融合预训练模型的端到端语音命名实体识别

兰天伟
北京理工大学，计算机学院
北京，100081
lantianwei0818@qq.com

郭宇航
北京理工大学，计算机学院
北京，100081
guoyuhang@bit.edu.cn

摘要

语音命名实体识别(Speech Named Entity Recognition, SNER)旨在从音频中识别出语音中命名实体的边界、种类和内容，是口语理解中的重要任务之一。直接从语音中识别出命名实体，即端到端方法是SNER目前的主流方法。但是语音命名实体识别的训练语料较少，端到端模型存在以下问题:(1)在跨领域识别的情况下模型的识别效果会有大幅度的下降。(2)模型在识别过程中会因同音词等现象对命名实体漏标、错标，进一步影响命名实体识别的准确性。针对问题(1)，本文提出使用预训练实体识别模型构建语音实体识别的训练语料。针对问题(2)，本文提出采用预训练语言模型对语音命名实体识别的N-BEST列表重打分，利用预训练模型中的外部知识帮助端到端模型挑选出最好的结果。为了验证模型的领域迁移能力，本文标注了少样本口语型数据集MAGICDATA-NER，在此数据上的实验表明，本文提出的方法相对于传统方法在F1值上有43.29%的提高。

关键词： 语音命名实体识别 ； 融合预训练模型 ； 外部知识 ； 跨领域识别 ； 少样本训练

End-to-End Speech Named Entity Recognition with Pretrained Models

Tianwei Lan
Beijing Institute of Technology
School of Computer Science
Beijing 100081, China
lantianwei0818@qq.com

Yuhang Guo
Beijing Institute of Technology
School of Computer Science
Beijing 100081, China
guoyuhang@bit.edu.cn

Abstract

Speech Named Entity Recognition (SNER) aims to recognize the boundary, type and content of named entities in speech from audio, which is one of the important tasks in spoken language understanding. Recognizing named entities directly from speech, that is, the end-to-end method is the current mainstream method of SNER. However, the training corpus for speech named entity recognition is less, and the end-to-end model has the following problems: (1) The recognition effect of the model will be greatly reduced in the case of cross-domain recognition. (2) During the recognition process, the model may miss or mislabel named entities due to phenomena such as homophones, which further affects the accuracy of named entity recognition. Aiming at problem (1), this paper proposes to use a pre-trained entity recognition model to

第二十二届中国计算语言学大会论文集，第174页-第185页，哈尔滨，中国，2023年8月3日至5日。
(c) 2023 中国中文信息学会计算语言学专业委员会

construct a training corpus for speech entity recognition. For problem (2), this paper proposes to use the pre-trained language model to re-score the N-BEST list of speech named entity recognition, and use the external knowledge in the pre-trained model to help the end-to-end model select the best result. In order to verify the domain migration ability of the model, we labeled the MAGICDATA-NER data set with few samples. The experiment on this data shows that the method proposed in this paper has an improvement of 43.29% in F1 value compared with the traditional method.

Keywords: Speech Named Entity Recognition , Fusion of Pre-trained Models , External Knowledge , Cross-domain Recognition , Few-shot Training

1 引言

命名实体识别(Named Entity Recogination, NER)是自然语言处理中的一项重要任务，传统的命名实体识别旨在将文本中的命名实体信息进行抽取，用特定的类别进行分类，如人名（PER）、机构名（ORG）、地名（LOC）等。任务要求同时确定命名实体的边界以及类别信息。基于文本的命名实体识别研究目前已经有了较为丰富的研究成果(Zhang and Yang, 2018; Huang et al., 2015; Gui et al., 2019a; Gui et al., 2019b)，和比较多的进展(Chiu and Nichols, 2016; Lample et al., 2016; Nadeau and Sekine, 2007)。近年来，语音命名实体识别(Speech Named Entity Recognition, SNER)作为一项口语理解任务得到了越来越多的关注(Caubrière et al., 2020)，在包括屏蔽音频医疗记录中患者姓名(Cohn et al., 2019)等隐私保护领域以及线上直播、视频会议等场景中都有着较大的应用价值。

传统的语音命名实体识别采用级联模型完成(Hatmi et al., 2013)，即首先使用语音识别(Automatic Speech Recogination, ASR)模型对音频识别出对应的文本(Li et al., 2020)，然后使用针对文本的模型在此基础上进行命名实体识别。这类级联模型的识别过程如图1所示。

