融合领域词汇扩充的低资源法律文书命名实体识别

帕尔哈提·吐拉江[ab], 孙媛媛*[a], 蔡艾辰[a], 王艳华[c], 林鸿飞[a]
[a]大连理工大学，计算机科学与技术学院，大连，116024
[b]新疆师范大学，计算机科学技术学院，乌鲁木齐，830054
[c]中国人民解放军空军通信士官学校，大连，710038
(prht,18941991166)@mail.dlut.edu.cn,(syuan,hflin)@dlut.edu.cn
wangyanhua818@foxmail.com

摘要

目前基于预训练语言模型的司法领域低资源法律文书命名实体识别研究主要面临两个问题：(1)在低资源语言中，如维吾尔语，法律文书相关的语料极其有限，这种语料资源稀缺限制了基于预训练语言模型的训练和性能。(2)法律文书中使用的专业术语不仅复杂且特定，新的法律术语和概念的出现使得现有的模型难以适应。针对上述问题，本文基于多语言预训练模型mBERT，通过领域词汇扩充及模型微调的方法，提升了模型在维吾尔语法律文书命名实体识别任务的性能。本文首先整理并构建维吾尔语司法领域专业词汇列表，并将其添加到mBERT模型的词汇表中。随后，在人工标注的维吾尔语法律文书命名实体数据集UgLaw-NERD上进行模型微调，验证了该方法的有效性。实验结果表明，相比于仅使用mBERT进行微调的基线模型，融合领域词汇扩充的模型在命名实体识别任务上F1得分提升至89.72%，较基线提高了7.39%。此外，本文还探讨了不同领域词汇扩充量对模型命名实体识别性能的影响，结果显示，领域词汇扩充增强了预训练模型在处理维吾尔语任务中的表现。这些结论为其他低资源语言在司法领域开展基于预训练模型的自然语言处理研究提供了有益的参考。

关键词： 司法领域； 词汇扩充； 预训练语言模型； 低资源； 维吾尔语； 命名实体识别

Named Entity Recognition for Low-Resource Legal Documents Using Integrated Domain Vocabulary Expansion

PAERHATI Tulajiang[ab],Yuanyuan Sun* [a], Aichen Cai[a], Yanhua Wang[c], Hongfei Lin[a]
[a]School of Computer Science and Technology,Dalian University of Technology,Dalian,116024
[b]College of Computer Science and Technology, Xinjiang Normal University, Urumqi,830054
(prht,18941991166)@mail.dlut.edu.cn,(syuan,hflin)@dlut.edu.cn
wangyanhua818@foxmail.com

Abstract

Current research in named entity recognition for low-resource legal documents, based on pre-trained language models, primarily faces two challenges: (1) In low-resource languages, such as Uyghur, relevant legal document corpora are extremely scarce, limiting the training and performance of models based on pre-trained language models. (2) The use of complex and specific legal terminology in legal documents, along with the emergence of new legal terms and concepts, makes it difficult for existing models to adapt. In response to these challenges, this paper leverages the multilingual pre-trained model mBERT and enhances its performance on the Uyghur legal document named entity recognition task through domain vocabulary expansion and model fine-tuning.

* 通讯作者

第二十三届中国计算语言学大会论文集，第431页-第441页，太原，中国，2024年7月25日至28日。
卷1：主会论文

This study begins by compiling and constructing a list of Uyghur legal domain-specific vocabulary and integrating it into the mBERT model's vocabulary. Following this, the model is fine-tuned on the manually annotated Uyghur legal document named entity dataset, UgLaw-NERD, validating the effectiveness of this approach. Experimental results indicate that the domain vocabulary-expanded model achieves an F1 score of 89.72%, an increase of 7.39% over the baseline model that only utilizes fine-tuning with mBERT. Additionally, this paper explores the impact of varying amounts of domain vocabulary expansion on model performance in named entity recognition, revealing that domain vocabulary expansion enhances the pre-trained model's capability in processing Uyghur language tasks. These findings offer valuable insights for other low-resource languages embarking on natural language processing research in the legal field based on pre-trained models.

Keywords: Legal Domain , Vocabulary Expansion , Pre-trained Language Models , Low-resource , Uyghur , Named Entity Recognition

1 引言

在司法领域，命名实体指法律文书中具有特定司法属性的名词和短语(李春楠et al., 2021; Chen et al., 2020)，主要包括涉案时间、地点、相关人员(如犯罪嫌疑人、被害人等)、组织机构、罪名以及涉案金额等。针对司法命名实体的识别研究不仅是法律文书案件情节关系抽取、构建司法知识图谱的基础工作，而且对法律问答、刑期预测、类案推荐等司法应用的实际落地具有至关重要的作用。

基于统计的机器学习方法在司法领域命名实体识别研究中，Leitner等人(Leitner et al., 2019)手工设计了多组特征，并将这些特征与CRF模型结合使用，以完成法律文书的命名实体识别。虽然CRF模型能根据词汇之间的依赖关系及其与标签的关系来建模概率分布，但在处理法律文书时受到上下文考虑受限的限制，通常只能处理局部上下文而难以捕捉长距离依赖关系。并且模型性能高度依赖于手工设计的特征工程，难以做到全面和准确。随着深度学习的发展，很多研究(Xu et al., 2021; Bonifacio et al., 2020)已经将BERT等预训练语言模型（PLM）应用到了司法领域，极大地提高了法律文本的特征编码质量，以及对法律文本的理解和信息抽取能力。然而，目前基于预训练模型针对司法领域命名实体识别任务的探索主要集中在资源丰富的语言上，对维吾尔语等低资源语言的研究还处于初级阶段，公开的多语言预训练语言模型不能很好地应用于低资源语言命名实体识别任务上。

