doi:10.11835/j.issn.1000.582X.2024.02.006

基于伪标签和迁移学习的双关语识别方法

姜思羽[1,2a],张智恒[1],姜立标[3a],马　乐[3b],陈博远[2b],王连喜[1],赵　亮[4]
(1. 广东外语外贸大学 信息科学与技术学院,广州 510006; 2. 华南理工大学 a. 软件学院; b. 机械与汽车工程学院,广州 510000; 3. 广州城市理工学院 a. 机械工程学院; b. 工程研究院,广州 510800; 4. 广东轻工职业技术学院 继续教育学院, 广州 510300)

摘要:针对双关语样本短缺问题,研究提出了基于伪标签和迁移学习的双关语识别模型 (pun detection based on Pseudo-label and transfer learning)。该模型利用上下文语义、音素向量和注意力机制生成伪标签;然后,迁移学习和置信度结合挑选可用的伪标签;最后,将伪标签数据和真实数据混合到网络中进行训练,重复伪标签标记和混合训练过程。一定程度上解决了双关语样本量少且获取困难的问题。使用该模型在 SemEval 2017 shared task 7 以及 Pun of the Day 数据集上进行双关语检测实验,结果表明模型性能均优于现有主流双关语识别方法。

关键词:双关语检测;伪标签;迁移学习
中图分类号:TP391.1　　**文献标志码:**A　　**文章编号:**1000-582X(2024)02-051-11

Pun detection basd on pseudo-label and transfer learning

JIANG Siyu[1,2a], *ZHANG Zhiheng*[1], *JIANG Libiao*[3a], *MA Le*[3b], *CHEN Boyuan*[2b], *WANG Lianxi*[1], *ZHAO Liang*[4]
(1. School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006, P. R. China; 2a. School of Software; 2b. School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510000, P. R. China; 3a. School of Mechanical Engineering; 3b. Engineering Research Institute, Guangzhou City University of Technology, Guangzhou 510800, P. R. China; 4. College of Further Education, Guangdong Industry Polytechnic, Guangzhou 510300, P. R. China)

Abstract: To address the problem of shortage of the pun samples, this paper proposes a pun recognition model based on pseudo-label speech-focused context (pun detection based on pseudo-label and transfer learning). Firstly, the model uses contextual semantics, phoneme vector and attention mechanism to generate pseudo-labels. Then, it combines transfer learning and confidence to select useful pseudo-labels. Finally, the pseudo-label data and real data are used for network theory and training, and the pseudo-label labeling and mixed training procedures are repeated. To a certain extent, the problem of small sample size and difficulty in obtaining puns has been solved.

收稿日期:2021-06-25　　**网络出版日期:**2022-01-05
基金项目:广州市科技计划资助项目(202102020637,202002030227);广东外语外贸大学师生合作资助项目(21SS10)。Supported by Guangzhou Science and Technology Plan Project (202102020637, 202002030227) and Teacher-Student Joint Research Project on Guangdong University of Foreign Studies (21SS10).
作者简介:姜思羽(1992—),女,博士, 主要从事迁移学习、机器学习方向研究,(E-mail) jsy2008 @126.com。
通信作者:姜立标(1965—),男,博士, 硕士生导师, 主要从事智能网联与自动驾驶技术方向研究,(E-mail)jlb620620@163.com。

By this model, we carry out pun detection experiments on both the SemEval 2017 shared task 7 dataset and the Pun of the Day dataset. The results show that the performance of this model is better than that of the existing mainstream pun recognition methods.

Keywords: pun detection; pseudo-label; transfer learning

随着社交媒体不断发展,人们在网络上创作了大量幽默内容。幽默的结构往往十分复杂,且依赖真实世界知识。在自然语言中,常见的修辞方法双关语是幽默的一种重要表现形式。双关语是将词语的真正含义模糊化,使同一个句子有2种或者多种释义,使文本产生不同程度的敏感性。双关语是著名文学、广告和演讲中幽默来源的标准修辞手法。例如,它常常作为一种幽默手段被用于广告中,引发听众联想双关语中的潜在表达,既能引人注意又能产生联想,加深记忆[1],有益于判断文本的情感倾向。因此,双关语自动识别被认为是传统语言学和自然语言处理领域认知科学中重要的研究课题,具有广泛应用价值。

双关语的经典分类是谐音双关语和语义双关语[2]。语义双关语,即指同词多义,如表1所示中的“What's the longest sentence in the world? Life sentence.”属于语义双关,“Life sentence”中的“sentence”还有徒刑的意思,故“Life sentence”表示为无期徒刑的意思。谐音双关语,2个不同的词语符合相同语境,即指同音不同词,表1中的“A bicycle can't stand on its own because it is two-tyred”中的“two-tyred”根据读音可被人联想为“too-tired”,使句子具有完全不同意思。理解双关语对于深入理解复杂语义有重要意义。

表1 双关语样例

Table 1 Examples of puns

双关语样例
语义双关语
1. What's the longest sentence in the world? Life sentence.
2. Better late than the late.
谐音双关语
3. Seven days without water makes one weak(week).
4. A bicycle can't stand on its own because it is two-tyred(too tired)

