文章编号：1003-0077(2021)06-0131-10

CDCPP：跨领域中文标点符号预测

刘鹏远[1]，王伟康[1]，邱立坤[2]，杜冰洁[1]

（1. 北京语言大学 国家语言资源监测与研究平面媒体中心，北京 100083；
2. 闽江学院 计算机控制与工程学院，福建 福州 350108）

摘 要：在中文文本特别是在社交媒体及问答领域文本中，存在非常多的标点符号错误或缺失的情况，这严重影响对文本进行语义分析及机器翻译等各项自然语言处理的效果。当前对标点符号进行预测的相关研究多集中于英文对话的语音转写文本，缺少对社交媒体及问答领域文本进行标点符号预测的相关研究，也没有这些领域公开的数据集。该文首次提出跨领域中文标点符号预测任务，该任务首先利用标点符号基本规范正确的大规模新闻领域文本，建立标点符号预测模型；然后在标点符号标注不规范的社交媒体及问答领域，进行跨领域标点符号预测。随后，构建了新闻、社交媒体及问答三个领域的相应数据集。最后还实现了一个基于 BERT 的标点符号预测基线模型并在该数据集上进行了实验与分析。实验结果表明，直接利用新闻领域训练的模型，在社交媒体及问答领域进行标点符号预测的性能均有所下降，在问答领域下降较小，在微博领域下降较大，超过 20%，说明跨领域标点符号预测任务具有一定的挑战性。

关键词：中文标点符号预测；跨领域；数据集

中图分类号：TP391 **文献标识码**：A

CDCPP: Cross-Domain Chinese Punctuation Prediction

LIU Pengyuan[1], WANG Weikang[1], QIU Likun[2], DU Bingjie[1]

(1. Language Resources Monitoring and Research Center Print Media Language Branch, Beijing Language and Culture University, Beijing 100083, China;
2. School of Computer and Control Engineering, Minjiang University, Fuzhou, Fujian 350108, China)

Abstract: Punctuation errors or omissions in s in Chinese texts seriously affects various natural language processing such as semantic analysis and machine translation. Existing researches on punctuation prediction are mostly focused on the speech transcribed text of English conversations, rather than texts in social media and question answering domain. This paper proposes a cross domain Chinese punctuation prediction task, i.e. punctuation prediction for the fields of social media and question answering via large-scale news texts with correct punctuation marks. Corresponding data sets in the fields of news, social media and question answering are then constructed. A BERT-based punctuation prediction baseline model is implemented. The experimental results show that the performance of punctuation prediction in social media and question answering domains decreases by directly using the model trained in the news domain. The decline in question answering domain is much less than that in Weibo domain(more than 20%). The task of cross domain punctuation prediction is challenging.

Keywords: Chinese punctuation prediction; cross-domain; dataset

收稿日期：2021-02-09 **定稿日期**：2021-03-09

基金项目：北京市自然科学基金(4192057)；教育部人文社会科学研究规划基金(18YJA740030)；北京语言大学校级项目(中央高校基本科研业务费专项资金)(17PT05)

0 引言

汉语书面语中，标点符号有着不可或缺的地位。《辞海》[①]中把标点解释为“书面语里用来表示停顿、语调以及语词的性质和作用的符号，是书面语的有机组成部分”。它可以帮助人们确切地表达思想感情和理解书面语言。

近年来，随着社交媒体（如微博），问答（如百度知道）及应用（如问答机器人）的活跃兴起，对社交媒体及问答领域文本的处理显得越来越重要。但这两类文本常常出现标点符号使用错误、缺失甚至完全不使用标点符号的情况。图 1 是标点符号标注错误及正确实例对语义分析[②]及机器翻译[③]影响的对比。

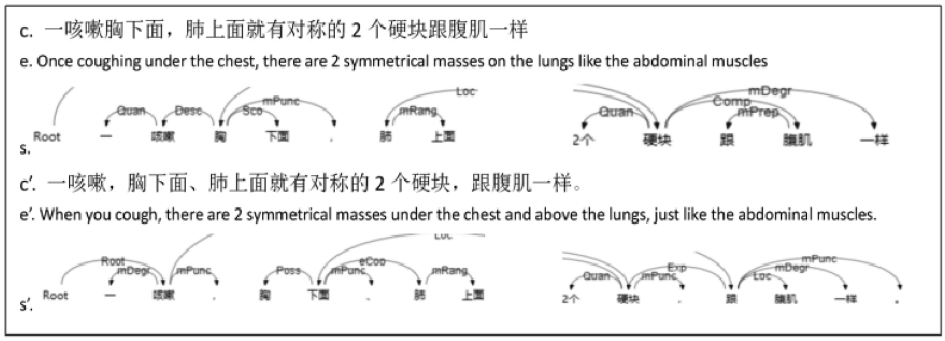

图 1 标点符号标注错误对依存语义分析及机器翻译结果的影响

其中，c 及 c′分别为百度知道[④]中的实例及人工对其进行标点重新标注后的文本；e 及 e′是分别对 c 及 c′用谷歌翻译的结果；s 及 s′是对两个中文文本分别利用 LTP 平台进行语义依存标注的结果（限于篇幅，仅截取了部分）。对比英文译文，可见标点符号错误不但引起局部翻译错误，也影响整句的翻译质量。对比语义依存自动分析的结果，出现标点错误的地方，均导致自动语义分析标注的结果产生错误。本文随机抽取了新浪微博文本 100 条，并对其中的标点符号进行了人工排查，发现其中有 82 条标点符号缺失或使用错误。这将对 NLP 任务的处理如句法分析、语义分析及机器翻译等各项自然语言处理任务的效果带来很大影响。为社交媒体及问答等领域中的文本标注正确的标点符号，具有重要意义和应用价值。