命名实体识别技术在不同语言环境中的效果各不相同，尤其是在资源匮乏的低资源语言和语言结构复杂的非主流语言中面临显著挑战。以维吾尔语为例，早期的维吾尔语命名实体识别始于基于词典和规则(Collins and Singer, 1999)、基于统计机器学习(Isozaki and Kazawa, 2002)的方法的探索，主要应用基于规则和条件随机场(CRF)(Lafferty et al., 2001)方法对人名、地名、组织机构等实体类型进行识别。随着深度学习技术的快速进步，研究和应用逐渐转向基于深度学习的端到端方法。吾守尔等人(吾守尔et al., 2018)基于BiLSTM-CNN-CRF的神经网络模型，在资源稀缺环境下提高了维吾尔语命名实体识别的性能。AzmatAnwar等人(Anwar et al., 2020)应对在资源匮乏的维吾尔语上进行神经网络命名实体识别的挑战，提出了基于神经机器翻译的命名实体标注转移方法。虽然基于深度神经网络的方法(王路路et al., 2019)在维吾尔语命名实体识别任务上达到了90%以上的F1分数，但这些任务主要针对通用领域的命名实体类型（如人名、地名和组织机构）的识别。相对于通用领域，司法领域命名实体识别任务所关注的实体类型粒度小、分类多，包含法律专业术语，再加上维吾尔语等低资源语言司法领域语料的稀缺性和语法规则的复杂性，给低资源语言法律文书的命名实体识别研究带来了额外的挑战。

对低资源语言，哈工大讯飞联合实验室Yang等人(Yang et al., 2022)以多语言预训练模型XLM-R为基座，训练了第一个面向中国少数民族语言的预训练语言模型CINO，并在文本分类任务上进行微调，提高了模型文本分类任务上的能力。Deng等人(Deng et al., 2023)提出的预

训练模型MiLMo在CINO-base-v2模型的基础上，进一步提升了模型性能。虽然多语言预训语言练模型在低资源语言分类任务上展现出了优异的性能，但以上研究都是面向通用领域的分类任务，并没有对特定领域的命名实体识别任务开展相关研究。

从上述研究工作来看，基于预训练语言模型的面向司法领域低资源语言法律文书命名实体识别研究主要面临两个问题：(1)在低资源语言中，如维吾尔语，法律文书相关的语料极其有限，这种语料资源稀缺限制了基于预训练语言模型的训练和性能。(2)法律文书中使用的术语不仅复杂且特定。法律法条的新增，修订、修正和修改，使得之前针对法律文本训练的预训练模型因缺少新增法律术语和概念的专业领域词汇，可能会造成下游任务的性能下降。

鉴于上述问题，本文提出了一种融合领域词汇扩充和预训练模型微调的方法，实现低资源语言法律文书的命名实体识别。本文采用了多语言预训练模型mBERT(Pires et al., 2019)作为基线，融合词汇扩充和模型微调的策略，提高维吾尔语法律文书中命名实体识别的性能。具体地，首先从中国裁判文书网[1]收集维吾尔语法律文书，经过数据处理，构建专门针对司法领域的维吾尔语词汇表。然后对预训练模型的词汇表进行扩充，并对模型的嵌入层进行调整。最后，通过数据增强和人工标注构建的维吾尔语法律文书命名实体数据集UgLaw-NERD对模型进行微调，实现对维吾尔语法律文书命名实体的识别。本文主要贡献总结如下:

（1）本文构建了专门针对司法领域的维吾尔语法律文书命名实体识别数据集UgLaw-NERD[2]，该数据集包含9个细粒度司法实体类型、3269句子和96745个token。

（2）提出了一种融合预训练模型的词汇扩充策略。通过结合词汇扩充与模型微调，该方法提高了预训练模型在维吾尔语法律文书命名实体识别的表现。

（3）本文在UgLaw-NERD数据集上进行实验，与现有方法相比，本文提出的方法展现了性能的提升，也证实了UgLaw-NERD数据集在维吾尔语法律文书命名实体识别领域的应用价值。

本文其余部分组织如下：第2章主要介绍低资源语言命名实体识别相关研究。第3章介绍数据集的开发过程和注释指南。第4章介绍本文提出的方法。第5章主要介绍实验设置、实验结果及参数分析。第6章为总结与未来研究方向的探讨。

2 相关工作

现代维吾尔语中有32个字母，每个单词都由字母拼写而成。维吾尔语单词的词形结构复杂，词尾通过接不同的词缀来实现其语法功能。针对维吾尔语的命名实体识别的研究中，木合塔尔等人(木合塔尔·艾尔肯et al., 2013)通过手动建立基于新疆维吾尔自治区的地名词典库等，提出了基于规则的维吾尔语地名识别技术。艾斯卡尔等人(艾斯卡尔·肉孜et al., 2013)实现了基于条件随机场的维吾尔人名识别方法。加日拉等人(加日拉·买买提热衣木et al., 2014)采用统计和规则相结合的混合策略，提出了一种维吾尔人名的自动识别方法。麦合甫热提等人(麦合甫热提et al., 2014)则从维吾尔语的语言特点出发，对维吾尔语中机构名称进行分类和形式化表示，并设计了基于状态转移原理的维吾尔文机构名识别算法。买合木提·买买提等人(买合木提·买买提et al., 2019)通过分析维吾尔文地名，提出了一种结合条件随机场和规则的识别方法。