随着深度神经网络的发展,现有双关语识别模型算法大多基于神经网络:例如,刁宇峰等[3]提出了英文字典编码的搭配注意网络模型(word Net-Encoded collocation-attention network, WECA),该模型以基于英文词典“WordNet“来理解和编码嵌入作为输入,结合上下文权重,使用神经注意力网络,捕捉语义双关语中的多义性。但此类基于神经网络方法学习模型存在的缺陷是:1)现有模型依赖大量有标签数据。现实中双关语收集较为困难,一般需要具有丰富相关知识的人进行准确判定和分类,Miller等[4]公布了SemEval 2017 shared task 7(SemEval 2017)数据集中一共包含4 030个双关语样例,反应出对于双关语的收集和标记有一定难度;2)在少样本学习中,如何提升模型的泛化能力是一个富有挑战性的问题。

笔者提出一种基于伪标签和迁移学习的双关语识别模型(pun detection based on pseudo-label and transfer learning,PDPTL)。利用未标签数据重叠信息在同类数据中寻找更为通用的特征,使用迁移学习和置信度结合挑选可用的伪标签,重复伪标签数据与混合训练过程,一定程度缓解双关语数据样本稀缺和模型泛化能力的问题。经过实验,PCPRPL在公开数据集的预测效果获得比较明显提高,且优于目前已知方法。

1 相关工作

双关语任务涉及到双关语识别与生成,研究主要运用伪标签和迁移学习技术为解决双关语任务提供新方法。

1.1　双关语识别与生成

Pedersen 等[5]利用词义消歧技术（word sense disambiguation technique，WSD）[6]识别语句中词语的合理释义，进而达到识别双关语的目的。Dieke等[7]利用外部数据库，例如英文词典“WordNet”，对双关语的词义进行判断。上述2种方法各有缺点，前者不能处理谐音词，因为谐音词具有不同拼写，后者知识库只包含有限词汇。为解决上述2个问题，Mikolov 等[8]和 Pennington 等[9]使用词嵌入技术(word embedding techniques, WET)为双关语提供了灵活表示。在实际情景中一个词语根据它所在文本的上下文可能有多种释义，词语的罕用含义也可能应用于创造双关语，使上述静态词嵌入技术，难以胜任动态变化。为解决上述问题，Zhou等[10]提出语音注意语境双关识别模型（pronunciation-attentive contextualized pun recognition，PCPR）将上下文语义向量和语音嵌入向量2种特征同时应用于双关语识别，取得不错效果。Xiu等[11]基于词汇网络以及词嵌入技术训练了无监督模型。该模型只依赖语义来检测异义双关语，忽略了语音中蕴含的丰富信息。Doogan Samuel等[12]拼接发音字符串利用词嵌入和语音信息，但单采用拼接方法效果有限，利用长短期记忆网络（long - short memory，LSTM）和条件随机场（conditional random fields，CRF）的标签联合检测和定位双关语。

1.2　伪标签

Lee等[13]在2013年实现了简单有效的半监督式学习方法，叫做“伪标签（pseudo-label）”，这个想法是在一批有标签和无标签的图像上，同时训练一个模型。有监督方式使用有标签数据和无标签数据训练模型，预测一批无标签数据生成伪标签，最后使用有标签数据和伪标签数据训练新模型。

Google AI 的 Qizhe Xie等[14]提出一种受知识蒸馏（knowledge distillation）启发的半监督方法“嘈杂学生（noisy student）”。核心思想是训练2种不同的模型，即“老师（teacher）”和“学生（student）”。教师模型首先对标签图像进行训练，对未标记图像进行伪标签推断。然后，将有标记和未标记的图像组合在一起，并根据这些组合的数据训练学生模型。再将学生模型作为新的教师模型进行迭代，研究使用的无标签数据大部分不属于目标数据集的分布。上述伪标签方法大多被应用于图形处理领域。

1.3　迁移学习

迁移学习（transfer learning）旨在通过迁移包含在不同但相关源域中的知识提高目标学习者在目标域上的表现，减少构建目标学习器对大量目标域数据的依赖[15]。根据领域之间差异，迁移学习可分为两类：同构迁移学习和异构迁移学习[16]。1）在同构迁移学习中，一些研究通过校正样本选择偏差[17]或协变量偏移[18]调整域的边缘分布。然而，这个假设在很多情况下并不成立，如在情感分类问题中，一个词在不同领域有不同意义倾向，这种现象也称为上下文特征偏差，为解决这个问题，一些研究进一步适应了条件分布；2）异构迁移学习是指域具有不同特征空间情况下的知识迁移过程。除了分布适应，异构迁移学习还需要特征空间适应[19]，这使得它比同构迁移学习更复杂。笔者主要针对相似特征空间的双关语数据集进行处理，因此属于同构迁移学习方法。