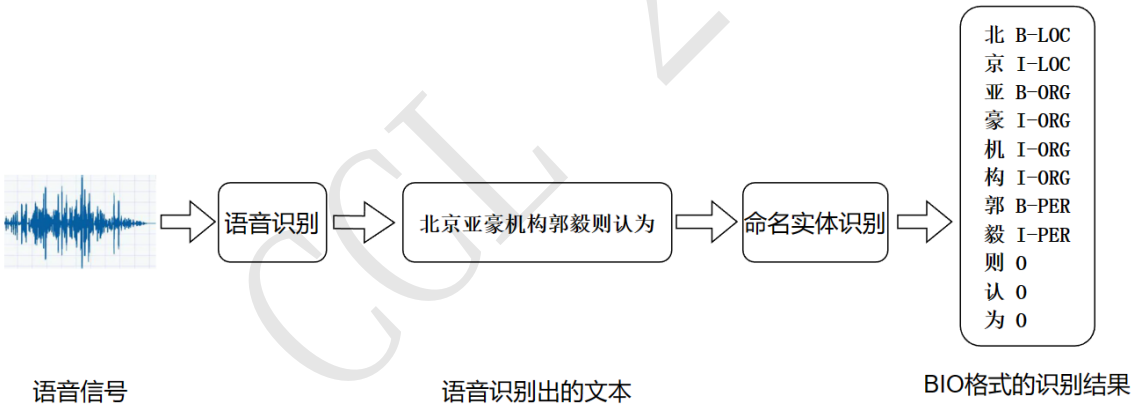

图1.级联模型进行语音命名实体识别的过程

这种方法有一定的缺点。首先，存在误差的级联传播问题(Jannet et al., 2015)。ASR过程中产生的错别字，语义不通顺等问题会增加NER模型识别命名实体的难度，使得这种情况下的准确率相比于在没有错误的文本上进行识别的准确率有所降低。其次，对ASR系统的评价指标一般是字/词错误率，而针对NER系统的评价指标一般是准确率、召回率以及F1值，评价指标的不同使得模型的两个部分无法进行统一的调优，同时也使得模型的训练变得愈发复杂。

已有的一些工作(Ghannay et al., 2018; Yadav et al., 2020; Chen et al., 2022)研究了针对语音命名实体识别的端到端模型,提出了实体感知的语音识别的概念。实体感知的语音识别指的是这些工作普遍将用于识别命名实体的标签直接加入到与音频相对应的文本中来体现对命名实体的标记，例如用方括号[]表示人名，圆括号()表示地点，尖括号⟨⟩表示机构名。在识别模型的训练过程中，用于训练的文本就是已经打好标签的文本，所以模型可以学习音频与打出的标签之间的对齐，直接针对输入的音频输出带有标签的文本，从而完成端到端的语音命名实体识

别。目前这类方法已经被成功应用到了法语、英语以及中文的数据集上(Ghannay et al., 2018; Yadav et al., 2020; Chen et al., 2022)。端到端模型的识别过程如图2所示。

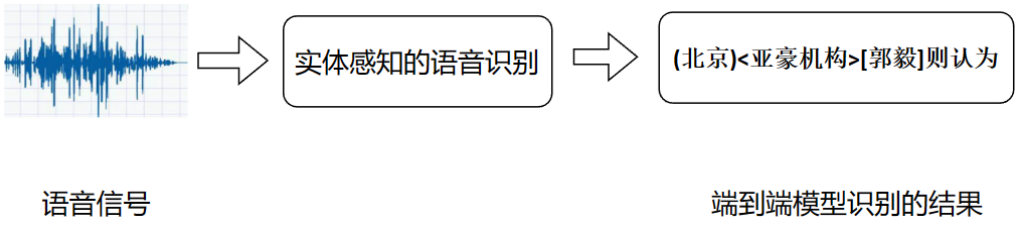

图2.端到端模型进行语音命名实体识别的过程

然而端到端语音命名实体识别仍然面临一些问题。

一方面，语音命名实体识别领域面临标注语料较少的问题。在单一数据集上训练出的语音命名实体识别的模型，往往不能直接用于其他领域数据的识别，在此情况下准确率会有大幅度的下降，模型的跨领域性能较差，而重新标注数据集有着较高的人工成本。在这种情况下，模型的泛化能力是一项重要的评价指标。如果模型在融合实体识别预训练模型中的知识之后或者在一个领域的大规模语料上训练之后，仅在另一领域的少量数据上进行训练就可以产生质量较高的识别结果，从而成为一个通用的语音命名实体识别模型，便可以在很大程度上减少标注成本和训练成本。

另一方面，在端到端场景下，不仅要求正确识别出命名实体的边界以及类别信息，还要正确识别命名实体的内容，所以识别系统输出的错别字以及不准确识别的实体也会导致错误。

表1列举了几种因为错别字以及实体识别原因而导致的错误。在第一个例子中虽然模型同时识别正确了命名实体的边界和类型，但是因为把“荆州”错误的识别成了“金州”，所以依然没有识别正确。第二个例子中错别字直接导致识别出的文本与正确答案有很大的语义差别，没有识别出人名实体。第三个例子中，端到端模型没有识别出和后面的地名实体并列的前一个实体。综合观察这几个错误，本文发现，受限于训练语料规模的不足，使用本地语料训练的端到端模型，在进行实体识别的时候对于一些特定的地理名词以及人名会产生错别字问题，这导致即使在正确识别实体边界和类型的情况下也不能得出正确的答案。

表1.端到端模型的错误识别示例

模型输出结果	正确识别结果
(湖北省)<**金州市安监局**>召开<安良百货>电梯事故情报通报会	<**荆州市安监局**>
出清就是<中国女排>的核心	[**朱婷**]
(崇礼县)发展较成熟的**万龙滑雪场**和(云顶滑雪场)	(**万龙滑雪场**)

(Chen et al., 2022)提出将预训练模型融入到级联模型的识别过程中并取得了结果的提高，但是目前融合了预训练模型的端到端系统并不常见。本文提出一种将预训练模型融入端到端语音命名实体识别模型来提高识别效果的方法。

