文章编号:1671-6833(2025)03-0128-08

面向关键词预测的动态对比表示增强方法

耿雪莲[1,2], 宋明阳[1,2], 冯 毅[1,2], 景丽萍[1,2], 于 剑[1,2]

(1.北京交通大学 计算机与信息技术学院,北京 100044;2.北京交通大学 交通数据分析与挖掘北京市重点实验室,北京 100044)

摘 要:关键词预测通常难以充分利用文本结构中的复杂层次和语义信息,针对该问题,提出了一种新型语义表示增强的关键词预测方法(ACL-KP),即利用动态对比学习增强关键词预测。该方法首先通过引入自适应权重机制,动态调整样本权重,解决在对比学习过程中难以区分真实样本与噪声样本的问题,减少误识别噪声样本的影响,优化空间表示。此外,为了提高训练数据的多样性,引入高斯白噪声,自动生成一些具有挑战性的虚拟样本,从而增强文档和关键词的语义表示。在关键词预测领域的多个公开数据集上进行的实验结果表明:模型在 $F1@5$ 和 $F1@M$ 指标上相较于当前先进模型提升了 2%~17%,与序列到序列模型和统一模型相比,展现出了更显著的性能优势。

关键词:自然语言处理; 关键词预测; 多目标优化; 对比学习; 嵌入表示

中图分类号:TP391 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2025.03.004

关键词预测主要是指从文档中获取一组能够表示文章主要思想或者主题的单词或短语。作为自然语言处理的基本任务之一,关键词预测可以应用于信息检索[1]、文档聚类[2]、推荐[3]等任务。

早期关键词预测的研究方法主要以关键词抽取为主,传统关键词抽取方法有基于 TF-IDF[4] 等词频统计指标抽取关键词、将文档转化为图结构[5]获取关键词等。但有些关键词可能并不在原文中出现。如图 1 所示,部分关键词并没有直接出现在原文中,不能简单地通过提取的方式获取关键词,而是需要对原文进行深入的语义理解。

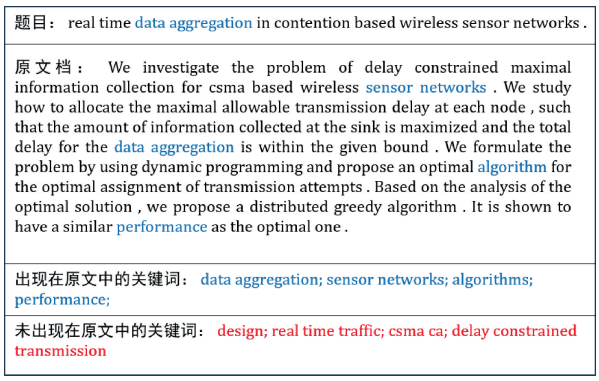


图 1 关键词预测的示例

Figure 1 Example of the keyphrase prediction

随着深度学习技术的发展,有研究者开始将 Seq2Seq 模型应用到关键词领域,利用编码器获取输入文本中的信息并形成特征向量,传递给解码器,解码器利用复制机制[6]来获取概率最大的字符作为生成的关键词。现有的生成模型仍存在一些局限性。这些模型常常通过最大似然估计[7]方法预测概率最高的词,但概率最高的关键词未必与目标关键词一致。而且在处理复杂文本或长文本时,由于文本的复杂性和上下文的丰富性,抽取和生成关键词变得更加困难。

随着预训练模型的发展,利用获取嵌入的方法可以捕捉上下文关系。通过将关键词映射到连续向量空间中,可以捕捉到关键词的语义信息和语义相似性。对比学习可以通过比较样本之间的相似度来学习更具区分性的关键词嵌入表示。因此,利用对比学习优化并拉近关键词和文本在嵌入空间中的表示,受 Focal Loss[8] 的启发,鼓励模型学习了更具辨别性的句子表示,提出了一种动态的对比学习方法

收稿日期:2024-11-20;**修订日期**:2024-12-25

基金项目:国家自然科学基金资助项目(62176020);国家重点研发计划项目(2020AAA0106800);北京市自然科学基金资助项目(L211016);中央高校基本科研业务费专项资金(20119jbz110)

通信作者:景丽萍(1978—),女,河南邓州人,北京交通大学教授,博士,博士生导师,主要从事机器学习及其在人工智能领域的应用研究,E-mail:lpjing@bjtu.edu.cn。

引用本文:耿雪莲,宋明阳,冯毅,等. 面向关键词预测的动态对比表示增强方法[J]. 郑州大学学报(工学版),2025,46(3):128-135. (GENG X L, SONG M Y, FENG Y, et al. Dynamic contrastive representation enhancement approach for keyphrase prediction[J]. Journal of Zhengzhou University (Engineering Science),2025,46(3):128-135.)