随着深度学习技术的快速进步，研究和应用已逐渐转向基于深度学习的端到端方法。利用卷积神经网络(CNN)(Chen, 2015)和循环神经网络(RNN)(Zaremba et al., 2014)等结构，从文本序列中抽取隐含特征，再通过应用条件随机场(CRF)来确定最优实体序列。该方法直接对原始数据进行深度学习建模和特征提取，不仅避免了传统方法中依赖复杂的人工特征工程，还有效解决了数据稀疏性导致的维度灾难问题。王路路等人(王路路et al., 2018)针对维吾尔文命名实体识别无法利用未标注预料的问题，提出了基于CRF和半监督学习的方法，并进一步结合注意力机制与BiLSTM-CRF(王路路et al., 2019)，提升了维吾尔语命名实体识别识别率。

在构建维吾尔语命名实体标注语料方面，艾斯卡尔等人(艾斯卡尔·肉孜et al., 2013)根据维吾尔人名特点收集了5258个句子，形成了维吾尔人名语料库。汪昆等人(汪昆et al., 2017)在人名、地名、机构名的一体化识别任务中构建了共有11257个标注句子的维吾尔语命名实体识别语料。Kahaerjiang等人(Abiderexiti et al., 2017)构建了第一个命名实体关系识别UyNeRel数据集。孙祥鹏等人(孔祥鹏et al., 2020)研究基于迁移学习的维吾尔语命名实体识别方法中构建了基于新闻预料标注的维吾尔语命名实体识别数据集。此外，除了通用领域的人名、地名、机构

[1]https://wenshu.court.gov.cn/
[2]https://github.com/PaErHaTi-DUTiR/UgLaw-NERD

名等实体类型之外，阿迪来等人(阿迪来et al., 2017)在对维吾尔语音乐实体识别研究中，构建了2400句含有音乐实体的数据集。维吾尔语作为低资源语言，只有少量公开的语料库可获取。

词汇扩充是指在现有预训练语言模型的基础上，增加特定领域或语言的词汇，以解决模型原有词汇表不能充分覆盖特定用途语言元素的问题。这一过程涉及将新词汇和表达方式整合到模型的词汇表中，从而扩展模型的语言理解和生成能力。通过词汇扩充，模型能够更好地处理领域特定的术语和表达，增强其在特定任务和环境中的应用性能。Chung等人(Chung et al., 2020)探讨了如何通过扩大词汇表来优化预训练的多语言模型，避免了使用大词汇表重新训练深度模型的需要。Tai等人(Tai et al., 2020)引入了一种通过在有限的训练资源下添加新的附加词汇来扩展BERT预训练模型的方法。Tanaka等人(Tanaka and Shinnou, 2022)提出为预训练的BERT模型添加词嵌入来扩展复合词词汇表的方法。

本文主要为解决在低资源环境中维吾尔语特定司法领域专业术语难以被预训练语言模型有效识别的问题，引入了领域词汇扩充的策略，采用多语言预训练语言模型在司法领域数据集上进行微调，在深层次上捕捉维吾尔语法律文本的特定语义和结构特点，优化模型的命名实体识别能力。本文融合领域词汇扩充和预训练模型微调的方法在低资源语言法律文书的命名实体识别研究，能够在一定程度上填补对这一领域的研究空白和资源建设的需求。同时，为其他低资源语言在司法领域的自然语言处理研究提供有益借鉴和参考。

3 数据集构建

3.1 数据来源

本文数据来源于中国裁判文书网，这是一个公开提供各类法律文书的权威平台。截止目前，中国裁判文书网已收入文书总量达6414万份，访问总量高达228亿次，访客来自全球210多个国家和地区，成为全球最大的裁判文书公开平台。中国裁判文书网公开提供了藏语、蒙古语、维吾尔语、哈萨克语和朝鲜语5种我国少数民族语言的法律文书，这为在司法领域开展多语言自然语言处理研究提供了数据资源。基于研究的需求和目标，本文收集整理了3921篇维吾尔语的法律文书，涵盖了民事、刑事、赔偿、执行以及行政等多种案件类型，覆盖了维吾尔语司法领域的广泛案例。本文以盗窃类案件作为研究对象，通过数据增强和人工标注，建立一个专用的维吾尔语法律文书命名实体数据集UgLaw-NERD，以支持对司法领域维吾尔语命名实体识别技术的研究，也为深入探索维吾尔语在司法领域的文本处理提供基础。

3.2 数据预处理

中国裁判文书网提供的民族语言裁判文书均是PDF文件，需要将PDF文档转化为可处理的文本。本文应用科大讯飞多语种印刷文字识别API[3]，实现PDF转换和OCR识别，完成了文本数据的清洗。另外还采用翻译增强，将中文法律文书翻译成维吾尔语扩展了数据集并增加多样性。通过规则识别和提取文书中的案件描述，最终，将内容划分为3269个独立的盗窃类事件提及，每个描述都是一个完整的信息单元，详细反映了案件的具体情况。通过以上的处理过程，从大量维吾尔语法律文书中提取出关键信息段落，为后续的命名实体识别和数据分析准备了数据资源。

