文章编号:1672-3961(2024)04-0067-09 DOI:10.6040/j.issn.1672-3961.0.2023.112

融合数据增强和知识迁移的汉维跨语言命名实体识别

葛一飞[1],艾孜尔古丽[1,2*],陈德刚[1]
(1.新疆师范大学计算机科学技术学院, 新疆 乌鲁木齐 830054; 2.国家语言资源监测与研究少数民族语言中心,北京 100081)

摘要:针对维吾尔语命名实体识别任务数据匮乏的问题,提出汉维跨语言命名实体识别零样本迁移方法。采用一种简单有效的序列标记翻译方式,将源语言训练数据翻译为目标语言数据,避免词序变化和实体跨度不确定等问题,结合源语言数据和翻译后得到的数据,引入一种基于相似度计算的实体增强方法,可以有效提高文本生成质量,进一步增加样本的多样性。通过一系列广泛的试验,这些增强数据使少数民族预训练语言模型(Chinese minority pre-trained language model, CINO)能够更好地实现知识迁移目标语言的特定语言特征和多语言的语言独立特征,在多语言数据增强跨语言知识迁移模型上 F_1 值达到 86.50%,相比于基线模型提升 7.42%,证明融合数据增强和知识迁移的汉维跨语言命名实体识别的可行性。

关键词:汉维跨语言; 命名实体识别; 数据增强; 知识迁移; CINO
中图分类号:TP391 **文献标志码**:A
引用格式:葛一飞,艾孜尔古丽,陈德刚. 融合数据增强和知识迁移的汉维跨语言命名实体识别[J].山东大学学报(工学版),2024,54(4):67-75.
GE Yifei, Azragul, CHEN Degang. Chinese-Uyghur cross-lingual named entity recognition by fusing data augmentation and knowledge migration [J]. Journal of Shandong University (Engineering Science), 2024, 54(4):67-75.

Chinese-Uyghur cross-lingual named entity recognition by fusing data augmentation and knowledge migration

GE Yifei[1], Azragul[1,2*], CHEN Degang[1]
(1. College of Computer Science and Technology, Xinjiang Normal University, Urumqi 830054, Xinjiang, China; 2. National Language Resource Monitoring & Research Center of Minority Languages, Beijing 100081, China)

Abstract: A zero-sample migration method for Chinese-Uyghur cross-lingual named entity recognition was proposed to address the problem of data scarcity for the Uyghur named entity recognition task. A simple and effective sequence-tagged translation method was used to translate the source language training data into the target language data, avoiding problems such as word order variation and entity span uncertainty. A similarity calculation-based entity augmentation method was introduced by combining the source language data and the translated data, which could effectively improve the quality of text generation and further increase the diversity of samples. Through a series of extensive experiments, these augmented data enabled the Chinese minority pre-trained language model (CINO) to better knowledge transfer the language-specific features of the target language and the language-independent features of multiple languages, reached an F_1 value of 86.50% on the multilingual data augmented cross-lingual knowledge transfer model, an improvement of 7.42% compared to the baseline model, which demonstrated that Chinese-Uyghur cross-lingual named entity recognition by fusing data augmentation and knowledge migration was feasible.

Keywords: Chinese-Uyghur cross-lingual; named entity recognition; data augmentation; knowledge migration; CINO

收稿日期:2023-05-24
基金项目:新疆维吾尔自治区创新环境(人才、基地)建设专项-自然科学计划(少数民族科技人才特殊培养)资助项目(2022D03001);国家自然科学基金资助项目(61662081);国家社会科学基金资助项目(14AZD11);新疆师范大学青年拔尖人才资助项目(XJNUQB2022-22)
第一作者简介:葛一飞(1998—),男,江苏宿迁人,硕士研究生,主要研究方向为自然语言处理。E-mail:1453259830@qq.com
*通信作者简介:艾孜尔古丽(1987—),女,新疆乌鲁木齐人,副教授,硕士生导师,博士,主要研究方向为自然语言处理。
E-mail:Azragul2010@126.com

0 引言

随着信息技术的普及和快速发展,自然语言处理技术在面对互联网上产生的众多非结构化文本数据集对文本的处理、理解和应用发挥着关键作用。命名实体识别(named entity recognition, NER)是自然语言处理任务中一项关键性技术[1],用来识别非结构化文本中具有指定意义的实体单元,主要包括人名(person, PER)实体、地名(local, LOC)实体、机构名(organization, ORG)实体等,不仅可以作为信息提取过程中的独立工具,而且在自然语言文本处理的各个研究领域(如文本摘要、关系抽取、知识图谱和知识库构建等)都发挥着重要作用。

基于深度神经网络的 NER 模型在许多命名实体识别任务中都取得了不错的效果[2],但其成功很大程度上依赖于带有标签的大规模训练数据。对于一些广泛使用的语言,手动标记数据的获取可能相对容易。然而,除了一些资源丰富的语言(如英语、汉语)外,大多数其他语言的训练集仍然非常有限。维吾尔语句法结构比较复杂,属于黏着语,由"主语-宾语-谓语"构成,而且维吾尔语命名实体识别只有少量的注释语料库,没有公开的用于 NER 的语料,因此,需要新的思路和方法进一步提高维吾尔语命名实体识别的准确率。