2　基于适应伪标签领域的语音专注语境的双关语识别模型

构建研究模型：基于伪标签和迁移学习的双关语识别模型PDPTL。

2.1　任务概述

遵循Zhou等对于任务的定义，对于一段含有N个词的文本$\{t_1,t_2,...,t_N\}$。每个词t_i具有M_i个音素，根据发音，可表示为$\boldsymbol{H}(t_i)=\{h_{i,1},h_{i,2},...,h_{i,M_i}\}$，$h_{i,j}$表示文本中第$i$个词的第$j$个音素。这些音素是由CMU发音字典（CMU pronouncing dictionary）[16]提供。双关语检测模型的任务是一个二分类问题，目的是检测输入文本是否包含双关语。

2.2　PDPTL模型框架

基础模型：PDPTL选用PCPR作为基础模型。模型使用BERT[20]生成词语的上下文语义向量$\boldsymbol{TC}_i$（$\boldsymbol{D}_C$维的向量），以及文本的总体语义$\boldsymbol{TC}_{[CLS]}$。

对于词语t_i的每个音素h_{ij}使用Keras的Embedding层投影为$\boldsymbol{D}_p$维向量$\boldsymbol{p}_{i,j}$，之后通过局部注意力机制（local-attention mechanism）[21]进行加权生成语音嵌入向量$\boldsymbol{TP}_i$（pronunciation embedding vector）

$$\boldsymbol{e}_{i,j}=\tanh\left(\boldsymbol{F}_{\mathrm{P}}\left(\boldsymbol{p}_{i,j}\right)\right), \tag{1}$$

$$\boldsymbol{\alpha}_{i,j}^{\mathrm{P}}=\frac{\boldsymbol{e}_{i,j}^{\mathrm{T}}\boldsymbol{e}_{s}}{\sum_{k}\boldsymbol{e}_{i,k}^{\mathrm{T}}\boldsymbol{e}_{s}}, \tag{2}$$

$$\boldsymbol{TP}_{i}=\sum_{j}\boldsymbol{\alpha}_{i,j}\boldsymbol{e}_{i,j}, \tag{3}$$

式中：$\boldsymbol{F}_{\mathrm{P}}(\cdot)$是输出$\boldsymbol{D}_{\mathrm{a}}$维向量的全连接层；$\boldsymbol{\alpha}_{i,j}^{\mathrm{P}}$是$\boldsymbol{p}_{i,j}$的重要分数；$\boldsymbol{e}_{\mathrm{s}}$是用来评估每个语音嵌入重要性的$\boldsymbol{D}_{\mathrm{a}}$维向量，$\boldsymbol{D}_{\mathrm{a}}$是模型定义的局部注意力机制的大小。

通过拼接上下文语义向量$\boldsymbol{TC}_{i}$和语音嵌入向量$\boldsymbol{TP}_{i}$（pronunciation embedding vector）生成$\boldsymbol{TJ}_{i}$（$\boldsymbol{D}_{\mathrm{j}}=\boldsymbol{D}_{\mathrm{a}}+\boldsymbol{D}_{\mathrm{P}}$维向量）并运用自注意机制（Self-attention Mechanism）[22]加权得到自注意向量$\boldsymbol{TJ}_{[\mathrm{ATT}]}$（self-attention embedding vector）

$$\boldsymbol{TJ}_{i}=[\boldsymbol{TC}_{i};\boldsymbol{TP}_{i}], \tag{4}$$

$$\boldsymbol{F}_{\mathrm{S}}(\boldsymbol{T})=\mathrm{Soft\,max}\left(\frac{\boldsymbol{TT}\hat{}\boldsymbol{T}}{\sqrt{a}}\right)\boldsymbol{T}, \tag{5}$$

$$\boldsymbol{\alpha}_{i}^{\mathrm{S}}=\frac{\exp\left(\boldsymbol{F}_{\mathrm{S}}\left(\boldsymbol{TJ}_{i}\right)\right)}{\sum_{j}\exp\left(\boldsymbol{F}_{\mathrm{S}}\left(\boldsymbol{TJ}_{j}\right)\right)}, \tag{6}$$

$$\boldsymbol{TJ}_{[\mathrm{ATT}]}=\sum_{i}\boldsymbol{\alpha}_{i}^{\mathrm{S}}\cdot\boldsymbol{TJ}_{i}, \tag{7}$$

式中：$\boldsymbol{F}_{\mathrm{S}}(\boldsymbol{T})$是用来估算注意力的函数；$\boldsymbol{\alpha}_{i}^{\mathrm{S}}$是每个单词$t_{i}$的重要分数；$a$是一个缩放系数，为了避免过小的梯度。最后拼接$\boldsymbol{TJ}_{[\mathrm{ATT}]}$与$\boldsymbol{TC}_{[\mathrm{CLS}]}$生成输入文本的整体特征即语音联合上下文语义向量$\boldsymbol{TJ}_{[\mathrm{CLS}]}$

$$\boldsymbol{TJ}_{[\mathrm{CLS}]}=[\boldsymbol{TC}_{[\mathrm{CLS}]};\boldsymbol{TJ}_{[\mathrm{ATT}]}], \tag{8}$$

预测标签由采用softmax激活函数的全连接层给出

$$\hat{\boldsymbol{y}}_{i}^{L}=\arg\max\boldsymbol{F}_{\mathrm{D}}\left(\boldsymbol{TJ}_{[\mathrm{CLS}]}\right)_{k},k\in\{0,1\}, \tag{9}$$

