徐嘉成. 基于对比学习的明喻修辞识别[J]. 智能计算机与应用,2025,15(7):162-167. DOI:10.20169/j.issn.2095-2163.250724

基于对比学习的明喻修辞识别

徐嘉成
(贵州大学 计算机科学与技术学院, 贵阳 550025)

摘 要: 明喻句识别是文本数据挖掘相关研究工作的重要组成部分,其任务主要包括:明喻句判定和明喻组件识别两项子任务,其中明喻句的判定容易受到带喻词的非明喻句等噪声语句的影响,造成模型判定错误,进而影响识别率。为了解决该问题,现阶段研究人员期望通过增强训练数据或构造复杂大模型的方式来提升识别效果,但该解决方案会产生较大的人工标注及算力成本。本文提出了一种基于对比学习的明喻识别算法,在不进行复杂建模和额外的人工标注成本的情况下,提升明喻句识别效果。首先,引入对比学习算法,通过特定的数据增强方法构造正负例样本,对大量易获得的无标签比喻相关句子进行预训练,在高维空间优化句子表征,再将此表征迁移到下游明喻识别任务;其次,在下游任务中拼接了 BERT 特征提取器的字、句嵌入,增强模型对明喻句的语义理解能力。在公开的中文明喻数据集 CSR 上的实验结果表明,本文所提算法识别长句子的性能优于现有的明喻识别 SOTA,与多个算法的比较中取得了最高的精确率,对明喻句判断任务有较好的效果。

关键词: 明喻识别; 对比学习; BERT
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 2095-2163(2025)07-0162-06

Simile recognition based on contrastive learning

XU Jiacheng
(College of Computer Science and Technology, Guizhou University, Guiyang 550025, China)

Abstract: Simile recognition is an important component of research related to text data mining. Its tasks mainly include two sub tasks: simile sentence judgment and simile component recognition. The judgment of simile sentences is easily affected by noise statements such as non simile sentences with metaphorical words, resulting in model judgment errors and thus affecting recognition rate. In order to solve this problem, current researchers hope to improve recognition performance by enhancing training data or constructing complex large models, but the above solutions will incur significant manual annotation and computational costs. This article proposes a simile recognition algorithm based on contrastive learning, which improves the recognition effect of simile sentences without complex modeling and additional manual annotation costs. This model first introduces a contrastive learning algorithm, constructs positive and negative example samples through specific data augmentation methods, trains a large number of easily obtained unlabeled metaphorical sentences, optimizes sentence representation in high-dimensional space, and then transfers this representation to downstream simile recognition tasks. Secondly, in downstream tasks, the BERT feature extractor's word and sentence embedding are concatenated to enhance the model's semantic understanding ability of simile sentences. The experimental results show that on the publicly available Chinese simile dataset CSR, the proposed method outperforms the existing simile recognition SOTA in both small sample training and long sentence recognition, and has good performance in simile sentence judgment tasks.

Key words: simile recognition; comparative learning; BERT

0 引 言

明喻句是比喻句的一种类型,存在“像”、“如”等喻词连接两个对象(本体和喻体)。带常见喻词的句子见表 1。通常明喻识别有两个子任务:判断一个句子是否是明喻句和明喻组件抽取(本体和喻体),本文针对第一个任务做算法研究。明喻识别可以帮助汉语学习者理解书本或小说里面出现的句子,对于教学中的文本鉴赏是有帮助的,研究中文明喻句的识别对于对话系统、情绪分析、语言理解、文学研究和教育等领域都具有重要的意义[1]。

对比学习近年来受到了广泛的关注,主要是为了解决在有监督深度学习算法中,受标注样本制约模型效果的问题,属于无监督学习的一种[2]。现有的明喻数据样本数较少,模型难以充分学习到类间