如何实现低资源语言的命名实体识别已成为研究领域的一个重要问题,该问题促进了跨语言自然语言处理任务的发展。近年来,跨语言迁移任务发展迅速,得益于多语言预训练语言模型的快速发展。多语言变换器的双向编码器表示技术(multilingual bidirectional encoder representations from transformers, MBERT)[3]与单语言变换器的双向编码器表示技术(bidirectional encoder representations from transformers, BERT)[4]以相同的方式进行预训练,在 104 种具有共享词汇表语言(不含维吾尔语)的维基百科页面上进行训练,由于缺乏维吾尔语语料数据,在维吾尔语自然语言处理任务上效果较差。跨语言模型-鲁棒优化的 BERT 预训练方法(cross-lingual language model-robustly optimized BERT pre-training approach, XLM-R)使用 100 种语言、2.5 TB 文本数据进行训练[5],在分类、序列标记和问题回答等方面都优于 MBERT 和跨语言模型(cross-lingual language model, XLM)[6]等多语言模型,可以显著提高跨语言迁移任务的性能。

基于实例的传输方法为训练集增加了有限的语义多样性,只将实体和相应的上下文翻译到不同的语言中。相比之下,数据增强是解决数据稀缺性问题的一种成功方法[7-8]。面对低资源跨语言命名实体识别的挑战,更好地利用多语言预训练模型的跨语言泛化能力,本研究提出一种融合数据增强和知识迁移的汉维跨语言命名实体识别方法。通过试验验证,采用少数民族预训练语言模型(Chinese minority pre-trained language model, CINO)可以更好地评估汉维跨语言的知识迁移性能,同时翻译得到的更多语言也可以作为一种有效的数据增强方法,有助于提高源语言和目标语言的性能。

1 相关工作

随着神经网络的快速发展,神经网络已经广泛应用于 NER 任务中,例如:文献[9]提出一种基于卷积神经网络-双向长短时记忆-条件随机场[10](convolutional neural network-bidirectional long short-term memory-conditional random field, CNN-BLSTM-CRF)的神经网络模型,利用卷积神经网络训练字向量和词向量作为输入,在 2 个生物医学语料上 F_1 值分别为 89.09%和 74.40%;文献[11]提出一种用于深度学习框架的字词联合方法,将字特征和词特征统一结合起来,避免了词特征分词错误蔓延和字典稀疏问题;文献[12]通过嵌入 BERT 预训练语言模型,构建 BERT-BiGRU-CRF 模型用于表征语句特征,F_1 值分别达到 95.43%和 94.18%;文献[13]提出基于 Affix-Attention 的命名实体识别语义补充方法,通过融合文本信息和词缀信息,达到语义补充的效果;文献[14]通过引入注意力机制,加强关键特征,弱化无用特征;文献[15]为了提高增强样本精度,对传统数据增强方法进行改进,有效提高数据增强样本生成质量;文献[16]使用机器翻译系统将句子翻译成维吾尔语,通过机器翻译使源实体和目标实体对齐,F_1 值提高 3.79%。

相比于汉语、英语等语料较为丰富的语种,维吾尔语命名实体识别中存在特殊语义形态及语料数据较少等问题,影响维吾尔语命名实体识别任务的发展。为了解决数据匮乏、标注成本高昂等问题,零样本跨语言迁移任务引起了越来越多的兴趣[17-18]。跨语言迁移方法可大致分为基于数据迁移的方法和基于模型迁移的方法两大类:基于数据迁移的方法通常将源语言训练句子翻译成目标语言;基于模型迁移的方法通常学习语言无关的特

征,在源语言的标注语料上训练模型直接用于目标语言。文献[19]开发了一种新颖的无监督短语边界恢复预训练任务,增强了多语言边界检测能力;文献[20]研究了基于上下文词表示的具有模型迁移的无监督跨语言命名实体识别,可以更好地捕捉源语言和目标语言之间的关系;文献[21]引入一种基于生成的多语言数据增强方法,通过生成多种语言的合成标记数据进一步增加多样性,能够更好地泛化语言特征;文献[22]为了有效地从未标记数据中提取弱监督信号,提出一种基于半监督学习和强化学习思想的新方法;文献[23]提出掩蔽实体语言模型作为低资源 NER 的新型数据增强框架,通过生成高质量增强数据提高命名实体识别性能。

这些方法已经证明了零样本跨语言命名实体识别性能,但其中大多数都假定源语言可以提供大量的训练数据,当减少训练数据量时,性能显著下降。对于基于实例的传输,减少训练集大小也放大了机器翻译和标签投影引入噪声的负面影响。对于基于模型的迁移,尽管大规模的预训练多语言模型在许多跨语言迁移任务上取得了先进的性能,但只要在小训练集上进行微调就容易过拟合。

本研究提出一种融合数据增强和知识迁移的汉维跨语言命名实体识别方法,采用占位符对 NER 语料数据进行数据迁移,有效避免了许多与标签投影相关的问题;对训练数据进行数据增强,进一步增加样本的多样性;通过 CINO 对源语言数据进行知识迁移,更好地提高维吾尔语零样本跨语言知识迁移性能。

2　汉维跨语言命名实体识别

本研究所提方法利用了基于 CINO 的跨语言网络传输的优势,采用一种新的标记序列翻译方法,将标记的训练数据从源语言 S 翻译为一组目标语言 $T=\{T_1, T_2,\cdots, T_n\}$。在 $\{D_S, D_{T1},\cdots, D_{Tn}\}$ 数据上进行实体增强,进一步增加样本的多样性,其中 D_S 为源语言训练数据,D_{Ti} 为语言 T_i 中的翻译数据。对增强数据进行后处理和过滤,训练汉维跨语言命名实体识别模型,以便对目标语言测试集进行实体识别。

