计算机科学
COMPUTER SCIENCE

基于预训练模型的多音字消歧方法

高贝贝, 张仰森

引用本文

高贝贝, 张仰森. 基于预训练模型的多音字消歧方法[J]. 计算机科学, 2024, 51(11): 273-279.

GAO Beibei, ZHANG Yangsen. Polyphone Disambiguation Based on Pre-trained Model[J]. Computer Science, 2024, 51(11): 273-279.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230900006

基于预训练模型的多音字消歧方法

高贝贝　张仰森
北京信息科技大学智能信息处理研究所　北京 100192
(beibgao@163.com)

摘　要　字音转换是中文语音合成系统(Text-To-Speech,TTS)的重要组成部分,其核心问题是多音字消歧,即在若干候选读音中为多音字选择一个正确的发音。现有的方法通常无法充分理解多音字所在词语的语义,且多音字数据集存在分布不均衡的问题。针对以上问题,提出了一种基于预训练模型 RoBERTa 的多音字消歧方法 CLTRoBERTa(Cross-lingual Translation RoBERTa)。首先联合跨语言互译模块获得多音字所在词语的另一种语言翻译,并将其作为额外特征输入模型以提升对词语的语义理解,然后使用判别微调中的层级学习率优化策略来适应神经网络不同层之间的学习特性,最后结合样本权重模块以解决多音字数据集的分布不均衡问题。CTLRoBERTa 平衡了数据集的不均衡分布带来的性能差异,并且在 CPP(Chinese Polyphone with Pinyin)基准数据集上取得了 99.08%的正确率,性能优于其他基线模型。

关键词:多音字消歧;预训练模型;字音转换;跨语言互译;层级学习率;样本权重
中图分类号　TP391

Polyphone Disambiguation Based on Pre-trained Model

GAO Beibei and ZHANG Yangsen
Institution of Intelligent Information Processing,Beijing Information Science and Technology University,Beijing 100192,China

Abstract　Grapheme-to-phoneme conversion(G2P) is an important part of the Chinese text-to-speech system(TTS). The key issue of G2P is to select the correct pronunciation for polyphonic characters among several alternatives. Existing methods usually struggle to fully grasp the semantics of words that contain polyphonic characters,and fail to effectively handle the imbalanced distribution in datasets. To solve these problems,this paper proposes a polyphone disambiguation method based on the pre-trained model RoBERTa,called cross-lingual translation RoBERTa(CLTRoBERTa). Firstly,the cross-lingual translation module generates another translation of the word containing the polyphonic character as an additional input feature to improve the model's semantic comprehension. Secondly,the hierarchical learning rate optimization strategy is employed to adapt the different layers of the neural network. Finally,the model is enhanced with the sample weight module to address the imbalanced distribution in the dataset. Experimental results show that CLTRoBERTa mitigates performance differences caused by uneven dataset distribution and achieves a 99.08% accuracy on the public Chinese polyphone with pinyin(CPP) dataset,outperforming other baseline models.

Keywords　Polyphone disambiguation,Pre-trained model,Grapheme-to-phoneme conversion,Cross-lingual translation,Hierarchical learning rate,Sample weight

1　引言

随着人工智能迅猛发展,语音合成系统(TTS)的应用领域不断扩展,涵盖了智能家居、语音助手等多个领域。作为一种象形文字,汉语在转化为语音之前需要经历汉字序列到拼音序列的转换,这一过程被称为"字音转换"(Grapheme-to-Phoneme conversion,G2P),字音转换在 TTS 中具有重要地位。

在汉语中,拼音的组成为声母、韵母和声调,本文将拼音的表示格式定为"声母＋韵母＋声调对应的数字"(其中"轻声"用数字"5"表示),如 liǎo 表示为 liao3,轻声 le 表示为 le5。汉语是一种表意文字,其中包含大量多音字,这些多音字在不同语境中可能有不同的含义和发音。如表 1 所列,在"他对问题的了解更加透彻"中,"了"的拼音为"liao3",所在词语"了解"的含义可以用英文翻译"understand"来体现;而在"他除了写作没有别的爱好"中,"了"的拼音为"le5",所在词语"除了"的含义可以用英文翻译"except"来体现,"角"字同理。可以发现,同一个字在不同语境下可能有不同的拼音,其所在词语的含义也有所不同,这一点可以通过其英文翻译直观体现出来。在进行字音转换时,一旦多音字的拼音选择出