标点符号预测（punctuation prediction，PP）或标点符号恢复（punctuation pestoration，PR）指利用计算机对无标点文本进行标点预测，使得预测之后的文本符合自身语义和标点使用规范。因为语音识别出的序列中没有标点符号，故而现有标点符号预测的相关研究工作集中在语音识别领域，主要是面向对话领域对语音转写文本进行标点符号预测[1-5]。目前常用公开的数据集 IWSLT[6] 是针对语音领域的英文文本。迄今为止，尚无公开的社交媒体领域的相关数据集，这对在该领域进行标点符号预测研究工作产生了很大限制。由于社交媒体及问答领域文本中标点符号缺失或使用错误较多，直接利用社交媒体及问答领域文本进行模型训练再进行标点符号预测意义不大，而人工标注一个大规模社交媒体及问答领域标点符号预测数据集又非常费时费力。与此同时，新闻领域中的文本标点符号用法基本规范，可认为是标点符号使用正确的实例，建立 PP/PR 任务的数据集较为容易，但面向新闻领域文本进行标点符号预测，应用价值较低。

基于以上现状，本文提出跨领域中文标点符号预测任务（CDCPP：cross-domain Chinese punctuation prediction），该任务利用标点符号基本规范的大规模新闻领域文本，建立标点符号预测模型，然后在标点符号标注不规范的社交媒体及问答领域进行跨领域标点符号预测。本文构建了新闻、社交媒体及问答三个领域的数据集[⑤]。在新闻领域，提供了测试集，共 10 000 条。新闻领域的文本较容易获得且标点符号使用非常规范，因此在大规模新闻文本上进行训练并进行标点符号预测，可视为各种方法在其他领域预测性能的上限（upper bound）。在社交媒体和问答领域，本文分别提供了人工标注的测

① http://chlb.cishu.com.cn/

② 采用哈工大语言技术 LTP 平台：http://ltp.ai/index.html

③ 采用谷歌翻译：https://translate.google.cn

④ http://zhidao.baidu.com

⑤ https://github.com/NLPBLCU/Cross-Domain-Chinese-Punctuation-Prediction

试集各 1 200 条。除测试集外，本任务没有提供(但不禁止使用)社交媒体和问答领域这两个目标领域内的任何数据。

鉴于近年来预训练语言模型 BRET(bidirectional encoder representation from transformers)[7] 在 NLP 领域各项任务中的优良表现，我们实现了一个基于 BERT 的序列标注模型作为数据集的基线模型，对本任务及数据集进行初步评估，同时还使用了 Focal Loss[8] 作为训练过程中的损失函数来缓解类别不平衡问题。在本数据集上的实验结果表明，直接利用新闻领域训练的模型，在社交媒体及问答领域进行标点符号预测的性能均有所下降，特别是在微博领域下降较大，说明跨领域标点符号预测任务具有一定的挑战性。

1　数据集构建

1.1　数据准备

新闻领域选择人民日报 2018 年全年语料。该语料共 574 332 条文本。我们首先按照本文标点符号标签表对该文本进行处理：去除无关的一些噪声标点和符号，对标点进行全半角和重复符号归一化处理，再将冒号替换为逗号，感叹号和省略号替换为句号。然后在该语料中随机抽取 10 000 条作为测试集，另随机抽取 100 000 条作为训练集。

社交媒体领域选择新浪微博①，新浪微博是目前中文影响力最大的社交媒体，基于新浪微博进行自然语言处理的研究非常广泛[9-12]。本文随机爬取微博语料共 120 250 条，语料经数据预处理之后统计得平均文本长度(含标点)为 66.41、标准差为 55.64，分布不太均衡。为考察不同文本长度对模型的影响，我们随机选取文本长度在 65～67 之间(中等长度文本)及文本长度在 100～110 之间(长文本)的两类，去掉其中所有标点及空格，作为社交媒体领域测试集待标注语料，并分别记为：微博(中)及微博(长)。

问答领域语料来源于中文问答匹配数据集 LCQMC[13]，该数据集被广泛用于中文问答特别是问句匹配(question matching)中。该数据集中的语料来源于百度知道，共 260 068 对句子，经过去重后，统计得平均文本长度为 10.73、标准差为 4.0，平均文本长度较微博更短，文本长度分布也相对均衡，因此考察不同文本长度对模型的影响意义较小。通过对语料的观察我们发现，在文本长度小于 15 的文本中，标点符号相对较少，因此我们随机抽取文本长度在 15 以上的文本，去掉原语料中所有标点及空格，作为问答领域测试集待标注语料。