3.3 数据标注

本文除了通用领域命名实体识别任务中常用的实体类型，如人名、时间、地点之外，增加了司法领域命名实体识别任务中对案件理解十分重要的特有实体类型。本文定义了9种司法领域命名实体识别类型，以更好的抽取维吾尔法律文书种的案件要素。实体类型及定义如表1所示。

数据标注过程中，本文采用BIO(Begin-Inside-Outside)(Thompson and Lascarides, 1999)标注方案来标注命名实体。该方案分为三种标签：B 代表实体的开始，I 代表实体的内部，而O则表示非实体部分。相比IOBES(Uchimoto et al., 2000)方案，BIO标注方案在本文实验过程中，提供了一个较好的平衡点，既能有效识别大部分实体类型，更适用于维吾尔语法律文书的实体识别任务，也便于模型学习和实现。图1展示了标注数据案例。

3.4 人工评估

四名母语为维吾尔语的法学专业学生经过接受为期两周的命名实体数据标注培训后，被分

[3]https://www.xfyun.cn/services/xf-printed-word-recognition

标签	定义
NHCS	案件中出现的嫌疑人，被告人的姓名等表示身份的名称。
NT	案件中所涉及的时间段和某一时刻的时间表达式。
NS	案件所涉及的地理位置信息，包括行政区地名，街巷名。
NCGV	案件中所涉及的被盗物品价值，形式包括贵金属货币，纸币，电子货币等。
NHVI	案件中的受害人，被害人的姓名等表示身份的名称。
NCSM	案件中所涉及的被盗的现金数额价值，包括¥，人民币，元等字符。
NASI	案件中被害人的财产损失，被盗物品等。
NO	案件涉及的行政机构、企业机构及民间团体组织。
NATS	嫌疑人在作案过程中使用的工具。

表 1: 命名实体标签及定义

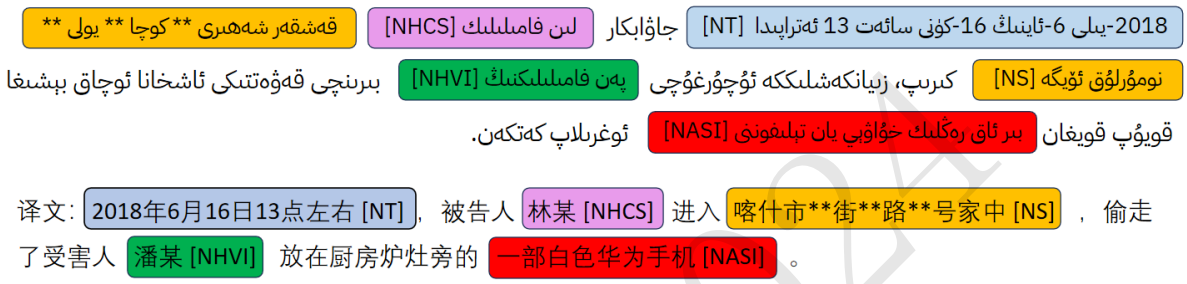

图 1: 标注数据案例

为两组，对同一数据进行两轮标注：第一轮作为预标注，标注完成后对遇到的问题进行整理和分析，根据情况调整和细化标注方案；第二轮则为正式标注，期间使用使用开源的数据标注平台Label-Studio[4]来监控整个过程，确保数据标注的一致性和质量。为了保证标注数据的准确性，本文从数据集中随机抽取了50个样本，并手动检查样本中的实体是否与标签相符。结果如表2所示，匹配率(Matching Acc)表示在人工评估下有多少个实体与其标签相匹配，总的匹配率达到了86.89%。

	NHCS	NT	NS	NCGV	NHVI	NCSM	NASI	NO	NATS
样本数量	72	12	53	33	35	25	8	23	6
正确标注	67	10	48	23	30	21	7	21	5
匹配率	93.06%	83.33%	90.57%	69.70%	85.71%	84.00%	87.50%	91.30%	83.33%

表 2: 标注实体人工评估结果

4 模型

4.1 问题描述

司法领域命名实体识别任务的具体目标可以描述为：给定低资源语言法律文书的NER训练集$D = \{(x_i, y_i)\}_{i=1}^{N}$，其中$x_i = (w_1, w_2, ..., w_m)$ 表示输入的法律文本序列，m 为序列中的单词总数。$y_i \in \{$司法实体类型集合$\}$ 为相应的实体标签，N 为训练集中的样本总数。该任务的目标是学习一个映射函数$f : D \rightarrow Y \in \{$司法实体类型集合$\}$，以便准确地预测输入文本序列中的法

