DOI: 10.13718/j.cnki.xdzk.2024.04.015
王坤，盛鸿宇. 一种提高跨语言理解的 NLP 迁移学习 [J]. 西南大学学报 (自然科学版)，2024，46(4)：153-163.

一种提高跨语言理解的 NLP 迁移学习

王坤[1]，　盛鸿宇[2]

1. 四川信息职业技术学院，四川 广元 628017；2. 北京联合大学 机器人学院，北京 100101

摘要：随着互联网信息的发展，如何有效地表示不同语言所含的信息已成为自然语言处理(Natural Language Processing，NLP)领域的一项重要任务. 然而，很多传统的机器学习模型依赖在高资源语言中进行训练，无法迁移到低资源语言中使用. 为了解决这一问题，结合迁移学习和深度学习模型，提出一种多语言双向编码器表征量(Multi-lingual Bidirectional Encoder Representations from Transformers，M-BERT)的迁移学习方法. 该方法利用 M-BERT 作为特征提取器，在源语言领域和目标语言领域之间进行特征转换，减小不同语言领域之间的差异，从而提高目标任务在不同领域之间的泛化能力. 首先，在构建 BERT 模型的基础上，通过数据收集处理、训练设置、参数估计和模型训练等预训练操作完成 M-BERT 模型的构建，并在目标任务上进行微调. 然后，利用迁移学习实现 M-BERT 模型在跨语言文本分析方面的应用. 最后，在从英语到法语和德语的跨语言迁移实验中，证明了本文模型具有较高的性能质量和较小的计算量，并在联合训练方案中达到了 96.2%的准确率. 研究结果表明，该文模型实现了跨语言数据迁移，且验证了其在跨语言 NLP 领域的有效性和创新性.

关　键　词：自然语言处理；多语言双向编码器表征量；迁移学习；跨语言；深度学习

中图分类号：TP393　　**文献标志码**：A

文 章 编 号：1673-9868(2024)04-0153-11

开放科学(资源服务)标识码(OSID)：

An NLP Migration Learning for Improving Cross-lingual Understanding

WANG Kun[1]，　SHENG Hongyu[2]

1. *Sichuan Vocational College of Information Technology，Guangyuan Sichuan 628017，China*；
2. *College of Robotics，Beijing Union University，Beijing 100101，China*

Abstract: With the development of internet-based information, effectively representing the information contained in different languages has become an important task in the field of Natural Language Processing (NLP). However, many traditional machine learning models rely on training in high-resource languages and cannot be used in low-resource languages. To address this issue, this paper proposes a migration learning method called Multi-lingual Bidirectional Encoder Representations from Transformers (M-BERT)

收稿日期：2023-06-27
基金项目：国家自然科学基金项目(12104289).
作者简介：王坤，硕士，副教授，主要从事计算机应用技术研究.
通信作者：盛鸿宇，硕士，教授级高级工程师.

that combines migration learning with deep learning models. This method utilizes M-BERT as a feature extractor to transform features between the source language domain and the target language domain, thereby reducing the differences between different language domains and improving the generalization ability of the target task across domains. First, the BERT model was constructed. Then, the construction of the M-BERT model was completed through pre-training operations such as data collection and processing, training setup, parameter estimation, and model training. Fine-tuning was performed on the target task. Finally, migration learning was employed to apply the M-BERT model in cross-lingual text analysis. The cross-lingual migration experiments from English to French and German demonstrated that the model proposed in this paper exhibited high performance quality and required minimal computational effort, achieved an accuracy of 96.2% in the joint training scheme. The research results indicate that this model achieved cross-lingual data migration, validated its effectiveness and innovation in the field of cross-lingual NLP.

Key words: NLP; M-BERT; migration learning; cross-lingual; deep learning

随着全球化不断推进和信息技术的迅猛发展，跨语言理解在自然语言处理(Natural Language Processing，NLP)领域扮演着重要的角色[1]. 自然语言处理(NLP)是数据挖掘的一个前沿方向，融合了机器学习与统计学、数学、语言学等学科，近年来发展迅速[2-4]. 通过统计和机器学习方法，计算机能够快速地处理、分析并运用文本的深层语义信息，像人类一样理解并生成自然语言. “自然语言”的含义是自然进化形成的人类语言，如中文、英文、拉丁语等，有别于Java、C++等程序语言.

在互联网时代，人们可以通过网络轻松地获取来自不同语言和文化背景的信息，这使得跨语言NLP任务变得尤为关键. 例如，机器翻译、情感分析和命名实体识别等任务都需要处理多种语言之间的转换和理解. 然而，不同语言之间存在结构、词汇、语法和文化等方面的差异，这给跨语言理解带来了巨大的挑战.

传统的机器学习方法在处理跨语言任务时往往需要大量的人工特征工程和领域知识[5-6]，这些方法通常依赖手工设计特征来捕捉不同语言之间的差异和共性，然后使用分类器或回归模型进行训练和推理. 然而，这种方法面临多个问题. 首先，人工特征工程耗时耗力，并且对不同语言之间的差异和数据稀缺性处理困难. 其次，这种方法可能无法充分利用深层语义和上下文信息，导致在跨语言理解任务中的性能不尽如人意.

神经网络和深度学习模型在NLP领域取得了显著的突破和成功[7-9]，相比于传统的机器学习方法，神经网络和深度学习模型具有更强的表示能力和泛化能力，能够从大规模数据中自动学习特征和模式，并且能够处理复杂的语言结构和语义关系. 在跨语言NLP领域，神经网络和深度学习模型的应用也取得了一定的成果，通过迁移学习和跨语言数据的利用，能够有效地解决语言差异和数据稀缺性带来的挑战. 其中，BERT(Bidirectional Encoder Representations from Transformers)是一种最具代表性的深度学习模型. BERT是跨语言NLP领域中基于模型的迁移学习方法，它在多个跨语言NLP任务上取得了最先进的性能，并成为了跨语言NLP任务的基准模型.