作者简介: 徐嘉成(1998—),男,硕士,主要研究方向:自然语言处理,作文自动评分。Email:xujiacheng5334@163.com。

收稿日期: 2023-11-22

关系，本文引入对比学习算法在高维空间上优化句子表示，聚合同类数据，拉远异类数据，降低数据敏感性，提高模型鲁棒性。

表 1　包含喻词的句子

Table 1　Sentences containing metaphorical words

1 天空中的片片雪花，像微风拂絮	明喻句
2 蜘蛛像往常一样织蜘蛛网。	常规句
3 他像他爸爸一样高了。	比较句

本文设计了基于对比学习的明喻识别算法。在前置任务中利用预训练模型 BERT（Bidirectional Encoder Representation from Transformer），通过对比学习在大量无标签的明喻句和非明喻句上进行训练，帮助模型学习到明喻句和非明喻句之间的差异和共性特征，让特征提取器 BERT 能够获得更好的句子表征向量，使模型具有良好的感知困难负样本的能力，提高模型的抗干扰能力，削弱困难负样本对模型的判断能力带来的影响，再将学习到的信息迁移到下游明喻识别任务。

1　相关工作

1.1　明喻识别

明喻句识别可以视为文本特征由本体、喻体、喻词（触发词）、喻解构成的分类任务。存在以下两点挑战，第一是明喻句和带喻词的非比喻句句子的语法结构和语义结构非常相似，难以将标准的自然语言理解（NLU）技术（如句法依存分析和语义依存分析）应用上来；第二是喻词虽然可以为明喻识别任务提供一些提示，但是在非比喻句里面，一些喻词如“像”、“如”也会被频繁使用，给明喻句识别任务带来大量的噪声。

目前针对明喻句已有研究，根据特征提取方法的不同，分为手工特征和深度学习特征。2008 年，李斌[3]提出了基于命名实体识别（CRF）的明喻识别，需要对数据集进行大量的人工标注，对中文比喻句自动识别进行了初步的探讨和研究；Niculae[4]使用自定义的规则对比喻句进行成分解析，通过分析文本的主谓宾结构和词汇之间的从属关系，提出了基于语法结构的比喻识别算法。随着深度学习不断融入到自然语言处理领域，2016 年，Liu[5]首次使用双向长短期记忆网络（Bi-LSTM）进行名语句识别，采取了多任务学习的方法联合优化 3 个任务，让明喻识别和组件抽取共享了句子表征信息，在深度学习领域进行了明喻识别的探索，并贡献了中文明喻数据集 CSR（Chinese - Simile - Recognition）；穆婉青[6]使用卷积神经网络（CNN）作为特征提取器进行实验，在比喻句识别任务上准确率达到了 94.7%；沿着这个思路，Zhang[7]在 Liu[5]的基础上融合了词性信息和自注意力机制，在嵌入层使用了 word2vec 技术，由于中文词语的多义性，任务性能没有较好的效果；Zeng[8]提出了一个结合 BERT 的循环多任务学习模型，两个子任务和一个额外的语言建模子任务堆叠成循环任务来互相促进明喻句的识别。2022 年，Wang[9]将异构图建模方法应用到明喻识别中，基于词语的词性、句法结构、词语外部知识构建异构图，丰富编码端信息，进一步提升了明喻识别的效果，但是该方法需要构建复杂的异构体神经网络模型，人工标注额外的输入特征。

1.2　对比学习

对比学习是 2019 年末在计算机视觉（Computer Vision，CV）领域开始兴起的预训练方法。2020 年，Facebook[10]发布了 MoCo（Momentum Contrastive Learning）模型，谷歌[11]发布了 SimCLR 框架（Contrastive Learning of Visual Representations），这些模型基于类孪生神经网络的网络架构，训练过程中所用的图像对为同一幅图像分别增广后得到的图像对（正样本对）或不同图像分别增广后构成的图像对（负样本对）。对比学习通过大量的正负样本对间的比对计算，使得神经网络模型能够对数据自动提取到更好的特征表达，自此对比学习开始在表征预训练领域大放光彩[12]。

2021 年陈丹琦团队[13]发布了 SimCSE 算法，使用对比学习的训练框架，在维基百科语料上微调 BERT，将句子级表征向量提升到一个新的高度，成为句子相似度任务的 SOTA；2022 年奚琰[14]利用对比学习来优化模型对人脸表情图片的表征能力，实现了高效的遮挡状态下的人脸表情识别；高怡[15]在原型网络的基础上引入对比学习，从高维空间上优化句子表示，提高小样本事件检测的鲁棒性。本文探索利用自监督学习从大量易获得的无标签文本数据提升特征提取器的明喻句表征能力，并将其用于明喻识别任务。

