机制具备通用性,兼容不同结构的深度语言模型。实验结果表明,融合多粒度对比学习机制可以有效提升模型性能。

1　相关工作

1.1　新闻主题文本分类

新闻主题文本分类是指概括总结和判断新闻文本所蕴含的主题类型,现有的新闻主题文本分类主要分为基于机器学习的方法[6-8] 和基于深度学习的方法两类。

基于机器学习的方法参数量较小,训练速度快,但难以学习到新闻文本较深层次的特征信息,因此模型整体泛化能力较差。随着深度学习的快速发展,研究人员广泛使用深度神经网络模型来解决新闻主题文本分类任务[9-10],例如 TextCNN、RCNN 等模型学习文本的特征语义,但这些方法使用了静态词向量技术,无法解决一词多义的问题。预训练模型的出现,利用动态词向量技术,文本特征得到了更好的表示,在多个 NLP 任务上效果显著,促进了新闻文本分类任务的研究发展,例如曾诚等[11] 融合 ALBERT(a lite BERT)与卷积循环神经网络(convolutional recurrent neural network, CRNN),通过 ALBERT 提取文本中每个词向量并输入 CRNN 模型中,有效提升了文本分类效果。张海丰等[12] 结合 BERT 和特征投影网络,提升了新闻主题文本分类模型的性能,针对 BERT 模型忽略了遮掩词位置之间的依赖关系。Yang 等[13] 提出了 XLNet 模型,用自回归的特点弥补 BERT 模型的不足,在新闻文本分类任务中表现良好,因此本文采用 XLNet 模型学习文本的特征向量表达。

1.2　预训练语言模型 XLNet

预训练语言模型分为自回归(auto regression, AR)和自编码(auto encoder, AE)两类。AR 模型用于生成类 NLP 任务,GPT[14] 和 ELMO[15] 便是其中的代表,但无法同时利用上下文信息。AE 模型可以建模双向语义信息,但会导致预训练和微调阶段不一致。XLNet 模型结合 AR 和 AE 模型各自的优点,通过引入循环传递机制和编码相对位置,克服 AR 模型的缺点,在学习语境中上下文信息的同时,更好地表征词语的多义性。为了与微调阶段保持一致,XLNet 模型引入排列组合的方式来重构输入文本。

与 BERT 模型的随机打乱机制相比,XLNet 采用排列组合方式学习双向上下文信息,避免了有效

信息的丢失。另外,为保持句子顺序和获得动态向量表示,XLNet 使用了双流自注意力机制记录位置信息。在处理长文档时,XLNet 融合了 Transformer-XL 框架[16],并利用段循环机制拼接状态信息以实现信息传递依赖,同时引入相对位置编码来解决分段机制造成的位置信息丢失问题。XLNet 的模型结构如图 1 所示。

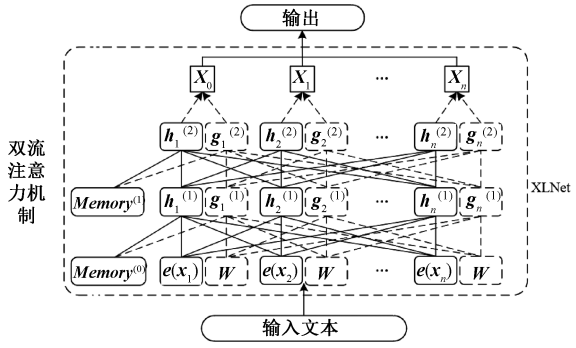


图 1 XLNet 模型结构

Figure 1 XLNet model structure

尽管 XLNet 在文本分类任务上表现良好,但是传统的新闻主题分类通常只考虑了词粒度或者句粒度中的一种粒度向量,无法充分利用新闻主题所精练的文字信息。事实上,对比学习可以很好地解决这个问题,其可以学习到更好的语义表示,因此本文引入对比学习机制来进行不同粒度向量的学习,提升模型的整体分类性能。

1.3 对比学习

对比学习的主要思想是将相似的文本拉近,将不同的文本推开,从文本自身学习出所蕴含的语义信息。如图 2 所示,充分学习同类别文本的相似性和其他类别文本的差异性,将文本进行聚类。