1.2　标注规范

标注规范采用 2011 年由中国国家标准化管理委员会发布，2012 年实施的《标点符号用法》(以下简称《用法》)文件②中的标准。《用法》将标点符号分为点号和标号。其中，点号的作用是点断，主要表示停顿和语气。而标号的作用是标示某些成分的特定性质和作用。本次主要标注点号，即句号、问号、叹号、逗号、顿号、分号、冒号。在这些点号中，叹号主要用在句末表达情感，句子的情感时常因人而异，有时可用句号替代。冒号用于句中，表示语段中提示下文或总结上文的停顿，其部分功能与逗号类似。由本文任务的目的出发，我们选择最终标注的标点符号只有五种：逗号、顿号、分号、句号、问号。

1.3　标注过程

整个标注由 3 名语言学在读硕士研究生作为标注员共同完成。首先是研读规范并进行试标注，对不一致的地方进行讨论。待 3 名标注员均熟悉规范和标注后再进行正式标注。

对于给定的问答领域及微博领域的测试集待标注语料，由两名标注员分别进行标注。在标注过程中，如果遇到难以理解的句子，由于很难对其进行正确的标点符号标注，标注员就直接抛弃该条文本。对于语料中出现的其他非标点符号，如表情符号等，均手动删除。对于新浪微博语料，标注每种文本长度的语料各 600 条，共 1 200 条。对于百度知道语料，标注共 1 200 条。标注完成后，由第三名标注员进行审核。审核的标准是排除标点符号使用不符合《用法》的原则性错误。审核后的文本，如果两名标注员标注结果一致，则作为金标准文本保留，对于标注不一致的文本，由第三名标注员进行仲裁。

由于《用法》并没有对标点符号进行严格的标注规定，特别是句中的成分之间是否停顿因使用场景、使用习惯等存在差异，例如：

(1) 突击停产后，企业为了抢回时间满足订单

① http://weibo.com

② http://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=22EA6D162E4110E752259661E1A0D0A8

生产,往往会匆忙复工。这一停一开,可能危机四伏。

(2) 突击停产后,企业为了抢回时间满足订单生产往往会匆忙复工,这一停一开可能危机四伏。

两段文本的标注都不存在原则性错误,只是个人语感不同,语块切分的大小不同,这样对同一文本进行标点符号标注,可能会出现多种正确的标注。因此,在仲裁时,首先由第三名标注员保留仲裁得到的标注结果,作为金标准;然后,3名标注员对标注的不同结果进行讨论,讨论后确定符合《用法》的标注保留下来,作为可选标准文本,附加在金标准文本后,以空格分开。

1.4 统计分析

最终形成的数据集共包含问答领域、微博中等长度及微博长文本数据各1 200、600及600条。由于每条数据可能包含多个正确的标注结果,所以最终得到问答领域、微博(中)及微博(长)数据各1 328、803及779句。问答领域、微博(中)及微博(长)数据集的标注一致率分别为0.975 1,0.957 2及0.973 1[①]

数据集的基本统计情况及标点符号的分布情况分别见表1及表2。此处仅对金标准文本进行统计,而没有将可选标准文本包含在内。实际上,由于备选的正确标注文本不多,因此在平均长度、平均标点个数,平均文本长度与标点个数之比及标点符号分布几个方面,所得到的结果差异不大。

表1 数据集的基本统计情况

数据集	条目数	平均长度(字)/条	平均标点个数/条	句长/标点数
新闻	10 000	70.16	5.61	**12.51**
问答	1 200	26.28	2.64	9.95
微博(中)	600	58.14	6.18	9.41
微博(长)	600	**94.14**	**9.04**	10.41

从表1可知,微博(长)的平均长度最长,包含标点个数最多,问答领域的平均文本长度最短,包含标点个数最少。从平均文本长度与平均标点个数之比可知,微博(中)的标点符号"密度"最大,平均9.41个字就有一个标点,而新闻领域标点符号"密度"最小,平均12.51个字才有一个标点。

从表2可知,各领域中逗号都是最常用标点符号,分号都是最不常用的标点符号。根据《用法》的规定,顿号常用于重复的词语或成分之间,而分号则多用于并列的分句之间,因此在层级关系上分号的层级要大于顿号,因此顿号的使用比分号多。在问答领域,由于文本普遍较短,没有出现分号。问号与句号在各领域中使用的情况比较复杂,新闻领域中问号比例较低,符合新闻文体特点,这符合我们的认知。在问答领域,提问较多,因此问号的比例总体高于微博领域。句号在新闻、问答及微博(中)的分布相对接近,且在新闻中的使用相对更多。

表2 数据集中各种标点分布情况

标点	新闻	问答	微博(中)	微博(长)
逗号/%	**55.13**	49.13	52.62	55.01
问号/%	0.61	**23.48**	8.22	23.73
句号/%	**26.06**	25.21	25.99	15.37
顿号/%	**17.00**	2.16	12.83	4.96
分号/%	**1.19**	0.00	0.32	0.91