[4]https://labelstud.io/

律命名实体及其类别。在此过程中，本文重点关注如何有效地扩充词汇并利用有限的标注数据来提升模型对低资源语言法律文书的命名实体识别能力。

4.2 多语言预训练模型mBERT

BERT Multilingual Base(简称为mbert-base)(Pires et al., 2019)是一个由Google 开发的多语言版自然语言处理模型，基于原始的BERT(Bidirectional Encoder Representations from Transformers)(Devlin et al., 2019)架构。此模型在包括维吾尔语在内的104种语言上进行预训练，能够处理并理解多种语言的文本，适用于多语种自然语言处理任务。mbert-base 采用了Transformer 架构，具有12层的双向编码器，隐藏层大小为768，以及12个自注意力头。这种结构使得模型能够捕捉文本中的复杂语境和依赖关系。此外，模型使用了512个位置嵌入来处理不同长度的输入序列。在预训练阶段，mbert-base主要通过两种方式进行训练：Masked Language Model(MLM) 和Next Sentence Prediction(NSP)。MLM任务通过随机遮盖输入中的单词并预测其身份来增强模型对上下文的理解能力。NSP任务则训练模型预测两个句子是否在原始文本中顺序相连，这有助于模型捕获句子级的语义关系。本文基于维吾尔语法律文书构建的微调数据集，因此采用Hugging Face Transformers框架中mbert-base模型,该模型包含约1.1亿参数。本文将mBERT模型微调20个epoch，初始学习率设置为1e-5，权重衰减率设置为0.01，batch size设置为16，500步的warmup进行训练。

4.3 词汇扩充

多语言预训练语言模型mBERT无法识别维吾尔语法律文书中的字符。这意味着当模型预测维吾尔语法律文书中的命名实体时，可能会将很多实体标记为未知[UNK]，从而影响模型命名实体识别的性能。

分词及词表构建。维吾尔语从右到左书写，单词之间使用空格隔开。每个维吾尔语单词都可以作为一个特征项。本文对维吾尔语法律文本的分词处理，采用预训练模型内置的WordPiece分词方法。这是一种基于词级别的分词策略，适用于处理像维吾尔语这样的黏着型语言。通过WordPiece方法获得175257个tokens，通过数据清洗，去掉标签符号、去掉重复字词，最终得到包含的13118个tokens的维吾尔语法律文书专用词汇列表。

词表合并和嵌入矩阵扩展。为了整合新词汇，本文采用了一种扩展词嵌入矩阵的方法。mBERT模型原始词汇表包含119547个tokens。分词模型生成的新词汇表与预训练模型的原始词汇表合并时，只增加预训练模型原始词汇表中没有的字词，扩展后mBERT模型的词汇列表大小为131912个token。接下来，为使模型能处理新增词汇，原始的词嵌入矩阵从$V \times D$被扩展至$V' \times D$，其中V 是初始词汇表的大小，D 是嵌入向量的维度，而V' 表示更新后的词汇表大小。此过程中，新的行被添加到矩阵中，为新词汇提供数值化的词向量表示。这些新增词向量以均值0和方差0.02的高斯分布进行初始化，为新词汇提供了一个合适的数值起点，以便在接下来的训练中进行调整，以此改善新词汇在模型中的语义表征:

$$\vec{v}_{new} \sim \mathcal{N}(0, 0.02I) \tag{1}$$

其中，$\vec{v}_{\text{new}}$ 表示新词嵌入的词向量，$\vec{v}_{\text{new}} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ 表示该向量遵循均值为0，方差为σ^2的正态分布，$\mathbf{I}$ 代表单位矩阵。

嵌入向量调整。为了让初始化的新词汇嵌入向量有效地融入到mBERT的语义空间中，本文利用维吾尔语法律文书的标注数据集UgLaw-NERD进行微调训练，目的是调整新词汇的嵌入向量，使之更准确地反映词汇的语义特性和上下文相关性。嵌入向量调整通过优化以下目标函数来实现:

$$L(\theta) = -\sum_{(x,y)\in D}\sum_{c\in C} y_c \log(f_c(x;\theta)) \tag{2}$$

$L(\theta)$ 是关于模型参数θ 的损失函数，D 是标注数据集，x 是输入文本，y 是与x 对应的实体标注向量，C 是所有可能的类别集合，y_c 是真实标注中类别c 的指示(0或1)，$f_c(x;\theta)$是mBERT模型预测输入文本x 属于类别c 的概率。这个损失函数计算了对于数据集中每个样本真实标注和模型预测之间的差异总和。

5 实验与分析

在本节中，从数据集与评价指标、对比实验、消融实验和参数分析四个部分介绍实验过程和细节。

5.1 数据集与评价指标

实验采用本文构建的维吾尔法律文书命名实体数据集UgLaw-NERD，本数据集包含9种司法实体类型，共96745个实体。实验中按照8:1:1的比例将数据集划分为训练集、验证集和测试集，数据集详细信息如表3所示：

数据集	NHCS	NT	NS	NCGV	NHVI	NCSM	NASI	NO	NATS
训练集	7458	12186	25256	6149	3518	2718	17589	2525	650
验证集	863	1416	2939	724	417	298	1984	340	67
测试集	913	1472	3194	807	391	398	2094	288	91
UgLaw-NERD	9234	15074	31389	7680	4326	3414	21667	3153	808

表 3: 维吾尔语法律文书命名实体数据集统计

本文用精确率P(Precision)、召回率R(Recall)和综合评价指标F1(F1-Measure)作为实验结果的评价指标，应用序列标注任务的seqeval脚本来评估UgLaw-NERD数据集在模型命名实体识别任务上的性能。

5.2 对比实验

本文采用过去研究中主要用于维吾尔语命名实体识别的以下基线模型进行对比实验：

- **CRF**:通过经典CRF模型提取维吾尔语法律文本特征进行命名实体识别。
- **Bi-LSTM**:利用可以更好的捕捉双向语义依赖关系的Bi-LSTM模型。
- **Bi-LSTM+CNN+CRF**:采用Bi-LSTM模型结合CNN和CRF进行命名实体识别。
- **mBERT-base(F.t)**:多语种预训练语言模型mBERT在数据集上微调进行维吾尔语法律文书的命名实体识别。
- **mBERT-base(F.t.voc)**:采用融合词汇扩充的多语种预训练语言模型mBERT在数据集上微调。