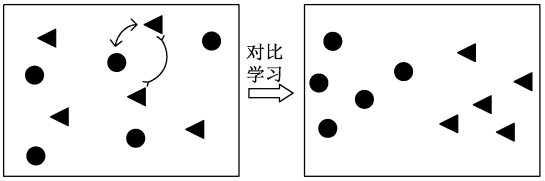


图 2 对比学习

Figure 2 Contrast learning

Chen 等[17]提出自监督对比学习 SimCLR 框架,通过对原始图像进行数据增强,包括裁剪、旋转、缩放和颜色变换等方式,学习同一图像的不同表现形式,充分挖掘出图像自身的特征信息。SimCLR 是图像视觉领域中对比学习研究的重要里程碑。随后,Gao 等[18]提出 SimCSE 方法,采用简单的 Dropout 方法构造正例和负例,在文本表示方面取得了巨大的成功。Wu 等[19]提出有监督学习下对比学习思路 R-Drop,利用两次 Dropout 在不改变原有结构的情况下增加 KL 散度损失来进行模型训练,在各类任务中都有显著的提升。因此,本文引入 R-Drop 对比学习方法,对新闻主题文本的各粒度向量进行特征表示学习。

在新闻文本分类任务中引入对比学习的关键是通过对比学习思想,学习新闻主题所蕴含的关键信息,通过挖掘不同粒度之间的潜在信息,使模型既能完成对比学习任务,又能完成分类任务。

2 XLNet-MGCL(multi-granularity contrastive learning)新闻主题文本分类方法

基于 XLNet 和多粒度对比学习的新闻主题分类方法的主体思想如图 3 所示,输入文本经过两个权重共享的 XLNet 模型分别获得文本的词向量、句向量表示,利用句向量完成模型分类任务,模型优化的对比任务中采用词向量和句向量进行多粒度对比学习。模型训练过程包括以下主要步骤。

步骤 1 数据预处理。新闻主题文本数据集中存在着部分不规范或对结果产生影响的特殊字符,因此需要对数据进行正则化处理并剔除噪声数据。对于处理好的实验数据本文随机打乱并按照 8∶1∶1 的比例划分为训练集、测试集和验证集。

步骤 2 文本特征生成将已处理好的数据,利用 XLNet 模型进行特征提取,得到文本的词向量和句向量表示。输入文本为

$$x_i = <[CLS], c_1, c_2, \cdots, c_n, [SEP]>,$$

其中:x_i 表示训练批次为 $\{(x_i, y_i)\}_{i=1}^{N}$ 中的第 i 条样本;c_n 表示第 n 个词的序列化表示。利用 XLNet 模型得到最后一层中[*CLS*]位置向量,其包含文本全局语义信息,即句向量特征 $\boldsymbol{hs}_i$,同时每个 *token* 也充分学习到了上下文语境,从而获得词向量特征为 $\boldsymbol{ht}_i$。如式(1)所示,本文利用句向量特征 $\boldsymbol{hs}_i$ 作为分类任务的特征表示。

$$\hat{y}_i = \mathrm{Softmax}((\boldsymbol{W})^{\mathrm{T}}\boldsymbol{hs}_i + b), \tag{1}$$

其中:$\boldsymbol{W}$ 为参数矩阵;b 为偏置项;$\hat{y}_i$ 为当前输入文本 x_i 的预测值。

步骤 3 对比学习机制。对于该批次训练样本 $\{(x_i, y_i)\}_{i=1}^{N}$,通过 R-Drop 策略生成正样本集合。R-Drop 核心思想是对于同一样本,经过两次模型输出,在随机失活神经元的机制下会得到两个不同但差异很小的概率分布。因此,利用 R-Drop 策略可以生成该批次样本的词向量和句向量正样本集合 $\{(\boldsymbol{hs}_i^{+}, \boldsymbol{ht}_i^{+})\}_{i=1}^{N}$,在有监督对比学习模式下,负样