针对语音命名实体识别训练语料较少、跨领域识别困难的问题，本文提出使用命名实体识别预训练模型帮助构建语音实体识别的训练语料，利用较易获得的语音识别平行语料解决了模型的跨领域问题同时保持了端到端模型的简洁性。针对同音词等现象导致的对命名实体的漏标、错标问题，采用预训练语言模型对语音命名实体识别的N-BEST列表进行重打分，利用预训练模型中的外部知识帮助端到端模型挑选出最好的结果。

为了验证模型的跨领域识别能力，本文标注了少样本数据集MAGICDATA-NER并设计了零样本实验来验证预训练模型对整体模型的泛化能力的提升效果。在AISHELL-NER为训练语料，MAGICDATA-NER为测试语料的场景下，相比于传统方法，本方法的F1值可以提升43.29%。

在AISHELL-NER上的中文实验结果表明，当融合了预训练语言模型BERT和GPT2重排序之后，F1值从74.31%分别提升至了77.27%和79.26%。在DATA2上的英文实验中，F1值从87.0%分别提升至了88.9%和88.7%。

以上实验均证明了本文所提方法的有效性。

本文的贡献总结如下：

(1)用预训练命名实体识别模型构建语音命名实体识别训练语料，解决了语音命名实体识别训练语料较少的问题，利用较易获得的语音识别平行语料解决了模型的跨领域问题。

(2)通过重打分的方式将预训练语言模型融合到语音命名实体识别模型当中，在不同语种的多个语料中取得了比之前更好的结果，并通过各项实验结合实例分析了取得进步的原因。

(3)提出了少样本数据集MAGICDATA-NER，证明了结合预训练模型能够提升模型整体的泛化性能，为今后降低语音命名实体识别模型训练的人工标注成本以及计算成本提供了参考。

2 融合外部知识的语音命名实体识别

2.1 融合声学模型和预训练实体识别模型

在单一数据集上训练出的语音命名实体识别的模型，往往不能直接用于其他领域数据的识别，在此情况下准确率会有大幅度的下降，模型的跨领域性能较差。因此本文提出使用预训练的命名实体识别模型来对较为容易获得的语音识别平行语料的转录文本部分进行实体标记，使之成为可以用于训练语音命名实体识别的伪语料，最终用这份语料来训练语音命名实体识别模型。

在实际训练阶段，先用经过人工标记的语料训练模型，在将模型应用于某个具体的领域时，使用以上方法利用该领域内的语音识别平行语料将预训练实体识别模型中的知识蒸馏到语音命名实体识别模型中，完成领域迁移的工作。如此训练出的模型在应用阶段可以在不借助语言模型的情况下取得较好的识别效果，保持了端到端模型的简洁性同时完成了领域迁移，相关实验结果在将第四部分展示。

2.2 融合声学模型和预训练语言模型

本文采用的方法如图3和图4所示。在实体识别阶段，本文依然采用端到端方法中实体感知语音识别的思想，在输出时直接在命名实体的左右打好标签。所不同的是，这一次模型输出的不是贪婪搜索得到的最好的结果，而是采用束搜索算法得到的N-BEST列表，这个列表是实体识别模型产生的一个候选者列表，通过设定束搜索宽度，可以控制得到的列表长度，即候选句子的个数，同时也可以得到相应的声学模型的打分S_{AM}。接着，本文将获取的预训练语言模型在带括号的文本语料上进行微调，以便让模型学习打标签的相关知识。最后用包含外部知识的语言模型对N-BEST列表中的句子进行重打分，得到语言模型的分数S_{LM}，将两个得分相结合，最终挑选出得分最高的结果输出。通过以上过程，既避免了传统级联模型中误差累积的问题，也改善了端到端模型中因训练语料不足以及无法融合预训练模型而导致的错别字问题。在接下来的部分中，本文详细介绍了模型各阶段所使用的方法。

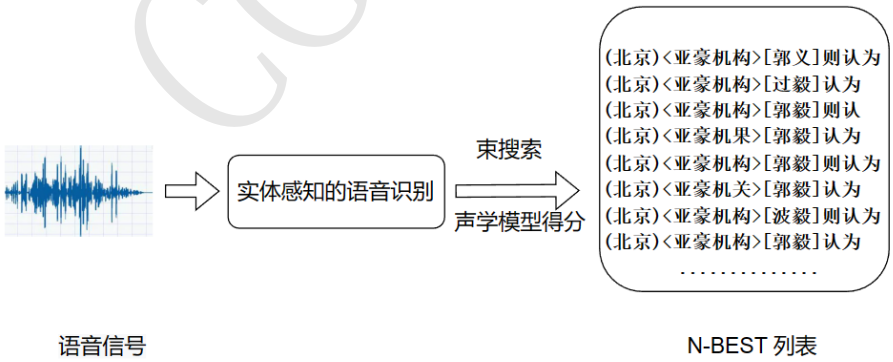

图3.使用束搜索算法得到N-BEST列表

2.3 实体感知的语音识别过程

在语音识别的过程中，本篇文章使用端到端语音命名实体识别所采用的方法，用添加过标签的文本数据对实体识别模型进行训练，另外向词表中添加相应的标签，来让实体识别模型在解码时能够直接得出已经打好标签的文本结果。