在跨语言学习领域，Vázquez等[10]通过重用多语言神经机器翻译的编码器进行零样本二元情感分类，他们使用特定于任务的分类器组件扩展了该编码器，并用新语言执行文本分类. Verma等[11]提出了ULMFiT模型，该模型通过在通用领域语料库上预训练通用语言模型，使用判别式微调对目标任务数据上的模型进行微调，从而应用于任何NLP任务. Gu等[12]使用针对特定任务训练的双向LSTM，通过查看整个句子来呈现词嵌入中词的上下文敏感表示. 还有一些学者研究生成了两种基于Transformer的语言模型，分别是OpenAI GPT和BERT[13]. OpenAI GPT是一种单向语言模型；而BERT是第一个深度双向、无监督的语言表示模型，仅使用纯文本语料库进行预训练[14-15]. Pelicon等[16]使用BERT通过在斯洛文尼亚语中训练分类器，并使用其他语言的文本进行推理来执行情感分类. Kim等[17]将预训练好的嵌入向量迁移到LSTM结构建立了Docbert模型，在实现参数压缩的同时保持了BERT在文本分类任务中的准确性. 贾明华等[18]定性地研究了BERT-Large中每层Transformer在不同NLP任务中的贡献.

基于以上研究，本文结合迁移学习和深度学习模型，提出一种M-BERT迁移学习模型，用于解决跨语言NLP任务中的关键问题. 通过实验对比了本文提出的方法与其他先进算法在多个跨语言NLP任务上的

性能差异，结果证明本文方法在各项任务中都取得了显著优于现有算法的结果，具有较高的性能. 本文的目的与研究意义旨在探索一种基于 NLP 迁移学习的方法，结合深度学习模型(M-BERT 模型)，用于提高跨语言理解的性能. 通过迁移学习，可以利用源语言的丰富资源和知识来改善目标语言的学习能力，从而解决数据稀缺和语言差异带来的问题. 同时，通过引入深度学习模型，可以利用其强大的表示学习能力和上下文理解能力，进一步提高跨语言理解的准确性和泛化能力.

1　BERT 模型构建

1.1　输入嵌入

输入分 3 个阶段处理：标记化、将标记映射为数字表示、词嵌入. 在标记化之后，每个标记都被映射到语料库词汇表的不同整数，称为映射标记. 每个标记都获得一个唯一的数字表示. 此外，还需要填充以确保批次中输入序列的长度相同. 标记化、映射和词嵌入都是将词转换为向量的过程，与神经词嵌入的完成方式类似. 本文给出下面这个小型句子：“卢森堡是一个了不起的国家”. 首先，将其标记化：

“Luxembourg is an amazing country.”(卢森堡是一个了不起的国家.)得到的标记为——[“Luxembourg，”“is，”“an，”“amazing，”“country，”“.”(“卢森堡”“是”“一个”“了不起的”“国家”“.”)].

然后进行映射，在语料库的词库中每个标记被分配成一个唯一的整数. 例如

[“Luxembourg，”“is，”“an，”“amazing，”“country，”“.”]→[34，90，15，684，55，193]. 接着，得到序列中每个单词的嵌入. 序列中每个短语都与一个嵌入维向量有关，模型将在整个学习过程中发现该向量，并可以把它看作是对每个记号的向量查找. 这些向量作为模型参数处理，通过反向传播进行调整，与其他权重的优化方式相同.

因此，本文搜索与每个标记相关的向量. 例如：

$$34 \rightarrow E[34]=[123, 0.32, \cdots, 94, 32]$$
$$90 \rightarrow E[90]=[83, 34, \cdots, 77, 19]$$
$$15 \rightarrow E[15]=[0.2, 50, \cdots 33, 30]$$
$$684 \rightarrow E[684]=[289, 433, \cdots, 150, 92]$$
$$55 \rightarrow E[55]=[80, 46, \cdots, 23, 32]$$
$$193 \rightarrow E[193]=[41, 21, \cdots, 74, 33]$$

然后通过堆叠每个向量来生成一个尺寸为：(输入长度)×(嵌入维度)的矩阵 K，如图 1 所示. 其中 d_{emb} 表示嵌入向量的维度(embedding dimension).