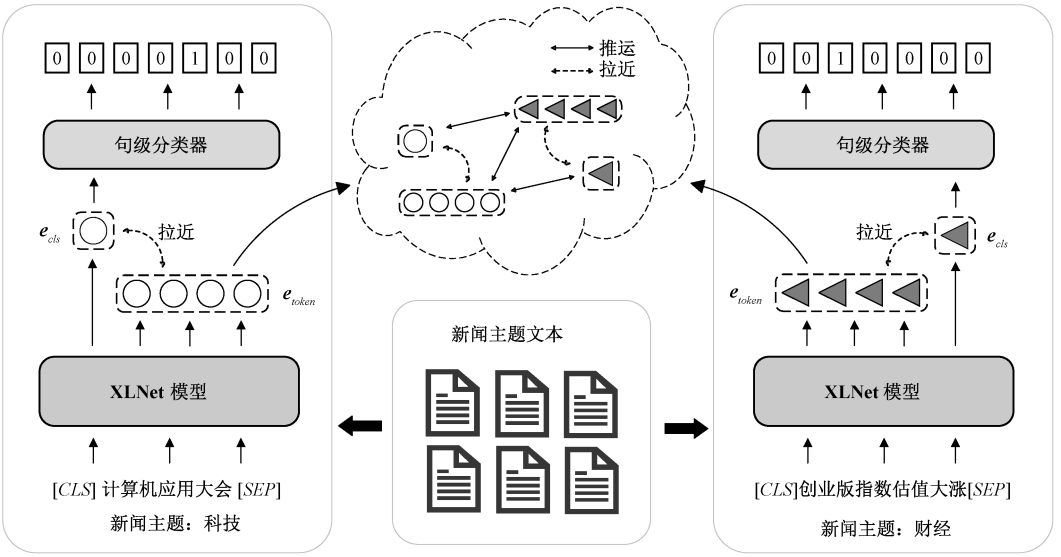


图 3　XLNet-MGCL 方法结构

Figure 3　XLNet-MGCL method structure

本的构造方式则为从该批次中随机挑选其他类别样本集合,负样本集合定义为 $\{(\boldsymbol{hs}_i^-,\boldsymbol{ht}_i^-)\}_{i=1}^N$。

关于正负样本对的构造,本文考虑利用不同粒度向量的信息构建出如下样本对:词-词向量 $(\boldsymbol{ht}_i^+,\boldsymbol{ht}_i^-)$;词-句向量 $(\boldsymbol{ht}_i^+,\boldsymbol{hs}_i^-)$ 和 $(\boldsymbol{hs}_i^+,\boldsymbol{ht}_i^-)$;句-句向量 $(\boldsymbol{hs}_i^+,\boldsymbol{hs}_i^-)$。在不同粒度对比下,分别在字符属性和语句属性的维度空间上拉近同类别文本,推开不同类别文本,充分挖掘出新闻主题所包含的特征信息,使模型能够学习到更好的向量表示。对比学习损失函数为

$$\ell_d = -\log\frac{\mathrm{e}^{\mathrm{sim}(\boldsymbol{h}_i^{(0)},\boldsymbol{h}_i^{(1)})/\tau}}{\sum_{j=1}^{N}\mathrm{e}^{\mathrm{sim}(\boldsymbol{h}_i^{(0)},\boldsymbol{h}_i^{(1)})/\tau}}, \tag{2}$$

$$\mathrm{sim}(\boldsymbol{h}_1,\boldsymbol{h}_2)=\frac{\boldsymbol{h}_1^{\mathrm{T}}\boldsymbol{h}_2}{\|\boldsymbol{h}_1\|\cdot\|\boldsymbol{h}_2\|}, \tag{3}$$

其中:温度系数 τ 是超参数,用于调整模型将重点聚集到困难负例的惩罚程度。将构造好的不同粒度正负样本对传入到 ℓ_d 损失函数中,得到多粒度对比学习损失函数

$$\begin{aligned}\mathcal{L}_{\mathrm{MGCL}} &= \mathcal{L}_{cl}(\boldsymbol{ht}_i^+,\boldsymbol{hs}_i^+)+\mathcal{L}_{cl}(\boldsymbol{ht}_i^-,\boldsymbol{hs}_i^-)+\\ &\mathcal{L}_{cl}(\boldsymbol{ht}_i^+,\boldsymbol{ht}_i^-)+\mathcal{L}_{cl}(\boldsymbol{ht}_i^+,\boldsymbol{hs}_i^-)+\\ &\mathcal{L}_{cl}(\boldsymbol{hs}_i^+,\boldsymbol{ht}_i^-)+\mathcal{L}_{cl}(\boldsymbol{hs}_i^+,\boldsymbol{hs}_i^-)。\end{aligned} \tag{4}$$