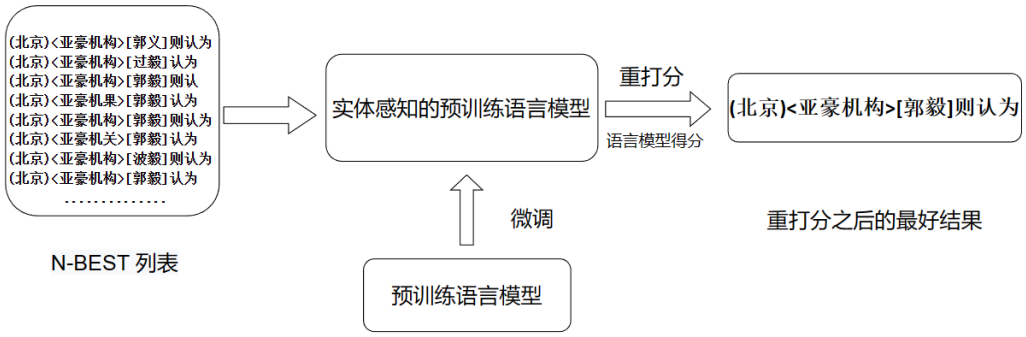

图4.使用预训练模型重打分

同时，为了用预训练的语言模型给语音识别模块输出的N-BEST列表进行重打分，以融合外部知识，语音识别的过程共分为两步。第一步，经过特征提取的音频信息输入到语音识别模型进行编码，每段经过采样的音频会产生一个长度为词表大小的向量表示，这些向量拼接在一起形成了将要进行束搜索的矩阵。可以将这个矩阵看作一个全连接的Lattice图，通过束搜索算法计算两个元素之间的转移概率，就可以按照概率大小解码出进行重打分需要的N-BEST列表。第二步，应用束搜索算法在该矩阵上进行搜索，通过设置束搜索的宽度可以确定最终获得的N-BEST列表的长度，即在重打分的过程中候选者的数目。同时在进行束搜索时，由转移概率可以计算出相应的得分，这个得分就是声学模型部分的打分S_{AM}。

2.4 预训练语言模型重打分

2.4.1 GPT2

本篇文章所使用的预训练模型GPT2由多层Transformer decoder堆叠而成，是一个单项模型。采用GPT2对语音识别过程所获得的N-BEST列表中的每个句子进行打分，以获得相应的语言模型分数S_{LM}。打分时GPT2基于给定的单词来计算下一个单词出现的对数似然概率，将一个句子中每个单词的对数似然概率相加，就得到了一个给定句子的得分，如公式(1)所示。其中W表示给定的句子，W_t表示句子中的第t个字，$W_{<t}$表示句子的前t-1个字，Θ表示用来打分的语言模型的参数。

$$Score_{LM}(\mathbf{W}) = \sum_{t=1}^{|W|} \log P_{LM}(\mathbf{w}_t|\mathbf{W}_{<t};\Theta) \tag{1}$$

2.4.2 BERT

本篇文章所使用的预训练模型BERT由多层Transformer encoder堆叠而成，与GPT2不同的是，BERT是一个双向模型。在给句子进行打分的时候，BERT会同时考虑到一个字左边和右边的字，结合之前和之后的情况进行打分，如公式(2)所示。其中W表示给定的句子，W_t表示句子中的第t个字，$W_{\backslash t}$表示句子中除第t个字以外的其他字，Θ表示用来打分的语言模型的参数。

$$Score_{\mathrm{LM}}(\mathbf{W}) = \sum_{t=1}^{|\mathbf{W}|} \log P_{\mathrm{LM}}(\mathbf{w}_t|\mathbf{W}_{\backslash t};\Theta) \tag{2}$$

2.4.3 重打分

公式(1)和(2)的计算方式会使得长度较长的句子获得较低的得分，因为每一个对数似然概率都是一个负数，句子长度较长意味着有更多的负数相加并获得一个较低的得分。而在语音识别的过程中有可能更长的句子中包含了更多的正确信息，所以在实际应用过程中，本文首先获取语言模型对整句的打分(Salazar et al., 2019)，然后用整句的分数除以每个句子的长度以得到每个单词的平均对数似然概率，来当作语言模型部分句子的得分，如公式(3)所示。其中T表示给定的句子的长度。

$$S_{LM}(\mathbf{W}) = \frac{1}{T} Score_{LM}(\mathbf{W}) \tag{3}$$

同时，由于预训练模型本身的训练文本中不包含用括号标记处的命名实体，未经过微调的预训练模型在进行打分的时候可能会在计算括号标签的转移概率的时候输出一个过低的数值。所以为了取得更好的打分效果，还需要在域内数据上对预训练模型进行微调。

2.5 融合声学得分和语言模型得分

N-BEST列表中每一个句子最终得分的计算方法如公式(4)所示。

$$\mathrm{Score}(\mathbf{W}) = (1-\lambda) \cdot \mathrm{S_{AM}}(\mathbf{W}) + \lambda \cdot \mathrm{S_{LM}}(\mathbf{W}) \tag{4}$$