针对微博长句中问号使用比句号更频繁的现象,我们进一步分析文本的内容,发现微博长句中存在不少连续发问的现象,例如:

"渣完基三的直接后果就是,为毛我没有轻功?为毛我没有内力?为毛我只是想去个我想个地方要做交通工具神行不了?为毛下个楼还要等电梯或走楼梯?不能轻功跳过去?桑不起。于是我要睡觉了,梦里好好调戏内功。"[②]

2 任务

2.1 形式化

标点符号预测可视为一个序列标注任务,即给定一个文本输入序列:$X=\{x_1,x_2,x_i,\cdots,x_n\}$,需要得到一个标点符号标签序列$Y=\{y_1,y_2,y_i,\cdots,y_n\}$。模型所需要预测的标签集合如表3所示。标签集合中,0为无标点(space),1为逗号,2为句号,3为问号,4为顿号,5为分号。

表3 标点标签

标点	space	,	。	?	、	;
标签	0	1	2	3	4	5

① 计算标注一致率时包含标点符号类型的一致和断句的一致,所以五个标点符号以及空符号都包含在内。

② 选自微博(长)标准数据集。

2.2 设置

严格设置 仅以数据集每个条目的金标准文本作为正确的标注结果，既每个待预测文本，其正确的标点符号唯一。此种设置下，各数据集分别命名为：问答—严格，微博(中)—严格，微博(长)—严格。

宽松设置 将数据集中每个条目的所有标注文本作为正确的标注结果，即包含金标准文本也包含可选标准文本，对每个待预测文本，部分标注的结果可以不唯一，但都视为正确的标注结果。此种设置下，各数据集分别命名为：问答—宽松，微博(中)—宽松，微博(长)—宽松。

此外，由于数据集中每个条目均可以由多句组成，因此在本文的标点标签体系中，句中标点标签有6种可能(含无标点)，但句末标点仅有两种可能：问号/句号，模型会相对容易地学到句末标点的信息。考虑到这个影响，针对句末，使用以下两个设置：

(1) 含句末。在测试时，将文本中所有标点符号考虑在内(包含句末标点)。

(2) 无句末。在测试时，不将文本中的句末标点考虑在内。

3 模型与实验

3.1 模型

基于预训练语言模型 BRET 的方法在 NLP 领域各项任务上均取得了很好的性能，本文也基于 BERT 建立了一个简单的标点符号预测基线模型。向模型输入一段文本，$X=\{x_1,x_2,x_3,\cdots,x_n\}$，BERT 首先把模型的输入转化为词嵌入矩阵，再经过一个线性变换层将最后一维的词嵌入维度转换为标签，最后经过 softmax 层输出序列 Y，$Y=\{y_1,y_2,y_3,\cdots,y_n\}$，代表每个字后面的标签序列。

由表 2 可知，本数据集中标点符号的分布很不平衡。实际上，文本中的标点符号数量比文字的数量少得多，因此上述模型输出大部分的标签都是无标点的“0”标签，直接采用交叉熵作为损失函数会导致模型在训练时更倾向于输出无标点类别，使得模型学习不到足够的标点特征。为解决这个问题，我们将原来的交叉熵损失改为 Focal Loss 损失[14]，该损失函数调整了样本在训练中所占的权重。原本的 Focal Loss 是在二分类中实现的，这里将它扩展到多分类问题当中。原 Focal Loss 公式如式(1)所示。

$$L_{fl}=\begin{cases}-\alpha(1-y')^{\gamma}\log y', & y=0\\ -(1-\alpha)y'^{\gamma}\log(1-y'), & y=1\end{cases} \tag{1}$$

其中，α 和 γ 是两个可以调节的超参数。

我们将二分类的 Focal Loss 拓展到多分类中，如式(2)所示。

$$L_{fl}=-\alpha_i(1-y_i)^{\gamma}\log(y_i) \tag{2}$$

其中，α_i 代表第 i 个标签的调节因子，y_i 代表第 i 个标签的预测概率。

3.2 参数设置

本文使用的 BERT 模型是 Google 公开的 bert-base[①]，模型由 12 层的 Transformer Encoder 预训练而成，自注意力头数为 12，隐藏层维度为 768，总参数量为 110 MB。训练时，我们设置的学习率大小是 5e-5，批大小是 64，Dropout 设置为 0.25，训练轮数为 15 轮。

3.3 评价指标

本文使用分类问题的评价指标精确度 P(Precision)、召回率 R(Recall)和 F_1 值来评价模型整体性能，以 F_1 值作为主要评价指标，具体如式(3)所示。

$$F_1=\frac{2*\text{Precison}*\text{Recall}}{\text{Precison}+\text{Recall}} \tag{3}$$

3.4 实验结果

首先利用在第 1 节中整理好的人民日报训练语料进行训练。然后，在本文建立数据集的问答及微博领域进行测试。实验结果详见表 4。该表列出了本文各领域数据集在各项任务设置下的模型性能。基线模型在新闻领域中的性能并无宽松/严格设置，列在最后一行。其中，“含”指包含句末标点，“无”表示不包含句末标点。