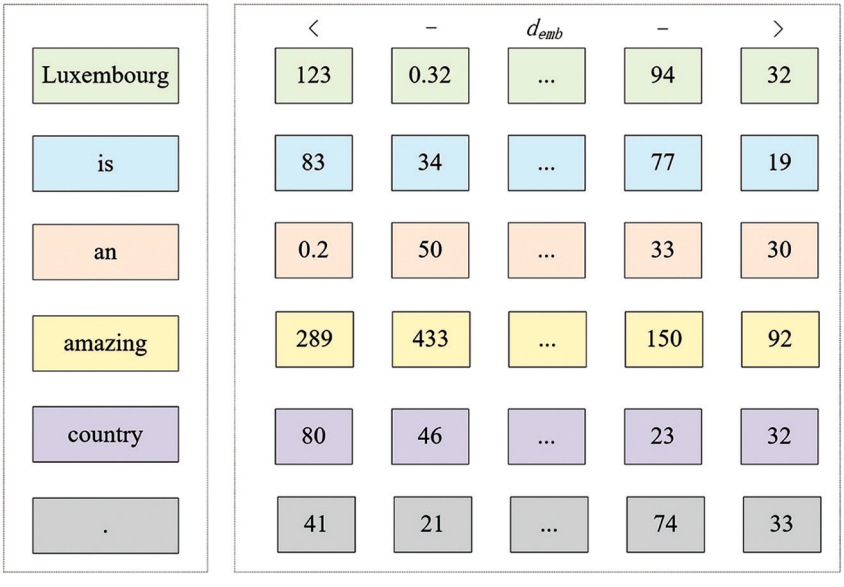

图 1　尺寸为(输入长度)×(嵌入维度)的代表矩阵 K

最后，需要使用填充来确保批处理中所有输入序列的长度相同. 因此，本文通过包含“pad”标记来延长一些序列. 将第9个长度进行填充后的序列为：[“〈pad〉，”“〈pad〉，”“〈pad〉，”“Bangladesh，” “is，”“a，” “beautiful，”“country，”“.”]→[5，5，5，34，90，15，684，55，193].

1.2 位置编码

BERT算法的优势通过学习位置嵌入获得，生成的文本序列被表示为矩阵，尽管这些表示没有考虑单词存在于不同位置的事实，但它能够根据单词的位置来改变单词的表征含义，目的并不是要改变这个词的完整表达，而是稍微改变它，以编码其位置.

该分析采用一种策略，使用不可学习的正弦函数将[−1，1]之间的数字添加到标记嵌入中. 编码器其余部分根据词的位置(即使是同一个词)以略微不同的方式表示这个词，另外一些词则处于同一序列中不同的特定位置. 我们希望网络既能理解绝对位置，也能理解相对位置. 本文选择的正弦函数可以将位置表示为彼此的线性组合，从而使系统能够学习标记位置之间的相关关系. 将具有位置编码的矩阵 U 添加到 K 合并该信息，就变成了 $U+K$.

BERT采用正弦函数合成. 从数学角度来讲，标记在序列中的位置用 x 表示，嵌入特征的位置用 y 表示. 正弦函数用如下公式描述.

$$u_{x,y}=\begin{cases}\sin\left(\dfrac{x}{10\ 000\dfrac{y}{d_{\text{emb}}}}\right) & \text{if } j \text{ is even}\\ \cos\left(\dfrac{x}{10\ 000\dfrac{y-1}{d_{\text{emb}}}}\right) & \text{if } y \text{ is odd}\end{cases}\tag{1}$$

给定文本 U 的位置嵌入矩阵如图2所示. 与学习的位置表征相比，这种确定性方法具有许多明显的优势. 例如，输入长度参数可以无止尽地增加，因为函数可以在任意位置计算. 此外，需要学习的参数更少，因此可以更快地训练模型. 得到的矩阵是 $I=K+U$，大小是(输入长度)×(嵌入维度)，它是第一个编码器块的输入.

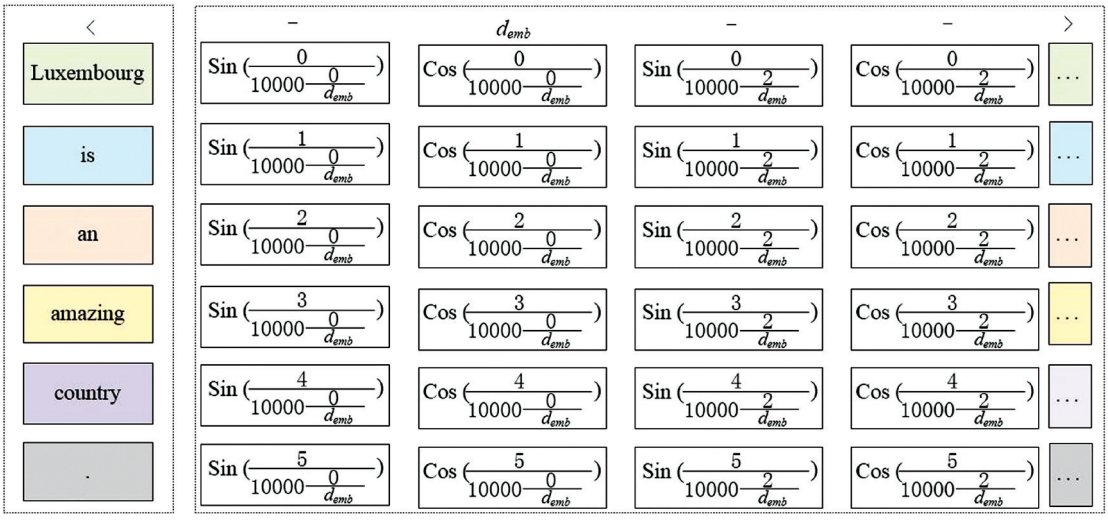

图2 位置嵌入矩阵

1.3 编码器块

BERT编码器是一种基于注意力机制和前馈神经网络相结合的变压器编码方法，编码器由多个编码器块叠加而成，每个编码器块包括两个前馈层和一个双向的自注意力层.

当数据通过编码器块时，对于一个给定的输入序列，通过位置编码产生的位置信息会返回一个尺寸为(输入长度)×(嵌入维度)的矩阵. 一个特定的块负责建立输入表示之间的关系，并在输出中对其进行编码.

1.4　多头注意力机制

编码器架构是围绕多头注意力构建的，它使用各种权重矩阵多次计算注意力 b，然后将结果串联起来. 头部是每一个注意力平行计算的结果，下标 x 被用来表示一个特定的头和它相应的权重矩阵. 计算完所有头后，将它们进行连接，产生一个尺寸为(输入长度)×x(b×d_q)的矩阵. 最终，添加了由维度(b×d_q)×(嵌入维度)的权重矩阵 M^0 组成的线性层，产生了维度为(输入长度)×(嵌入维度)的最终输出.

$$Multihead\ (V, Z, Q) = Concat(head_1, \cdots, head_b)M^0 \tag{2}$$

式(2)中，$head_x = Attention\ (VM_x^V, ZM_x^Z, QM_x^Q)$. 在这种情况下，$V$，$Z$ 和 Q 为各种输入矩阵的占位符.