采用线性插值的方法将两个部分的得分融合在一起，λ 是本文定义的一个超参数，用来调节两个模型之间的权重比例。最终本文挑选N-BEST列表中得分最高的句子作为整个语音命名实体识别系统的输出。

如此，因为在解码阶段同时考虑了声学模型部分和语言模型部分的打分，本文在原有模型的基础上融合了预训练模型中的外部知识，这些外部知识可以帮助模型在解码时减少因为错别字而造成的错误，用以获得更加合理的输出结果，这样的效果在中文命名实体识别中应当较为明显，因为中文中有很多同音字。本文预计，通过结合预训练模型，可以使输出句子中的错别字更少，更加通畅，进而提高结果的精确率（Precision）。而通过对预训练模型进行微调，可以使结果中做出预测的数量变得更多，进而提高结果的召回率（Recall），在第四部分的实验中，本文的猜想也得到了进一步的证实。

3 MAGICDATA-NER数据集

为了检验本文的方法在零样本以及少样本情况下的表现，即预训练语言模型对模型整体的跨领域识别能力的提升效果，本文标注了数据集MAGICDATA-NER。MAGICDATA-NER基于开源的语音识别数据集MAGIADATA，该数据集包含755个小时的语音数据，由1080名来自中国大陆的说话人用普通话朗读。与AISHELL-NER中包含财经、科技、体育、娱乐和新闻的较为正式的语音数据不同，该数据集的语料取自日常生活对话。本文在MAGICDATA中抽取部分语音以及对应的文本用以标注本文的跨领域少样本数据集MAGICDATA-NER。MAGICDATA-NER的语音以及原始的转录文本部分抽取自MAGICDATA的验证集，共包含11793条，约14个小时的语音数据，这部分数据将会在论文录用后公开。通过比较在零样本以及少样本训练情况下模型融合大规模预训练语言模型前后的命名实体识别效果，来检验预训练语言模型对整体的跨领域识别能力是否有提升效果。

3.1 标注过程

在标注过程中本文采用了和以往工作(Chen et al., 2022)类似的标注方法，标注了三个种类的命名实体，即人名(PER)、地名(LOC)、机构名(ORG)。首先本文获取了一个用于在中文文本上进行命名实体识别的预训练模型，该模型取自Hugging Face(Wolf et al., 2020)。然后把该模型在MSRA数据集上进行微调，并在测试集上取得了91.59%的F1值。最后本文用微调过后的模型标注从MAGICDATA中抽取的转录文本，以此作为人工标注的起点。

在人工标注阶段，将数据集中的语句分为2000条一组，针对每一组数据，两个学习了命名实体标注规则的研究人员会分别对机器标注的数据进行检查并对机器标注错误的实体进行修改以及标记。之后对比两组经过人工查验的数据，当两组数据的一致性达到95%以上时才接受其为语料数据，并对其中不一致的地方进行进一步探讨。当一致性低于95%时，更换研究人员进行重新标记。最终形成本文标注的跨领域少样本数据集MAGICDATA-NER。

3.2 语料数据

表2中展示了MAGICDATA-NER中各个集合中命名实体的情况。在将数据集分割成训练集、开发集、测试集时，为在进行验证的时候避免因集合过小而造成的偶然性因素，本文将验证集和测试集的大小控制在2000条语料以上。同时作为少样本语料，本数据集的训练集相对较小，用以检验模型在少样本以及零样本情况下的命名实体识别能力。各集合的实体类型分布基本一致，保证了数据分布的一致性。

表2.MAGICDATA-NER各个集合中命名实体的情况

数据集	句子数量	人名实体数量	地名实体数量	机构实体数量	实体总数量
MAGICDATA-NER	11793	2036	770	242	3048
训练集	6538	1135	391	146	1672
验证集	2310	365	155	55	575
测试集	2945	536	224	41	801

4 实验

本文分别进行了在AISHELL-NER上的中文实验，在DATA2上的英文实验，和在MAGICDATA-NER上的跨领域实验，以下是对实验过程以及结果的详细说明。

4.1 模型架构

中文实验的实体识别模块使用的是ESPNET中的Conformer模型，模型的基本设置与原文(Chen et al., 2022)相同,在AISHELL-NER数据集上进行了实验。在MAGICDATA-NER数据集上的跨领域以及少样本实验同样使用此模型，只在训练过程中需要在相应的数据集上进行后续的训练和验证。为了验证本篇文章所提出的方法对于不同数据集以及不同语种的有效性，本文用这篇文章(Yadav et al., 2020)使用的DeepSpeech2框架，在其提出的英文数据集DATA2上同样做出了验证实验。

实验中所用到的预训练模型均获取自Huggingface Transformers(Wolf et al., 2020)。中文实验使用的GPT2语言模型是uer/gpt2-chinese-cluecorpussmall(Radford et al., 2019; Zhao et al., 2019)，使用的BERT语言模型是bert-base-chinese(Devlin et al., 2018)。少样本实验使用的预训练模型与中文实验相同，只在微调语料上不同。英文实验使用的GPT2语言模型是gpt2(Radford et al., 2019),使用的BERT语言模型是bert-base-uncased(Devlin et al., 2018)。各模型均在对应的训练语料上进行5个epoch的微调，然后用于重打分。预训练实体识别模型使用的是ckiplab/bert-base-chinese-ner(Ckiplab, 2020)。

