CDCPP：跨领域中文标点符号预测

刘鹏远　　王伟康　　邱立坤　　杜冰洁
北京语言大学信息科学学院
国家语言资源监测与研究平面媒体中心
闽江学院计算机与控制工程学院
liupengyuan@pku.edu.cn　978955719wwk@gmail.com
qiulikun@pku.edu.cn　blcudbj@gmail.com

摘要

标点符号对文本理解起很大作用。但目前，在中文文本特别是在社交媒体及问答领域文本中的标点符号使用存在非常多的错误或缺失的情况，这严重影响对其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预测的相关研究，也没有这些领域公开的数据集。本文首先提出跨领域中文标点符号预测任务，该任务是要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上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%，跨领域标点符号预测任务具有一定的挑战性。

关键词： 中文标点符号预测任务 ； 跨领域 ； 数据集

CDCPP: Cross-Domain Chinese Punctuation Prediction

PengYuan Liu　　WeiKang Wang　　LiKun Qiu　　BingJie Du
Beijing Language and Culture University, School of Information Science
Language Resources Monitoring and Reserch Center
Minjiang University,School of Computer and Control Engineering
liupengyuan@pku.edu.cn　978955719wwk@gmail.com
qiulikun@pku.edu.cn　blcudbj@gmail.com

Abstract

Punctuation marks play a important role in text understanding. But at present, there are many errors or lacks in Chinese texts, especially in social media and Q&A texts, which seriously affects the effect of various natural language processing such as semantic analysis and machine translation. The current research on punctuation prediction is mostly focused on the speech transcribed text of English conversations. There is a lack of research on punctuation prediction in social media and Q&A texts, and there is no public dataset in these fields. This paper first proposes a cross-domain Chinese punctuation prediction task. The task is to build a punctuation prediction model using a large-scale news domain text that is basically standardized and correct punctuation, and then conduct punctuation predict in the social media and Q&A fields where punctuation is not standardized. Subsequently, corresponding datasets in the three fields of news, social media and Q&A were constructed. Finally, a BERT-based punctuation prediction baseline model was implemented and experiments were

performed on this dataset. The experimental results show that directly using the model trained in the news field, the performance of punctuation prediction in the social media and Q&A fields has decreased, and the decrease in the Q&A field is small, and the decrease in the Weibo field is large, exceeding 20% . Cross-domain punctuation prediction task has certain challenges.

Keywords: Chinese Punctuation Prediction Task , Cross-Domain , Dataset

1 引言

汉语书面语中，标点符号有着不可或缺的地位。《辞海》[0]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，随着社交媒体领域（如微博）及问答领域（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变得愈来愈重要。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[1]及机器翻译[2]影响的对比。其中：c及c’分别为百度知道[3]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。我们还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产生很大影响。为社交媒体及问答等领域中的文本标注正确的标点符号，具有较高的意义和应用价值。

标点符号预测（Punctuation Prediction，PP）或标点符号恢复（Punctuation Restoration，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，故而标点符号预测相关研究工作集中在语音识别领域，主要是面向对话领域的语音转写文本进行标点符号预测(Beeferman et al., 1998; Liu et al., 2006; Lu and Ng, 2010; Peitz et al., 2011; Tilk and Alumäe, 2015)。目前常用公开的数据集为IWSLT(Federico et al., 2012)，是 ˆ 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相关数据集，这对在该领域进行标点符号预测进行研究产生很大限制。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自动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，该任务是要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。本文构建了新闻、社交媒体及问答三个领域的数据集[4]。在新闻领域，提供了测试集，共10000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试集各1200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。

为对本任务及数据集进行初步评估，鉴于近年来预训练语言模型BRET（Bidirectional Encoder Representation from Transformers）(Devlin et al., 2018)在NLP领域各项任务中的优良

[0]http://chlb.cishu.com.cn/
[1]采用哈工大语言技术LTP平台：http://ltp.ai/index.html
[2]采用谷歌翻译：https://translate.google.cn
[3]http://zhidao.baidu.com
[4]https://github.com/NLPBLCU/Cross-Domain-Chinese-Punctuation-Prediction