步骤 4　模型参数更新。模型的训练目标是完成对比学习任务和分类任务,其中对比学习任务是约束模型将同类别文本尽可能聚类在一起,以获得更好的特征向量表示;分类任务则是利用训练好的文本特征,通过 Softmax 函数得到类别预测概率,利用交叉熵(cross entropy,CE)损失函数约束模型将文本往正确类别学习。为此,本文设计的目标函数为

$$\mathcal{L}_{ce} = -\sum_{j=1}^{m} g_j\log(k_j), \tag{5}$$

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha\,\mathcal{L}_{\mathrm{MGCL}}, \tag{6}$$

其中:$\mathcal{L}_{\mathrm{MGCL}}$ 为多粒度对比学习损失;$\mathcal{L}_{ce}$ 表示交叉熵损失;k_j 为预测类别属于类别 j 的概率;g_j 是指示变量;α 为超参数,用于调节对比学习任务和分类任务的权重。

3　实验与分析

3.1　数据集

为验证本文方法在新闻主题文本分类任务上的有效性,采用了三个新闻主题的数据集 THUCNews、Toutiao 和 SHNews 进行实验,实验数据按照 8∶1∶1 的比例划分为训练集、验证集和测试集,数据集详细信息如下。

1) THUCNews 数据集是根据新浪新闻 RSS 订阅频道 2005—2011 年间的历史数据筛选过滤生成,本文对原始数据集进行数据清洗并重新整合,划分出财经、股票、科技、社会、时政、娱乐共计 6 个候选分类类别,每个类别数据约 1 万条,平均长度约为 20。

2) Toutiao 数据集来源于今日头条客户端,本文对原始数据集进行数据预处理,从中挑选出体育、财经、房产、汽车、科技、旅游共计 6 个分类类别,每个类别数据约 5 000 条,平均长度约为 25。

3) SHNews 数据集来源于搜狐新闻整理的开源数据,包含娱乐、财经、房地产、旅游、科技、体育、健

康、教育、汽车、新闻、文化、女人共 12 个分类类别，每个类别数据约 2 800 条，平均长度约为 20。

3.2 评价指标

本文采用准确率(Acc)和 $F1$ 值对分类结果进行评价，计算公式如下：

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \tag{7}$$

$$P = \frac{TP}{TP + FP}, \tag{8}$$

$$R = \frac{TP}{TP + FN}, \tag{9}$$

$$F1 = \frac{2 \times P \times R}{P + R}, \tag{10}$$

其中：P、R 分别为精确率和召回率；TP 表示正样本预测为正；FP 表示负样本预测为正；TN 表示正样本预测为负；FN 表示负样本预测为负。

3.3 对比实验

为验证所提方法 XLNet-MGCL 的有效性，本文选择在新闻主题文本分类上表现较好的以下方法作为对比实验。

1) TextRCNN。利用双向门控单元 BiGRU 对文本进行双向训练，加入池化层对特征输出进行最大池化操作。

2) BERT。利用 BERT 模型得到文本的句向量表示，采用随机掩码机制(masked language model, MLM)和下句预测任务(next sentence prediction, NSP)充分学习句子中上下文的语境信息，并通过全连接层进行分类。