到稿日期:2023-09-04　返修日期:2024-02-08
基金项目:国家自然科学基金(62176023)
This work was supported by the National Natural Science Foundation of China(62176023).
通信作者:张仰森(zhangyangsen@163.com)

现错误，则可能会导致对句子意义的误解。因此，为多音字选择正确的拼音，即多音字消歧，成为 TTS 的一项重要任务。

表 1 不同语境中的多音字发音和英语翻译

Table 1 Pronunciation of polyphonic characters and their English translations in different contexts

包含多音字的句子及其分词	多音字	多音字的拼音	多音字所在分词的翻译
他/p 对/p 问题/n 的/p 了解/v 更加/ad 透彻/a	了	liao3	understand
他/p 除了/p 写作/v 没有/v 别的/p 爱好/n	了	le5	except
他/p 可以/v 从/p 新奇的/a 角度/n 看待/v 问题/n	角	jiao3	angle
他/p 很/ad 喜欢/v 这个/p 角色/n	角	jue2	role

近年来，深度学习在许多自然语言处理(NLP)任务上取得了显著的成果。其中，预训练模型已成为研究的热点，在多音字消歧任务中也被广泛应用。首先，在应用预训练模型的方法如 BERT 中，面临着一些挑战。BERT 以字符为基本单元，无法充分理解汉语词汇的词语级特征，包括词语词性特征和词语语义特征。在词语词性特征方面，研究人员进行了大量实验，其中应用最广泛的是词性标注。然而，即使利用了词性标注，仍然无法很好地对多音字进行词义挖掘。如表 1 所列，多音字“角”的读音 1(jiao3)和读音 2(jue2)所在词语“角度”和“角色”的词性相同，均为名词(n)，但语义却完全不同；在词语语义特征方面，Bruguier 等[1]和 He 等[2]讨论了如何利用外部知识来扩充信息，但是检索过程相对复杂。其次，在预训练模型与传统神经网络(如 LSTM，CNN 等)结合的方法中，大多数方法都忽视了模型不同层级之间的信息流动，即层间特性。神经网络的不同层可以捕获不同级别的句法和语义信息，应当对其设置不同的参数。最后，实际应用中，多音字存在严重的不均衡问题，这不仅存在于不同多音字之间，还存在于单个多音字的不同发音之间。例如，在 CPP 数据集中，“了”字有 202 条数据，“跄”字有 8 条数据；而在“了”字中，“le5”有 200 条数据，“liao3”有 2 条数据。

为了解决上述问题，本文提出了一种基于预训练模型的多音字消歧方法 CLTRoBERTa。其采用 RoBERTa 作为编码器，结合了跨语言互译模块，引入多音字所在词语的另一种语言翻译，以挖掘多音字的词语语义信息。此外，针对 CLTRoBERTa 的层间特性，本文为不同层级应用不同的参数，设计了层级学习率优化策略。最后，为了解决多音字分布不均衡的问题，引入了样本权重参数，根据数据集中的不同多音字和同一多音字的不同拼音的相对比例来衡量每一条样本的权重。本文的主要贡献包括：

1)首先，将跨语言互译用于多音字消歧任务，深度挖掘多音字所在词语的语义信息。模型联合跨语言互译模块输出的不同语言的信息，能够更准确地预测其发音。

2)针对多音字消歧任务的独特特点以及模型的层级特性，采取了层级学习率优化策略。不同层级的神经网络在处理句法和语义信息时具有差异，因此针对 CLTRoBERTa 各个层级的参数进行了优化调整。通过为不同层级设定不同参数，实现了对模型的精准调控。

3)为了解决数据集中存在的“字与字”以及“音与音”之间的不平衡问题，引入了样本权重参数。这一策略使得模型能够在训练过程中更加关注少数类别，可以有效地防止不同多音字之间以及同一多音字不同发音之间的不均衡数据分布造成的不良影响。