表 4 实验结果

数据集	P		R		F_1	
	含	无	含	无	含	无
问答-严格	0.822 3	0.748 3	0.834 8	0.769 7	0.828 5	0.758 9
微博(中)-严格	0.679 9	0.629 5	0.666 1	0.613 8	0.672 9	0.621 6
微博(长)-严格	0.682 3	0.647 8	0.688 9	0.654 3	0.685 6	0.651 0

① https://github.com/google-research/bert

续表

数据集	P		R		F_1	
	含	无	含	无	含	无
问答-宽松	0.835 0	0.760 3	0.847 4	0.788 4	0.841 2	0.774 1
微博(中)-宽松	0.698 1	0.650 9	0.693 5	0.644 7	0.695 8	0.647 8
微博(长)-宽松	0.693 0	0.659 2	0.702 9	0.669 5	0.697 9	0.664 3
新闻	0.883 4	0.864 7	0.879 4	0.860 0	0.881 4	0.862 4

可以发现,模型在问答领域的性能较好,明显高于微博领域的性能。因为问答领域文本较短,微博领域的文本更不规范,这也比较符合实际情况与我们的预期。同时,宽松设置均优于严格设置,含句末设置均优于不含句末设置。

图2是对比新闻领域严格-宽松两种任务设置下性能下降的柱状图。可以看出,对比新闻领域,基线模型迁移到问答及微博领域后,标点符号预测的性能在所有设置下,均有不同程度的下降,微博领域的下降更多,微博(中)下降的幅度大于微博(长),微博(中)-严格下降得最为明显,超过了20%。跨领域标点标注任务具有一定挑战性,模型性能还有较大提升空间。

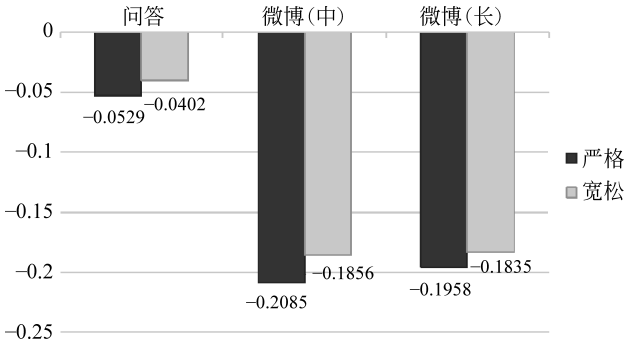

图2 各领域较新闻领域性能下降柱状图

图3是无句末设置比含句末设置时模型性能下降的柱状图。在所有领域,“无句末”均较“含句末”有不同程度的下降,其中问答领域下降幅度最高(近7%),微博次之,新闻领域下降最少。问答领域下降幅度最高的原因在于,问答领域中的问句(对应句末问号)或答句(对应句末句号)比较典型,模型相对容易判断。

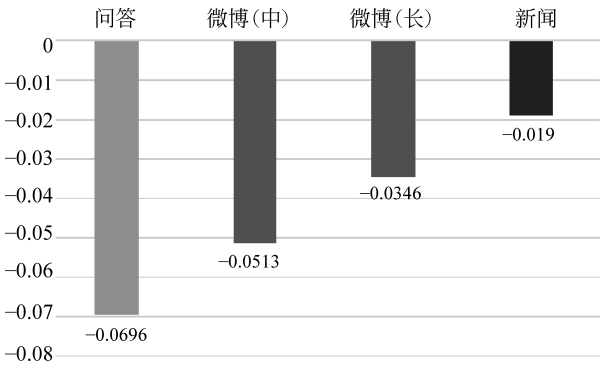

图3 “无句末”比“含句末”性能下降柱状图

4 错误分析与讨论

标点符号的预测错误可分为三种类型:①漏标,即本应有标点的地方,预测为无标点;②多标,即本应没有标点的地方,预测出现标点;③错标,即预测的标点类型与数据集人工标注的标点不一致。

表5为基线模型预测错误类型分布。是否包含句末标点只影响错标的分布,且对整体预测错误的分布影响较小。注意,此处没有列出不包含句末标点的情况①。

表5 基线模型预测错误类型分布

数据集	模型预测		
	漏标/%	多标/%	错标/%
问答-严格	39.02	24.97	36.00
微博(中)-严格	25.73	44.03	30.23
微博(长)-严格	29.74	42.74	27.52
问答-宽松	39.16	24.43	36.41
微博(中)-宽松	26.25	42.97	30.79
微博(长)-宽松	29.46	42.35	28.19

表5中,漏标、多标是断句错误,错标是标点使用错误。可以看出,严格和宽松两个领域错误类型分布差别不大。问答领域的漏标错误要明显高于微博领域,相反多标的情况要少于微博领域。受到领域特征的影响,问答领域的文本长度普遍较短,而基线模型是在新闻领域中训练的,新闻领域中平均12.51个字才有一个标点(表1),因此,问答领域中