(adaptive contrastive learning for keyphrase prediction,ACL-KP)预测关键词。在不同数据集上进行实验,结果表明,所提模型性能优于最近几年的大部分关键词预测模型。

本文的主要贡献包括以下 3 个方面:①提出了一种自适应的方式来动态调整样本信息,减少噪声样本对模型的影响;②在训练过程中引入了高斯白噪声,自动创建一些困难样本,增加训练数据的多样性;③公开数据集上的实验结果表明,所提方法比其他方法在关键词预测性能上表现更出色。

1　相关工作

1.1　关键词抽取

关键词抽取旨在获取源文本中出现过的关键词。现有的抽取方法大致可分为序列标记模型和两阶段模型。

序列标记模型主要通过对文本句子中的词进行标注来实现。通过不同的标记,可以确定每个单词是否是关键词的一部分。目前常用的标记方法包括 BIO 和 BIOES。

两阶段模型将抽取模型分为两部分进行,首先使用不同的启发式规则从文本中确定一组候选短语,然后通过算法对这些候选短语进行重要度打分和排序。打分模型通常由支持向量机[9]等监督算法或者基于图排序[5]等无监督算法获得。受益于预训练语言模型的发展,获取候选短语嵌入和文档嵌入之间的相关性,捕获上下文关系。Bennani-smires 等[10]通过估计候选短语嵌入与文档之间的相似性来对短语进行排序和提取。Liang 等[11]通过边界感知的中心性增强短语和文档相关性,捕获上下文关系。Zhang 等[12]基于 mask 策略,利用源文档和被屏蔽文档的嵌入之间的相似性对候选文本进行排序,解决了在进行关键词选取的时候存在偏向选取长关键词的问题。

虽然关键词抽取能够有效地提取出原文中存在的关键词,但是无法预测原文中不存在的关键词。

1.2　关键词生成

为了解决关键词抽取中存在的问题,研究者们开始考虑如何预测未出现在文本中的关键词。Meng 等[6]首次提出了 CopyRNN,这是一个具有注意力和复制机制的序列到序列的框架,用于 One2One 范式下生成每个文档的单个关键短语,但该模型必须固定预测的关键词数量。后续研究者提出了一种基于 One2Seq 的改进模型,利用半监督学习、强化学习等方法来改进关键词生成。随后,研究发现,在关键词生成过程中引入关键词抽取可以有效提升关键词生成的质量。Chen 等[13]提出了一种联合模型 CorrRNN,将抽取式模型和生成式模型进行简单组合,获取源文档中已有的短语,对文档覆盖率进行显式建模,从而使生成的关键短语可以覆盖更多主题。

1.3　对比学习

对比学习通过最大化正样本对之间的相似度和最小化负样本对之间的相似度来学习数据的表示。通过对比损失,模型可以更有效地区分语义相似和不相似的句子或文档。Gao 等[14]提出的 SimCSE 方法通过最小化正样本对之间的距离和最大化负样本对之间的距离来学习句子嵌入。在训练过程中,SimCSE 使用了 InfoNCE[15]损失函数,这是一种广泛用于对比学习的损失函数,它鼓励模型区分不同句子的细微差别。通过这种方式,SimCSE 能够学习到句子语义的嵌入表示。Xu 等[16]提出的 SimCSE++ 认为负样本对中的 dropout 噪声对模型性能有负面影响,提出了 off-dropout 的策略,该策略通过关闭负样本对中的 dropout 随机性来提升性能。Hou 等[17]引入了 Focal-InfoNCE 损失函数,该函数在对比目标中引入了自适应调制项,对容易的负样本的损失进行降权,并鼓励模型集中注意力在困难的负样本上。

部分研究已经将关键词信息与对比学习相结合,以丰富文本在向量空间中的表示。例如,Li 等[18]提出了一种基于聚类的对比学习方法 CCL,这种方法通过从聚类中精心挑选负样本,有效降低了噪声干扰,优化了主题短语的表示。Cai 等[19]通过融合对比学习和关键词提取,成功生成了融合关键词信息的 token 级别句子嵌入,增强了句子的语义表达能力。此外,Choi 等[20]首次在关键词预测领域引入对比学习,优化了关键词嵌入的空间表示,为关键词的生成和提取提供了新的视角,但其在对比学习过程中未充分考虑假负样本等噪声的干扰。

采用聚类的方法筛选负样本在很大程度上依赖聚类算法的性能,而且在为接近质心的实例分配伪标签上可能存在不确定性,因此在关键词嵌入的表示学习中,如何更有效地提升样本在空间中的区分度、减少假负样本的干扰仍然是一个值得深入探讨的问题。针对这一挑战,本文提出了一种利用自适应机制动态调整样本信息的方法,以应对假负样本的干扰,同时降低错误识别带来的直接影响。引入高斯白噪声对数据进行增强,在训练过程中自动创造一些虚拟的困难样本,以此来锻炼和增强模型的鲁棒性和表达能力。

2 研究方法

2.1 问题定义

每篇文档对应了一个包含若干关键词的集合,给定文档 $\boldsymbol{X}$,其包含 n 个单词 $\{x_i\}_{i=1}^{n}$,关键词预测的任务就是从文档中识别一组关键词 $y=\{y_i\}_{i=1,2,\cdots,|y|}$,其中 $|y|$ 为关键词的数量。这组关键词包括出现在文本中的关键词和未出现在文本中的关键词,分别用 $y_{\mathrm{p}}=\{y_i\}_{i=1,2,\cdots,|y_{\mathrm{p}}|}$ 和 $y_{\mathrm{a}}=\{y_i\}_{i=1,2,\cdots,|y_{\mathrm{a}}|}$ 表示。