式中，$\boldsymbol{F}_{\mathrm{D}}(\cdot)_{k}$生成二元分类中两类的值。

伪标签：先前的伪标签学习方法筛选伪标签的策略通常为选取高置信度的样本。策略的依据是聚类假设，即高置信度样本在相同类别的可能性较大。具体步骤为设定confidence_coefficient这一置信度阈值，只有生成的伪标签概率大于confidence_coefficient时，模型才会将其加入训练数据中。

概率由以下公式得出

$$\mathrm{confidence}=\mathrm{MAX}\left(\mathrm{Soft\,max}\left(\boldsymbol{F}_{\mathrm{D}}\left(\boldsymbol{TJ}_{[\mathrm{CLS}]}\right)\right)\right), \tag{10}$$

但这样的策略，一方面阈值的确定过于依赖人工实验，另一方面忽视了潜藏的危险“高置信度的陷阱”——模型所认为的高置信度样本并不一定可靠，最终导致高置信度的错误样本加入到了模型训练过程中。为了筛选出更加可靠的样本，模型在高置信度策略基础上结合迁移学习方法中的MMD（maximum mean discrepancy）[23]距离来评估伪标签样本的可靠性。

MMD是由Gretton等人提出，用于度量2个数据集分布的匹配程度，常用于检测双关样本问题。度量值代表2个数据集分布在再生希尔伯特空间（reproducing kernel Hilbert space，RKHS）中的距离，度量值越小，则距离越近，代表2个分布越相似，MMD的计算公式如下

$$\mathrm{MMD}\left(\boldsymbol{TD},\boldsymbol{PD}\right)=\left\|\frac{1}{n^{2}}\sum_{i}^{n}\sum_{i}^{n}k\left(\boldsymbol{TD}_{i},\boldsymbol{TD}_{i}\right)-\frac{2}{nm}\sum_{i}^{n}\sum_{j}^{m}k\left(\boldsymbol{TD}_{i},\boldsymbol{PD}_{j}\right)-\frac{1}{m^{2}}\sum_{j}^{m}\sum_{j}^{m}k\left(\boldsymbol{PD}_{j},\boldsymbol{PD}_{j}\right)\right\|_{\mathrm{H}}。 \tag{11}$$

本模型的伪标签样本筛选策略，给定置信度阈值confidence_coefficient一个初始值，置信度阈值以一定步幅（speed）增长，计算在当前置信度阈值下筛选得出的伪标签数据（Pseudo_label_data）与训练数据（labeled_data）的MMD距离，将其中MMD距离最小的阈值作为最终置信度阈值，由此筛选出最终伪标签数据（Pseudo_label_data），标记伪标签，加入训练。为了保证模型能尽可能学到正确知识及从有标签数据中学习到足够知识，笔者采用了加权损失函数，即在T_{start}批次前对带有伪标签的数据权重设置为零后慢慢增加，直到T_{End}批次保持不变为常数weight。

$$\text{weight}(t)=\begin{cases}0, & t<T_{\text{start}},\\ \dfrac{t-T_{\text{start}}}{T_{\text{End}}-T_{\text{start}}}\times\text{weight}, & T_{\text{start}}\leqslant t<T_{\text{End}},\\ \text{weight}, & T_{\text{End}}\leqslant t_{\circ}\end{cases}\tag{12}$$

损失函数为交叉熵损失函数,真实训练数据(labeled_data)和伪标签数据(Pseudo_label_data)将会分开计算损失值,最后如下加权合并得出最终损失 Loss

$$\text{Loss}=\text{loss}(\text{labeled}_{\text{data}})+\text{weight}(t)*\text{loss}(\text{Pseudo\_label\_data})_{\circ}\tag{13}$$