① 需要说明的是,由于标点使用不具有唯一性,所以实验得出的错误文本只是相对于标准数据集而言的,机器标注的某些文本虽然不在标准数据集中但其标注也可能是正确的,但这种情况非常少。

的文本，模型在预测时可能会完全不标注。所以问答领域的漏标情况会比微博领域更多。在微博领域中，多标的错误要比其他两种错误多，这是因为微博领域的文本更加口语化，通常还会有一些专有名词、缩略语、新词新语等，在自动标注的过程中，会出现固定搭配之间被插入标点符号的情况，最终导致微博领域的多标错误所占比例较大。

以下是不同领域中的错误案例，各种类型各三组文本，每组文本包含两条文本，第一条文本为自动标注得到的错句，句前有 * 号；后一句为标准数据集中的正确文本。三组文本分别来自问答、微博(中)及微博(长)。

(1) **漏标**

* 巴金的激流三部曲爱情三部曲分别是什么？

巴金的激流三部曲、爱情三部曲分别是什么？

* 余华、马原、李耳格非不敢和他们谈文学，但是私下找余华老师看了下牙，他说牙龈是可以自我恢复的，解我多年心头大惑。

余华、马原、李耳、格非，不敢和他们谈文学，但是私下找余华老师看了下牙，他说牙龈是可以自我恢复的，解我多年心头大惑。

* ……[①]会上，李公平局长强调，要严肃工作纪律，切实加强对填报志愿工作的督查和管理，确保今年网上填报志愿工作平稳有序顺利进行。

……会上，李公平局长强调，要严肃工作纪律，切实加强对填报志愿工作的督查和管理，确保今年网上填报志愿工作平稳、有序、顺利进行。

(2) **多标**

* 可以修改邮箱，请您提供其他邮箱，谢谢，您了。

可以修改邮箱，请您提供其他邮箱，谢谢您了。

* 光明能挺住吗？带不带？这么恐吓老百姓的？这个是不是夸张了点啊？…

光明能挺住吗？带不带这么恐吓老百姓的？这个是不是夸张了点啊？…

* ……人是衰，到了何种境界才能发生百年难遇的事？思绪凌乱了。

……人是衰到了何种境界才能发生百年难遇的事？思绪凌乱了。

(3) **错标**

* 八年级上学期期末考试，地理考不考下学期的内容。

八年级上学期期末考试，地理考不考下学期的内容？

* 体育满分，是我的体质变好了，能跑了，还是上大学的女生都颓废了，我五十米第一，排球第一。这是怎么了？我神灵附体了。

体育满分。是我的体质变好了，能跑了，还是上大学的女生都颓废了？我五十米第一，排球第一，这是怎么了？我神灵附体了？

* 犯了一个错，需要另外十个错误来掩盖啊，他们小区的监控镜头能证明他在家睡觉，还有这样负责的物管啊。……

犯了一个错，需要另外十个错误来掩盖啊。他们小区的监控镜头能证明他在家睡觉？还有这样负责的物管啊？……

综合两个领域的错误案例发现，在断句错误中，漏标常见于文本中的并列成分之间，漏标的标点多为顿号。多标常会造成语义内涵错误，多标的符号多是逗号或者句号。而标点符号类型错误则多见于句号和问号之间，造成疑问和陈述语气混淆。有些自动预测的错误是 OOV 识别造成的，如类型(1)中的第二组；有些预测错误较为明显，可能是人民日报语料中没有出现类似“考不考”这样的上下文，如类型(3)中的第 1 组；但更多的预测错误难以分析具体原因，通常需要对句子意义的精细把握与理解。

5　相关工作

国际上标点符号预测或标点符号恢复任务的相关研究主要集中在语音识别领域，使用的方法大都是基于机器学习或深度学习的方法，输入数据为听觉信息、文本信息或两者的结合。标点符号预测任务的目前主流研究方法可按目标问题分为以下两类：

第一类是将该任务视为序列标注问题[15-16]，模型要为每一个位置指定一个标点符号(或无)。一些研究[3,15,17]表明条件随机场(CRF)在标点符号预测任务上是比较有效的。近年来，随着神经网络的兴起，文献[18]提出了一种基于卷积神经网络的模型来进行标点预测，结果表明，基于神经网络的方法优于之前基于 CRF 的方法。文献[5]基于长短时记忆网络(LSTM)及带注意力机制的双向反馈神经网络模型(T-BRNN)进一步提高了标点符号预测的性能。文献[19]利用双向 LSTM 结合 CRF 模型

① 省略号(……)并非文本之中的符号，因文本较长，在此省略上下文。

(BiLSTM-CRF)以及一个以上的集成模型取得了基于序列标注方法目前的最佳性能。

第二类是将其视为单语机器翻译问题,源语言为不含有标点符号的文本,目标语为带标点符号的文本[4,20-21],或目标语为标点符号序列[22-23],提出了一个带注意力机制的编码器/解码器架构来解决标点符号预测。Kim[24]提出一种带逐层多头注意力的RNN网络进行标点符号预测,并取得了仅使用词汇特征方法的最好性能。受自注意力机制[25]在NLP任务中有效性的影响,文献[26]提出了一个利用自注意力机制的神经网络模型,可同时在文本和语音的嵌入基础上利用自注意力来获得更好的表示。