2.1　数据迁移

数据迁移方法主要针对目标语言不存在或存在少量有标签训练数据的情况,常见解决方法包括翻译[24]、半监督学习[25]、弱监督学习[26]和数据生成[27]等。然而,这些方法仍存在标签投影问题,如词序变化、跨度不确定等。避免标签投影问题的另一种方法是逐字翻译,但通常会牺牲翻译质量。

为了解决上述问题,本研究采用文献[21]提出的数据迁移方法,在句子翻译之前用上下文占位符替换命名实体,在翻译之后用相应的翻译实体替换翻译句子中的占位符。考虑到直接将实体插入语句中会影响上下文的连贯性,在进行数据迁移之前,把汉语 NER 数据中的实体提取构建汉维实体词典 1,通过实验室原有的汉维实体词典 2 进行实体匹配,如果词典中包含对应的实体,则采用汉维实体词典 2 对汉维实体词典 1 中的实体进行替换,如果匹配不成功,采用维吾尔语词干切分工具对汉维实体词典 1 中的维吾尔语实体进行词干切分,从而解决数据迁移上下文连贯性的问题。

假设一个句子 $X_S=\{x_1, x_2,\cdots, x_M\}\in D_S$ 和相应的 NER 标签 $\{y_1, y_2,\cdots, y_M\}$,其中 x_i 为句子中的每个字,y_i 为每个字对应的标签,M 为句子长度。令 $\{E_1, E_2,\cdots, E_N\}$ 表示预定义的命名实体类型。

第 1 步:将占位符 E_k 替换 $\{x_1, x_2,\cdots, x_M\}$ 中的所有实体。E_k 是重构的标记,以相应的实体类型 E 作为前缀,以实体 k 的索引作为后缀。假设 $\{x_i,\cdots, x_j\}$ 为源句中的第 k 个实体,对应的类型为 E_z,可以用占位符 E_zk 替换该实体得到 $\{\cdots, x_{i-1}, E_zk, x_{j+1},\cdots\}$。$X_S^*$ 表示用占位符替换所有实体后生成的句子,注入小牛翻译系统得到翻译 X_T^* 目标语言 T。这样的设计,占位符前缀 E 可以提供机器翻译模型实体相关的上下文信息,使模型翻译出质量好的句子。此外,观察到大多数占位符在翻译后没有变化,可以用来帮助定位实体。

第 2 步:用相应的上下文翻译每个实体。用方括号标记每个实体的跨度,并依次将其转换为目标语言,一次一个。例如,要翻译实体 $\{x_i,\cdots, x_j\}$,则提供 $\{\cdots, x_{i-1}, [x_i,\cdots, x_j], x_{j+1},\cdots\}$ 到小牛翻译系统,从翻译的句子中提取标记得到实体翻译。如果没有找到方括号,说明部分实体通过翻译未得到对应的标签,将直接删除此次翻译的句子。

第 3 步:将 X_T^* 中的占位符替换为相应的实体翻译(从第 2 步获得),并将占位符前缀复制为实体标签,生成目标语言的合成训练数据。

第 4 步:将第 3 步生成的目标语言数据处理成需要的数据标注格式,全程用小牛翻译系统进行翻译,获得高质量翻译数据。

2.2　数据增强

传统实体替换(mention replacement, MR)方法

如图 1 所示，利用汉维实体词典对训练集中相同的实体类型进行随机替换。但同类实体还是存在很大的语义差距，例如“清华大学”与“上海东方电视台”同为 ORG 实体，前者是学院 ORG 实体，而后者是企业 ORG 实体。

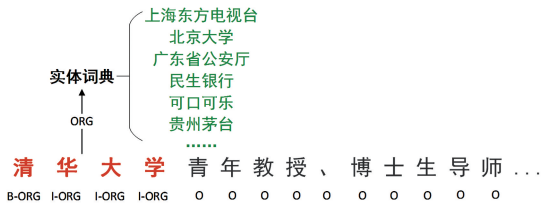

图 1 实体替换方法

Fig.1 Mention replacement method

为了减小实体替换引发的语义差异问题，采用句子-基于 BERT 的嵌入（sentence-BERT，SBERT）计算文本相似度[28]。SBERT 提供了现成的方法解决相似度问题，并在速度上更有优势，直接使用更方便，SBERT 的子网络使用 BERT 模型[4]，且 2 个 BERT 模型共享参数。当对比 A、B 文本相似度时，将 A、B 分别输入 BERT 网络，输出 2 组表征文本的向量 $\boldsymbol{u}$、$\boldsymbol{v}$，计算二者的余弦相似度分数 $\cos\ s_{\text{im}}(\boldsymbol{u},\ \boldsymbol{v})$。SBERT 相似度计算架构如图 2 所示。