1.5　缩放点积注意力

在缩放点积注意力机制中，每个头由 3 个不同的投影(矩阵乘法)定义：

$$\begin{aligned} &M_x^z \text{ with the dimensions } d_{emb} \times d_z \\ &M_x^v \text{ with the dimensions } d_{emb} \times d_v \\ &M_x^q \text{ with the dimensions } d_{emb} \times d_q \end{aligned} \tag{3}$$

输入矩阵 X 通过这些权重矩阵分别投射，计算头部.

$$\begin{aligned} IM_x^Z &= Z_x \text{ with the dimensions input\_length} \times d_z \\ IM_x^V &= V_x \text{ with the dimensions input\_length} \times d_v \\ IM_x^Q &= Q_x \text{ with the dimensions input\_length} \times d_q \end{aligned} \tag{4}$$

本文使用这些 Z_x，V_x 和 Q_x 来确定缩放后的点积注意力.

$$Attention\ (V, Z, Q) = softmax\left(\frac{VZ^N}{\sqrt{d_z}}\right)Q \tag{5}$$

式(5)中，Z_x 和 V_x 投影的点积可以用来量化标记投影的相似度. 考虑到 w_x 和 t_y 分别是第 x 个和第 y 个通过 Z_x 和 V_x 的投影，其点积为：

$$w_x t_y = \cos\ (w_x, y) \parallel w_x \parallel 2 \parallel t_y \parallel 2 \tag{6}$$

表示 t_x 和 w_y 之间方向上的相似性. 此后，矩阵被按元素除以 d_z 的平方根进行缩放，下一阶段将逐行实施 softmax. 因此，矩阵的行值会收敛到 0～1 的数值，从而将其加到 1. 最后，Q_x 将这个结果相乘，得到头部.

在先前的例子“Luxembourg is an amazing country.”中，“Luxembourg”的结果表示如图 3 所示.

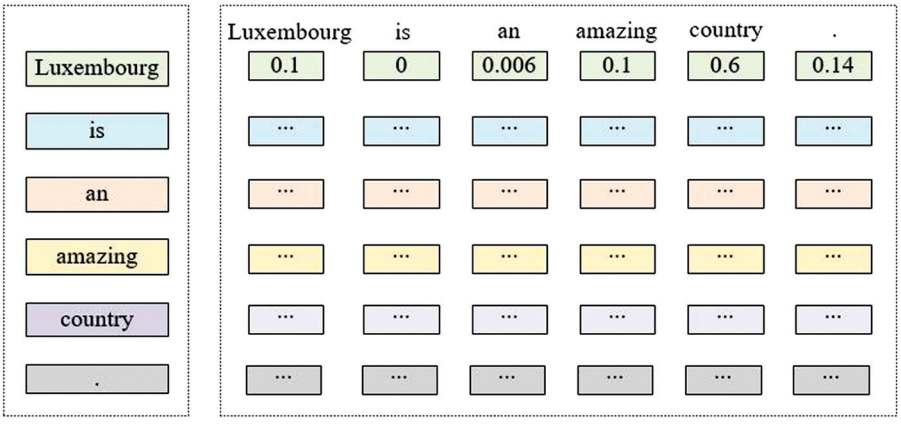

图 3　小数正数和为 1 的先前例子说明

然后将其乘以 q_x，得到图 4 的结果.

图 4 生成了一个矩阵，其中每一行都由通过 q_x 投射的标记表征组成，如图 5 所示.

图 5 特殊的头代表着“Luxembourg”和“country”的结合，我们可以计算出每个编码器块存储这些不同关系所需的 b 次(b 个头). 以前面的例子为第一头.

$$Q_{Luxembourg,1} = 0.1q_{Luxembourg} + 0.0q_{is} + 0.006q_{an} + 0.1q_{beautiful} + 0.6q_{country} + 0q \tag{7}$$

在此阶段，“Luxembourg”被表示为：

$$Concat\ (Q_1, Q_2, Q_3, \cdots, Q_b)M_0 \tag{8}$$

使用 b 个不同的学习投射将 b 个加权的标记表达变化串联起来，得到标记表示. 前馈神经网络(FNN)在位置基准下建立多层结构，每一层的输出通过以下方式进行计算：

$$FFN(i) = \max(0, iM_1 + h_1)M_2 + h_2 \tag{9}$$

式(9)中，M_1 和 M_2 分别是(嵌入维度)×(d_F)和(d_F)×(嵌入维度). 标记向量表示法不会相互“影响”，它等于逐行进行计算，然后将各行堆叠在一个矩阵中. 该步骤的输出尺寸为(输入长度)×(嵌入维度)，且输出被传递到丢弃层、添加层和规范层. 在位置感知前馈网络与丢弃层、添加层和规范层网络之间始终有一个名为子层的层. 子层是具有相同输入和输出的层(多头注意力或前馈)，每个子层之后以 10% 的概率应用丢弃层(子层(i)). 该结果应用于子层输入 i，产生 i + 丢弃层(子层(i)). 在多头注意层中通过一个标记 x 与其他标记的关系，用原始表示法来补充完成. 然后，利用每行的平均值和标准差构建一个标记行级的标准化，从而增加网络的稳定性.