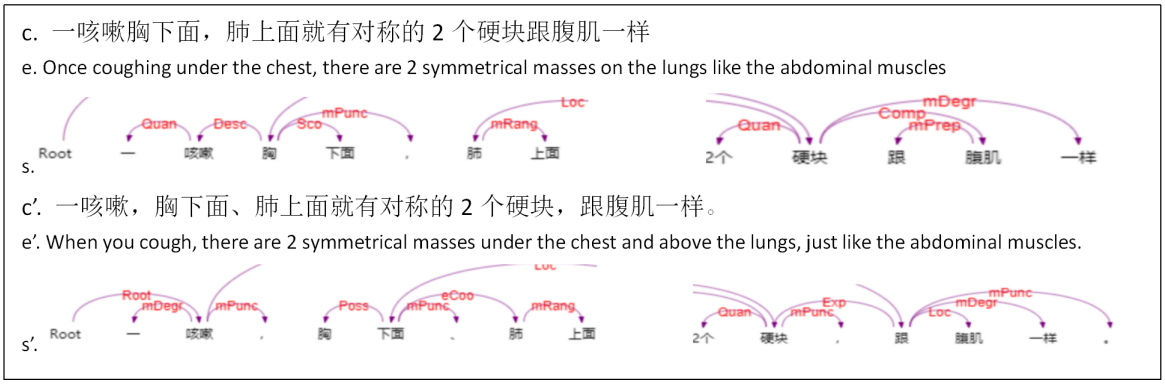


Figure 1: 标点符号标注错误对依存语义分析及机器翻译结果的影响。其中c及c'分别为百度知道中的实例及人工对其进行标点重新标注后的文本。e及e'是分别对c及c'用谷歌翻译的结果。s及s'是分别对c及c'利用LTP平台进行语义依存标注的结果，限于篇幅，仅截取了部分。

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，同时还使用了Focal Loss(Lin et al., 2017)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上进行标点符号预测的性能均有有所下降，特别是在微博领域下降较大，跨领域标点符号预测任务具有一定的挑战性。

2 数据集构建

2.1 数据准备

新闻领域选择人民日报2018年全年语料。该语料共574332条文本。我们首先按照本文标点符号标签表（见第三小节中的表3）对该文本进行处理：去除无关的一些噪音标点和符号，对标点进行全半角和重复的符号归一化处理，再将冒号替换为逗号，将感叹号和省略号替换为句号。然后，在该语料中随机抽取10000条作为测试集，另随机抽取100000条作为训练集。

社交媒体领域选择新浪微博[5]，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(谢丽星，周明，孙茂松, 2012; 古万荣，董守斌，曾之肇，何锦潮，刘崇, 2016; 贺敏，刘玮，刘悦，王丽宏，白硕，程学旗, 2017; 王志宏，过弋, 2019)。本文随机爬取微博语料共120250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(Liu et al., 2018)，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260068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，文本长度在15以内的文本中，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。

2.2 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[6]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

[5]http://weibo.com
[6]http://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=22EA6D162E4110E752259661E1A0D0A8

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。

2.3 标注过程

整个标注由3名语言学在读硕士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以及规范参考。待3名标注员均熟悉规范和标注后再进行正式标注。

对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。语料中出现的其他非标点符号，如表情符号等手动删除。对新浪微博语料，我们控制标注每种文本长度的语料各600条，共1200条。对百度知道语料，我们控制标注共1200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2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，如：

a.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。

b.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，只是各人语感不同，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，三名标注员对本条标注的不同结果进行讨论，对三人加以讨论后确定符合《用法》的结果也加以保留，作为可选标准文本，附加在金标准文本后，以空格分开。

2.4 统计分析

最终形成的数据集共包含问答领域、微博中等及微博长文本数据各1200、600及600条。由于每条可能包含有多个正确的标注结果，实际共包含问答领域、微博（中）及微博（长）数据各1328、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：0.9751，0.9572及0.9731[7]。

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而没有将可选标准文本包含在内进行统计。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，所得到的的结果差异不大。