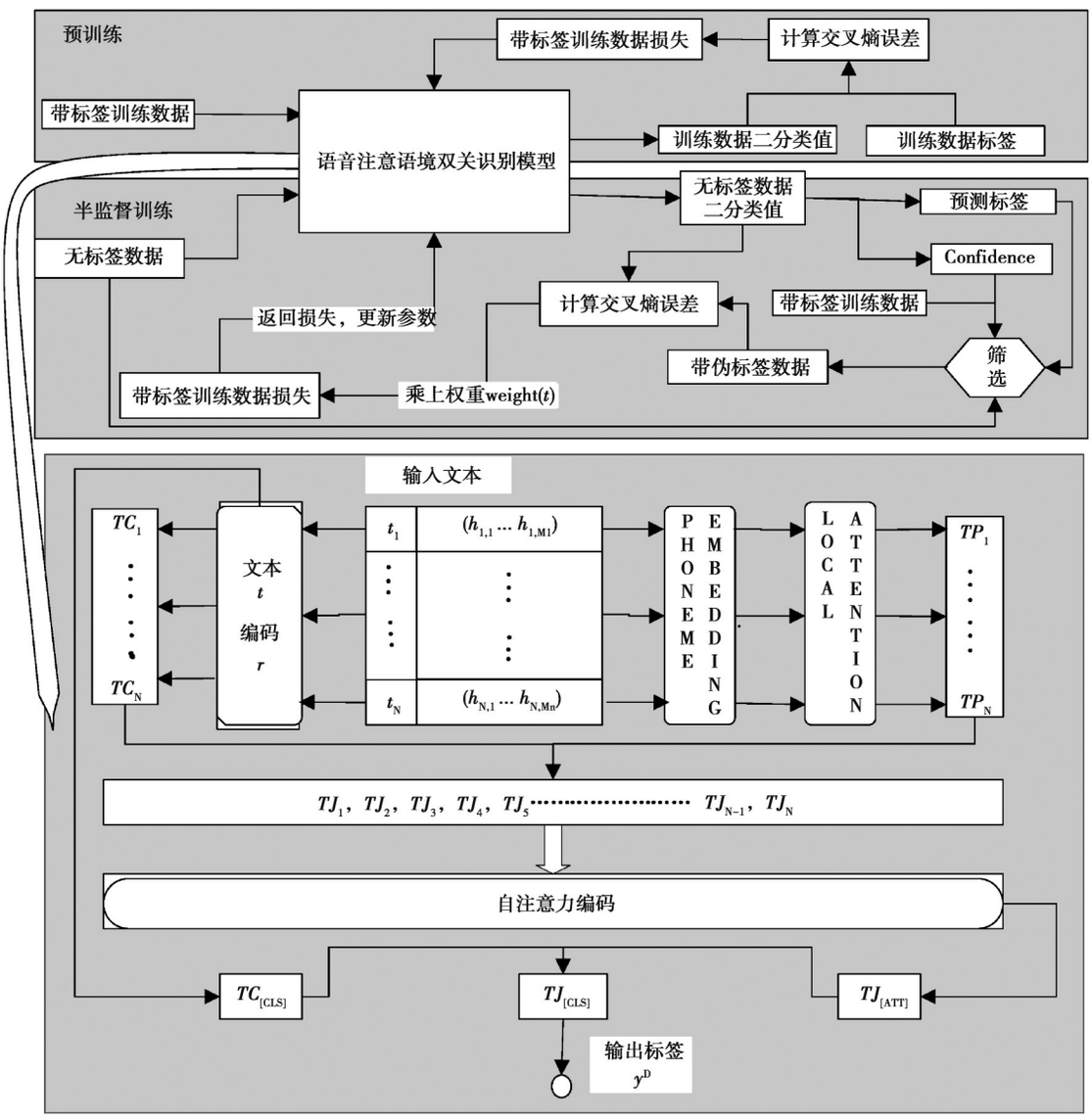


图 1　PDPTL 框架图

Fig. 1　The frame work of PDPTL

PDPTL:图 1 体现了 PDPTL 的整体框架。概括而言,模型分为 3 步:

1) 通过有标签数据训练基础模型,得到已训练模型;

2) 已训练模型对无标签数据进行预测获得带有伪标签的数据;

3) 将有标签数据和筛选后的伪标签数据混合取代有标签数据重新训练基础模型,进入下一轮。

根据以上阐述,算法 1 展示了 PDPTL 的总体流程。

算法 1

```
/*
times 循环更新 pseudo_labels 的次数
Base_Model 基础模型
num_train_epochs 模型训练批次
eval_data 无标签数据。
eval()评估函数输入模型和无标签数据输出伪标签数据
confidence_coefficient 初始阈值
Best_MMD 最小的 MMD 距离
Best_confidence_coefficient 最佳阈值
speed 阈值增加步幅
*/
for index<-0 to times:/*times 循环更新 pseudo_labels 的次数*/
{
    init Base_Model/*Base_Model 基础模型*/
    for epoch<-0 to num_train_epochs:/*num_train_epochs 模型训练批次*/
    {
        train Base_Model with train_data_with_label /*使用训练数据训练 Base_Model*/
    }
    data_with_pseudo_labels <- eval(Base_Model,eval_data)
    /*eval_data 无标签数据。eval()评估函数输入模型和无标签数据输出伪标签数据*/
    init train_data_with_label
    /*初始化训练数据,即去除上一轮加入的伪标签数据*/
  Now_confidence_coefficient = confidence_coefficient
  While Now_confidence_coefficient <= 1:
  {
    for data_with_pseudo_label in data_with_pseudo_labels:/*遍历每一条伪标签数据*/
    {
        if probability of data_with_pseudo_label larger than Now_confidence_coefficient:
/*判断的概率大于置信度 confidence_coefficient*/
            add data_with_pseudo_label to pseudo _data_with_label/*将伪标签数据加入伪标签数据集中*/
    }
    MDD = getMDD(train_data_with_label,pseudo _data_with_label)/*获取当前伪标签数据集与训练数据集
的 MDD*/
    if Now_ confidence_coefficient == confidence_coefficient:
        Best_MDD = MDD
    else:
        if Best_MDD < MDD:/*距离变小则更新*/
            {
                Best_MDD = MDD
                Best_confidence_coefficient = Now_confidence_coefficient/*更新最佳阈值和最佳伪标签数
据集*/
                best_pseudo_data_with_label =                pseudo_data_with_label
            }
    init pseudo _data_with_label/*初始化当前伪标签数据集,即清空
      Now_confidence_coefficient = Now_confidence_coefficient + speed/*按照 speed 递增*/
  }
  Add   best_pseudo_data_with_label to   train_data_with_label
}
```