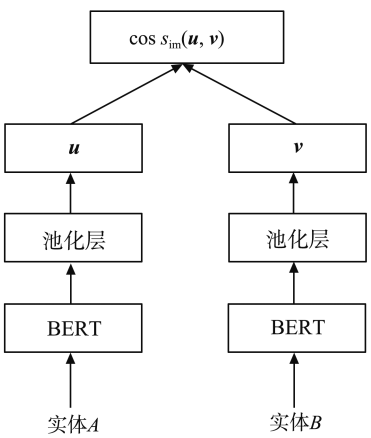

图 2 SBERT 相似度计算架构

Fig.2 SBERT similarity calculation architecture

由于少数民族语言与汉语、英语不同，SBERT 没有专门的少数民族语言预训练模型，为了更好地计算文本相似度，通过小牛翻译系统将汉语实体词典翻译成目标语言，将生成的目标语言实体词典与原来的汉语实体词典进行拼接，以“'\t'”为分隔符构成实体词典，通过汉语 BERT 预训练模型获取每个汉语实体的表征向量，做相似度计算。此外，考虑到两两计算的复杂度，事先获取训练集中每个实体对应的向量并保存下来，在生成新样本时，找到相似度最高的 5 个汉语实体的位置，从中随机选取 1 个，然后用对应位置的目标语言实体进行替换，如表 1 所示，以机构名“新疆师范大学”为例，通过 SBERT 与实体词典中的机构名实体进行相似度计算，从中随机抽取 1 个实体进行替换，可以很好地解决句子中的语义差异问题。

表 1 实体“新疆师范大学”相似度计算示例

Table 1 Example of similarity calculation for the entity “Xinjiang Normal University”

维吾尔语	中文	相似度
شىنجاڭ يېزا ئىگىلىك ئۇنىۋېرسىتېتى	新疆农业大学	0.956 4
شىنجاڭ تىببى ئۇنىۋېرسىتېتى	新疆医科大学	0.950 5
شىنجاڭ مالىيە ئىقتىسات ئۇنىۋېرسىتېتى	新疆财经大学	0.943 8
شىنجاڭ ئۇنىۋېرسىتېتى	新疆大学	0.932 4
شەرقىي شىمال پېداگوگىكا ئۇنىۋېرسىتېتى	东北师范大学	0.922 6

通过 stylecloud 库对实体“新疆师范大学”进行相似度可视化，如图 3 所示。

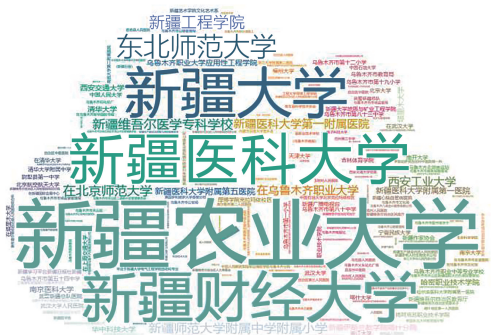

图 3 实体“新疆师范大学”相似度可视化

Fig.3 Visualization of the similarity of the entity “Xinjiang Normal University”

数据增强方法需要控制句子中被替换文本的占比，占比过高会导致整个语义发生变化，所以本研究提出的基于相似度计算的实体增强方法的增强概率为 20%，最多增强样本数为 1。同时为了移除增广数据中信息较少或带干扰信息的样本，本研究进行 2 步后处理：移除少于 10 个词的句子，因为语境较短，生成的新实体容易带有噪声；在原训练集数据上训练一个基线 NER 模型，用其标注增广数据，只保留基线标注标签与原标签一致的样本。

2.3 少数民族预训练语言模型

从源语言到目标语言的翻译可能会破坏词序，因此需要一个对齐模型将实体从源语言句子投影到目标语言，以便源语言的标签可以零样本迁移。XLM-R 模型使用来自上下文和另一种语言平行句子的信息预测被掩蔽的单词，使模型获得很好的跨语言和潜在对齐能力。

CINO 是基于多语言预训练模型 XLM-R 开发

的[29],在多种少数民族语言语料上进行二次预训练,提高了藏语、蒙古语(回鹘体)、维吾尔语、哈萨克语(阿拉伯体)、朝鲜语、壮语等少数民族语言的理解能力。CINO 和 XLM-R 模型的主要区别在于词嵌入和分词器,如图 4 所示,分别在藏语和蒙古语的单语预训练语料库上训练分词器,每个分词器的词汇量为 1.6×10^4,将藏语和蒙古语分词器中的词汇合并到原始的 XLM-R 分词器中,合并的分词器词汇量为 2.75×10^5,使用合并分词器对预训练语料库进行分词,并从合并分词器的词汇表和词嵌入矩阵中删除语料库中未出现的所有标记,得到词汇量为 1.35×10^5 的 CINO 分词器。

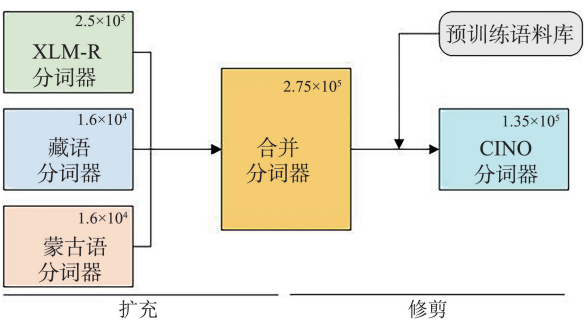

图 4　扩展 XLM-R 并构建 CINO 分词器

Fig.4　Extend XLM-R to construct the CINO tokenizer

XLM-R 模型架构如图 5 所示,XLM-R 是一种基于 Transformer 的预训练语言模型[30],由 Facebook AI Research 开发。XLM-R 模型使用了掩模语言模型(masked language modeling, MLM)和排列语言模型(permutation language modeling, PLM)2 种预训练模型,帮助 XLM-R 模型学习更多语言知识和语言结构。同时,XLM-R 模型使用了更大的数据集和更长的训练时间,使其在多种自然语言处理任务上取得显著突破。XLM-R 模型已经成为自然语言处理领域的重要研究方向,为跨语言自然语言处理任务提供了有力支持。