2.2 实验方法

本节将介绍如何利用动态对比学习优化关键词嵌入表示。模型的整体架构图如图 2 所示。

2.2.1 动态对比学习策略

按照文献[20]中获取候选短语的方法,使用 Stanford POSTagger 为每个单词分配词性标记,并使用 NLTK RegexpParser 将短语结构树分块为有效短语树,然后删除不符合语义规范的短语。

将文档 $\boldsymbol{X}$ 输入预训练语言模型[21]中,获取编码器的最后隐藏状态,嵌入向量为

$$[\boldsymbol{e}_0,\boldsymbol{e}_1,\cdots,\boldsymbol{e}_T]=\mathrm{Encoder}(\boldsymbol{X})。\tag{1}$$

式中:T 表示文本 token 的长度。然后利用求和池化操作获取其中的候选关键词的向量表示:

$$\boldsymbol{e}_k=\mathrm{SumPooling}([\boldsymbol{e}_i,\boldsymbol{e}_{i+1},\cdots,\boldsymbol{e}_j])。\tag{2}$$

式中:i 和 j 分别为候选关键词开始位置和结束位置的索引。将文档和候选短语嵌入送入线性层,然后进行非线性激活:

$$\begin{cases}\boldsymbol{z}_d=\tanh(\boldsymbol{W}_d\boldsymbol{e}_0+\boldsymbol{b}_d);\\ \boldsymbol{z}_k=\tanh(\boldsymbol{W}_k\boldsymbol{e}_k+\boldsymbol{b}_k)。\end{cases}\tag{3}$$

式中:$\boldsymbol{W}_d$、$\boldsymbol{W}_k$、$\boldsymbol{b}_d$、$\boldsymbol{b}_k$ 均为可学习的参数。

在数据集中存在一些噪声样本,它们可能对模型的训练产生负面影响[22]。通过动态调整样本信息,可以根据样本的重要性或可信度对样本进行加权,减少噪声样本的权重,从而降低它们对模型的影响。这种自适应的学习方式可以提高模型的鲁棒性和适应能力,进而增强关键词预测的性能。因此为了降低负样本采样偏差,减少假负样本在嵌入空间中的影响,将获取到的关键词与相应文档的嵌入表示通过对比学习进行语义对齐。将候选短语中真值及其对应的文档设置为"正对",而其余的候选词和文档则设置为"负对",去除负样本中的假负样本,将与原句特征语义相似度高于阈值 t 的前 k 个负样本视为假负样本,即

$$F_j=\{1>\mathrm{sim}(\boldsymbol{X},\boldsymbol{c}_{j_-})>t\cap\mathrm{sim}(\boldsymbol{X},\boldsymbol{c}_{j_-}),\mathrm{sim}(\boldsymbol{X},\boldsymbol{c}_{j_-})\in\mathrm{top}(\mathrm{sim}(\boldsymbol{X},\boldsymbol{c}_{j_-}),k)\}。\tag{4}$$

式中:$\mathrm{sim}(\boldsymbol{X},\boldsymbol{c}_{j_-})$ 为文档和关键词负样本的余弦相似度。当"正对"$(\boldsymbol{z}_d,\boldsymbol{z}_{k,i}^{+})$ 对应文档中有 N 个候选对时,训练目标定义为

$$\ell_i^{\mathrm{ad}}=-\log\frac{s(\boldsymbol{z}_d,\boldsymbol{z}_{k,i}^{+})}{\boldsymbol{S}+\sum_{s\in F_j}^{N}\alpha_{i,s}\cdot s(\boldsymbol{z}_d,\boldsymbol{z}_{k,s}^{-})}。\tag{5}$$

其中,

$$\boldsymbol{S}=s(\boldsymbol{z}_d,\boldsymbol{z}_{k,i}^{+})+\sum_{j=1,j\notin F_j}^{N}s(\boldsymbol{z}_d,\boldsymbol{z}_{k,j}^{-});\tag{6}$$

$$s(\boldsymbol{z}_d,\boldsymbol{z}_{k,i}^{+})=\mathrm{e}^{\mathrm{sim}(\boldsymbol{z}_d,\boldsymbol{z}_{k,i}^{+})/\tau};\tag{7}$$

$$\alpha_{i,s}=1-\frac{\exp(\mathrm{sim}(\boldsymbol{z}_i,\boldsymbol{z}_s^{+}))}{\sum_{j=1,j\neq i}^{N}\exp(\mathrm{sim}(\boldsymbol{z}_i,\boldsymbol{z}_j^{+}))},s\in F_j。\tag{8}$$

式中:τ 为温度参数;$\mathrm{sim}(\cdot,\cdot)$ 表示 2 个向量之间的余弦相似度;α 为使用自适应加权方法动态调整的相似性权重,该权重为检测到的假负样本为负样本的置信度分数。如果检测到的假负样本与原文的余弦相似度较高,则识别出的假负样本更有可能是正确的假负样本,将赋予其与原文之间的相似度较小的权重,降低假负样本带来的影响。相反,如果检测