数据集	UgLaw-NERD		
模型	P	R	F1
CRF	74.85%	72.53%	73.67%
Bi-LSTM	79.53%	81.83%	80.67%
Bi-LSTM+CNN+CRF	80.60%	81.76%	81.18%
mBERT-base(F.t)	81.19%	83.50%	82.33%
mBERT-base(F.t.voc)	**88.96%**	**90.48%**	**89.72%**

表 4: UgLaw-NERD数据集上实验结果，(F.t)表示在数据集上微调的模型，(F.t.voc)表示融合词汇扩充后在数据集上微调的模型，加粗表示最优实验结果。

实验结果如表4所示，从表中可以得到如下结论：(1)本文提出的融合预训练模型词汇扩充和微调的mBERT-base(F.t.voc)模型在UgLaw-NERD数据集上得到了最好的结果，其F1分数为89.72%，相较于现有的基线模型mBERT-base(F.t)提升了7.39%。这一显著的提升证明了通过词汇扩充进行模型微调的有效性，特别是在处理语言多样性和专业术语较多的维吾尔语法律

文书数据集上。(2)从表中可见，mBERT-base模型在与传统CRF和Bi-LSTM模型的比较中表现更佳。这种性能提升部分得益于Transformer架构的全局注意力机制，它优化了上下文信息的捕捉。同时，mBERT的预训练背景也为其性能增强提供了支持。(3)Bi-LSTM+CNN+CRF模型组合相比单独的Bi-LSTM模型在F1分数上提升了0.51%，这表明结合不同类型的神经网络可以提升模型对细节的捕捉能力，特别是在实体边界识别和语义特征提取方面。(4)对于CRF模型，虽然在性能上不及深度学习模型，但其在数据集上的表现仍显示出一定的基线竞争力。综上所述，通过将维吾尔语司法领域的专有词汇添加到mBERT-base模型词汇表中，显著提升了其对维吾尔语中低频词汇和特定法律术语的理解与处理能力。词汇扩充扩大了模型的语言覆盖范围，使得在命名实体识别任务中能够实现更准确的预测。这一改进有效增强了模型在维吾尔语法律文书中识别命名实体的能力。

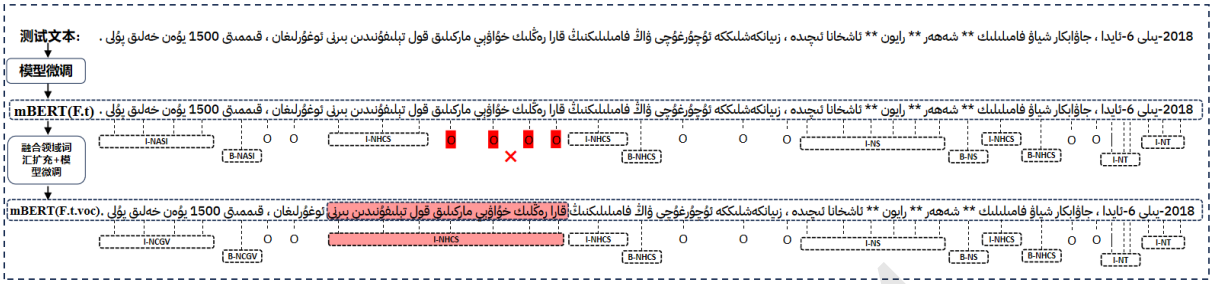

图 2: 样例图

图2给出了维吾尔语司法领域命名实体识别样例。在没有融合领域词汇扩充的情况下，仅对模型进行微调可能无法准确识别所有相关实体。这表明融合领域词汇扩充结合模型微调，能够有效提高维吾尔语法律文书命名实体识别的性能，不仅改善了模型的总体识别精度，还特别增强了对法律术语的识别能力。

5.3 消融实验

为了验证融合词汇扩充的预训练模型命名实体识别效果的有效性，本文在UgLaw-NERD数据集上进行消融实验，并设计以下模型变体：-N表示原始模型；(F.t)表示仅在数据集上微调的模型；(F.t.voc)表示融合词汇扩充和在数据集上微调的模型。

模型	P	R	F1
mBERT-base-N	0.37%	1.78%	0.61%
mBERT-base(F.t)	81.19%	83.50%	82.33%
mBERT-base(F.t.voc)	**88.96%**	**90.48%**	**89.72%**