4.2 数据集

中文实验使用的AISHELL-NER(Chen et al., 2022)数据集包含超过170小时的普通话语音数据，语料库涵盖五个领域:“财经”、“科技”、“体育”、“娱乐”和“新闻”。英文实验使用的DATA2(Yadav et al., 2020)数据集包含约150个小时的英文语音数据，共39769条语料。这些语料是在英语数据集Librispeech(Panayotov et al., 2015)、CommonVoice(WikipediaContributors, 2020)、Tedlium(Rousseau et al., 2012)和Voxforge(WikipediaContributors, 2019)的子集基础上进行标注的。少样本、零样本以及跨领域实验使用的是本篇文章所提出的数据集MAGICDATA-NER，共包含11793条，约14个小时的语音数据，内容为日常生活对话。

4.3 实验结果以及分析

4.3.1 预训练语言模型对于识别结果的影响

本实验采用conformer结构的中文语音命名实体识别作为基线模型，并对比了使用不同语言模型进行重打分对于识别结果的影响。其中transformer模型从头在域内数据上进行了15个epoch的训练，而GPT2和BERT则在域内数据上进行了5个epoch的微调。在进行实验时，为了确定语言模型的最佳权重λ，首先在验证集上进行实验，观察在不同权重条件下验证集上的F1值，如图5所示。

在获得结果后将在验证集上取得最好结果的权重值应用于测试集以获得最终的实验结果，如表3所示。Conformer+BERT-PT表示在先前工作(Chen et al., 2022)中级联模型取得的最好结果。

根据实验结果本文可以看出，通过结合Transformer模型，命名实体识别的F1值增长至75.32%，使用BERT让结果进一步提升至了77.27%，而结合微调过后的GPT2模型之后实验结果有了大幅度的增长来到了79.26%，并且在精确率以及召回率上都有较大幅度的提升。由此本文可以得出，结合语言模型对N-BEST列表进行重打分的策略对提升命名实体识别的准确

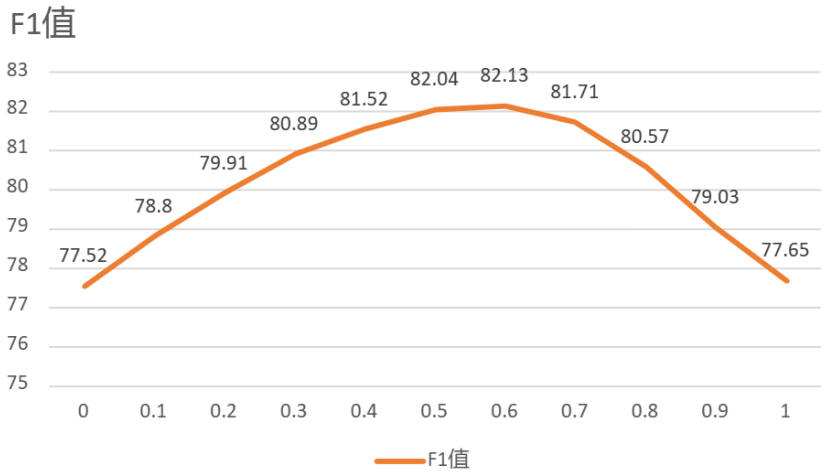

图5.验证集的F1值随λ的变化情况(以Conformer + GPT2_lm为例)

表3.增加预训练语言模型对于识别结果的影响

模型	精确率	召回率	F1值
Conformer(baseline)	77.56	71.32	74.31
Conformer+BERT-PT(Chen et al., 2022)	75.54	74.27	74.90
Conformer + Transformer_lm	78.37	72.50	75.32
Conformer + BERT_lm	80.21	74.54	77.27
Conformer + GPT2_lm	**82.55**	**76.13**	**79.26**

性有正向效果，而不同语言模型对于F1值的提升效果有所不同。Transformer模型只在域内的数据上进行了训练，碍于数据规模的大小，模型的能力要落后于先预训练后进行微调的预训练模型，这体现出了在预训练阶段模型中融合的大规模外部知识对于整个系统识别的指导作用，也验证了本篇文章所提出的方法的有效性。

4.3.2 选取不同候选列表长度对于识别结果的影响

在使用束搜索算法以及语言模型进行解码的过程中，可以通过设定束搜索宽度来决定候选名单的长度。候选的句子是根据声学模型打分从高到低来排列的，设定的候选列表长度越大，进行重打分的时候可以参考的句子数量就越多，下面的实验验证了在融合GPT2模型的情况下候选列表长度与最终结果之间的关系。实验结果如表4所示。

表4.选取不同候选列表长度对于识别结果的影响

候选列表长度	精确率	召回率	F1值
1	77.56	71.32	74.31
5	80.57	74.54	77.43
10	81.35	75.48	78.31
20	81.96	75.89	78.79
30	82.45	75.89	79.04
40	82.64	76.04	79.20
50	**82.55**	**76.13**	**79.26**

从实验结果可以看出，当候选列表长度越大，也就是候选的句子数目越多的时候，结果的得分就会越高，这是因为考虑的范围越广泛，挑选出最佳句子的机率就越大。但同时也要注意到的是，由于候选的句子是根据声学模型打分从高到低来排列的，越往后的句子的参考价值也会越小。从实验结果可以看出，增加候选句子的数量对于实验结果的增益效果是越来越小的，并在候选列表长度为50的时候基本收敛，此时模型的识别效果达到最优。