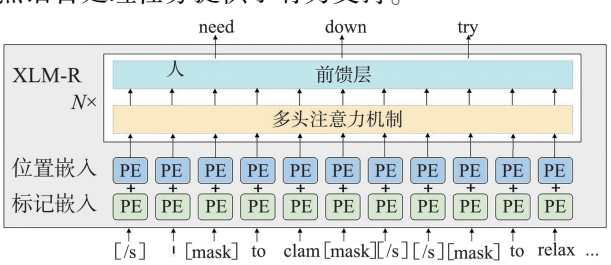

图 5　XLM-R 模型架构

Fig.5　XLM-R model architecture

3　试验结果与分析

本研究利用试验评估所提出的融合数据增强和知识迁移的汉维跨语言的有效性。在经过数据迁移后得到的少数民族语料上比较改进后的实体增强方法,在 CINO 上进行试验,测试少数民族语言数据的知识迁移能力,分别比较训练在单语言、双语言和多语言增强数据上模型的跨语言 NER 任务性能,评估基于数据增强和知识迁移的跨语言命名实体识别方法。

3.1　试验数据

本研究使用占位符的方法,对“1998 人民日报语料集实体识别标注集”的数据进行迁移,翻译成带有 NER 标签的维吾尔语(UG)语料数据。为了进行双语言和多语言的跨语言模型知识迁移性能测试,同时选取少数民族语言蒙古语(MO)和藏语(TI)进行数据迁移,因为这 2 种语言是 CINO 二次预训练的重要语料组成部分,可以更好地进行知识迁移。蒙古语存在自然的分隔符,和维吾尔语一样进行数据迁移;由于藏语没有分隔符,在进行数据迁移后需要进行分词,采用 TIP-LAS 工具进行藏语的分词[31],经过处理,合成藏语 NER 标注数据。为了保证机器翻译语料的质量和可靠性,分别招募维吾尔语、蒙古语和藏语等多名语言学研究生对数据质量进行评估,评估结果符合 NER 任务要求,具体数据信息如表 2 所示,其中 Train+MR 表示进行实体增强后的试验数据,Train 表示训练集,Dev 表示校验集,Test 表示测试集。

表 2　试验语料信息

Table 2　Information on the experimental corpus

试验数据	数据类型	句子数量	人名数量	地名数量	机构名数量
汉语	Train+MR	25 938	12 622	26 134	14 217
	Train	20 864	8 144	16 571	9 277
	Dev	2 318	984	1 951	884
	Test	4 636	1 864	3 658	2 185
维吾尔语	Train+MR	25 015	11 231	23 133	13 038
	Train	20 324	7 334	14 666	8 651
	Dev	2 257	822	1 658	928
	Test	4 525	1 679	3 197	2 075
蒙古语	Train+MR	24 733	11 019	22 902	12 934
	Train	20 063	7 182	14 718	8 485
	Dev	2 332	815	1 648	932
	Test	4 461	1 656	3 170	2 006
藏语	Train+MR	22 772	8 923	17 277	10 329
	Train	18 876	6 000	11 619	6 953
	Dev	2 080	639	1 288	718
	Test	4 167	1 294	2 493	1 618

3.2　标注方法和评价指标

本研究采用的命名实体识别标注方法为 BIO 标注法,其中 B 表示实体的开始部分,I 表示实体的

非开始部分,O 表示非实体部分。需要预测 7 种类型的标签,分别为:B-PER、I-PER、B-LOC、I-LOC、B-ORG、I-ORG 和 O。

本研究采用精确率 P、召回率 R 和 F_1 值作为模型的评价指标,计算公式分别为:

$$P=\frac{T_P}{T_P+F_P}\times 100\%, \tag{1}$$

$$R=\frac{T_P}{T_P+F_N}\times 100\%, \tag{2}$$

$$F_1=\frac{2PR}{P+R}\times 100\%, \tag{3}$$

式中,T_P 为正确识别实体的个数,F_P 为错误识别实体的个数,F_N 为未识别出的实体个数。

3.3 基线模型

为了测试 CINO 和 XLM-R 模型在少数民族语言命名实体识别任务上的效果,使用统一的训练轮数、学习率等参数。分别对有监督和无监督 NER 任务进行评估,结果如表 3 所示。其中,ZH-ZH 表示有监督学习,第 1 个 ZH 表示源语言为汉语,第 2 个 ZH 表示目标语言也为汉语;ZH-UG 表示无监督学习,即零样本跨语言的方法,第 1 个 ZH 表示源语言为汉语,第 2 个 UG 表示目标语言为维吾尔语,即本研究采用数据迁移的方法进行翻译后的测试数据。

表 3 CINO 和 XLM-R 模型评估结果

Table 3 CINO and XLM-R model evaluation results

源语言-目标语言	模型	F_1值/%	时间/s
ZH-ZH	XLM-R	92.44	3 099
	CINO	**92.82**	2 389
ZH-UG	XLM-R	60.27	2 100
	CINO	**75.63**	1 397

由表 3 可以看出:CINO 在有监督和无监督 NER 任务上的 F_1 值相比于 XLM-R 模型分别提升 0.38 百分点和 15.36 百分点,相比于 XLM-R 模型训练时间分别缩短 22%和 33%,CINO 可以更好提高少数民族语言零样本跨语言知识迁移性能。因此,本研究采用 CINO 评估汉维跨语言的理解能力。