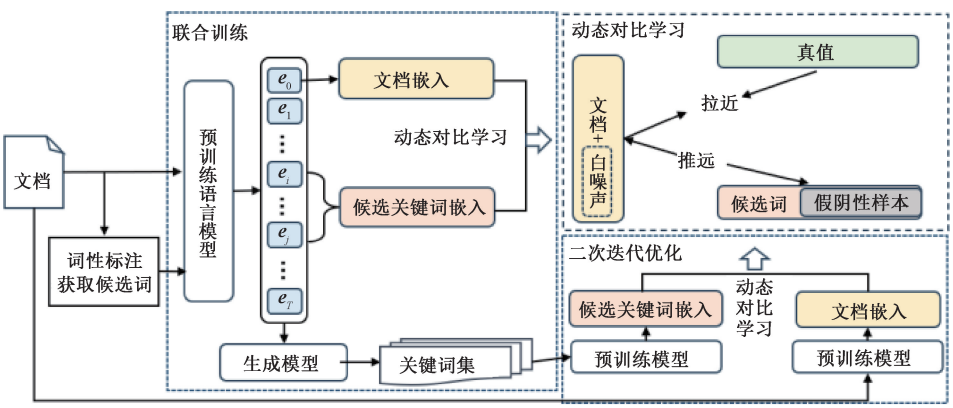


图 2 ACL-KP 模型整体框架

Figure 2 The overview of the framework of ACL-KP

到的假负样本与原文的余弦相似度较低,则该样本可能不是假负样本,赋予其与原文之间的相似性更大的权重。通过这种方式,α 权重帮助模型更加精确地区分真假负样本,优化学习过程,提高句子嵌入的质量和准确性。

为了进一步提高嵌入表示的作用,给数据增加一个高斯白噪声。首先生成一个初始的扰动 $\boldsymbol{\delta}_i$,这个扰动是从各向同性高斯分布中采样得到的高斯白噪声。优化扰动 $\boldsymbol{\delta}_i$ 来最大化实例判别任务的损失,从而找到最优的扰动 $\boldsymbol{\delta}_i^*$ 使模型在训练过程中自动创造一些虚拟的困难样本:

$$\boldsymbol{z}_d^{\delta} = \boldsymbol{z}_d + \boldsymbol{\delta}_i^* 。 \tag{9}$$

式中:$\boldsymbol{\delta}_i^*$ 为选取的最优的噪声。

最终的训练目标为

$$\ell_i^{\mathrm{ad}} = -\log \frac{s(\boldsymbol{z}_d^{\delta}, \boldsymbol{z}_{k,i}^{+})}{\boldsymbol{S}^{\delta} + \sum_{s \in F_j}^{N} \alpha_{i,s} \cdot s(\boldsymbol{z}_d^{\delta}, \boldsymbol{z}_{k,s}^{-})} 。 \tag{10}$$

给定文档的全部损失为

$$\ell_{\mathrm{CL}}^{\mathrm{ad}} = \sum_{i=1}^{|y_p|} \ell_i^{\mathrm{ad}} 。 \tag{11}$$

动态对比学习算法的基本流程如图 3 所示。

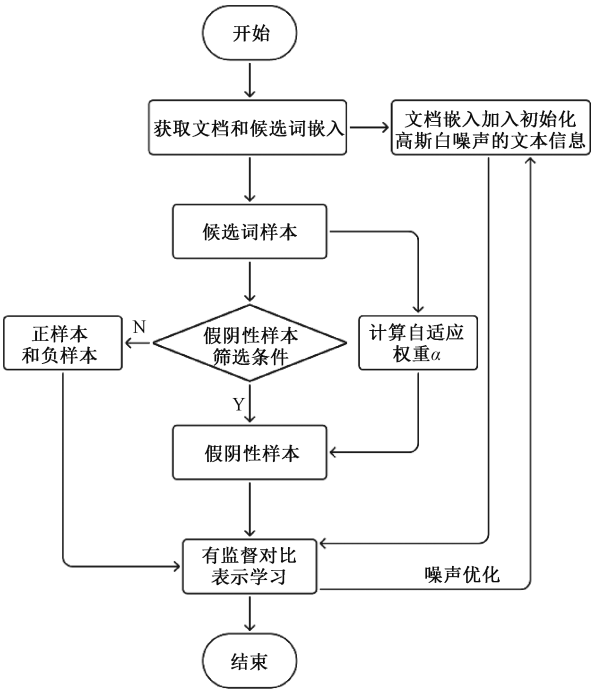


图 3　动态对比学习算法流程图

Figure 3　Flowchart of dynamic contrastive learning algorithm

2.2.2　多目标优化

对候选关键词和文本进行对比学习训练,同时通过学习文本数据中的关键字序列来生成关键词。在这个过程中,模型学习文本中关键词序列的概率分布,生成与原文内容相关的关键词。因此模型通过学习文本中的关键字序列 $y_a = \{y_i\}_{i=1,2,\cdots,|y_a|}$ 上的概率分布 $p_\theta(y_a)$ 来生成关键词。其中 θ 表示模型参数,反映了在给定模型参数 θ 下,序列 y_a 出现的概率。利用最大似然估计(maximum likelihood estimation,MLE)找到能够使关键词序列出现概率最大的参数值。对于生成关键字的任务 MLE 目标为

$$\ell_a = -\frac{1}{|y_a|} \sum_{t=1}^{|y_a|} \log p_\theta(y_a \mid y_{a<t}) 。 \tag{12}$$