2　基于对比学习的比喻识别算法

基于对比学习的明喻识别算法框架如图 1 所示，模块整体分成两个部分：

1）对比学习模块：在预训练阶段使用对比学习的策略，通过在网络中收集的大量比喻相关的句子进行无监督对比学习，从高维空间优化句子表示，将训练好特征提取器送入下游任务；

2)明喻识别模块:将经过对比学习训练的 BERT 模型的句嵌入[CLS]和字嵌入进行拼接,增强模型对比喻句的语义信息的理解。

BERT 模型最后一层有两个输出,一个是特殊符号[CLS],可作为文本的句子级表示,另一个是句子的序列输出,BERT 模型把句子按字分割之后,每个字以 768 维输出,可视为文本的字嵌入。将字嵌入通过 Bi-LSTM 融合上下文信息得到一个新的句嵌入,并将该语义信息通过多头注意力层(Multi-Head Attention)来捕捉句子中不同部分之间的关系,进一步加强语义信息,提升模型对句子的理解,和句嵌入[CLS]进行拼接操作,输入到输出层(Linear & Softmax)进行明喻句识别。

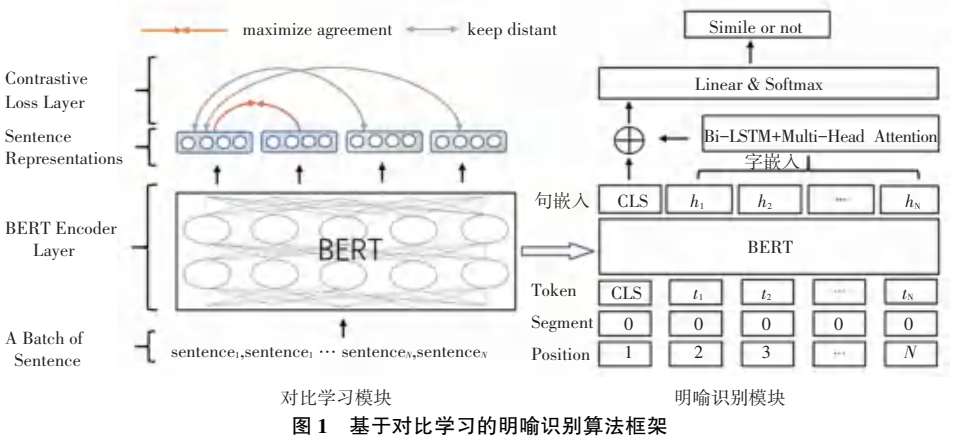

图 1 基于对比学习的明喻识别算法框架

Fig. 1 Architecture of simile recognition based on contrastive learning

2.1 对比学习模块

好的句向量表征有益于下游任务,对比学习不依赖标注数据,在不使用数据标签的情况下利用数据之间的差异进行学习,充分利用未标记数据。本文在从网络上收集到的大量明喻句和非明喻句上做无监督训练,从高维空间优化模型表征能力,得到更好的样本表示,使句嵌入更加均匀,进而应用到下游的明喻识别任务。

本文借鉴了 SimCSE 构造正负样例的方法,使用 Dropout 给文本输入加噪声来构成正负例。Dropout 是在神经网络中随机关闭掉一些神经元的连接,关闭的神经元不一样,模型最终的输出也就不一样,通过 Dropout 加噪声后的输入仍与原始输入在语义空间距离相近,将同一条样本送入 BERT 两次作为正样本,同一个 Batch 其他数据作为负样本。

如图 1 所示,假设在一个训练批 D 中有 N 条句子,记为 $D=X_n,\{n=1,\cdots,N\}$,对于 D 中的每条句子 X_i,都通过 Dropout 构造相应的正例对,得到属于 X_i 的一对正样本,分别记为 X_i 和 X_j,则该批次总共有 $2N$ 个样本,对每个样本对 X_i 和 X_j 利用 BERT 编码器提取特征向量 $\boldsymbol{h}_i$ 和 $\boldsymbol{h}_i^+$,采用目标函数 InfoNCE loss 进行训练,公式如下:

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

$$\mathrm{sim}(\boldsymbol{h}_i,\boldsymbol{h}_i^+)=\frac{\boldsymbol{h}_i\boldsymbol{h}_i^+}{\|\boldsymbol{h}_i\|\cdot\|\boldsymbol{h}_i^+\|} \tag{2}$$