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	10000	70.16	5.61	**12.51**
问答	1200	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

Table 1: 数据集的基本统计情况

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最短，包含标点个数最少。从平均文本长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

[7]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。

标点符号	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

Table 2: 数据集中各种标点分布情况(%)

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。

ˆ 对微博长句种问号使用比句号使用更频繁的现象，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，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[8]。”

3 任务

3.1 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:$X = \{x_1, x_2, x_i, \dots x_n\}$，需要得到一个标点符号标签序列$Y = \{y_1, y_2, y_i, \dots y_n\}$。模型所需要预测的标签集合如下表所示。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。

标点	标签	标点	标签
space	0	？	3
，	1	、	4
。	2	；	5

Table 3: 标点标签表

3.2 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，即每个待预测文本，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，对对每个待遇测文本，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，ˆ 对句末，使用以下两个设置：

1）**含句末。**即在测试时，将文本中所有标点符号考虑在内（包含句末标点）。

2）**无句末。**即在测试时，不将文本中的句末标点考虑在内。

4 实验

4.1 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，文本也基于BERT建立了一个简单的标点符号预测基线模型。如下图所示，模型输入一段文本，$X = \{x_1, x_2, x_i, \dots x_n\}$,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

[8]选自微博（长）标准数据集。

一维的词嵌入维度转换为标签，最后经过softmax层输出序列Y，$Y = \{y_1, y_2, y_i, \ldots y_n\}$，代表每个字后面的标签序列。

从本文第2节表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其中字的数量少得多，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现在的Focal Loss损失(Lin et al., 2017)，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 公式如下：

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

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：

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

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。

4.2 参数设置

本文使用的BERT模型是Google公开的bert-base[9]，模型是由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110MB。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。

4.3 评价指标

在本文的中文标点符号预测任务中，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，具体公式如下图所示：

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

4.4 实验结果

首先利用在本文第2小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下进行测试。实验结果详见表4。表4列出了本文各领域数据集在本文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。

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，因此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。

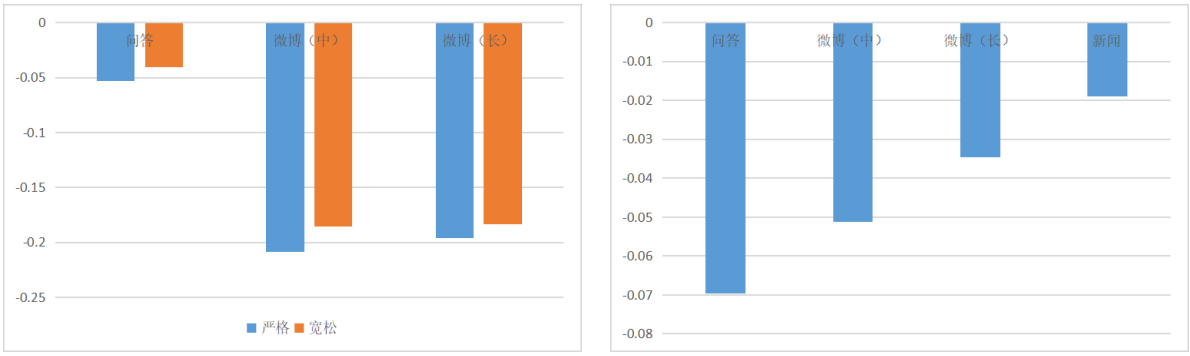
图2（a）是对比新闻领域，严格-宽松两种任务设置下性能下降的柱状图。可以看出，对比新闻领域，基线模型迁移到问答及微博领域后，标点符号预测的性能在所有设置下，均有不同程度的下降，微博领域的下降更多，微博（中）下降的幅度大于微博（长），微博（中）-严格下降的最为明显，超过了20%。跨领域标点标注任务具有一定挑战性，模型性能还有较大提升空间。

图2（b）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，无句末均较含句末均有有不同程度下降，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。