3) XLNet。使用 XLNet 模型提取文本特征，并连接池化层和全连接层进行分类。

4) XLNet-RCNN[20]。利用 XLNet 模型初步得到文本的特征表示，接着通过 RCNN 网络对文本特征进行双向训练，获得文本的深层语义。

5) XLNet-SimCSE。将 XLNet 模型和 SimCSE 对比学习策略融合，使用 SimCSE 思想生成正负样本对，以此提升文本特征表达能力。

6) XLNet-rDrop。对比学习方法采用 R-Drop 策略，在有监督学习下，相较于 SimCSE 分类效果更好。

同时，为说明本文所提出的多粒度对比学习机制具备通用性，使用预训练模型 BERT、RoBERTa[21] 替代本文所选分类模型进行实验。

3.4 实验参数设置

本文实验参数包括 XLNet 模型参数以及对比学习模块参数，其中 XLNet 采用哈工大讯飞联合实验室发布的中文自回归语言模型[22]，隐藏层尺寸为 768，隐藏层层数为 12，激活函数为 ReLU。对比学习模块包括对比学习损失函数中的温度系数 τ，以及对比损失函数和交叉熵损失函数权重平衡因子 α，其中学习率可选范围为[1e-5,2e-5,5e-5]，温度系数可选范围为[0.05,0.1,0.15]，经过多次迭代选择最佳训练结果。

实验使用的 Dropout 随机失活率为 0.5，优化策略选择效果较好的 Adam 优化器。模型经过多次训练选取的文本输入长度为 50，权重平衡因子为 0.3。

3.5 实验结果与分析

表 1 展示了不同模型在三个数据集上的表现(表中加粗数据为较佳数据)，本文方法 XLNet-MGCL 在 THUCNews 和 SHNews 数据集上的 $F1$ 值分别为 93.88%、87.35%，与 XLNet 融合学习的另外两种方法 XLNet-SimCSE 和 XLNet-rDrop 对比，本文所提的多粒度对比学习方法效果更好，在 THUCNews 和 SHNews 数据集上的 $F1$ 值相比 XLNet-SimCSE 模型分别提升了 0.36、0.3 个百分点，相比 XLNet-rDrop 模型分别提升了 1.19、0.68 个百分点。充分证明了对不同粒度向量的依赖关系进行学习有利于提升模型的整体表达能力。

由实验结果可知，TextRCNN 使用静态词向量技术 Word2Vec，整体表现处于低位，难以学习到新闻主题文本的内在信息。BERT、RoBERTa 和 XLNet 模型使用动态词向量技术，具有丰富的先验知识，整体表现更好，其中，XLNet 作为 BERT 模型的改进版，使用更多的语料信息以及更先进的算法策略，因此分类效果相比 BERT 更优。

同时为验证所提的多粒度对比学习机制的兼容性，本文使用其他预训练模型 BERT、RoBERTa 融合多粒度对比学习机制，实验结果可以看出，在多粒度对比学习机制下模型分类效果得到明显提升。在文本表示阶段，利用对比学习机制学习字符属性和语句属性的依赖关系，挖掘出词向量和句向量，可以对文本特征进一步聚类，同时实验结果表明在新闻主题文本分类工作中对文本的不同粒度向量进行深入研究的必要性。

为进一步说明多粒度对比学习机制的有效性，本文选取 SHNews 数据集中财经、健康和汽车三个类别，使用 t-SNE 算法(t-distributed stochastic neighbor embedding)对测试集进行可视化处理，如图 4 所示。图 4(a)为基准模型 XLNet 最终用于分类任务的句向量特征在二维空间上的表示，尽管不同类别之间有明显的分界线，但同类别文本在嵌入空间的

表 1　各方法在不同数据集上的结果

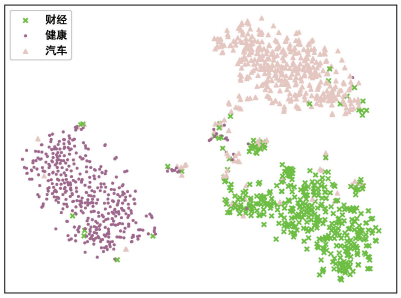
Table 1　Experimental results of different methods with different datasets　　单位:%