目标函数旨在最大化所有可能的关键字序列的概率,即模型参数应该选择使得所有观察到的关键字序列的概率最大的值。将对比损失与负对数似然损失相结合,训练模型来生成关键短语:

$$\ell = \ell_a + \lambda \ell_{\mathrm{CL}}^{\mathrm{ad}} 。 \tag{13}$$

式中:λ 为平衡损失的超参数。通过结合 MLE 目标和对比损失,训练一个能够生成与原文内容相关的关键字的模型。这种模型不仅可以学习文本中关键字的概率分布,还可以通过区分正负样本来提高生成关键字的质量。通过调整超参数,可以在生成关键字的准确性和区分能力之间找到平衡。

2.2.3　二次迭代优化

在进行关键词生成时,通常会采用较大的波束进行解码。然而,这可能导致未出现在原文中的关键词生成过程中产生许多重复和噪声较多的短语[23]。为了解决这个问题,删除重复的关键词并重新排序,以确保每个唯一的短语都被独立地发送到编码器进行处理,挑选与原文相似度更高的前 k 个关键词进行二次迭代优化。

为了对给定文档中的相关关键词进行高效排序,同样采用了前文所述的动态对比学习的方法生成关键词。将未出现在原文中的关键词表示与对应的文档进行语义对齐,训练模型学习它们之间的关联,通过对比学习的训练目标公式(式(10))来不断优化关键词的生成质量。

3　实验

3.1　数据集及预处理

采用关键词预测领域中公开的数据集进行实验,分别是 Inspec[24]、Krapivin[25]、NUS[26]、SemEval[27]和 KP20k[6]。数据集的统计结果如表 1 所示。将数据集中的标题和正文连接起来作为源文本,并使用最大的数据集 KP20k 作为模型的训练集,其中训练集包含了 530 809 个数据样本。

3.2　实验设置

本文实验所用服务器搭载 Ubuntu 20.04.6 操作系统,实验均由 8 张 NVIDIARTX A4000(16 GB

表1 数据集的统计数据

Table 1 Statistics of dataset

数据集	样本总数	平均长度	关键词平均数量	present关键词占比/%	absent关键词占比/%
Inspec	500	2.48	9.83	73.6	26.4
Krapivin	400	2.21	5.84	55.7	44.3
NUS	211	2.22	10.85	54.4	45.6
SemEval	100	2.15	14.97	44.3	55.7
KP20k	20 000	2.04	5.27	62.9	37.1

显存)显卡完成。使用PyTorch框架,编程语言采用Python3.8,CUDA版本为11.6。训练过程中*Batch-size*设为8,共迭代10轮,初始学习率设为5×10^{-5},利用BART作为编码器解码器模型,用AdamW优化权重以及调整超参数以最大化验证集上的$F1@M$。在实验过程中,候选短语的最大n-gram长度设置为6,并将λ设置为0.3平衡损失。在动态对比学习损失中,将k设置为3,高斯白噪声通过迭代的方式选取最优的白噪声。生成关键词时,使用光束搜索,光束大小为50。在推理过程中,如果预测的数量少于5个,检索最前面的相似短语,直到获得5个。

3.3 基准方法

大多数生成模型遵循catSeq[28],这是One2Seq范式下的Seq2Seq框架,对比分析catSeq及其变体catSeqTG[29]、catSeqTG-2RF1[30]的性能。统一模型结合提取和生成方法来预测关键词,对比分析最新的模型UniKeyphrase[31]、PromptKP[32]和SimCKP[20]等的性能。

3.4 评估指标

本文选择$F1@M$和$F1@5$作为评价指标。$F1@M$表示比较所有的预测关键词和真实关键词的$F1$分数。$F1@5$表示比较前5个预测关键词和真实关键词的$F1$分数,但如果模型预测的关键词少于5个,将随机添加不正确的关键词,直到得到5个预测关键词。

3.5 实验结果及分析

表2和表3分别为ACL-KP和对比方法在文本中存在的关键词和文本中缺失的关键词2种类型上的实验结果,其中,SimCKP*的数据结果是在相同服务器上的复现实验,结果与原文一致。

由表2和表3可知,ACL-KP在大多数数据集上的性能超过了对比模型。表2中,ACL-KP在Inspec、Krapivin、SemEval、KP20k数据集上的$F1@5$和$F1@M$指标均有所提升。

表2 原文本中存在关键词上的实验结果

Table 2 Results of present keyphrase prediction

方法	$F1@5$					$F1@M$				
	Inspec	Krapivin	NUS	SemEval	KP20k	Inspec	Krapivin	NUS	SemEval	KP20k
catSeq	0.225	0.269	0.323	0.242	0.291	0.262	0.354	0.397	0.283	0.367
catSeqTG	0.229	0.282	0.325	0.246	0.292	0.270	0.366	0.393	0.290	0.366
catSeqTG-2RF1	0.253	0.300	0.375	0.287	0.321	0.301	0.369	0.433	0.329	0.386
UniKeyphrase	0.260	—	0.415	0.302	0.347	0.288	—	0.443	0.322	0.352
PromptKP	0.260	—	0.412	0.329	0.351	0.294	—	0.439	0.356	0.355
SimCKP*	0.352	0.386	0.494	0.391	0.425	0.354	0.383	0.495	0.390	0.426
ACL-KP($k=1$)	0.355	0.399	0.487	0.391	0.426	0.357	0.400	0.487	0.389	0.427
ACL-KP($k=3$)	0.360	0.410	0.489	0.394	0.429	0.363	0.413	0.488	0.395	0.430
ACL-KP($k=5$)	0.359	0.408	0.489	0.385	0.428	0.365	0.408	0.490	0.385	0.429