3.4 参数设置

对于超参数调优,本研究进行以下调整:学习率设置为 5×10^{-5},权值衰减设置为 1×10^{-4},批处理大小为 16,优化器使用 AdamW, Dropout 设置为 0.5,训练轮数为 10。

3.5 数据增强性能测试

为了模拟低资源场景,试验分别从训练集中随机采样 1 000、2 000、4 000、8 000 条和完整的训练集进行试验,具体试验结果如表 4、5 所示。其中,ZH-ZH、MO-MO、TI-TI 和 UG-UG 分别表示汉语、蒙古语、藏语和维吾尔语有监督命名实体识别任务,ZH-MO、ZH-TI 和 ZH-UG 分别表示汉蒙、汉藏和汉维跨语言命名实体识别任务,方法中 Base 表示原始数据没有经过数据增强,MR 表示传统实体增强方法,Ours 表示本研究提出基于相似度计算的实体增强方法。

表 4 不同训练组合的有监督命名实体识别性能

Table 4 Performance of supervised named entity recognition for different training combinations

源语言-目标语言	方法	F_1 值/%				
		1 000 条训练集	2 000 条训练集	4 000 条训练集	8 000 条训练集	完整训练集
ZH-ZH	Base	75.77	82.67	86.49	**89.76**	92.97
	MR	76.86	82.48	87.02	89.58	93.12
	Ours	**78.26**	**83.75**	**87.30**	89.63	**93.37**
MO-MO	Base	80.12	84.22	88.55	90.59	92.88
	MR	80.59	84.67	88.47	90.44	92.93
	Ours	**80.69**	**85.09**	**89.02**	**90.71**	**93.52**
TI-TI	Base	71.60	77.78	83.81	**87.72**	91.16
	MR	71.49	79.27	83.87	87.52	91.12
	Ours	**73.40**	**79.43**	**85.02**	87.49	**91.74**
UG-UG	Base	86.90	88.83	91.28	92.40	94.43
	MR	85.99	**89.40**	91.03	92.45	94.12
	Ours	**87.34**	89.24	**91.42**	**92.61**	**94.50**

表 5 不同训练组合的无监督命名实体识别性能

Table 5 Performance of unsupervised named entity recognition for different training combinations

源语言-目标语言	方法	F_1 值/%				
		1 000 条训练集	2 000 条训练集	4 000 条训练集	8 000 条训练集	完整训练集
ZH-MO	Base	44.68	48.18	49.99	55.63	58.78
	MR	46.30	41.84	51.19	45.99	55.20
	ours	**52.12**	**53.71**	**55.73**	**58.36**	**60.84**
ZH-TI	Base	40.06	**35.57**	**39.51**	**40.07**	**38.08**
	MR	39.62	34.07	37.28	34.68	27.57
	ours	**43.24**	**40.91**	38.87	37.44	32.30
ZH-UG	Base	65.59	71.11	73.19	75.83	76.48
	MR	65.83	68.48	73.44	72.49	76.82
	ours	**70.59**	**72.39**	**75.08**	**76.52**	**79.08**

从表 4 可以看出:传统数据增强方法容易引发句子语义歧义问题,有时反而会降低模型性能;本研究提出的基于改进的实体增强方法通过对随机替换实体进行文本相似度计算,解决了传统实体增强同类实体的语义差异问题,提升了数据增强样本精度,相

比于 Base 模型,在汉语、蒙古语、藏语和维吾尔语完整数据集上 F_1 值分别提升 0.40、0.64、0.58 和 0.07 百分点,可以看出本研究提出的基于相似度计算的实体增强方法在命名实体识别有监督任务上可以作为一种有效的数据增强方法。

由表 5 可以看出:本研究提出的基于改进的实体增强方法在汉蒙和汉维跨语言命名实体识别任务上取得了很大的提升,相比于 Base 模型,F_1 值在完整数据集上分别提升 2.06 百分点和 2.60 百分点;然而在汉藏跨语言命名实体识别任务上,本研究提出的数据增强方法反而降低了任务性能,相比于 Base 模型,传统数据增强方法和本研究提出的数据增强方法的 F_1 值分别降低 10.51 百分点和 5.78 百分点,性能降低的原因可能是本研究在进行藏语数据迁移过程中,由于藏语没有明显的分隔符,采用 TIP-LAS 工具进行藏语分词的过程中可能对实体边界划分错误,使其上下文语境发生变化,导致标注质量和标注一致性可能存在问题,可以看出藏语命名实体识别任务发展相对困难,还有很长的路要走。总的来说,本研究提出的基于相似度计算的实体增强方法在汉蒙和汉维跨语言命名实体识别任务中也可以作为一种有效的数据增强方法。

3.6　试验结果分析

通过上述试验可知,在单语言(汉语)数据增强上的模型可以提高目标语言(维吾尔语)跨语言性能,本节进行训练双语言和多语言数据增强上模型的跨语言迁移性能测试,具体试验结果如表 6 所示。在源语言为汉语和藏语时,相比于 Base 模型,维吾尔语跨语言的 F_1 值提高 3.28 百分点;在源语言为汉语和蒙语时,相比于 Base 模型,维吾尔语跨语言的 F_1 值提高 6.44 百分点;在源语言为汉语、蒙语和藏语时,相比于 Base 模型,维吾尔语跨语言 F_1 值提高7.42百分点,取得了最优的效果。

表 6　多语言数据增强跨语言试验结果

Table 6　Results of cross-language experiments with multilingual data augmentation