3　实验

展示实验相关设置,将PDPTL模型与其他经典算法在2个公开数据集上进行性能比较。

3.1　实验设置

实验数据集:模型在 SemEval 2017 shared task 7 数据集(SemEval 2017)[4]以及 the Pun of The Day 数据集(PTD)[24]进行实验。SemEval 2017 task 7 数据集由 4 030 个双关语样例组成,且每个样例都被细分为语义双关语或者谐音双关语,表2详细统计了数据集。SemEval 2017 数据集包含了双关语和非双关语笑话、格言以及由专业幽默作家创作,或从网络上收集的短文。这个数据集是目前此研究领域中使用的最大公开数据集。

PTD 数据集包含 4 826 个样例。表3显示了 PDT 的统计信息。PTD 数据集则包含从双关语网站上筛选收集的双关语笑话和从美联社新闻、《纽约时报》、雅虎问答以及英文谚语中筛选摘取的非幽默文本。虽然 PTD 数据集原意是为识别幽默文本创建,但由于其上述特殊的内在构成,本模型也将在该数据集上进行实验。

表 2　SemEval数据集数据统计

Table 2　SemEval dataset statistics

数据集	SemEval		PTD
	语义	谐音	
包含双关语的样例	1 607	1 271	2 423
不包含双关语的样例	643	509	2 403

注:语义代表语义双关语,谐音代表谐音双关语。

表 3　PDT数据集统计

Table 3　PDT data set statistics

数据集	PTD
幽默文本	2 423
不幽默文本	2 403

评价标准:选择使用准确率(P),召回率(R)以及 $F1$ 值来比较PDPTL和基础模型以及其他基准模型的性能。其中 TP 代表被模型正确分类的包含双关语的样例数量,MP 代表了模型判断为包含双关语的样例的数量,TP 为真实包含双关语的样例数量。

$$P=\frac{TP}{MP}\ , \tag{14}$$

$$R=\frac{TP}{TR}, \tag{15}$$

$$F1=\frac{2RP}{R+P}。 \tag{16}$$

基准模型:在 SemEval 2017 数据集上,PDPTL 会与 Duluth , CRF [24], Joint [24], JU_CSE_ NLP[25], PunFields[26], Fermi[27]以及 CPR[10]7个基准模型比较。JU_CSE_ NLP 基于规则分类双关语。PunFields 使用同义词典识别双关语。Fermi 在监督学习的基础上使用 RNN 分类。CPR 即是 PCPR 模型去除语音特征,只使用语义特征。在 PDT 数据集上,模型会和 HAE[28], MCL[29], PAL [30]、HUR[31]、WECA[2]以及 CPR 5个基准模型进行比较。HAE[23]应用了基于 Word2Vec 和以人为中心的随机森林方法。MCL[24]利用带有多种文体特征的单词表示。PAL[29]运用 CNN 方法去自动学习基本特征。HUR[31] 在已有 CNN 模型基础上调整了过滤器的大小和

添加highway层。

实验细节设置：模型的超参数 weight=0.84，T_{start}=2，T_{End}=4，times=5，num_train_epochs=7，confidence_coefficient=0.999 7，speed=0.000 1。但在PDT数据集上，times=3，num_train_epochs=5。模型的实验环境：pytorch-pretrained-bert==0.6.1, seqeval==0.0.5，torch==1.0.1.post2，tqdm==4.31.1，nltk==3.4.5，GPU型号为Tesla V100-SXM2，实验在Goolgle的Colab平台运行。

3.2 实验结果

表4将PDPTL模型与其他经典模型在检测SemEval数据集上的语义双关语和谐音双关语性能方面进行比较。在SemEval 2017数据集上，PDPTL对比3个基准模型表现最优。在语义双关语上对比最优基准模型分别在准确率（P）、召回率（R）和$F1$值（$F1$）提高5.01%、2.93%、4.04%，在谐音双关语上对比最优的基准模型分别在准确率（P）、召回率（R）和$F1$值（$F1$）提高9.12%、3.77%、6.55%。

表4 模型与基准模型在SemEval 2017数据集中的双关语检测性能

Table 4 The pun detection performance of the model and the benchmark model in the SemEval 2017 data set %

模型	SemEval 2017 语义双关语			SemEval 2017 谐音双关语		
	准确率	召回率	$F1$	准确率	召回率	$F1$
JU_CSE_ NLP	72.51	90.79	68.84	73.67	94.02	71.74
PunFields	79.93	73.37	67.82	75.80	59.40	57.47
Fermi	90.24	89.70	85.33	_	_	_
Duluth	78.32	87.24	82.54	73.99	86.62	68.71
CRF	72.51	90.79	68.84	73.67	94.02	71.74
Joint	91.25	93.28	92.19	86.67	93.08	89.76
CPR	94.18	94.21	92.79	93.35	95.04	94.19
PDPTL	96.26	96.21	96.23	95.79	96.85	96.31