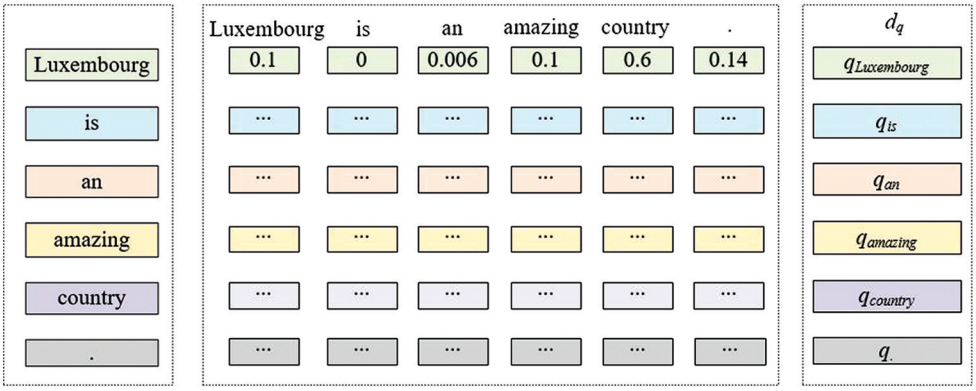

图 4 q_x 的乘积

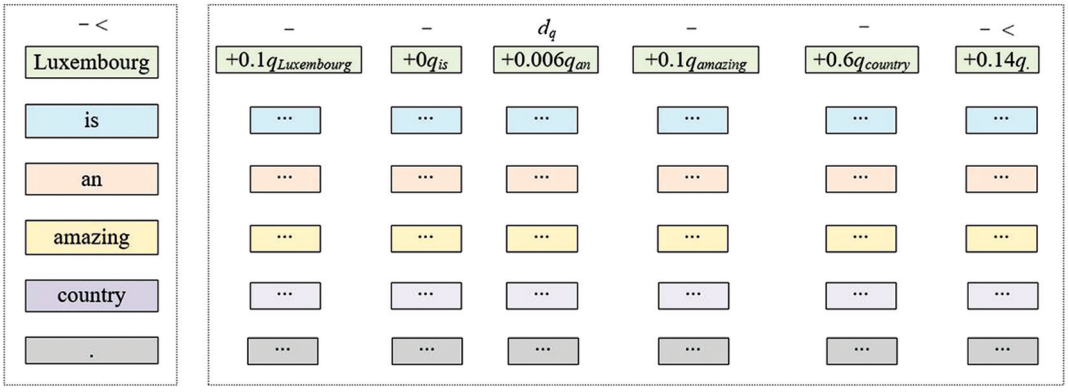

图 5 每行通过 q_x 投射的表征组成的矩阵

2 本文模型构建

M-BERT 与 BERT 具有类似的模型架构，是一种优化的 BERT 变体，可在跨语言 NLP 任务中实现最先进的性能. M-BERT 保留了 BERT 的模型结构，通过执行额外的预处理操作，确保该架构可以适合各种庞大的多语言数据集. 本文所提出的方法由两个阶段组成：第一阶段包括收集和处理与数据集相关的数据；第二阶段重点关注模型架构，主要包括 3 个任务，即训练设置、参数估计和模型训练.

2.1 数据收集与处理

训练 M-BERT 模型的第一步是构建合适的无标签文本语料库. 由于 BERT 是基于 Transformer 的机制，因此需要庞大的语料库才能完成训练. BERT 最初使用从庞大的英语维基百科和图书语料库中检索到的 33 亿个单词，用作训练 M-BERT 模型的输入.

训练 M-BERT 至关重要，因此本文根据原始数据对多语言数据集进行结构化. 多语言数据集提供 3 种变体：原始数据、预处理 V1 和预处理 V2. 我们使用预处理 V1 来预训练模型，使用预处理 V2 进行微调. 整个数据集的确切大小为 39 GB，包括 3 个版本，V1 和 V2 变体，每个版本包含约 2 000 万个观测值. 此外，训练语料库包含约 8.21 亿个单词和 170 万个特殊单词，主要处理不同长度字符串中的文本数据. 每个子语料库都采用了严格的清洁和过滤过程，而且噪音、表情符号、URL 标签、HTML 标签以及所有无意义的内容(例如电话/传真号码、电子邮件地址等)都已被消除. 任何高级语言操作(例如词干提取和词形还原)都没有应用于训练. 由于 BERT 是基于上下文并且具有句法能力，因此通过这些操作(词形还原和词干提取)将单词更改为词根会降低句法能力和上下文词义. 数据集中除英语外的所有外语都被删除，因为它们的出现率低于 0.01%，并且没有任何显著的影响. 与 V2 相比，预处理 V1 中的标点符号并未被删除，因为它有助于识别单词关系. 表 1 总结了从拟合到预训练模型 V2 之前数据集的属性.

表 1　从词义拟合到预训练模型 V2 之前数据集的属性

属性	具体值	属性	具体值
句子总数	19 853 487	表情符号	无
最小语句长度	2	URL 标签	无
最大语句长度	515	HTML 标签	无
总单词数	820 918 413	标点符号	无
特殊单词数	1，711，542	停顿词	有
总字符长度	5 358 612 845	词干	无
噪声	无	词组化	无

2.2　训练设置

单语言 BERT 过程在所有语言中几乎都是相同的. 预训练过程首先根据可用的语料库形成词汇表，然后字节对编码(BPE)主要用于生成有词组和无词组的词汇，正确执行这些步骤可以显著提高模型的性能. 如果句子被标记化(即每个单词分成的部分越少)，模型效果会更好，因为标记化的句子更准确. 本文的预培训过程分为两个基本活动：第一个是掩码语言建模，第二个是下一个句子预测. 我们使用交叉熵损失训练掩码语言模型(MLM)来预测随机掩码标记. 其中，80%的被选标记被替换成专属标记，10%的被选标记被替换成随机标记，10%的被选标记保持不动. 流程如图 6 所示.