还有学者引入其他相关任务来提升标点符号预测的性能,文献[27]提出一种联合句法分析的标点符号预测方法,该方法能利用丰富的句法标注信息,取得了很好的效果。此外,在训练CRF与神经网络时,由于发现词性信息可以作为有效提升标点符号标注性能的特征[28],文献[29]提出了一种基于BERT的对抗多任务学习方法,在标点符号预测任务外,额外训练词性标注任务,两者进行对抗,最终在标点符号预测任务上取得了很好的性能。

虽然有少数对越南语[30]及中文[31]的相关研究,以及针对中汉语断句研究[32-35]。但大多数研究基本都集中在英语上。绝大多数研究基于IWSLT数据集[6]。该数据集语料来源于TED公开演讲,主题十分广泛,转录质量很高。这个数据集[18]经重新组织整理,训练数据集来源于IWSLT2012英文翻译track,约210万个单词、14.4万个文本。开发集约29.6万个单词,2.1万个文本。有两个测试集及Ref和ASR,来源于IWSLT2011,包含约1.3万个单词、860个文本。数据集中有逗号、句号和问号三种标点符号,以及一个非标点标记"O"。

6 结论

本文提出一个领域迁移的标点恢复/标注任务,标注了一个包含问答领域、微博短文本和微博长文本领域的测试集集合,并用给定人民日报语料作为验证集。我们给出一个基于预训练语言模型BERT的基线模型,并使用Focal Loss来缓解标签不均衡问题。该模型在人民日报语料集上进行训练并在本数据集上进行了验证。在此基础上,向问答及微博两个领域进行迁移。实验结果表明,向问答领域的迁移效果较好,但是向社交媒体(微博)领域的迁移效果较差,且比源领域下降了20%。我们进一步对模型自动标注的结果进行了分析,发现漏标、多标与错标这几种错误的分布较为均衡;从领域比较来看,微博领域更容易多标,问答领域更容易漏标。有些自动标注的错误确实需要比较敏感的语感才能辨别。总体来说,跨领域标点符号迁移任务具有一定挑战性,特别是向微博领域迁移,各模型在这个任务上还有较大的提升空间,未来可以利用各种迁移学习或多任务学习的方法来尝试解决。

参考文献

[1] Beeferman D, Berger A, Lafferty J. Cyberpunc: A lightweight punctuation annotation system for speech [C]//Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998, 2: 689-692.

[2] Liu Y, Shriberg E, Stolcke A, et al. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(5): 1526-1540.

[3] Lu W, Ng H T. Better punctuation prediction with dynamic conditional random fields[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010: 177-186.

[4] Peitz S, Freitag M, Mauser A, et al. Modeling punctuation prediction as machine translation[C]//Proceedings of the International Workshop on Spoken Language Translation, 2011: 238-245.

[5] Tilk O, Alumäe T. LSTM for punctuation restoration in speech transcripts[C]//Proceedings of the 16th Annual Conference of the International Speech Communication Association, 2015: 683-687.

[6] Federico M, Cettolo M, Bentivogli L, et al. Overview of the IWSLT 2012 evaluation campaign[C]//Proceedings of the IWSLT-International Workshop on Spoken Language Translation, 2012: 12-33.

[7] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv: 1810.04805, 2018.

[8] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 2980-2988.

[9] 谢丽星，周明，孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报，2012，26(1)：73-83.

[10] 古万荣，董守斌，曾之肇，等. 基于微博用户模型的个性化新闻推荐[J]. 中文信息学报，2016，30(1)：93-101.

[11] 贺敏，刘玮，刘悦，等. 基于特征驱动的微博话题检测方法[J]. 中文信息学报，2017，31(3)：101-108.

[12] 王志宏，过弋. 微博谣言事件自动检测研究[J]. 中文信息学报，2019，33(6)：132-140.

[13] Liu X，Chen Q，Deng C，et al. LCQMC：A large-scale chinese question matching corpus[C]//Proceedings of the 27th International Conference on Computational Linguistics，2018：1952-1962.

[14] Lin T Y，Goyal P，Girshick R，et al. Focal loss for dense object detection[J]. arXiv preprint arXiv:1708.02002，2017.

[15] Ueffing N，Bisani M，Vozila P. Improved models for automatic punctuation prediction for spoken and written text[C]//Proceedings of the Interspeech，2013：3097-3101.

[16] Želasko P，Szymański P，Mizgajski J，et al. Punctuation prediction model for conversational speech[J]. arXiv preprint arXiv:1807.00543，2018.

[17] Hasan M，Doddipatla R，Hain T. Noise-matched training of CRF based sentence end detection models[C]//Proceedings of the 16th Annual Conference on the International Speech Communication Association，2015：349-353.

[18] Che X，Wang C，Yang H，et al. Punctuation prediction for unsegmented transcript based on word vector[C]//Proceedings of the 10th International Conference on Language Resources and Evaluation，2016：654-658.