表 5: 消融实验结果，加粗表示最优实验结果。

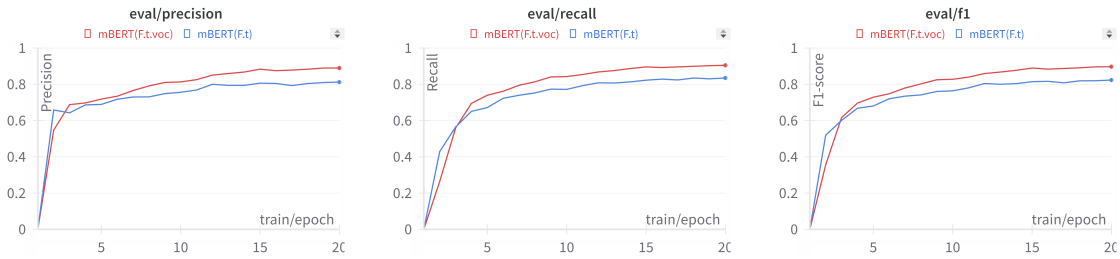

图 3: 模型在训练过程的表现，蓝色曲线代表mBERT-base(F.t)、红色曲线代表mBERT-base(F.t.voc)。

消融实验结果如表5和图3所示。从表和图中可以得到如下结论：(1)mBERT-base-N模型的性能极低，几乎没有实用价值，其精确率为0.37%，召回率为1.78%，F1分数为0.61%。这表明预训练模型直接应用于特定任务而不进行任何调整是不可取的，尤其是在语言特性复杂和语料资源有限的情况下。(2)当mBERT-base模型在UgLaw-NERD数据集上进行微调时，性能显著提升，F1分数达到82.33%。这种提升表明微调对预训练模型适应于低资源语言和特定任务是非常有效的。(3)在mBERT-base模型基础上进行词汇扩充并微调后，模型的性能进一步提升，F1分数提高到89.72%。这表明词汇扩充对于处理维吾尔语这种资源稀缺语言的命名实体识别任务尤其关键。词汇扩充帮助模型更好地理解和处理低资源语言中的低频词汇和复杂表达，从而提高了模型的识别精度和覆盖范围。

5.4 参数分析

为了探究词汇量大小对低资源语言法律文书命名实体识别模型性能的影响。本文通过调整模型中词汇量的大小，创建了不同的预训练模型变体，分别为mBERT-0、mBERT-10、mBERT-30、mBERT-50、mBERT-80和mBERT-100，其中数字代表扩充词汇量占本文构建的完整司法领域词汇表中词汇数量的百分比。

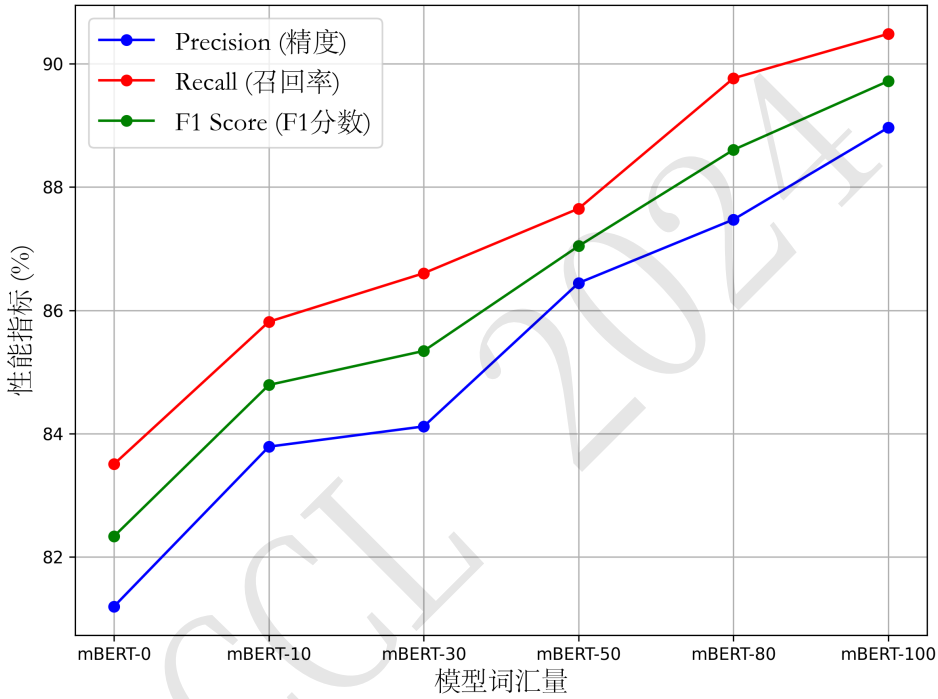

图 4: 新增词汇量大小对模型性能的影响

图4展示了不同词汇量大小对于模型性能的影响。从图中可以看出，词汇扩充量从0%增至100%，模型的精确度、召回率和F1分数均实现了逐步提升。具体来看，mBERT-0在这三个性能指标上的表现是最低的，而mBERT-100则在所有指标上达到了最高值。这一结果表明词汇扩充在增强模型对命名实体识别能力上的直接效果。F1分数作为精确度与召回率的综合指标，其显著的增长证实了词汇扩充策略在改善模型整体性能上的作用。在词汇扩充的各个级别中，性能提升与词汇扩充量呈现线性关系，特别是在词汇扩充量较高的阶段(mBERT-50至mBERT-100)，性能提升较为显著。综上所述，词汇扩充对低资源语言法律文书的命名实体识别模型有显著的正向作用。这表明在处理低资源语言的命名实体识别任务时，增加特定领域的词汇来扩充模型的词汇量是提高性能的一种有效策略。

6 总结与展望

针对现有多语言预训练语言模型在低资源语言法律文书命名实体识别任务中应用的局限性，本文提出了一种融合词汇扩充的改进策略。该方法通过收集维吾尔语司法领域词汇，动

态扩展了预训练模型的词汇表，并利用人工标注的维吾尔语法律文书命名实体数据集UgLaw-NERD上进行微调，提高了模型对维吾尔语法律文书中命名实体的识别能力。实验结果表明，与现有基准模型相比，本文提出的方法在性能上取得了明显提升。未来，本文将尝试把融合预训练模型词汇扩充的方法应用于更多低资源语言，进一步验证和提高方法的泛化性和实用性，从而为推进低资源语言处理技术的发展做出贡献。

参考文献

Kahaerjiang Abiderexiti, Maihemuti Maimaiti, Tuergen Yibulayin, and Aishan Wumaier. 2017. Construction of uyghur named entity relation corpus. *International Journal of Asian Language Processing*, 27(2):155–172.