模型	THUCNews		Toutiao		SHNews	
	Acc	*F*1	*Acc*	*F*1	*Acc*	*F*1
TextRCNN	89.58	89.58	84.27	84.20	78.73	78.69
BERT	91.25	91.20	88.05	88.02	85.15	85.18
RoBERTa	93.18	93.16	89.70	89.69	86.39	86.44
XLNet	93.68	93.66	88.10	88.09	86.39	86.44
XLNet-RCNN	93.80	93.79	**90.33**	**90.28**	87.19	87.26
XLNet-SimCSE	93.52	93.52	90.20	90.12	86.95	87.05
XLNet-rDrop	92.70	92.69	90.30	90.26	86.59	86.67
XLNet-MGCL	**93.88**	**93.88**	90.10	90.08	**87.29**	**87.35**
BERT-MGCL	92.60	92.58	88.50	88.47	86.12	86.36
RoBERTa-MGCL	93.70	93.67	90.28	90.20	87.10	87.28

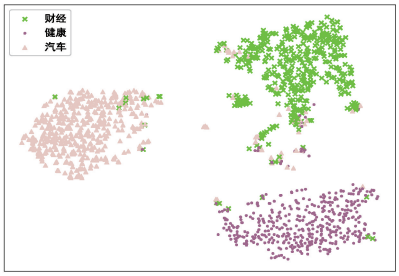
分布情况散乱,例如财经类别未能很好地聚合在一起,文本特征信息待充分挖掘。图 4(b)为使用多粒度对比学习机制优化后的模型,明显看出同类别文本之间聚合程度较高,离群文本减少,这说明多粒度对比学习机制对文本的向量特征表达进行了优化。

3.6　消融实验

为进一步验证所提方法的合理性,本文进行了消融实验,多粒度对比学习机制的核心是利用词、句向量的依赖关系来约束模型学习文本的深层特征表达,因此设计分别去除词-词向量、词-句向量、句-句向量对比模块所得模型作为实验对照组,其中,去除词-词向量的实验为 XLNet-MGCL(w/o w-w),去除词-句向量的实验为 XLNet-MGCL(w/o w-s),去除句-句向量的实验为 XLNet-MGCL(w/o s-s),消融实验结果如表 2 所示(表中加粗数据为较优数据)。从表中可以看到去除不同向量粒度对比模块后,模型的分类性能会有一定的下降,证明了融合多粒度对比学习来提升新闻主题文本分类的有效性。



(a) XLNet



(b) XLNet-MGCL(本文方法)

图 4　SHNews 测试集数据可视化

Figure 4　SHNews test set data visualization

表 2　消融实验

Table 2　Ablation experiment　　单位:%

方法	Toutiao		SHNews		THUCNews	
	Acc	*F*1	*Acc*	*F*1	*Acc*	*F*1
XLNet-MGCL	**93.88**	**93.88**	90.10	90.08	**87.29**	**87.35**
XLNet-MGCL(w/o s-s)	93.69	93.69	**90.64**	**90.61**	86.49	86.63
XLNet-MGCL (w/o w-s)	93.69	93.68	90.34	90.31	86.90	86.88
XLNet-MGCL (w/o w-w)	93.79	93.75	90.25	90.23	87.03	87.16

4　结语

本文针对如何高效利用新闻主题文本的精炼性、高度概括性等特性,提出了基于 XLNet 和多粒度对比学习的信息主题分类方法 XLNet-MGCL。该方法利用 XLNet 模型获得文本的词、句粒度的特征表示,并使用 R-Drop 策略构建出对比学习样本组合,学习出词-词向量、词-句向量和句-句向量不同粒度之间的潜在关系,充分利用新闻主题文本简短

却含义丰富的特点,以此对文本进行更好的特征表达。在三个公开的新闻主题文本数据集上的表现充分证明了所提方法的有效性。在下一步工作中,考虑使用更多策略对文本不同粒度进行学习,在自监督学习下充分利用新闻文本自身特性获得更优越性的向量表示,进一步提高新闻主题文本分类模型的效果。

参考文献:

[1] 杨朝强, 邵党国, 杨志豪, 等. 多特征融合的中文短文本分类模型[J]. 小型微型计算机系统, 2020, 41(7): 1421-1426.
YANG Z Q, SHAO D G, YANG Z H, et al. Chinese short text classification model with multi-feature fusion [J]. Journal of Chinese computer systems, 2020, 41(7): 1421-1426.

[2] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: ACM Press, 2013: 3111-3119.