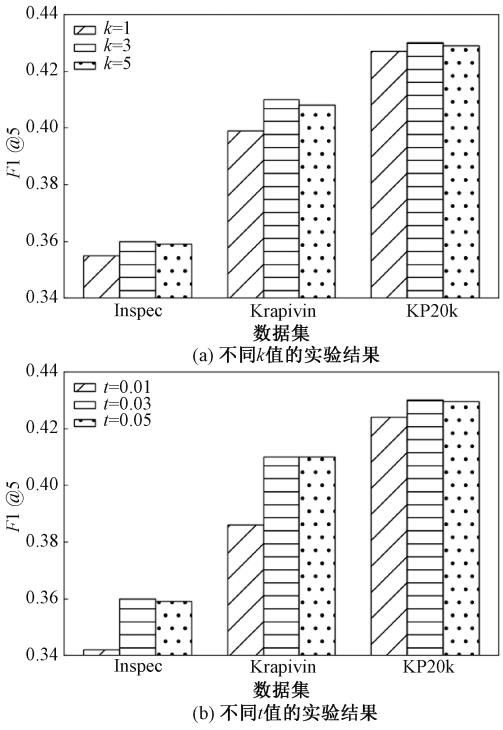
表3 原文本中缺失关键词上的实验结果

Table 3 Results of absent keyphrase prediction

方法	$F1@5$					$F1@M$				
	Inspec	Krapivin	NUS	SemEval	KP20k	Inspec	Krapivin	NUS	SemEval	KP20k
catSeq	0.004	0.018	0.016	0.016	0.015	0.008	0.036	0.028	0.028	0.032
catSeqTG	0.005	0.018	0.011	0.011	0.015	0.011	0.034	0.018	0.018	0.032
catSeqTG-2RF1	0.012	0.030	0.019	0.021	0.027	0.021	0.053	0.031	0.030	0.050
UniKeyphrase	0.026	—	0.045	0.045	0.046	0.036	—	0.056	0.052	0.068
PromptKP	0.017	—	0.036	0.028	0.032	0.022	—	0.042	0.032	0.042
SimCKP*	0.023	0.069	0.068	0.046	0.065	0.026	0.076	0.071	0.050	0.066
ACL-KP	0.027	0.065	0.084	0.041	0.074	0.028	0.073	0.092	0.051	0.079

在表 3 中,ACL-KP 在 Inspec、NUS、SemEval、KP20k 数据集上的实验结果也有所改善。在少数数据集上,ACL-KP 并未达到最优表现。这可能是因为不同数据集具有不同的分布特征,导致文本中存在关键词和缺失关键词的比例不同,从而影响了 ACL-KP 在不同数据集上的实验效果,而且不同数据集可能涵盖不同领域或主题,因此关键词的分布可能受到这些领域特征的影响。关键词的出现方式和上下文可能因数据集的不同而异。此外,数据集的规模和质量也可能对 ACL-KP 的表现产生影响。

对实验过程中的部分超参数进行分析,特别关注了动态对比学习中 k 值和阈值 t 对 ACL-KP 性能的影响,如图 4 所示。固定阈值 $t=0.03$,并观察了不同 k 值对性能的影响,如图 4(a)所示。k 值增大,意味着加权调整的样本数增多。k 值较小时,ACL-KP 效果会随 k 值增大而发生变化,但当 k 值过大时,ACL-KP 效果在大多数数据集上下降,如表 2 所示。因为 k 值过大会导致一定数量的负样本也进行加权调整,降低了负样本在训练过程中的作用,因此会导致 ACL-KP 性能下降。给定 $k=3$ 并设定 $t=0.01,0.03,0.05$,如图 4(b)所示。t 值增大,ACL-KP 的效果有所提升。$t=0.05$ 和 $t=0.03$ 时实验效果相似,这是因为当阈值过大时筛选的样本主要取决于当前 k 值。



(a) 不同k值的实验结果

(b) 不同t值的实验结果

图 4　不同参数设置对实验结果的影响

Figure 4　Impact of different parameter settings on experimental results

3.6　消融实验

本文采用消融实验验证动态对比学习和高斯白噪声的作用。表 4 为在样本数最多的 KP20k 数据集和样本数最少的 SemEval 数据集上的实验结果。

表 4　消融实验

Table 4　The ablation experiments

方法	$F1@5$	
	KP20k	Semeval
ACL-KP	**0.074**	**0.041**
w/o 假负样本消除	0.072	0.039
w/o 高斯白噪声	0.070	0.037