其中,τ 表示温度系数,用于控制模型对正负样本之间差异的敏感程度,越小的温度系数越关注于将样本和最相似的困难样本分开,去得到更均匀的表示。

对比学习预训练阶段不需要额外的人工标注数据成本,通过充分利用来自同源样本不同视野的相似性和不同实例(Instance)的差异性,最大化地挖掘和学习明喻句和非明喻句之间的语义差别,得到的特征提取器迁移到下游明喻识别任务,提高明喻识别的性能。

2.2 明喻识别模块

2.2.1 BERT 预训练模型

BERT 模型广泛应用于自然语言处理任务中。模型输入向量由位置编码(Position Embedding)、段编码(Segment Embedding)和字符级编码(Token Embedding)相加而成,输入句子 $S=\{t_1,t_2,\cdots,t_n\}$,经过双向 Transformer 得到句子的编码向量 $\boldsymbol{H}=\{[\mathrm{CLS}],\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_N\}$,其中[CLS]较完整地保存了句子的语义信息,一般作 NLP 任务句子级表征,且 BERT 编码器在中文任务中是以字作为分隔单位进行编码,$\boldsymbol{h}_i$ 可以视为句子的字嵌入。

2.2.2 Bi-LSTM 层

明喻识别任务包含的句子长度不一,为了在长

句子中准确识别句子是否是明喻句，本文采用 Bi-LSTM 编码文本的全局语义信息，解决文本长距离依赖问题。Bi-LSTM 同时在两个方向上进行文本语义表达，将 BERT 的字嵌入 $\boldsymbol{h}_i$ 作为 Bi-LSTM 的输入，捕获上下文语义，正向和反向的上下文信息分别为 $\overrightarrow{h_t}$ 和 $\overleftarrow{h_t}$，公式如下：

$$\overrightarrow{h_t}=\overrightarrow{\mathrm{LSTM}}(\boldsymbol{h}_i,\overrightarrow{h_{i-1}}) \tag{3}$$

$$\overleftarrow{h_t}=\overleftarrow{\mathrm{LSTM}}(\boldsymbol{h}_i,\overleftarrow{h_{i-1}}) \tag{4}$$

将 $\overrightarrow{h_t}$ 和 $\overleftarrow{h_t}$ 进行拼接可以得到包含文本两个方向的上下文特征，获得的最终输出为 $\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_N)$，$\boldsymbol{h}_i=[\overrightarrow{h_t};\overleftarrow{h_t}]$。

2.2.3　多头注意力机制

多头注意力机制（Multi-Head Attention）可以将一个文本序列内的不同部分联系起来，帮助模型捕获到各个字对句子的影响力，有利于分析每个字不同的语义特征，对识别喻词是否有比喻语义是有帮助的。本文用多头注意力机制来建立时序向量 $\boldsymbol{h}_i$ 和 Bi-LSTM 最后一个时刻的输出向量 $\boldsymbol{h}_N$ 的关系，增强模型对不同特征的关注度，获取更细粒度的特征，实现更准确的识别效果。多头注意力机制的公式如下：

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{W}_{\mathrm{q}}\boldsymbol{h}_N\\ \boldsymbol{K}=\boldsymbol{W}_{\mathrm{k}}\boldsymbol{h}_i\\ \boldsymbol{V}=\boldsymbol{W}_{\mathrm{v}}\boldsymbol{h}_i\end{cases} \tag{5}$$

其中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别表示查询向量（query）、键向量（key）、和值向量（value），$\boldsymbol{W}_{\mathrm{q}}$、$\boldsymbol{W}_{\mathrm{k}}$、$\boldsymbol{W}_{\mathrm{v}}$ 是注意力权重矩阵。

$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$ 是将每个键向量的权重乘以值向量，进行加权求和得到带有注意力的文本表征，公式如下：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}k^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V} \tag{6}$$

其中，d_{k} 是键向量的维度。

$head_i$ 为每个注意力头的输出，公式如下：

$$\mathrm{head}_i=\mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{\mathrm{q}},\boldsymbol{K}\boldsymbol{W}_i^{\mathrm{k}},\boldsymbol{V}\boldsymbol{W}_i^{\mathrm{v}}) \tag{7}$$