2 相关工作

在以往的研究中，多音字消歧方法通常可以被分为 3 个主要类别：基于规则的方法、基于统计的方法，以及基于深度学习的方法。基于规则的方法主要依赖于人工构建的发音词典和规则[3-5]，然而构建词典和规则需要耗费大量的人力和时间，并且随着规则数量的不断增加，可能会出现多音字在规则匹配过程中产生冲突的问题。总的来说，这种方法在处理之前未出现过的多音字时存在一定的局限性，泛化能力相对较差。

为了应对这些问题，基于统计的方法逐渐受到关注，例如决策树[6]、最大熵模型[7]等被应用于多音字消歧任务。尽管在这些方法中机器可以自动学习特征，但在构建特征工程时仍需要充足的语言学知识，因此其建模成本较高。此外，这类方法无法充分挖掘语义信息，通常借助经验阈值来作出决策，这从根本上削弱了模型的性能和灵活性。

近年来，随着深度学习和其他自然语言处理技术的不断发展，学者开始运用深度学习、预训练模型等方法，挖掘不同语境下多音字发音的规律。例如，Shan 等[8]和 Cai 等[9]将多音字消歧视为分类任务，并采用 BiLSTM 结合词性标注等附加信息，取得了显著的成果。此外，Cai 等[9]将多种特征嵌入作为输入，并使用 BiLSTM 和 Word2Vec 对多种特征进行融合。Zhang 等[10]将掩码模型应用于多音字消歧问题以解决模型预测当前多音字以外读音的问题，并采用经过修正的焦点损失函数以应对样本不平衡问题。Gehring 等[11]提出将卷积 seq2seq 模型在包含多音字的音频生成文本上进行训练。随着 Transformer 的发展，Zhang[12]将 FLAT 方法迁移到多音字消歧领域，与预训练模型相结合取得了显著效果。同时，预训练模型如 BERT[13]，RoBERTa[14]，Albert[15]，Electra[16]等，通过充分利用未标记的文本数据，在句子理解方面取得了重要进展。预训练模型通过强化字符表示能力，能够与传统的神经网络(如 LSTM 等)相结合，以支持各种下游任务。例如 Dai 等[17]将 BERT 与传统神经网络分类器相结合，首次将 BERT 应用于多音字消歧问题；Zhang 等[18]通过对 BERT 的词汇表进行调整和丰富，提出了一种多音字消歧 BERT；Shi 等[19]基于 Electra 模型提出了一种用于多音字消歧的半监督学习框架。这些方法均获得了比以往更优的效果，凸显了预训练模型在多音字消歧领域的出色潜力。

3 多音字消歧模型 CLTRoBERTa

本文提出的多音字消歧模型 CLTRoBERTa 如图 1 所示，多音字消歧模块作为模型的主要部分，融合了跨语言互译模块、层级学习率优化策略和样本权重计算模块。模型的输入包括原始句子、用于计算样本权重的多音字掩码(Phoneme Mask)，以及多音字的位置标识(Position id)。经过 RoBER-

Ta 编码后，输出 pool output 和 sequence output。将二者进行拼接后输入跨语言互译模块，产生多音字所在词语的英语翻译。随后，将多音字及其掩码输入样本权重计算层，根据多音字权重和发音权重，计算出当前样本的权重以衡量样本的相对比例，从而实现对分布不均衡的样本的处理。另一方面，pool output 和 sequence output 也被输入多音字预测模块的分类层。值得强调的是，利用 Softmax 函数进行预测时，同时考虑了当前样本的权重和多音字掩码，掩码的作用是防止模型预测该多音字之外的读音。最终，模型输出对应的预测读音。本章将深入介绍模型的每一个组成部分。