[9]https://github.com/google-research/bert

数据集	P		R		F_1	
	含句末	无句末	含句末	无句末	含句末	无句末
问答-严格	0.8223	0.7483	0.8348	0.7697	0.8285	0.7589
微博（中）-严格	0.6799	0.6295	0.6661	0.6138	0.6729	0.6216
微博（长）-严格	0.6823	0.6478	0.6889	0.6543	0.6856	0.6510
问答-宽松	0.8350	0.7603	0.8474	0.7884	0.8412	0.7741
微博（中）-宽松	0.6981	0.6509	0.6935	0.6447	0.6958	0.6478
微博（长）-宽松	0.6930	0.6592	0.7029	0.6695	0.6979	0.6643
新闻	0.8834	0.8647	0.8794	0.8600	0.8814	0.8624

Table 4: 基线模型在本文任务设置下各领域数据集上的性能结果



(a) 各领域较新闻领域性能下降柱状图　　(b) 无句末比含句末性能下降柱状图

Figure 2: 各模型性能下降柱状图

5 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。前两种类型是断句错误，最后一种类型是标点使用错误。表5为两个领域的标注错误类型分布情况[10]。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（见2.4节表1），因此，问答领域中的文本，模型在预测时可能会完全不标注。因此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要多，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，因此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，句前有错号，后一句为标准数据集中的正确文本，句前有对号。三组文本分别来自问答、微博（中）及微博（长）：

[10]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。

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	27.79
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

Table 5: 基线模型预测错误类型分布及原始标注错误类型分布（%）。由于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，此处没有列出不包含句末标点的情况。

a.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。

✗...[11]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。

b.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。

c.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

[11]以下省略号并非文本之中的符号，因文本较长，在此省略上下文。

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，如例a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，如例c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。

6 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在语音识别领域。主要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。标点符号预测或任务的目前主流研究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(Ueffing et al., 2013; Żelasko et al., 2018)，模型要为每一个位置指定一个标点符号（或无）。一些研究(Lu and Ng, 2010; Ueffing et al., 2013; Hasan et al., 2015)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，Che et al. (2016)首先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。Tilk and Alumäe (2015)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)进一步提高了标的符号预测的性能。Yi et al. (2017)利用双向LSTM结合CRF模型（BiLSTM-CRF）以及一个其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(Peitz et al., 2011; Driesen et al., 2014; Cho et al., 2012)，或目标语为标点符号序列如(Klejch et al., 2016; Klejch et al., 2017)，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 (2019)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Vaswani et al. (2017)在NLP任务中有效性，Yi and Tao (2019)提出了一个利用自注意力机制的神经网络模型，可同时在文本和声学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，Zhang et al. (2013)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，由于词性信息可以作为有效提升标点符号标注性能的特征Cho et al. (2015)，Yi et al. (2020)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(Pham et al., 2019)及中文(Zhao et al., 2012)的相关研究，以及 ˆ 对中文古文的古文断句研究如黄建年，侯汉清 (2008)；张开旭，夏云庆，宇航 (2009)；王博立，史晓东，苏劲松 (2017)及俞敬松， O 一，张永伟 (2019),。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(Federico et al., 2012)。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。这个数据集经Che et al. (2016)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”。

7 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、多标与类型错误这几种错误的分布较为均衡；从领域比较来看，微博更容易多标，问答更容易漏标。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。

致谢

本文受北京市自然科学基金资助项目（4192057）资助。感谢匿名评阅人的建议。

参考文献

Doug Beeferman, Adam Berger, and John Lafferty. 1998. Cyberpunc: A lightweight punctuation annotation system for speech. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 2, pages 689–692. IEEE.

Xiaoyin Che, Cheng Wang, Haojin Yang, and Christoph Meinel. 2016. Punctuation prediction for unsegmented transcript based on word vector. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 654–658.

Eunah Cho, Jan Niehues, and Alex Waibel. 2012. Segmentation and punctuation prediction in speech language translation using a monolingual translation system. In *International Workshop on Spoken Language Translation (IWSLT) 2012*.