4.3.3 识别结果实例分析

表5.识别结果实例分析

模型	识别结果
Conformer	(湖北省)<**金州市安监局**>召开<安良百货>电梯事故情报通报会 **出清**就是<中国女排>的核心 (崇礼县)发展较成熟的**万龙滑雪场**和(云顶滑雪场)
Conformer + GPT2_lm	(湖北省)<**荆州市安监局**>召开<安良百货>电梯事故情报通报会 [**朱婷**]就是<中国女排>的核心 (崇礼县)发展较成熟的**(万龙滑雪场)**和(云顶滑雪场)

在本部分，本文挑选了几个比较典型的例子来说明融合GPT2语言模型之后，外部知识对于命名实体识别的提升作用，如表5所示。第一个例子体现的是预训练模型中的外部知识对于同音字、近音字的区别作用。因为“金”和“荆”的发音相近，所以语音识别的过程中可能会因错别字而导致实体识别错误，而预训练模型因为有地名相关知识，即荆州才是湖北省下辖的一个市从而做出正确的选择。第二个例子中，因为在语音识别的结果中有噪声所以基线模型未能正确识别出人名“朱婷”。而“朱婷”是一个较为知名的公众人物的名字，预训练模型的大规模训练语料中会包括相关内容，所以在融合预训练模型之后模型能够识别出“朱婷”是一个人名。人名、地名的相关知识在语音识别中是一个比较难以解决的问题，单纯依靠训练本地模型因为受到数据规模大小的限制，难以完善的结合相关知识，而前两个例子证明通过结合预训练模型可以较好地改善这一问题。第三个例子中，基线模型漏掉的地名实体“（万龙滑雪场）”可以在融合GPT2之后正确的识别出，由此可以看出预训练模型提高了系统对于上下文的感知能力，意识到了万龙滑雪场是和云顶滑雪场并列的一个地名实体。

4.3.4 英文数据集实验

为了验证本篇文章所提出的方法对于不同数据集以及不同语种的有效性，使用DeepSpeech2框架，本文在英文数据集DATA2(Yadav et al., 2020)上同样做出了验证实验，实验结果如表6所示。其中DeepSpeech2 + NER tagger表示先使用DeepSpeech2进行语音识别，再进行命名实体识别的级联模型的结果，而其余三项实验中使用DeepSpeech2做端到端的语音命名实体识别。

表6.英文数据集DATA2实验结果

模型	精确率	召回率	F1值
DeepSpeech2+NER tagger(Yadav et al., 2020)	80.0	59.0	63.0
DeepSpeech2(Yadav et al., 2020)	96.0	80.0	87.0
DeepSpeech2 + GPT2_lm	**99.5**	80.0	88.7
DeepSpeech2 + BERT_lm	99.1	**80.6**	**88.9**

从实验结果可以看出，通过融合英文预训练模型，识别的准确性也获得了提高，这印证了在不同语种的不同数据集上，本文的方法都对端到端语音命名实体识别有提升效果。

4.3.5 零样本及少样本实验

在零样本实验中，模型在AISHELL-NER数据集上进行训练然后直接用于识别少样本数据集MAGICDATA-NER测试集中的语音。在少样本实验中，模型首先在AISHELL-NER数据集上进行训练而后在MAGICDATA-NER的训练集上进行小规模后续训练，最后用于识别少样本数据集MAGICDATA-NER测试集中的语音。应用于两个实验的语言模型均在MAGICDATA-NER训练集的文本上进行了微调。以下为实验结果。

从实验结果可以看出，在两种情况下融合预训练模型后识别结果均有明显提升。零样本实验中，识别结果的F1值从24.79%提升至了29.96%，少样本实验中，识别结果的F1值从51.34%提升至了60.98%。这说明在缺少训练样本乃至没有对应领域训练样本的情况下，预训练模型很好的帮助了实体识别过程，提升了模型的泛化能力。相比于中英文实验，预训练模型在零样本以及少样本实验中对于实验结果的提升更为明显，这体现了预训练模型在训练样本较

表7.零样本实验结果

模型	精确率	召回率	F1值
Conformer(baseline)	36.67	18.73	24.79
Conformer + GPT2_lm	45.06	22.22	29.77
Conformer + BERT_lm	**43.96**	**22.72**	**29.96**

表8.少样本实验结果

模型	精确率	召回率	F1值
Conformer(baseline)	53.84	49.06	51.34
Conformer + GPT2_lm	62.25	53.93	57.79
Conformer + BERT_lm	**62.75**	**59.30**	**60.98**

少情景中对于提升模型表现的显著作用。这一结果进一步说明使用本文提出的方法融合预训练模型，通过先大规模语料训练后少样本语料微调的方法，可以大幅提升模型在新领域的识别表现，进而帮助减少标注成本和训练成本，降低语音命名实体识别模型在实际应用中的开发难度。