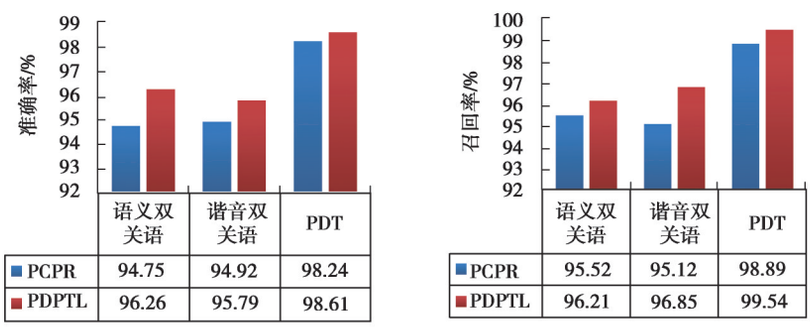
表5则是在PDT数据集上比较了模型的性能。在PDT数据集上，PDPTL对比最优的基准模型分别在准确率（P）、召回率（R）和$F1$值（$F1$）提高12.01%、5.54%、8.98%。

表5 模型与基准模型在PDT数据集中的双关语检测性能

Table 5 The pun detection performance of the model and the benchmark model in the PDT data set %

模型	Pun of the Day		
	准确率	召回率	$F1$
MCL	83.80	65.50	73.50
HAE	83.40	88.80	85.90
PAL	86.60	85.40	85.70
HUR	86.60	94.00	90.10
WECA	89.19	90.64	89.21
CPR	98.12	99.34	98.73
PDPTL	98.61	99.54	99.08

图 2 与图 3 为 PDPTL 与基础模型在 2 个数据集上的比较。在 SemEval 数据集上，对于语义双关语，PDPTL 模型相较于基础模型分别在准确率（*P*）、召回率（*R*）和 *F*1 值（*F*1）提高 1.51%、0.69%、1.10%。对于谐音双关语，PDPTL 模型对比基础模型分别在准确率（*P*）、召回率（*R*）和 *F*1 值（*F*1）提高 0.87%、1.73%、1.30%。在 PDT 数据集上，PDPTL 模型对比基础模型分别在准确率（*P*）、召回率（*R*）和 *F*1 值（*F*1）提高 0.37%、0.65%、0.52%。值得注意，PCPR 方法和 CPR 方法在 PDT 数据集上相比较结果相差无几。CPR 方法即是 PCPR 去除语音向量，仅依靠 BERT 生成的上下文语义向量及注意力机制。明显看出 PDPTL 方法在 PDT 数据集上提升效果不如在 SemEval 2017 数据集，PDT 数据集的样本数量是 SemEval 数据集单一子集数量的 2 倍，结果符合假设。



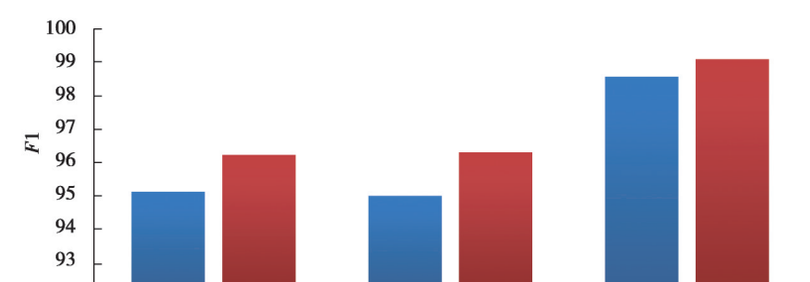
	语义双关语	谐音双关语	PDT
PCPR	94.75	94.92	98.24
PDPTL	96.26	95.79	98.61

	语义双关语	谐音双关语	PDT
PCPR	95.52	95.12	98.89
PDPTL	96.21	96.85	99.54

（a）准确率　　（b）召回率

图 2　模型与基础模型在各个数据集的准确率与召回率

Fig. 2　The accuracy and recall rate of the model and the basic model in each data set



	语义双关语	谐音双关语	PDT
PCPR	95.13	95.01	98.56
PDPTL	96.23	96.31	99.08

PCPR　PDPTL

图 3　模型与基础模型在各个数据集的 *F*1 值

Fig. 3　The F1 value of the model and the basic model in each dataset

4　结束语

针对现有的双关语数据集样本较少问题，提出利用伪标签技术辅助模型进行训练；考虑到伪标签数据和真实数据之间的特征分布差异，迁移学习技术和置信度相结合，提出一种新型双关语识别模型。使用该模型在 SemEval 2017 shared task 7 以及 Pun of the Day 数据集上进行双关语检测实验，表明了 PDPTL 模型可拉近伪标签和真实标签数据的特征分布，预测性能均优于现有的主流双关语识别方法。

参考文献

[1] 徐琳宏, 林鸿飞, 祁瑞华, 等. 基于多特征融合的谐音广告语生成模型[J]. 中文信息学报, 2018, 32(10): 109-117.
Xu L H, Lin H F, Qi R H, et al. Homophonic advertisement generation based on features fusion[J]. Journal of Chinese Information Processing, 2018, 32(10): 109-117.(in Chinese)

[2] Redfern W D. Guano of the mind: puns in advertising[J]. Language & Communication, 1982, 2(3): 269-276.