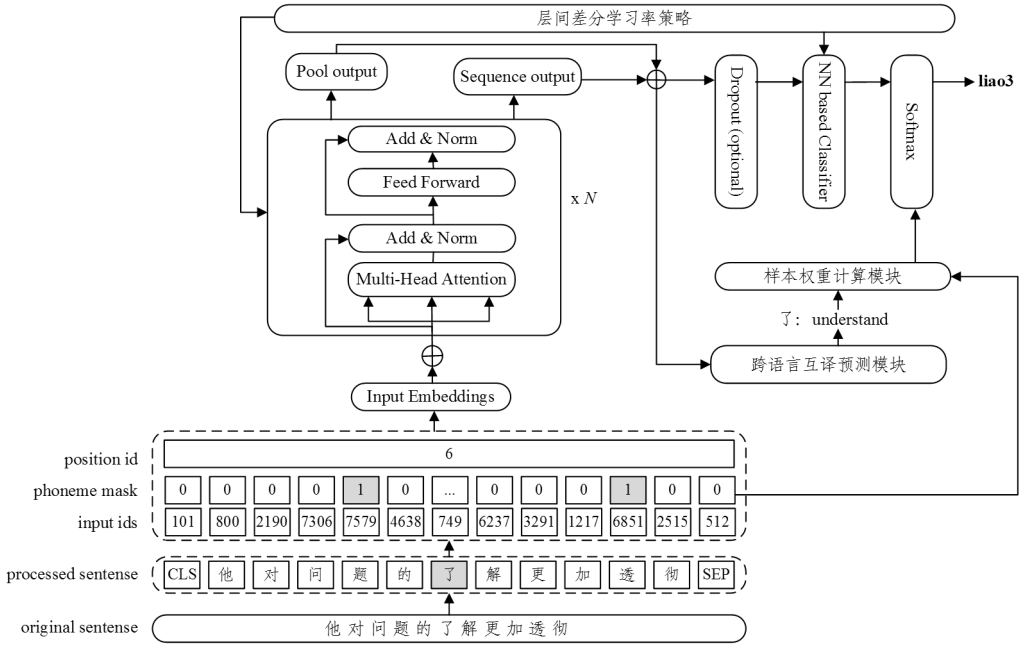

图 1 模型结构

Fig. 1 Model structure

3.1 跨语言互译

跨语言互译被应用于词义消歧任务中进行特征提取。目标语言中的词语往往能使用另一种语言的词语来解释和标注[20]。受这一思想的启发，本文认为多音字消歧本质上就是对多音字所在词语的词义消歧，由此设计了一个跨语言互译模块，如图 2 所示。其主要目标是使用目标语言的词语来解释和标注源语言的词语。该模块在多音字消歧任务中被设计为一种特征提取方法，以增强模型对多音字所在词语的语义理解。

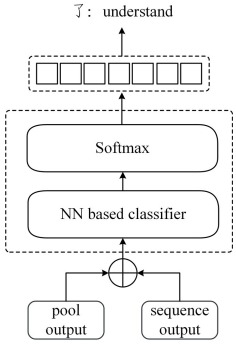

图 2 跨语言互译模块

Fig. 2 Cross-lingual translation module

通过引入跨语言互译模块，模型可以更加准确地捕捉多音字所在词语在不同语境下的语义信息。该模块能够将目标语言的语义信息与原始语言中的多音字关联起来，从而为多音字的消歧提供更丰富的背景知识。以“了”字为例，其对应的跨语言互译标签有 understand(了解)、except(除了)等。跨语言互译模块在训练过程中与多音字消歧模块共享同一个解码器(RoBERTa)，模型实现了两个模块之间的信息交互和共享：跨语言互译模块的结果可以作为额外特征辅助多音字消歧模块的训练，从而进一步提升模型的语义理解能力。

该模块和多音字消歧模块均采用交叉熵损失函数，如式(1)所示：

$$L=-\sum_{i=1}^{n} y_{\text{true}}\log(y_{\text{pred}}) \tag{1}$$

其中，n 表示类别的个数。

此外，模型的全局损失如式(2)所示：

$$L_{\text{global}}=\alpha_1 L_{\text{poly}}+\alpha_2 L_{\text{en}} \tag{2}$$

其中，α_1 和 α_2 为损失函数因子，通过调整多音字消歧模块的损失函数 L_{poly} 和跨语言互译模块的损失函数 L_{en} 之间的关系来调整模型对两个模块的关注度。第 4 章将详细讨论损失函数因子 α_1 和 α_2 的取值。