使用 Softmax 进行归一化，将多个注意力头得到的表征进行一次线性变换，得到最终的多头注意力输出 $\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$，公式（8）：

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}[\mathrm{head}_1,\mathrm{head}_2,\cdots,\mathrm{head}_n]\boldsymbol{W}_h \tag{8}$$

其中，$\boldsymbol{W}_h$ 为输出变换矩阵。

2.2.4　输出层

本文在输出层使用 Softmax 作为特征分类器，送进输出层的文本表征是 BERT 的句嵌入[CLS]和 Bi-LSTM 与多头注意力机制的计算结果拼接，融合 BERT 的字、句嵌入，通过 Softmax 将输入转换为明喻或非明喻句的概率。

3　实验结果与分析

3.1　数据集

（1）目标数据集 CSR（Chinese - Simile - Recognition）：由 Liu 团队[5]贡献的中文明喻数据集，是目前最广泛使用的明喻数据集，一共有 11 337 条样本，沿着之前的工作，本文使用 5 折交叉验证，把数据集均匀分成 5 个子集，依次把其中 4 个子集中的 80%用于训练，20%用于验证，剩下的一个子集作为测试集；

（2）对比学习数据集：本文从各大作文网站上爬取了网站精选的比喻修辞手法相关的句子 30 万条，所有的句子都带有常见喻词“像”、“如”、“似”，并且把这 30 万条数据拆分成 1 万、2 万、5 万、10 万、30 万条来作为对比学习模块训练数据，测试预训练数据集大小对算法性能的提升，需强调的是预训练数据集未进行任何数据标注。

3.2　实验设置及评价指标

本文使用 Pytorch 深度学习框架在 Tesla P40 显卡上进行训练，采用 BERT-base 作为特征提取器，隐藏层的维度为 768，最大编码长度为 128，使用 AdamW 作为优化器，权重衰减系数为 1E-4，采用 4 个注意力头；对比学习阶段损失函数的温度系数 τ 为 0.05，训练 5 个 Epoch，学习率为 3E-5；明喻识别阶段训练 3 个 Epoch，学习率为 4E-5。

训练过程中依次选择 5 个数据集子集中的 4 个子集作为训练集，一个子集作为测试集，在训练集训练模型，在验证集上得到最好的训练参数，最后在测试集上进行评估，这个过程重复 5 次，确保每个子集都被用作测试集一次，将 5 次测试的性能度量结果取平均值，得到模型的最终性能评估。评价指标采用精确率（*precision*）、召回率（*recall*）和 *F*1 值。

3.3　实验结果

3.3.1　实验一　不同算法的明喻识别性能

本节设置如下算法进行对比实验，验证本文算法的有效性。为了保证公平性，所有对比算法均在 CSR 数据集进行实验。

MTL：该模型是 Liu[5]提出的一种基于深度学习的模型，Embedding 层使用 word2vec 方法获取，通过 Bi-LSTM 对文本建模进行明喻句识别。

Self_Attn+Pos:Zhang[7]提出的明喻识别模型,通过 word2vec 获取词嵌入,结合了词语的词性信息,并使用自注意力机制进行明喻句识别。

Cyc-MTL:Zeng[8]在 MTL 算法的基础上,将语言模型任务、明喻识别任务进行首尾连接,每次将前一个任务的输出作为输入,通过循环任务来增益明喻句的识别。

HGSR:Wang[9]基于词语的词性、句法结构、词语外部知识构建复杂异构图模型进行明喻识别,是目前明喻识别任务的 SOTA。

不同算法在 CSR 数据集上的实验结果见表 2。由表 2 可以看出,本文提出的算法在精确率上取得了最好的结果,比 HGSR 提高了 2.38%,在 *F*1 值上略低于 HGSR,但是本文的算法未使用复杂网络结构,未添加额外的输入输出特征,依靠无监督对比学习模块优化了句子在高维空间的表征,帮助下游任务取得识别效果的提升,融合了字句嵌入的明喻识别模块可以深度地捕获句子的语义信息,从而获得较好的识别性能。

表 2 不同算法在 CSR 数据集上的实验结果比较

Table 2 Comparison of experimental results of different algorithm on CSR %

算法	精确率	召回率	*F*1 值
MTL	80.84	92.20	86.15
Self_Attn+Pos	80.44	91.69	85.70
Cyc-MTL	85.81	**94.43**	89.92
HGSR	89.04	94.39	**91.64**
本文算法	**91.42**	91.15	91.18