[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 (EMNLP). Stroudsburg: ACL Press, 2014: 1532-1543.

[4] 李舟军, 范宇, 吴贤杰. 面向自然语言处理的预训练技术研究综述[J]. 计算机科学, 2020, 47(3): 162-173.
LI Z J, FAN Y, WU X J. Survey of natural language processing pre-training techniques[J]. Computer science, 2020, 47(3): 162-173.

[5] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language un-derstanding[EB/OL]. (2018-10-11)[2023-05-10]. https://arxiv.org/pdf/1810.04805.pdf.

[6] DEY S, WASIF S, TONMOY D S, et al. A comparative study of support vector machine and naive Bayes classifier for sentiment analysis on Amazon product reviews[C]// 2020 International Conference on Contemporary Computing and Applications. Piscataway: IEEE Press, 2020: 217-220.

[7] WANG S D, MANNING C D. Baselines and bigrams: simple, good sentiment and topic classification[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. New York: ACM Press, 2012: 90-94.

[8] 许英姿, 任俊玲. 基于改进的加权补集朴素贝叶斯物流新闻分类[J]. 计算机工程与设计, 2022, 43(1): 179-185.
XU Y Z, REN J L. Naive Bayesian logistics news classification based on improved weighted complement [J]. Computer engineering and design, 2022, 43(1): 179-185.

[9] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 2014: 1746-1751.

[10] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015: 2267-2273.

[11] 曾诚, 温超东, 孙瑜敏, 等. 基于ALBERT-CRNN的弹幕文本情感分析[J]. 郑州大学学报(理学版), 2021, 53(3): 1-8.
ZENG C, WEN C D, SUN Y M, et al. Barrage text sentiment analysis based on ALBERT-CRNN[J]. Journal of Zhengzhou university (natural science edition), 2021, 53(3): 1-8.

[12] 张海丰, 曾诚, 潘列, 等. 结合BERT和特征投影网络的新闻主题文本分类方法[J]. 计算机应用, 2022, 42(4): 1116-1124.
ZHANG H F, ZENG C, PAN L, et al. News topic text classification method based on BERT and feature projection network [J]. Journal of computer applications, 2022, 42(4): 1116-1124.

[13] YANG Z, DAI Z, YANG Y M, et al. XlNeT: generalized autoregressive pretrai-ning for language understanding [EB/OL]. (2019-12-08) [2023-04-20]. https://dl.acm.org/doi/pdf/10.5555/3454287.3454804.

[14] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pretraining[EB/OL]. (2018-08-22)[2023-04-20]. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.

[15] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL Press, 2018: 2227-2237.

[16] DAI Z H, YANG Z L, YANG Y M, et al. Transformer-XL: attentive language models beyond a fixed-length context[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2019: 2978-2988.

[17] CHEN X L, FAN H Q, GIRSHICK R, et al. Improved

baselines with momentum contrastive learning[EB/OL]. (2020-03-09)[2023-04-20]. https://arxiv.org/pdf/2003.04297.pdf.

[18] GAO T Y, YAO X C, CHEN D Q. SimCSE: simple contrastive learning of sentence embeddings[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic. Stroudsburg: ACL Press, 2021: 6894-6910.

[19] WU L, LI J, WANG Y, et al. R-Drop:regularized dropout for neural networks[EB/OL]. (2021-06-28)[2023-04-20]. https://arxiv.org/pdf/2106.14448.pdf.

[20] 潘列, 曾诚, 张海丰, 等. 结合广义自回归预训练语言模型与循环卷积神经网络的文本情感分析方法[J]. 计算机应用, 2022, 42(4): 1108-1115.
PAN L, ZENG C, ZHANG H F, et al. Text sentiment analysis method combining generalized autoregressive pretraining language model and recurrent convolutional neural network[J]. Journal of computer applications, 2022, 42(4): 1108-1115.

[21] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL]. (2019-07-26)[2023-04-20]. https://arxiv.org/pdf/1907.11692.pdf.

[22] CUI Y M, CHE W X, LIU T, et al. Revisiting pretrained models for Chinese natural language processing[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg:ACL Press, 2020: 657-668.