Eunah Cho, Kevin Kilgour, Jan Niehues, and Alex Waibel. 2015. Combination of nn and crf models for joint detection of punctuation and disfluencies. In *Sixteenth annual conference of the international speech communication association*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Joris Driesen, Alexandra Birch, Simon Grimsey, Saeid Safarfashandi, Juliet Gauthier, Matt Simpson, and Steve Renals. 2014. Automated production of true-cased punctuated subtitles for weather and news broadcasts. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Marcello Federico, Mauro Cettolo, Luisa Bentivogli, Paul Michael, and Stüker Sebastian. 2012. Overview of the iwslt 2012 evaluation campaign. In *IWSLT-International Workshop on Spoken Language Translation*, pages 12–33.

Madina Hasan, Rama Doddipatla, and Thomas Hain. 2015. Noise-matched training of crf based sentence end detection models. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Seokhwan Kim. 2019. Deep recurrent neural networks with layer-wise multi-head attentions for punctuation restoration. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7280–7284. IEEE.

Ondřej Klejch, Peter Bell, and Steve Renals. 2016. Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 433–440. IEEE.

Ondřej Klejch, Peter Bell, and Steve Renals. 2017. Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5700–5704. IEEE.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. *arXiv e-prints*, page arXiv:1708.02002, August.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on audio, speech, and language processing*, 14(5):1526–1540.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. LCQMC:a large-scale Chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, August.

Wei Lu and Hwee Tou Ng. 2010. Better punctuation prediction with dynamic conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 177–186.

Stephan Peitz, Markus Freitag, Arne Mauser, and Hermann Ney. 2011. Modeling punctuation prediction as machine translation. In *International Workshop on Spoken Language Translation (IWSLT) 2011*.

Quang H Pham, Binh T Nguyen, and Nguyen Viet Cuong. 2019. Punctuation prediction for vietnamese texts using conditional random fields. In *Proceedings of the Tenth International Symposium on Information and Communication Technology*, pages 322–327.

Ottokar Tilk and Tanel Alumäe. 2015. Lstm for punctuation restoration in speech transcripts. In *Sixteenth annual conference of the international speech communication association*.

Nicola Ueffing, Maximilian Bisani, and Paul Vozila. 2013. Improved models for automatic punctuation prediction for spoken and written text. In *Interspeech*, pages 3097–3101.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jiangyan Yi and Jianhua Tao. 2019. Self-attention based model for punctuation prediction using word and speech embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7270–7274. IEEE.

Jiangyan Yi, Jianhua Tao, Zhengqi Wen, Ya Li, et al. 2017. Distilling knowledge from an ensemble of models for punctuation prediction.

Jiangyan Yi, Jianhua Tao, Ye Bai, Zhengkun Tian, and Cunhang Fan. 2020. Adversarial transfer learning for punctuation restoration. *arXiv preprint arXiv:2004.00248*.

Piotr Żelasko, Piotr Szymański, Jan Mizgajski, Adrian Szymczak, Yishay Carmiel, and Najim Dehak. 2018. Punctuation prediction model for conversational speech. *arXiv preprint arXiv:1807.00543*.

Dongdong Zhang, Shuangzhi Wu, Nan Yang, and Mu Li. 2013. Punctuation prediction with transition-based parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 752–760.

Yanqing Zhao, Chaoyue Wang, and Guohong Fu. 2012. A crf sequence labeling approach to chinese punctuation prediction. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 508–514.

俞敬松，０一，张永伟. 2019. 基于bert的古文断句研究与应用. 中文信息学报, 33(11):57–63.

古万荣，董守斌，曾之肇，何锦潮，刘崇. 2016. 基于微博用户模型的个性化新闻推荐. 中文信息学报, 30(1):93–100.

张开旭，夏云庆，宇航. 2009. 基于条件随机场的古汉语自动断句与标点方法. 中文信息学报, 49(10):1733–1736.

王博立，史晓东，苏劲松. 2017. 一种基于循环神经网络的古文断句方法. 中文信息学报, 53(2):255–261.

王志宏，过弋. 2019. 微博谣言事件自动检测研究. 中文信息学报, 33(6):132–139.

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

贺敏，刘玮，刘悦，王丽宏，白硕，程学旗. 2017. 基于特征驱动的微博话题检测方法. 中文信息学报, 31(3):101–107.

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