去除这两部分模块会导致模型性能不同程度地下降,从而证明了 ACL-KP 方法的有效性。在文本嵌入表示中,去除假负样本消除机制可能导致模型错过一些重要的正样本,从而削弱模型识别相关文本的能力。移除高斯白噪声会使模型对数据中的微小变化更加敏感,因为模型缺乏学习如何忽略或适应噪声的机会,这可能导致模型在测试数据上表现不佳。实验结果进一步表明,假负样本消除和高斯白噪声在提升文本嵌入表示性能方面具有关键作用。假负样本消除确保模型能够捕捉所有重要信息,而高斯白噪声则通过增强模型对噪声的抗干扰能力来提高其泛化性能。

4　案例分析

为了更好地展示实验结果,给出了测试集中的一个案例。如图 5 所示,其结果包含了出现在原文中的关键词和未出现在原文中的关键词。由图 5 可知,减少了噪声样本的干扰,模型更好地识别了“percolative phase transition”,并非 SimCKP 方法预测的“phase transition”。

原文档: Social percolation and the influence of mass media. <eos> In the marketing model of Solomon and Weisbuch, people buy a product only if their neighbours tell them of its quality, and if this quality is higher than their own quality expectations. Now we introduce additional information from the mass media, which is analogous to the ghost field in percolation theory. The mass media shift the percolative phase transition observed in the model, and decrease the time after which the stationary state is reached.

目标关键词:
原文中出现的: stationary state; social percolation; ghost field; quality expectations; percolative phase transition
原文中未出现的: cinema; solomon-weisbuch marketing model; customers; external field

SimCKP:
present: Social percolation; percolation; field; marketing; phase transition;
absent: sociophysics; mass media influence; econophysics; monte carlo simulation; cybernetics;

ACL-KP:
present: field; percolation; Social percolation; marketing; percolative phase transition;
absent: mass media influence; marketing models; cybernetics; e-commerce; percolative process

图 5　案例文本的关键词预测结果对比

Figure 5　Comparison of keyphrase prediction results for a sample case

5 结论

本文提出了一种动态对比增强表示嵌入方法提高关键词的预测性能。在对比学习过程中通过自适应的方式动态调整样本信息,并减少噪声样本的影响以此增强文本表示。在关键词数据集上的实验结果表明,本文方法在原文中存在关键词的抽取和原文中缺失关键词的生成上均优于近年来的最优基准方法。对比学习是本文方法的核心组成部分,下一步,将考虑改进对比学习策略,在对比学习过程中探索更多的策略和技巧。例如,引入更复杂的样本选择机制、设计更有效的对比损失函数或结合强化学习等方法,以进一步提升模型的性能,并将其应用于更广泛的关键词相关任务中。

参考文献:

[1] ZHAI C X, LAFFERTY J. A study of smoothing methods for language models applied to ad hoc information retrieval [J]. ACM SIGIR Forum, 2017, 51(2): 268-276.

[2] HAMMOUDA K M, MATUTE D N, KAMEL M S. CorePhrase: keyphrase extraction for document clustering[C]//LectureNotes in ComputerScience. Berlin: Springer, 2005: 265-274.

[3] BAI H L, CHEN Z B, LYU M R, et al. Neural relational topic models for scientific article analysis[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. New York: ACM, 2018: 27-36.

[4] WANG Z H, WANG D, LI Q. Keyword extraction from scientific research projects based on SRP-TF-IDF[J]. Chinese Journal of Electronics, 2021, 30(4): 652-657.

[5] MIHALCEA R, TARAU P. Textrank: bringing order into text[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2004: 404-411.

[6] MENG R, ZHAO S Q, HAN S G, et al. Deep keyphrase generation[EB/OL]. (2021-05-31)[2024-06-05]. http://arxiv.org/abs/1704.06879.

[7] ZHAO G Z, YIN G S, YANG P, et al. Keyphrase generation via soft and hard semantic corrections[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2022: 7757-7768.

[8] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2999-3007.

[9] HUANG S J, CAI N G, PACHECO P P, et al. Applications of support vector machine (SVM) learning in cancer genomics[J]. Cancer Genomics & Proteomics, 2018, 15(1): 41-51.

[10] BENNANI-SMIRES K, MUSAT C, HOSSMANN A, et al. Simple unsupervised keyphrase extraction using sentence embeddings[EB/OL]. (2018-09-05)[2024-06-05]. http://arxiv.org/abs/1801.04470.

[11] LIANG X N, WU S Z, LI M, et al. Unsupervised keyphrase extraction by jointly modeling local and global context[EB/OL]. (2021-09-15)[2024-06-05]. http://arxiv.org/abs/2109.07293.

[12] ZHANG L H, CHEN Q, WANG W, et al. MDERank: a masked document embedding rank approach for unsupervised keyphrase extraction[EB/OL]. (2023-02-28)[2024-06-05]. http://arxiv.org/abs/2110.06651.

[13] CHEN J, ZHANG X M, WU Y, et al. Keyphrase generation with correlation constraints[EB/OL]. (2018-08-22)[2024-06-05]. http://arxiv.org/abs/1808.07185.

[14] GAO T Y, YAO X C, CHEN D Q. SimCSE: simple contrastive learning of sentence embeddings[EB/OL]. (2022-05-18)[2024-06-05]. http://arxiv.org/abs/2104.08821.