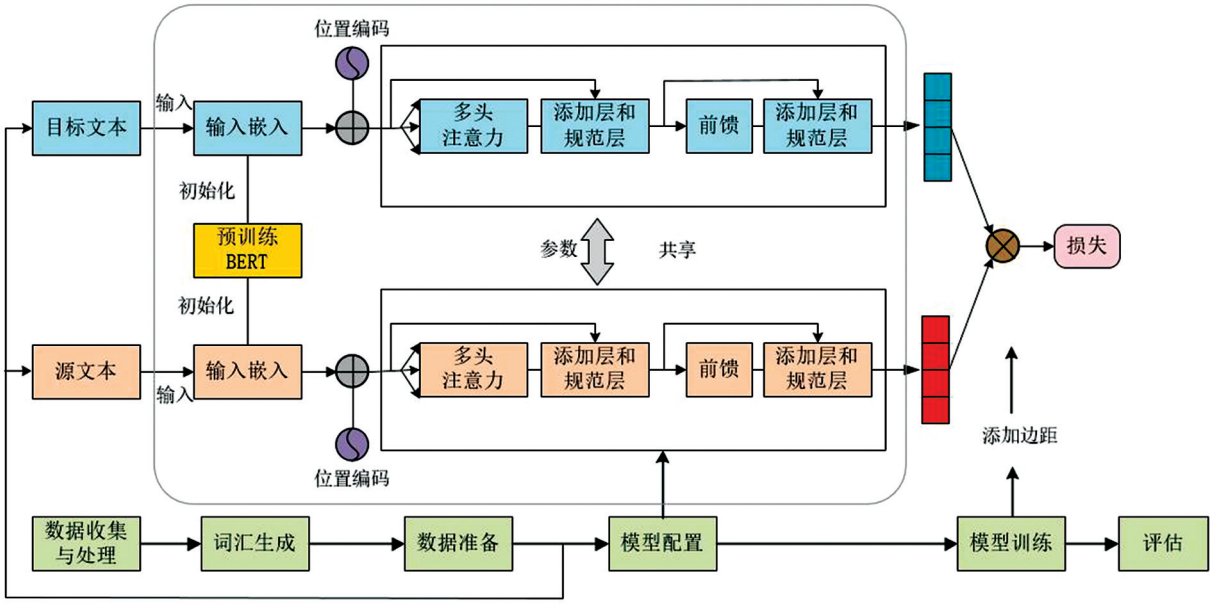

图 6　M-BERT 编码器结构和权重共享机制

该流程中，右边的编码器用来开发M-BERT预训练模型，并使用多语言无监督数据集. 左边的编码器则接受来自预训练模型(右边的编码器)的训练参数，并被用作下游任务的微调.

2.3 参数估计

模型设置对于获得所需的输出至关重要，这也是为何需要仔细选择模型配置值的原因. 前馈层大小，即中间大小为3 072，本文将pad id设置为0. 编码器和池化器的非线性激活函数(函数或字符串)为gelu，用于初始化所有权重矩阵的截断正态初始化器的标准差为0.02. 将use_cache设置为True来指示模型是否必须提供模型的最新键/值注意力.

2.4 模型训练

由于本文模型基于BERT架构，因此在训练设置中主要使用原始的BERT配置和技术. 此任务通过保持接近原始实施的参数，确保配置设置产生与主BERT相同的性能，并且需要经常设置原始实施和超参数值. 本文对所需的模型进行了修改，将模型设置为12个编码器块、768个维度和12个注意力头，包含102 k的庞大词汇量，几乎是原始BERT的3倍. 这种设置可以使得模型更加稳健，并且在计算上更具挑战性. 本文模型的基础建立在拥抱面变压器版本4.2.2之上. 我们将值1e-12设置为分母，以便在归一化层中保持数值稳定性. 由于向后传递速度较慢，本文选择梯度检查点为false节省内存. 对于注意力概率以及嵌入、编码器和池化器中的所有全连接层，我们将层下降率定为0.1，并使用自适应矩估计(ADAM)优化器对目标函数进行了优化，这非常适合涉及大量数据或参数的情况，与M-BERT一样. 选择学习率为1e-6，$\beta_1=0.900$，$\beta_2=0.999$，采用1e-6ε保持数值稳定性. 预训练全部在Google Cloud TPU V3上进行，用时120 h.

3 实验设计

3.1 实验设置与评估指标

使用4个标准评估指标(准确率、精确率、召回率和$F1$值)来衡量本文模型的性能质量. 为了进行计算，配备了Intel i9-CPU、64G-RAM和单个NVIDIA GeForce RTX3070 GPU的PC机.

3.2 数据集

(1) JRC-Acquis多语言数据集

JRC-Acquis数据集是一个较小的数据集，包含20种语言的并行文档. 该数据集与EU-RLEX57K数据集重叠并包含其他文档，使用EuroVoc的描述符进行标记. 本文实验选择了英语、法语和德语的文件.

(2) EURLEX57K多语言数据集

本文收集了与EURLEX57K数据集中文档相同的德语和法语文档，使用原始EURLEX57K数据集中的CELEX ID将数据分为训练集、开发集和测试集. 并行语料库中的文档，按照EURLEX57K数据集的分割方式进行. 因此，最终数据集包含3种语言(英语、法语和德语)的并行文本.

在所有这些实验中，本文使用IRC Acquis英语训练数据模型，并使用法语和德语测试集进行测试.