单位:%

源语言-目标语言	P	R	F_1
ZH-UG	82.84	75.64	79.08
ZH+MO-UG	83.01	81.73	82.36
ZH+TI-UG	86.20	84.85	85.52
ZH+MO+TI-UG	**86.75**	**86.25**	**86.50**

从表 6 可以看出,添加额外的翻译数据可以持续提高维吾尔语跨语言 NER 任务性能,通过试验,相比于单语言模型,多语言数据增强跨语言知识迁移模型可以更好地提高零样本跨语言知识迁移性能。

3.7　错误分析

错误分析是训练和微调模型重要的方法之一。当模型性能比预期差很多时,错误分析可以深入了解模型的优点和不足之处,是了解模型优缺点的有力工具。

对输入标记进行分组,将每个标记的数量、均值及总和进行汇总,按照损失的总和对聚合数据进行降序排序,结果如表 7 所示。由表 7 可以看出:空白标记“ـــ”的数量最多,总损失最高,但是平均损失要比其他标记低很多,这意味着模型可以很好地对其进行分类;“ي”“يا”“ئېلى”“جۇڭگو”等词总损失较大,这是因为它们经常与实体一起出现,有时是实体的一部分,导致汉维跨语言模型识别错误。

表 7　输入标记损失统计表

Table 7　Enter a table of marker loss statistics

输入标记	数量	均值	总和
ـــ	7 934	0.14	1 111.63
ـــئېلى	112	3.92	438.51
ـــيا	383	0.62	236.97
ـــي	619	0.36	222.92
ـــجۇڭگو	523	0.30	155.35
ـــخەلق	264	0.55	145.44
ـــئى	448	0.32	141.47
ـــقۇر	361	0.38	137.87
ـــبەش	123	1.11	136.95
ـــمەركىزى	161	0.84	134.78

对标签 ID 进行分组,将每个标记的数量、均值和总和进行汇总,按照均值对聚合数据进行降序排序,结果如表 8 所示。由表 8 可以看出:I-LOC 的损失最高,主要是样本占比较少,模型无法充分学习到实体的特征。

表 8　标签 ID 损失统计表

Table 8　Tag ID loss statistics table

标签	数量	均值	总和
I-LOC	968	1.32	1 277.41
B-ORG	2 075	0.75	1 549.65
B-LOC	3 197	0.55	1 752.96
I-ORG	4 444	0.51	2 251.43
B-PER	1 679	0.46	774.18
I-PER	1 116	0.29	323.70
O	106 881	0.04	4 674.26

通过绘制混淆矩阵(图 6)可以看出:汉维跨语言模型最容易混淆 B-ORG 和 I-LOC 两个实体标记;从混淆矩阵的近对角线性质可以看出,汉维跨语言模型对其余实体进行分类时表现得非常好。

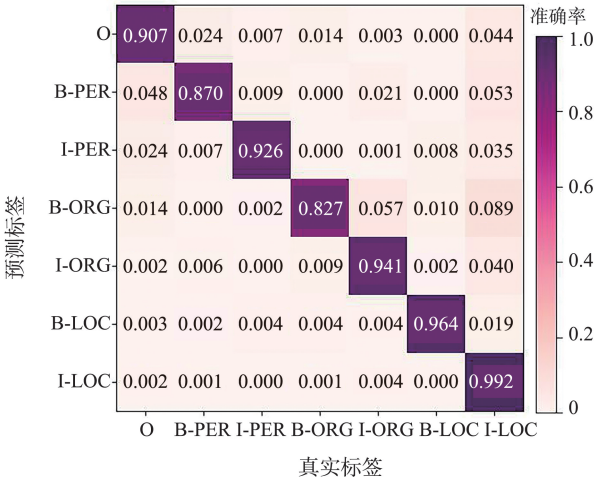

图 6 混淆矩阵

Fig.6 Confusion matrix

4 结论

本研究针对维吾尔语 NER 数据匮乏的问题,提出一种融合数据增强和知识迁移的汉维跨语言命名实体识别方法,该方法采用占位符对 NER 标注数据进行数据迁移,有效避免了标签投影相关问题,通过构建多语言词典,使用 SBERT 进行实体相似度计算,有效解决了实体替换引发的语义歧义问题。同时,本研究选择了 2 种多语言词汇表征,分别是 CINO 和 XLM-R 模型,通过在有监督和无监督维吾尔语 NER 任务上进行试验,CINO 维吾尔语跨语言迁移性能更优。另外,在源语言数据中添加额外的翻译数据,即藏语和蒙古语。试验表明,本研究提出的多语言数据增强跨语言迁移方法可以更好地提高维吾尔语零样本跨语言 NER 知识迁移性能。

通过进行简单的错误分析,本研究发现了汉维跨语言模型和数据集中的一些弱点,在未来工作中将迭代这一步,清理数据集,重新训练模型并分析新的错误,从而更好地提升维吾尔语 NER 任务性能。

参考文献:

[1] DONG C H, ZHANG J J, ZONG C Q, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition[C]// Natural Language Understanding and Intelligent Applications. Kunming, China: Springer, 2016: 239-250.

[2] MCCULLOCH W S, PITTS W. A logical calculus of the ideas immanent in nervous activity[J]. The Bulletin of Mathematical Biophysics, 1943, 5(4): 115-133.

[3] PIRES T, SCHLINGER E, GARRETTE D. How multilingual is multilingual BERT? [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2020: 4996-5001.

[4] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA: ACL, 2019: 4171-4186.

[5] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, USA: ACL, 2020: 8440-8451.