[15] VAN DEN OORD A, LI Y Z, VINYALS O. Representation learning with contrastive predictive coding[EB/OL]. (2018-09-05)[2024-06-05]. http://arxiv.org/abs/1807.03748.

[16] XU J H, SHAO W, CHEN L H, et al. SimCSE++: improving contrastive learning for sentence embeddings from two perspectives[EB/OL]. (2023-10-20)[2024-06-05]. https://arxiv.org/abs/2305.13192.

[17] HOU P Y, LI X Y. Improving contrastive learning of sentence embeddings with Focal-InfoNCE[EB/OL]. (2023-10-20)[2024-06-05]. http://arxiv.org/abs/2310.06918.

[18] LI J C, SHANG J B, MCAULEY J. UCTopic: unsupervised contrastive learning for phrase representations and topic mining[EB/OL]. (2022-02-27)[2024-06-05]. http://arxiv.org/abs/2202.13469.

[19] CAI H, CHEN W H, SHI K H, et al. Keyword extractor for contrastive learning of unsupervised sentence embedding[C]//Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing. New York: ACM, 2022: 88-93.

[20] CHOI M, GWAK C, KIM S, et al. SimCKP: simple contrastive learning of keyphrase representations[EB/OL]. (2023-10-12)[2024-06-05]. http://arxiv.org/abs/2310.08221.

[21] LEWIS M, LIU Y H, GOYAL N, et al. BART: denois-

ing sequence-to-sequence pre-training for natural language generation, translation, and comprehension[EB/OL]. (2019-10-29)[2024-06-05]. https://arxiv.org/abs/1910.13461.

[22] CHUANG C Y, ROBINSON J, YEN-CHENL, et al. Debiased contrastive learning[EB/OL]. (2020-10-21)[2024-06-05]. http://arxiv.org/abs/2007.00224.

[23] ZHAO G Z, YIN G S, YANG P, et al. Keyphrase generation via soft and hard semantic corrections[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi: EMNLP, 2022: 7757-7768.

[24] HULTH A. Improved automatic keyword extraction given more linguistic knowledge[C]//Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing Not Known. Stroudsburg: Association for Computational Linguistics, 2003: 216-223.

[25] KRAPIVIN M, AUTAEU A, MARCHESE M. Large dataset for keyphrases extraction[EB/OL]. (2008-05-01)[2024-06-05]. https://iris.unitn.it/bitstream/11572/358576/1/disi09055-krapivin-autayeu-marchese.pdf.

[26] NGUYEN T D, KAN M Y. Keyphrase extraction in scientific publications[C]//International Conference on Asian Digital Libraries. Berlin: Springer, 2007: 317-326.

[27] KIM N S, MEDELYAN O, KAN M Y, et al. SemEval-2010 task 5?: automatic keyphrase extraction from scientific articles[J]. Language Resources and Evaluation, 2010,47(3): 21-26.

[28] YUAN X D, WANG T, MENG R, et al. One size does not fit all: generating and evaluating variable number of keyphrases[EB/OL]. (2020-05-12)[2024-06-05]. http://arxiv.org/abs/1810.05241.

[29] CHEN W, GAO Y F, ZHANG J N, et al. Title-guided encoding for keyphrase generation[EB/OL]. (2019-01-16)[2024-06-05]. https://arxiv.org/abs/1808.08575.

[30] CHAN H P, CHEN W, WANG L, et al. Neural keyphrase generation via reinforcement learning with adaptive rewards[EB/OL]. (2019-06-10)[2024-06-05]. http://arxiv.org/abs/1906.04106.

[31] WU H Q, LIU W, LI L, et al. UniKeyphrase: a unified extraction and generation framework for keyphrase prediction[EB/OL]. (2021-08-31)[2024-06-05]. http://arxiv.org/abs/2106.04847.

[32] WU H Q, MA B, LIU W, et al. Fast and constrained absent keyphrase generation by prompt-based learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 11495-11503.

Dynamic Contrastive Representation Enhancement Approach for Keyphrase Prediction

GENG Xuelian[1,2], SONG Mingyang[1,2], FENG Yi[1,2], JING Liping[1,2], YU Jian[1,2]

(1. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China; 2. Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Keyphrase prediction often fail to fully utilize the complex hierarchies and semantic information within text structures. To address this issue, a novel keyphrase prediction method that enhanced semantic representation, called adaptive contrastive learning for keyphrase prediction (ACL-KP) was proposed. This method introduced an adaptive weighting mechanism to dynamically adjust sample weights, to solve the problem of distinguishing true samples from noise samples during contrastive learning, thereby reducing the impact of misidentifying noise samples and optimizing spatial representation. Additionally, to increase the diversity of training data, Gaussian white noise was incorporated to automatically generate some challenging virtual samples, thus enhancing the semantic representation of documents and keywords. Experimental results on multiple public datasets in the keyphrase prediction field showed that the model improved performance by 2% to 17% in $F1@5$ and $F1@M$ metrics compared to current state-of-the-art models. Compared to sequence-to-sequence models and unified models, the proposed model demonstrated a more significant performance advantage.

Keywords: natural language processing; keyphrase prediction; multi-objective optimization; contrastive learning; embedding representation