3.3 结果评估

3.3.1 不同模型对比评估

将本文模型与其他3种自然语言处理模型(Alpaca、Transformer-XH、XLM-RoBERTa)使用JRC-Acquis多语言数据集的英语部分在法语和德语上并行测试，联合训练方案的结果如表2所示.

表3为4种NLP模型在联合训练方案中使用EURLEX57K多语言数据集的英语部分，在法语和德语测试集上训练的结果.

由表2和表3法语和德语测试集上的结果可知，在联合训练方案中本文模型具有最高性能，最高达到96.2%的准确率. 然而，当使用联合训练方案时，法语和德语的结果存在一定的差异，表明相对于从英语表示迁移到德语表示，多语言模型更擅长从英语表示转移到法语表示.

表 2　JRC-Acquis 多语言数据集上的评估结果　　%

语言	模型	准确率	精确率	召回率	*F*1 值
法语	Alpaca	84.3	78.5	74.2	61.8
法语	Transformer-XH	89.4	81.2	79.7	70.2
法语	XLM-RoBERTa	91.7	88.7	80.5	88.3
法语	本文模型	95.9	93.6	89.4	91.7
德语	Alpaca	83.2	77.4	73.1	60.7
德语	Transformer-XH	88.3	80.1	78.6	69.1
德语	XLM-RoBERTa	90.6	87.6	79.4	87.2
德语	本文模型	95.7	92.5	88.3	90.6

表 3　EURLEX57K 多语言数据集上的评估结果　　%

语言	模型	准确率	精确率	召回率	*F*1 值
法语	Alpaca	86.5	80.7	76.4	64.0
法语	Transformer-XH	90.6	83.4	81.9	72.4
法语	XLM-RoBERTa	92.9	90.9	82.7	90.5
法语	本文模型	96.2	95.8	91.6	93.9
德语	Alpaca	85.4	79.6	75.3	62.9
德语	Transformer-XH	89.5	82.3	80.8	71.3
德语	XLM-RoBERTa	91.8	89.8	81.6	89.4
德语	本文模型	96.0	94.7	90.5	92.8

3.3.2　模型性能评估

图 7 显示了本文模型在不同文档方面的表现，除了在中等文档中观察到的 *F*1 值分数较低外，总体预测得分普遍较高，且小文档与大文档之间的得分情况比较相似. 所得分数表明，文件大小作为一个因素与性能质量无关，意味着文档大小只影响计算性能. 准确率得分呈现的平均值为 93%，这是因为该模型具有较高的 *TN*(真阴性)比率，而不是 *TP*(真阳性)比率.

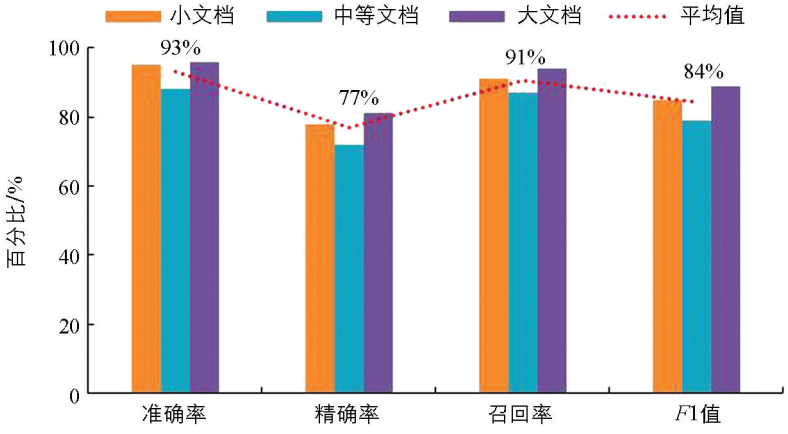

图 7　本文模型对不同文档的处理性能

如模型构架中所述，本文使用向量参数减少处理大文档时的计算成本，并进行了多次实验来分析模型的质量和计算性能，如表 4 所示. 结果显示，本文模型在 *F*1 值指标上优于所有先进的 NLP 处理模型；其次是 XLM-RoBERTa 模型，它使用最少的数据进行训练，但却体现出了仅次于最佳模型的性能. 同等实验

条件下，就执行时间而言，本文模型的预计计算时间为975.26 s，比XLM-RoBERTa模型的计算时间减少了199.59 s，比Transformer-XH模型的计算时间减少了420.46 s，比Alpaca模型的计算时间减少了900.17 s. 实验结果表明，本文模型不管在质量性能还是计算性能上都具有最佳表现.

表4 不同模型的质量及计算性能比较

模型	语言	$F1$值	预计时间/s
Alpaca	多种语言	0.514	1 875.43
Transformer-XH	多种语言	0.637	1 395.72
XLM-RoBERTa	多种语言	0.764	1 174.85
本文模型	多种语言	0.864	975.26

3.3.3 人工评估

为了进一步评估本文模型的性能，本文通过对54个文本文档进行人工评估来进行定性分析(表5). 这些文档从16 281份文档中随机抽取，请3位语言学专家使用五点李克特量表根据3个标准评估这些文档. 每位专家都获得18份样本文档，以便与3份原始文档(黄金标准参考文献)进行比较. 每个样本中的平均单词数约为115个.

评估的3个标准为：① 充分性(原文含义的保留程度)，语境合理性(上下文语句是否健全合理)和语法性(生成语句的正确性). 根据双尾独立样本t检验，在$p<0.05$时，带有+的基础预训练模型与本文的++模型有显著区别. 星号*表示最佳结果.