[6] CONNEAU A, LAMPLE G. Cross-lingual language model pretraining [C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, USA: NIPS, 2019: 634.

[7] DAI X, ADEL H. An analysis of simple data augmentation for named entity recognition[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain: COLING, 2020: 3861-3867.

[8] DING B, LIU L, BING L, et al. DAGA: data augmentation with a generation approach for low-resource tagging tasks[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Seattle, USA: ACL, 2020: 6045-6057.

[9] 李丽双, 郭元凯. 基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别[J]. 中文信息学报, 2018, 32(1): 116-122.
LI Lishuang, GUO Yuankai. Biomedical named entity recognition with CNN-BLSTM-CRF [J]. Journal of Chinese Information Processing, 2018, 32(1): 116-122.

[10] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, USA: ACM, 2001: 282-289.

[11] 张海楠, 伍大勇, 刘悦, 等. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报, 2017, 31(4): 28-35.
ZHANG Hainan, WU Dayong, LIU Yue, et al. Chinese named entity recognition based on deep neural network [J]. Journal of Chinese Information Processing, 2017, 31(4): 28-35.

[12] 杨飘, 董文永. 基于 BERT 嵌入的中文命名实体识别方法[J]. 计算机工程, 2020, 46(4): 40-45.
YANG Piao, DONG Wenyong. Chinese named entity recognition method based on BERT embedding [J]. Computer Engineering, 2020, 46(4): 40-45.

[13] 宋佳芮, 陈艳平, 王凯, 等. 基于 Affix-Attention 的命

名实体识别语义补充方法[J].山东大学学报(工学版), 2023, 53(2): 70-76.
SONG Jiarui, CHEN Yanping, WANG Kai, et al. Semantic supplement method for named entity recognition based on Affix-Attention [J]. Journal of Shandong University (Engineering Science), 2023, 53(2): 70-76.

[14] GE Y, CHEN D, LI K, et al. Uyghurlanguage recognition method based on BIGRU-IDCNN-ATT-CRF [C]//Proceedings of the 2021 7th International Symposium on System and Software Reliability (ISSSR). Chongqing, China: IEEE, 2021: 146-151.

[15] GE Y, YUSUP A, CHEN D, et al. UGDA: data augmentation methods for Uyghur language named entity recognition [C]//Proceedings of the 2022 9th International Conference on Dependable Systems and Their Applications (DSA). Urumqi, China: IEEE, 2022: 926-932.

[16] ANWAR A, LI X, YANG Y, et al. Constructing Uyghurnamed entity recognition system using neural machine translation tag projection[C]//China National Conference on Chinese Computational Linguistics. Hainan, China: CCL, 2020: 247-260.

[17] 梁世宁. 零样本跨语言序列标注关键技术研究[D]. 长春: 吉林大学, 2022.
LIANG Shining. Research on key techniques in zero-shot cross-lingual sequence labeling [D]. Changchun: Jilin University, 2022.

[18] 佘琪星. 面向低资源的跨语言命名实体识别方法[D]. 哈尔滨: 哈尔滨工业大学, 2021.
SHE Qixing. Cross-lingual named entity recognition in a low resource setting [D]. Harbin: Harbin Institute of Technology, 2021.

[19] LIANG S, SHOU L, PEI J, et al. CalibreNet: calibration networks for multilingual sequence labeling [C]//Proceedings of the 14th ACM International Conference on Web Search and Data Mining. Jerusalem, Israel: WSDM, 2021: 842-850.

[20] YAN H, QIAN T, XIE L, et al. Unsupervised cross-lingual model transfer for named entity recognition with contextualized word representations[J]. Plos One, 2021, 16(9): e0257230.

[21] LIU L, DING B, BING L, et al. MulDA:a multilingual data augmentation framework for low-resource cross-lingual NER [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Bangkok, Thailand: ACL-IJCNLP, 2021: 5834-5846.

[22] LIANG S, GONG M, PEI J, et al. Reinforcediterative knowledge distillation for cross-lingual named entity recognition[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Singapore: KDD, 2021: 3231-3239.

[23] ZHOU R, LI X, HE R, et al. MELM:data augmentation with masked entity language modeling for low-resource NER[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland: ACL, 2022: 2251-2262.

[24] JAIN A, PARANJAPE B, LIPTON Z C. Entity projection via machine translation for cross-lingual NER [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: ACL, 2019:1083-1092.

[25] ZHU S, CAO R, YU K. Dual learning for semi-supervised natural language understanding [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1936-1947.

[26] SHOU L, BO S, CHENG F, et al. Mining implicit relevance feedback from user behavior for web question answering[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. San Diego, USA: KDD, 2020: 2931-2941.

[27] HOU Y, CHEN S, CHE W, et al. C2C-GenDA: cluster-to-cluster generation for data augmentation of slot filling [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2021: 13027-13035.

[28] REIMERS N, GUREVYCH I. Sentence-BERT:sentence embeddings using siamese BERT-networks [C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: ACL, 2019: 3982-3992.

[29] YANG Z, XU Z, CUI Y, et al. CINO: a Chinese minority pre-trained language model [C]//Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Korea: COLING, 2022: 3937-3949.

[30] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: NIPS, 2017: 6000-6010.

[31] LI Y, JIANG J, JIA Y J, et al. TIP-LAS: an open-source toolkit for Tibetan word segmentation and POS tagging[J]. Journal of Chinese Information Processing, 2015, 29: 203-207.

(编辑:孙亚彤)