[19] Yi J，Tao J，Wen Z，et al. Distilling knowledge from an ensemble of models for punctuation prediction[C]//Proceedings of the Interspeech，2017：2779-2783.

[20] Driesen J，Birch A，Grimsey S，et al. Automated production of true-cased punctuated subtitles for weather and news broadcasts[C]//Proceedings of the 15th Annual Conference of the International Speech Communication Association，2014：2146-2147.

[21] Cho E，Niehues J，Waibel A. Segmentation and punctuation prediction in speech language translation using a monolingual translation system[C]//Proceedings of the International Workshop on Spoken Language Translation，2012：252-259.

[22] Klejch O，Bell P，Renals S. Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches[C]//Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE，2016：433-440.

[23] Klejch O，Bell P，Renals S. Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features[C]//Proceedings of the 2017 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2017：5700-5704.

[24] Kim S. Deep recurrent neural networks with layer-wise multi-head attentions for punctuation restoration[C]//Proceedings of the International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2019：7280-7284.

[25] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need[J]. arXiv preprint arXiv：1706.03762，2017.

[26] Yi J，Tao J. Self-attention based model for punctuation prediction using word and speech embeddings[C]//Proceedings of the International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2019：7270-7274.

[27] Zhang D，Wu S，Yang N，et al. Punctuation prediction with transition-based parsing[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics，2013：752-760.

[28] Cho E，Kilgour K，Niehues J，et al. Combination of NN and CRF models for joint detection of punctuation and disfluencies[C]//Proceedings of the 16th Annual Conference of the International Speech Communication Association，2015：3650-3654.

[29] Yi J，Tao J，Bai Y，et al. Adversarial transfer learning for punctuation restoration[J]. arXiv preprint arXiv:2004.00248，2020.

[30] Pham Q H，Nguyen B T，Cuong N V. Punctuation prediction for Vietnamese texts using conditional random fields[C]//Proceedings of the 10th International Symposium on Information and Communication Technology，2019：322-327.

[31] Zhao Y，Wang C，Fu G. A CRF sequence labeling approach to Chinese punctuation prediction[C]//Proceedings of the 26th Pacific Asia Conference on Language，Information，and Computation，2012：508-514.

[32] 黄建年，侯汉清. 农业古籍断句标点模式研究[J].中文信息学报，2008，22(4)：31-38.

[33] 张开旭，夏云庆，宇航. 基于条件随机场的古汉语自动断句与标点方法[J]. 清华大学学报(自然科学版)，2009，49(10)：1733-1736.
[34] 王博立，史晓东，苏劲松. 一种基于循环神经网络的古文断句方法[J]. 北京大学学报（自然科学版)，2017，53(2)：255-260.
[35] 俞敬松，魏一，张永伟. 基于 BERT 的古文断句研究与应用[J]. 中文信息学报，2019，33(11)：57-63.

刘鹏远(1974—)，通信作者，博士，副研究员，主要研究领域为情感分析、关系抽取、阅读理解。
E-mail：liupengyuan@pku.edu.cn

邱立坤(1979—)，教授，博士生导师，主要研究领域为人机对话、知识图谱与语言资源建设。
E-mail：qiulikun@pku.edu.cn

王伟康(1996—)，硕士，主要研究领域为自然语言处理。
E-mail：978955719wwk@gmail.com

第二十届中国计算语言学大会(CCL 2021)开放注册

会议介绍

中国计算语言学大会(The China National Conference on Computational Linguistics，CCL)创办于1991 年，是中国中文信息学会(CIPS)的旗舰会议。经过近三十年的发展，CCL 被广泛认为是最权威的、全国最具影响力、规模最大的 NLP 会议之一。CCL 聚焦于中国境内各类语言的智能计算和信息处理，为研讨和传播计算语言学最新学术和技术成果提供了高层次交流平台。

会议信息

第二十届中国计算语言学大会(The Twentieth China National Conference on Computational Linguistics，CCL 2021) 将于 2021 年 8 月 13 日—8 月 15 日在内蒙古自治区呼和浩特市召开。会议组织单位为中国中文信息学会计算语言学专业委员会，承办单位为内蒙古大学。今年会议邀请了西安交通大学徐宗本院士、北京语言大学语言科学院冯胜利教授、清华大学基础科学讲席教授刘嘉教授、京东探索研究院陶大程院长、清华大学计算机系副主任唐杰教授五位知名专家做特邀报告。本次讲习班邀请到了中国人民大学高瓴人工智能学院长聘副教授赵鑫、清华大学自动化系助理教授黄高、复旦大学大数据学院副教授魏忠钰、南京理工大学计算机学院教授夏睿四名学者。

时间：2021 年 8 月 13 日—15 日　**地点**：内蒙古开元名都大酒店　**地址**：内蒙古自治区呼和浩特市赛罕区呼伦贝尔南路 119 号　**会议官网**：http://cips-cl.org/static/CCL2021/index.html　**注册参会**：http://reg.cipsc.org.cn/ccl2021/　**会议订房**：https://jinshuju.net/f/zh0MnY

会议官网二维码　　会议注册二维码