3.3.2 实验二 消融实验

为了测试模型对明喻句的理解程度,本文为模型的各个部分编号:

①明喻识别模块的句嵌入部分:使用 BERT 模型的[CLS]作为句子的表征;②明喻识别模块的字嵌入部分:通过 Bi-LSTM 来获取句子的长距离依赖,用多头注意力机制获取句子不同部分对明喻句识别的权重;③对比学习预训练模块:对比学习的策略在高维空间优化句子表征。消融实验结果见表 3,其中"-"代表删去这个模块。

(1)明喻识别模块的句嵌入部分:移除模块①后模型 *F*1 值下降了 0.58%左右,可见句嵌入[CLS]能一定程度的表示句子表征,移除之后模型会失去部分句子的全局特征,*F*1 值会出现下降;

(2)明喻识别模块的字嵌入部分:移除模块②后 *F*1 下降了 2.05%左右,可见 Bi-LSTM 整合的上下文信息加上多头注意力机制精准地加权不同部分的特征,能深度捕获句子语义,该部分所包含的语义信息相对于模块①更多,移除后下降的 *F*1 更高;

(3)对比学习预训练模块:模块①加上模块③后 *F*1 值增加了 1.62%,模块②加上模块③后准确率略微下降,模块①加上模块②加上模块③后 *F*1 值上升了 0.91%。通过对比学习优化句子表征后,能够提高下游任务的性能,可见对比学习预训练模块通过学习同源样本不同表征下的相似性和不同样本之间的差异性,能提升明喻识别的能力。

表 3 消融实验

Table 3 Ablation experiment %

模块	精确率	召回率	*F*1 值
-①句嵌入	88.35	88.19	88.22
-②字嵌入	89.92	89.67	89.69
①句嵌入+②字嵌入	90.12	90.34	90.27
①句嵌入+③对比学习	90.98	90.58	89.84
②字嵌入+③对比学习	89.72	89.34	89.38
①+②+③(本文算法)	**91.42**	**91.15**	**91.18**

消融实验结果表明本文提出的使用对比学习进行预训练的方法是有效的,在模型对明喻句语义表征足够完整的情况下,进一步提升了明喻句识别的能力;直接使用 BERT 编码器进行 NLP 任务可以得到较好的结果,单独使用某个模块可能会造成 BERT 性能下降,可见基于对比学习的明喻识别算法的性能是各个模块共同作用的结果。

3.3.3 实验三 对比学习预训练数据规模对下游任务的影响

本实验测试在不同规模大小的无监督数据集训练对比学习模块对下游任务的影响如图 2 所示。从图 2 可以看到,通过加入对比学习模块,明喻句的识别性能均有提升,验证了使用对比学习方法的有效性。(0 表示没有添加对比学习模块,仅有明喻识别模块)

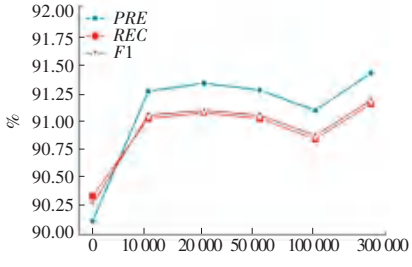

图 2 对比学习训练数据集规模对下游任务的影响

Fig. 2 Impact of contrastive learning training dataset size on downstream tasks

3.3.4　实验四 针对不同句子长度算法识别性能对比

为了验证解决句子长距离依赖时使用 Bi-LSTM 和多头注意机制的有效性，本文把数据集中的句子按照字符长度划分为 6 组，分别是 0~10 个字符，10~20 个字符，30~40 个字符，40~50 个字符，50 个字符以上，本文所提算法与 HGSR 算法在不同句子长度下的明喻识别性能对比，如图 3 所示。从图 3 可以看出，随着句子长度增加，本文的算法保持高效的识别性能，并且优于 HGSR。