3.2 层级学习率优化策略

模型的不同层级可以捕获不同类型的信息，因此在对预训练模型微调时区分不同层级的特性非常重要。针对这一特征，引入判别微调策略[21]。该策略针对模型结构中的不同层级设计了不同的参数，如式(3)所示，本文所采用的层级学习率优化正是基于这一思想。该策略旨在有效地平衡模型在不同层级之间的参数更新速度，从而更好地捕捉不同层级特征的信息，提升多音字消歧任务的性能。

$$\theta_t=\theta_{t-1}-\eta\cdot\nabla_\theta J(\theta) \tag{3}$$

其中，η 表示学习率，$\nabla_\theta J(\theta)$ 是相对于目标函数的梯度。在判别微调中，$\theta=\{\theta^1,\cdots,\theta^L\}$，$L$ 表示模型的层数，θ^l 中包含模型第 l 层的参数。

在多音字消歧任务中，RoBERTa 模型作为基础模型能够学习到丰富的语义信息，而全连接层则用于将 RoBERTa 的特征映射到最终的多音字预测结果。不同层级使用不同参数时具有不同的学习特性，因此将相同的学习率应用于所有层级可能导致性能下降。“层级学习率优化策略”的引入解决了这一问题，其允许对模型不同层级设置不同的学习率。

模型在 RoBERTa 内部采用学习率层级线性衰减的方式，全连接层另外指定学习率，如图 3 所示。在神经网络中，不同层级设置不同的学习率可以使网络更有针对性地进行参数更新。通常，底层的参数（靠近输入）可能需要较低的学习率，而顶层的参数（靠近输出）可能需要较高的学习率。这样可以平衡不同层级之间的训练速度。本文将模型的层级学习率表示为 $\eta=\{\eta^1,\cdots,\eta^L\}$，层级线性衰减如式(4)所示：

$$\eta^{k-1}=\zeta\cdot\eta^k \tag{4}$$

其中，$\zeta<1$，为衰减因子，它控制了模型的层级学习率衰减速度。第 4 章将详细讨论衰减因子 ζ 的取值。

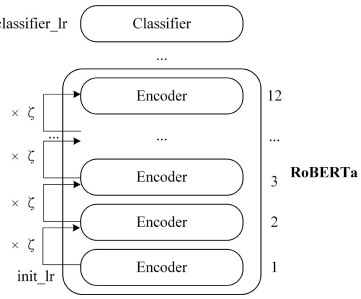

图 3 层级学习率衰减

Fig. 3 Hierarchical learning rate decay

通过为 CLTRoBERTa 的不同层级分别设定适当的学习率，该模块能够实现以下几个方面的作用：

1）提升特征学习效率。通过为 RoBERTa 模型的底层设置相对较低的学习率，模块能够更充分有效地学习底层特征，使其更快地适应特定任务。而对于全连接层等较浅层参数，适度提高学习率有助于稳定模型的收敛（注意：只需要在初始学习率的基础上适度提高，太高会使模型难以收敛）。这种策略旨在在不同层级上实现平衡，以优化整体性能。

2）优化特征表示。针对模型的不同层级，可以通过不同的学习率设置，更精细地调整模型的参数更新，从而更好地捕获多音字的上下文和语义信息，提高模型对多音字的判断准确度。

3）避免过拟合。通过提高全连接层等较浅层的学习率，“层级学习率优化策略”有助于防止模型在训练过程中对少量数据过于敏感，减少过拟合的风险。

3.3 样本权重计算

3.3.1 掩码 Softmax

拼音集指多音字数据集中所有多音字的所有发音的汇总，用 N 表示总发音数量。在每个多音字的预测过程中，模型都会有一个候选拼音集（候选集）。候选集包括了正在预测的多音字的所有发音选项，例如，“了”字有两个发音选项“le5”和“liao3”，因此其候选集大小为 2。对于每个多音字而言，其候选集的数量远远少于整个拼音集的发音数量 N。在这种情况下，如果在模型的预测中使用标准的 Softmax 函数，那么拼音集中的每个拼音都会被分配一个非零的概率，而实际上，模型只需关心多音字的候选集中的拼音。为了解决这个问题，引入了掩码 Softmax 技术[10]，计算方法如式(5)所示。其中，输入 Softmax 的向量表示为 $\boldsymbol{V}=\{v_1,v_2,\cdots,v_k\}$，$v_i$ 表示向量 $\boldsymbol{V}$ 的第 i 个元素；布尔值 $\boldsymbol{m}_i$ 表示掩码向量，用于确定是否遮蔽 Softmax 中对应的元素 v_i。通过这种方法，模型的预测值将被限制在候选集内，例如，“了”字的预测值只会在“le5”和“liao3”中选择，从而解决了模型预候选集以外发音的问题。

$$p_{\text{weighted}}=\frac{\boldsymbol{m}_i\times e^{v_i}}{\sum_{j=1}^{k}\boldsymbol{m}_j\times e^{v_j}} \tag{5}$$