Azmat Anwar, Xiao Li, Yating Yang, Rui Dong, and Turghun Osman. 2020. Constructing uyghur named entity recognition system using neural machine translation tag projection. In *Chinese Computational Linguistics: 19th China National Conference, CCL 2020, Hainan, China, October 30–November 1, 2020, Proceedings 19*, pages 247–260. Springer.

Luiz Henrique Bonifacio, Paulo Arantes Vilela, Gustavo Rocha Lobato, and Eraldo Rezende Fernandes. 2020. A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 648–662. Springer.

Yanguang Chen, Yuanyuan Sun, Zhihao Yang, and Hongfei Lin. 2020. Joint entity and relation extraction for legal documents with legal feature enhancement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1561–1571.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. 2020. Improving multilingual models with language-clustered vocabularies. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4536–4546, Online, November. Association for Computational Linguistics.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora*.

Junjie Deng, Hanru Shi, Xinhe Yu, Wugedele Bao, Yuan Sun, and Xiaobing Zhao. 2023. Milmo: minority multilingual pre-trained language model. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 329–334. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*, pages 272–287. Springer.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online, November. Association for Computational Linguistics.

Hirotaka Tanaka and Hiroyuki Shinnou. 2022. Vocabulary expansion of compound words for domain adaptation of BERT. In Shirley Dita, Arlene Trillanes, and Rochelle Irene Lucas, editors, *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 379–387, Manila, Philippines, October. Association for Computational Linguistics.

Henry S Thompson and Alex Lascarides. 1999. Ninth conference of the european chapter of the association for computational linguistics. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

Kiyotaka Uchimoto, Qing Ma, Masaki Murata, Hiromi Ozaku, and Hitoshi Isahara. 2000. Named entity extraction based on a maximum entropy model and transformation rules. In *proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 326–335.

Wenwen Xu, Mingzhe Fang, Li Yang, Huaxi Jiang, Geng Liang, and Chun Zuo. 2021. Enabling language representation with knowledge graph and structured semantic information. In *2021 International Conference on Computer Communication and Artificial Intelligence (CCAI)*, pages 91–96. IEEE.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A Chinese minority pre-trained language model. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

买合木提·买买提, 王路路, 吐尔根·依布拉音, 艾山·吾买尔, and 卡哈尔江·阿比的热西提. 2019. 基于条件随机场的维吾尔文机构名识别. 计算机工程与设计, 40:273–278.

加日拉·买买提热衣木, 吐尔根·依布拉音, and 艾山·吾买尔. 2014. 基于统计和规则混合策略的维吾尔人名识别研究. 新疆大学学报*(*自然科学版*)*, 31:319–324.

吾守尔, 斯拉木, 帕丽旦, 杨文忠, et al. 2018. 基于bilstm-cnn-crf 模型的维吾尔文命名实体识别. 计算机工程, 44(8):230–236.

孔祥鹏, 吾守尔, 斯拉木, 杨启萌, and 李哲. 2020. 基于迁移学习的维吾尔语命名实体识别. 东北师大学报*:*自然科学版, 52(2):58–65.

木合塔尔·艾尔肯, 艾斯卡尔·艾木都拉, and 地里木拉提·吐尔逊. 2013. 基于规则的维吾尔地名识别. 通信技术, 46:103–105.

李春楠, 王雷, 孙媛媛, and 林鸿飞. 2021. 基于bert 的盗窃罪法律文书命名实体识别方法. 中文信息学报, 35(8):73–81.

汪昆, 帕力旦, 吐尔逊, et al. 2017. 统计与规则相结合的维吾尔语人名识别方法. 自动化学报, 43(4):653–664.

王路路, 艾山·吾买尔, 买合木提·买买提, 卡哈尔江·阿比的热西提, and 吐尔根·依布拉音. 2018. 基于crf和半监督学习的维吾尔文命名实体识别. 中文信息学报, 32:16–26+33.

王路路, 艾山, 吾买尔, 吐尔根, et al. 2019. 基于深度神经网络的维吾尔文命名实体识别研究. 中文信息学报, 33(3):64–70.

艾斯卡尔·肉孜, 宗成庆, 姑丽加玛丽·麦麦提艾力, 热合木·马合木提, and 艾斯卡尔·艾木都拉. 2013. 基于条件随机场的维吾尔人名识别方法. 清华大学学报*(*自然科学版*)*, 53:873–877.

阿迪来, 冯向萍, et al. 2017. 基于条件随机场的维吾尔语音乐实体识别. 智能计算机与应用, 7(2):59–62.

麦合甫热提, 米日姑·肉孜, 麦热哈巴·艾力, and 吐尔根·依布拉音. 2014. 基于语法语义知识的维吾尔文机构名识别. 计算机工程与设计, 35:2944–2948.