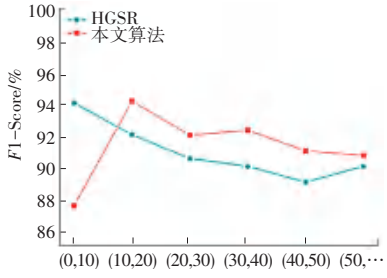

图 3　不同句子长度的识别性能对比

Fig. 3　Comparison of recognition performance for different sentence lengths

3.3.5　实验五 少样本下的算法性能对比

本实验测试了在少量样本的情况下所提算法识别明喻句的性能，分别选取了数据集中的 20%、40%、60%、80%、100%数据量进行训练，少样本下的算法性能对比结果如图 4 所示。由图 4 可以看出，与 HGSR 相比，在训练级数据量较少的时候，所提算法仍有 75%以上的 $F1$ 值，在 40%训练集数据量的时候达到一个较高的识别性能，表明本文所提算法在数据匮乏的时候依然有较好效果，通过利用比喻领域数据集进行无监督预训练，让特征提取器在明喻识别任务数据量较少的情况下进行微调就能实现高效的明喻识别。

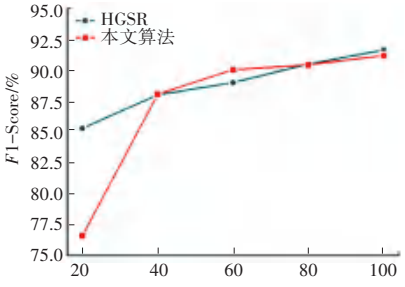

图 4　少样本下的算法性能对比

Fig. 4　Comparison of algorithm performance with few shot

4　结束语

本文提出了基于对比学习的明喻修辞识别算法，在不进行复杂建模以及耗费额外的人工标注成本的前提下，通过无监督对比学习来优化高维空间句子表征，并且通过对句子的字、句嵌入的融合，增强了特征提取器对句子的语义理解能力，提高了模型的明喻识别性能。实验证明了在明喻识别中引入对比学习的有效性，将来可以探索对比学习模块中不同的数据增强方法对明喻识别的提升，并将该思路推广到更多的中文修辞手法的识别中。

参考文献

[1] 赵琳玲，王素格，陈鑫，等. 基于词性特征的明喻识别及要素抽取方法[J]. 中文信息学报，2021，35(1)：81-87.

[2] 朱慧春. 基于对比学习和先验知识传播的深度半监督学习算法研究[D]. 上海：东华大学，2022.

[3] 李斌，于丽丽，石民. 基于 CRF 的汉语动词“像”的比喻义识别[C]//第四届全国学生计算语言学研讨会会议论文集. 2008：202-207.

[4] NICULAE V，YANEVA V. Computational considerations of comparisons and similes[C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013：89-95.

[5] LIU L，HU X，SONG W，et al. Neural multitask learning for simile recognition[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018：1543-1553.

[6] 穆婉青. 基于修辞格识别的鉴赏类问题解答方法研究[D]. 太原：山西大学，2018.

[7] ZHANG P，CAI Y，CHEN J，et al. Combining part-of-speech tags and self-attention mechanism for simile recognition[J]. IEEE Access，2019，7：163864-163876.

[8] ZENG J，SONG L，SU J，et al. Neural simile recognition with cyclic multitask learning and local attention[C]//Proceedings of AAAI Conference on Artificial Intelligence. AAAI，2020：9515-9522.

[9] WANG X，SONG L，LIU X，et al. Getting the most out of simile recognition[J]. arXiv preprint arXiv，2211.05984，2022.

[10] HE K，FAN H，WU Y，et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2020：9729-9738.

[11] CHEN T，KORNBLITH S，NOROUZI M，et al. A simple framework for contrastive learning of visual representations[C]// Proceedings of International Conference on Machine Learning. IMLS，2020：1597-1607.

[12] 张重生，陈杰，李岐龙，等. 深度对比学习综述[J]. 自动化学报，2023，49(1)：15-39.

[13] GAO T，YAO X，CHEN D. Simcse：Simple contrastive learning of sentence embeddings[C]// Proceedings of EMNLP. 2021：6894-6910.

[14] 奚琰. 基于对比学习的细粒度遮挡人脸表情识别[J]. 计算机系统应用，2022，31(11)：175-183.

[15] 高怡，纪焘，吴苑斌，等. 基于标签增强和对比学习的鲁棒小样本事件检测[J]. 中文信息学报，2023，37(4)：98-108.