表5 不同模型的人工评估结果

模型	充分性	语境合理性	语法性	平均值	p值(t值)
Alpaca	1.89	2.65	2.23	2.26	0.003 2(6.36)+
Transformer-XH	2.78	3.17	3.56	3.17	0.002 5(7.03)+
XLM-RoBERTa	3.45	3.74	3.24	3.48	0.468 5(0.81)
本文模型	4.16*	3.66	3.83*	3.88	++

表5的实验结果证实，本文模型在3个标准评估中均取得了不错的结果，特别是在充分性和语法性上获得了最高分，分别是4.16和3.83，平均分数为3.88，优于所有其他对比模型. 该实验还证明，为跨语言NLP任务微调一个全面的预训练多语言模型(如本文模型)比使用现有的NLP处理模型更直接，并且可以产生更好的性能.

4 结论

跨语言NLP任务面临着语言差异和数据稀缺的挑战. 为了解决这些挑战，研究者们提出了一系列方法，包括基于迁移学习和基于神经网络的方法. 本文结合迁移学习和深度学习模型，提出一种新的方法来提高跨语言理解效果. 首先，构建BERT模型；然后，通过一系列预训练操作完成M-BERT模型构建，并在目标任务上进行微调；最后，为了将所学到的知识应用于目标语言任务并提高目标语言的理解能力，本文采用迁移学习策略，将M-BERT作为特征提取器用于源语言领域和目标语言领域之间的特征转换. 这种迁移学习的方式能够在不同语言之间实现知识共享和迁移，提高目标语言任务的性能. 在实验部分，我们选择多个常见的跨语言NLP任务，并与其他先进算法进行了比较. 实验结果表明，本文提出的方法在这些任务上优于所有对比的先进算法. 通过对比实验和定量评估实验，验证了本文方法的可行性和优越性，该方法能够将源语言上学到的语义知识迁移到目标语言上，从而弥补目标语言中的数据稀缺性. 下一阶段将深入研究在M-BERT中的多层次语言提取编码，以便正确理解和分析该模型对不同信息的获取. 此外，还将评估其他BERT架构，如DeeBERT、MobileBERT、SpanBERT和AlBERT，进一步探究更先进的NLP处理模型，从而更好地应用于跨语言研究领域.

参考文献：

[1] KHURANA D，KOLI A，KHATTER K，et al. Natural Language Processing：State of the Art，Current Trends and Challenges [J]. Multimedia Tools and Applications，2023，82(3)：3713-3744.

[2] 赵京胜，宋梦雪，高祥，等. 自然语言处理中的文本表示研究 [J]. 软件学报，2022，33(1)：102-128.

[3] 张博，董瑞海. 自然语言处理技术赋能教育智能发展——人工智能科学家的视角 [J]. 华东师范大学学报(教育科学版)，2022，40(9)：19-31.

[4] 江洋洋，金伯，张宝昌. 深度学习在自然语言处理领域的研究进展 [J]. 计算机工程与应用，2021，57(22)：1-14.

[5] 陆金梁，张家俊. 基于多语言预训练语言模型的译文质量估计方法 [J]. 厦门大学学报(自然科学版)，2020，59(2)：151-158.

[6] 鲍小异，姜晓彤，王中卿，等. 基于跨语言图神经网络模型的属性级情感分类 [J]. 软件学报，2023，34(2)：676-689.

[7] SORIN V，BARASH Y，KONEN E，et al. Deep Learning for Natural Language Processing in Radiology-Fundamentals and a Systematic Review [J]. Journal of the American College of Radiology：JACR，2020，17(5)：639-648.

[8] WU L F，CHEN Y，SHEN K，et al. Graph Neural Networks for Natural Language Processing：a Survey [J]. Foundations and Trends © in Machine Learning，2023，16(2)：119-328.

[9] ZHANG W E，SHENG Q Z，ALHAZMI A，et al. Adversarial Attacks on Deep-Learning Models in Natural Language Processing：a Survey [J]. ACM Transactions on Intelligent Systems and Technology，11(3)：1-41.

[10] VÁZQUEZ R，RAGANATO A，CREUTZ M，et al. A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation [J]. Computational Linguistics，2020，46(2)：387-424.

[11] VERMA V K，PANDEY M，JAIN T，et al. Dissecting Word Embeddings and Language Models in Natural Language Processing [J]. Journal of Discrete Mathematical Sciences and Cryptography，2021，24(5)：1509-1515.

[12] GU Y，TINN R，CHENG H，et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing [J]. ACM Transactions on Computing for Healthcare，2021，3(1)：1-23.

[13] 岳增营，叶霞，刘睿珩. 基于语言模型的预训练技术研究综述 [J]. 中文信息学报，2021，35(9)：15-29.

[14] MOON J，PARK G，JEONG J. POP-ON：Prediction of Process Using One-Way Language Model Based on NLP Approach [J]. Applied Sciences，2021，11(2)：864-882.

[15] AGGARWAL A，CHAUHAN A，KUMAR D，et al. Classification of Fake News by Fine-Tuning Deep Bidirectional Transformers Based Language Model [J]. ICST Transactions on Scalable Information Systems，2018，27(7)：163973.

[16] PELICON A，PRANJIĆ M，MILJKOVIĆ D，et al. Zero-Shot Learning for Cross-Lingual News Sentiment Classification [J]. Applied Sciences，2020，10(17)：5993-6013.

[17] KIM B，YANG Y，PARK J S，et al. Machine Learning Based Representative Spatio-Temporal Event Documents Classification [J]. Applied Sciences，2023，13(7)：4230-4241.

[18] 贾明华，王秀利. 基于 BERT 和互信息的金融风险逻辑关系量化方法 [J]. 数据分析与知识发现，2022，6(10)：68-78.

责任编辑　夏娟