4.3.6 融合预训练实体识别模型的跨领域实验

在本实验中，模型首先在AISHELL-NER上训练至收敛，然后对MAGICDATA的测试集进行识别，取得未经过跨领域训练的实验结果。而后使用经过预训练实体识别模型标记的MAGICDATA的训练集进行下一步训练，同样对MAGICDATA的测试集进行识别，取得跨领域训练后的实验结果，用以检验预训练模型对跨领域的实体识别是否有帮助，实验结果如下。其中Conformer(baseline)表示只在AISHELL-NER上训练未融合预训练实体识别的模型，GPT2_LM和BERT_LM表示使用了语言模型重打分的方法，fusion表示使用了融合预训练实体识别模型的方法。

表9.融合预训练实体识别模型的跨领域实验

模型	精确率	召回率	F1值
Conformer(baseline)	41.74	22.80	29.49
Conformer+GPT2_LM	49.97	27.03	35.09
Conformer+BERT_LM	48.85	27.79	35.42
Conformer+fusion	75.91	62.76	68.71
Conformer+fusion+GPT2_LM	79.12	64.63	71.14
Conformer+fusion+BERT_LM	**78.84**	**67.59**	**72.78**

本文采用消融实验的方式验证了语言模型重打分和构造训练语料两种方法对于提升识别效果的作用，从实验结果可以看出，通过重打分，F1值从29.49%提升至了35.42%，通过融合预训练实体识别模型，在跨领域的情况下语音实体识别的准确率有了大幅的提升，F1值提升至了68.71%，并基本接近了在AISHELL-NER上使用人工标注的数据进行训练的识别准确率。在重打分的同时结合预训练语言模型的情况下，模型识别的F1值还可以进一步提升至72.78%。这说明在跨领域的情况下，两种方法对于提升识别效果都有正向作用，为尽可能提高识别的准确率可以使模型同时结合预训练语言模型和预训练实体识别模型，以达到最好的识别效果。

5 结论

针对端到端语音命名实体识别模型训练语料较少，在跨领域识别情况下效果大幅下降以及识别过程中的错标、漏标问题，本篇文章提出了通过融合预训练实体识别模型构建训练语料并用预训练语言模型重打分的方法，提升了端到端模型的识别效果，弥补了以往端到端方法未融合预训练模型的不足。为了验证模型的跨领域识别能力，本文标注了少样本数据集MAGICDATA-NER并设计实验来验证预训练模型对整体模型的泛化能力的提升效果。实验结果表明，使用本文的方法融合预训练模型，在跨领域的情况下本文的方法对提升命名实体识别效果有显著作用，同时在中英文数据集上语音命名实体识别的F1值都获得了提升。本篇文章针对实验结果进行了分析，并且验证了所提出方法在不同语种、不同数据集上的有效性。本文提出的方法为今后语音命名实体识别模型在跨领域情况下的训练以及在训练过程中降低人工标注以及计算成本提供了参考，同时新的数据集也可以为其他口语理解以及语音识别任务的跨领域实验提供帮助。在未来的工作中，可以通过在更多的数据集上进行实验以验证本方法对于效果的提升作用，同时也可以探索在融合预训练模型之后提高模型计算效率的方法。

致谢

本研究受科技创新2030-“新一代人工智能”重大项目(2020AAA0106600)资助。

参考文献

Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2020. Where are we in named entity recognition from speech? In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4514–4520, Marseille, France, May. European Language Resources Association.

Boli Chen, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Meishan Zhang, and Fei Huang. 2022. Aishellner: Named entity recognition from chinese speech. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8352–8356.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Ckiplab. 2020. Ckip bert base chinese. Website. https://github.com/ckiplab/ckip-transformers.

Ido Cohn, Itay Laish, Genady Beryozkin, Gang Li, Izhak Shafran, Idan Szpektor, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. 2019. Audio de-identification: A new entity recognition task. *arXiv preprint arXiv:1903.07037*.

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.

S. Ghannay, A. Caubrière, Y. Estève, N. Camelin, E. Simonnet, A. Laurent, and E. Morin. 2018. End-to-end named entity and semantic concept extraction from speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019a. Cnn-based chinese ner with lexicon rethinking. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4982–4988. International Joint Conferences on Artificial Intelligence Organization, 7.

Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuanjing Huang. 2019b. A lexicon-based graph neural network for Chinese NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1040–1050, Hong Kong, China, November. Association for Computational Linguistics.

Mohamed Hatmi, Christine Jacquin, Emmanuel Morin, and Sylvain Meignier. 2013. Named entity recognition in speech transcripts following an extended taxonomy. In *SLAM@INTERSPEECH*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv e-prints*, page arXiv:1508.01991, August.

Mohamed Ameur Ben Jannet, Olivier Galibert, Martine Adda-Decker, and Sophie Rosset. 2015. How to evaluate asr output for named entity recognition? In *Interspeech*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June. Association for Computational Linguistics.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842, Online, July. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30:3–26.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659*.

WikipediaContributors. 2019. Voxforge wikipedia, the free encyclo- pedia. Website. https://en.wikipedia.org/w/index.php?title=VoxForge&oldid=913093799.

WikipediaContributors. 2020. Common voice wikipedia, the free encyclopedia. Website. https://en.wikipedia.org/w/index.php?title=CommonVoice&oldid=939008593.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. *arXiv preprint arXiv:2005.11184*.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564, Melbourne, Australia, July. Association for Computational Linguistics.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.