[3] Diao Y F, Lin H F, Wu D, et al. WECA: a WordNet-encoded collocation-attention network for homographic pun recognition [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 2507–2516.

[4] Miller T, Hempelmann C F, Gurevych I. Semeval-2017 task 7: detection and interpretation of english puns[C]//Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Vancouver, Canada: Association for Computational Linguistics, 2017: 58-68.

[5] Pedersen T. Puns upon a midnight dreary, lexical semantics for the weak and weary[C]//Proceedings of the 11th International Workshop on Semantic Evaluation. Vancouver, Canada: Association for Computational Linguistics, 2017: 416-420.

[6] Ranjan Pal A, Saha D. Word sense disambiguation: a survey[J]. International Journal of Control Theory and Computer Modeling, 2015, 5(3): 1-16.

[7] Dieke O, Kilian E. Global vs. local context for interpreting and locating homographic english puns with sense embeddings[C]// Proceedings of the 11th International Workshop on Semantic Evaluation. Vancouver, Canada: Association for Computational Linguistics, 2017: 444-448.

[8] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[EB/OL]. [2021-06-10].. https://arxiv.org/abs/1310.4546.pdf.

[9] 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). Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1532–1543.

[10] Zhou Y C, et al. The boating store had its best sail ever: pronunciation-attentive contextualized pun recognition[EB/OL].[2021-06-10]. https://arxiv.org/pdf/2004.14457.pdf.

[11] Xiu Y L, et al. Using supervised and unsupervised methods to detect and locate english puns[C]//Proceedings of the 11th International Workshop on Semantic Evaluation. Vancouver, Canada：Association for Computational Linguistics, 2017: 453-456.

[12] Samuel D, Aniruddha G, Hanyang C, et al. Detection and interpretation of english puns[C]//Proceedings of the 11th International Workshop on Semantic Evaluation. Vancouver, Canada：Association for Computational Linguistics, 2017: 103-108.

[13] Lee D. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks[C]//In Workshop on Challenges in Representation Learning, Atlanta, Georgia: International Conference on Machine Learning, 2013.

[14] Xie Q Z, Luong M T, Hovy E, et al. Self-training with noisy student improves ImageNet classification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020. Seattle, WA, USA：IEEE, 2020: 10684-10695.

[15] Zhuang F Z, Qi Z Y, Duan K Y, et al. A comprehensive survey on transfer learning[J]. Proceedings of the IEEE, 2021, 109(1): 43-76.

[16] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[17] Huang J Y, Smola A J, Gretton A, et al. Correcting sample selection bias by unlabeled data[M]//Advances in Neural Information Processing Systems 19. US: MIT Press, 2007: 601-608.

[18] Sugiyama M, Suzuki T, Nakajima S, et al. Direct importance estimation for covariate shift adaptation[J]. Annals of the Institute of Statistical Mathematics, 2008, 60(4): 699-746.

［19］Day O, Khoshgoftaar T M. A survey on heterogeneous transfer learning[J]. Journal of Big Data, 2017, 4(1): 1-42.

［20］Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2021-06-10].. https://arxiv.org/abs/1810.04805.pdf.

［21］Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2021-06-10]. https://arxiv.org/abs/1409.0473.pdf.

［22］Ashish V, Noam S, Niki P, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017: 5998-6008.

［23］Gretton A, Borgwardt K, Rasch M, et al. A kernel two-sample test[J]. Journal of Machine Learning Research, 2012(13): 723-773.

［24］Yanyan Z, Wei L. Joint detection and location of english puns[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minnesota: Association for Computational Linguistics, 2019: 2117-2123.

［25］Aniket P, Dipankar D. Employing rules to detect and interpret english puns[C]// Proceedings of the 11th International Workshop on Semantic Evaluation. Vancouver, Canada: Association for Computational Linguistics, 2017: 432-435.

［26］Mikhalkova E, Karyakin Y. Pun fields at SemEval-2017 task 7: employing roget's thesaurus in automatic pun recognition and interpretation[C]//Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Vancouver, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017.

［27］Indurthi V, Oota S R. Fermi at SemEval-2017 task 7: detection and interpretation of homographic puns in English language[C]// Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Vancouver, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 457-460.

［28］Yang D Y, Lavie A, Dyer C, et al. Humor recognition and humor anchor extraction[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015: 2367-2376.

［29］Mihalcea R, Strapparava C. Making computers laugh: investigations in automatic humor recognition[C]//Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05. October 6-8, 2005. Vancouver, ColumbiaBritish, Canada. Morristown, NJ, USA: Association for Computational Linguistics, 2005.

［30］Chen L, Lee C M. Predicting audience's laughter using convolutional neural network[EB/OL]. [2021-06-10].https://arxiv.org/abs/1702.02584.pdf.

［31］Chen P Y, Soo V W. Humor recognition using deep learning[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 113-117.

（编辑　侯　湘）