3.3.2 多音字权重和发音权重

在多音字数据分布中，存在着“字与字”以及“音与音”之间的不均衡情况。以 CPP[22]（Chinese Polyphone with Pinyin）数据集为例，有时样本数量在不同多音字之间的差异很大。例如，“了”字的样本数为 202 条，而“跄”字的样本数仅有 8 条。在“了”字的样本中，发音为“le5”的样本有 200 条，而发音为“liao3”的样本仅有 2 条。受到 Gao 等[23]的启发，本文引入了多音字权重和发音权重的概念。在后续的描述中，对涉及的参数进行了如下定义：假设总共有 x 个多音字，第 i 个多音字拥有 n_i 个样本，m_i 个发音，其中第 j 个发音含有 k_j 个样本。针对当前多音字 c 以及当前发音 p，其多音字权重计算如式(6)所示，发音权重计算如式(7)所示：

$$w_c=\max(n_i)\times\frac{(1/n_c)^{\gamma_1}}{\sum_{i=1}^{x}(1/n_i)^{\gamma_1}} \tag{6}$$

$$w_p=m_c\times\frac{(1/k_p)^{\gamma_2}}{\sum_{j=1}^{m_c}(1/k_j)^{\gamma_2}} \tag{7}$$

其中，γ_1 和 γ_2 为样本权重因子，分别用来调整模型对“字与字”和“音与音”不平衡情况的关注程度。第 4 章将详细讨论损失函数因子 γ_1 和 γ_2 的取值。

以“了”字和“跄”字为例，样本数 $n_{了}>n_{跄}$，所以 $w_{了}<w_{跄}$，从而样本较少的“跄”字将获得更大的多音字权重。以“了”字中的两个发音为例，$k_{\text{le5}}>k_{\text{liao3}}$，所以 $w_{\text{le5}}<w_{\text{liao3}}$，从而样本较少的“liao3”的样本将获得更大的发音权重。

3.3.3 样本权重

为了增强模型在训练过程中对样本的针对性关注，本文提出了样本权重模块，旨在为每一条样本分配相应的权重。这个模块的计算方法由 3 个部分组成：掩码、多音字权重，以及发音权重。如式(8)所示：

$$p=p_{\text{weighted}}*w_c*w_p \tag{8}$$

首先，掩码技术有效避免了对候选集外不相关发音的预测。其次，多音字权重加入模型训练过程中，能够更加准确地调整模型在不同多音字之间的关注度。最后，发音权重为多音字的不同发音赋予不同的权重，从而更好地挖掘样本的分布情况。

样本权重模块能够有效地引导模型关注多音字的候选集，同时在训练过程中更加关注少数样本，从而提高模型在多音字消歧任务中的性能和鲁棒性。

4 实验

4.1 实验数据集

本文所用的数据集是 CPP[22] 数据集，该数据集共有 623 个多音字，99 264 条句子样本，每个句子中仅标有一个多音字。如表 2 所列，其中包含 2 个发音的多音字有 553 个，样本条数占总数的 88.2%；包含 3 个发音的多音字有 60 个，样本条数占总数的 10.2%；包含大于等于 4 个发音的多音字有 10 个，样本条数占总数的 1.6%。

表 2 CPP 数据集中不同发音数量的多音字的样本数量

Table 2 Number of sample polyphones with different pronunciations in CPP dataset

发音数量	多音字个数	样本条数	样本条数占比/%
总数	623	99264	100
2	553	87584	88.2
3	60	10162	10.2
大于等于 4	10	1518	1.6

数据集共有 3 个部分：训练集、验证集和测试集，如表 3 所列。其中，训练集有 79 117 条句子样本，验证集有 9 893 条句子样本，测试集有 10 254 条句子样本。

表 3 CPP 数据集分布

Table 3 Distribution of CPP dataset

	总数	训练集	验证集	测试集
句子样本数量	99264	79117	9893	10254
多音字数量	623	623	623	623

4.2 实验设置

实验采用的 RoBERTa-wwm 预训练语言模型的基础架构为 12 层堆叠双向 Transformer，具有 768 个隐藏状态，在训练过程中利用 Adam 优化器进行梯度优化更新参数。实验参数设置如下：输入句子长度 *sequence_len* 设为 32，训练集的 *batch_size* 为 256，测试集的 *batch_size* 为 256，训练的初始学习率为 5×10^{-5}，全连接层的学习率为 1×10^{-4}。为了防止训练过程中出现过拟合现象，使用 dropout 技术时将值设为 0.5。设定模型总共进行 5 000 次迭代，每 10 次迭代之后以 256 的批量对验证集进行一次验证，在此过程中保存在验证集上表现最好的模型。最后使用上述模型对测试集进行测试。下面对第 3 章涉及的超参数因子进行不同取值的实验，并选定使模型效果最佳的取值。

4.2.1 损失函数因子 α_1 和 α_2 的选择

对于损失函数因子的选择，α_1 为多音字预测模块的损失因子，α_2 为跨语言互译模块的损失因子。首先将 α_1 定为 1.0，依次验证 α_2 为 0.01，0.05，0.1，0.5 时的模型性能和跨语言互译模块的性能。如表 4 所列，随着损失因子 α_2 的增加，跨语言互译模块的正确率逐渐提升，这说明模型在训练时随着 α_2 的增加而给予了跨语言互译模块更高的关注度。$\alpha_2=0.1$ 时，跨语言互译模块的正确率为 89.94%，与其最优性能相差 1.45%，但此时模型在多音字预测上取得最好效果 99.08%。因为模型的主要任务是多音字预测，所以最终选择牺牲一部分跨语言预测模块的性能，选定 $\alpha_2=0.1$。

表 4 不同损失函数因子 α_1 和 α_2 的选择

Table 4 Evaluation with different α_1 and α_2

	α_1	α_2	验证集正确率/%	测试集正确率/%	跨语言预测正确率/%
α_2 的选择	1.0	0.01	99.06	99.07	50.15
	1.0	0.05	99.03	98.97	85.65
	1.0	0.1	99.02	**99.08**	89.94
	1.0	0.5	98.97	99.03	**91.39**
α_1 的选择	0.8	0.1	99.02	99.04	**90.67**
	0.9	0.1	99.05	99.03	88.40
	1.0	0.1	99.02	**99.08**	89.94
	1.2	0.1	99.09	98.99	90.44

将 α_2 确定后，对 α_1 进行调整，依次验证 α_1 为 0.8，0.9，1.0，1.2 时的模型性能。如表 4 所列，当 $\alpha_1=1.0$ 时，模型取得最好效果 99.08%。最终确定损失函数因子 $\alpha_1=1.0$，$\alpha_2=0.1$。

4.2.2 学习率衰减因子 ζ 的选择

对于学习率衰减因子 ζ 的选择，依次验证 ζ 为 0.90，0.95，0.98 时的模型性能。如表 5 所列，当 $\zeta=0.95$ 时，模型取得最好效果 99.08%。最终确定学习率衰减因子 $\zeta=0.95$。

表 5 不同学习率衰减因子 ζ 的选择

Table 5 Evaluation with different ζ

(%)

ζ	验证集正确率	测试集正确率
0.90	99.03	99.03
0.95	99.02	**99.08**
0.98	98.99	98.94

4.2.3 样本权重因子 γ_1 和 γ_2 的选择

对于样本权重因子的选择，γ_1 为多音字权重因子，γ_2 为发音权重因子。参考 Zhang 等[24] 的重加权参数选择，权重因子大于 1 时取得的效果较好，因此首先将 γ_1 定为 1.0，再依次验证 γ_2 为 1.0，1.2，1.4 时的模型性能。如表 6 所列，当 γ_2 为 1.0 时，模型取得最好效果 99.08%，所以选定 γ_2 为 1.0。

表 6 不同样本权重因子 γ_1 和 γ_2 的选择

Table 6 Evaluation with different γ_1 and γ_2

	γ_1	γ_2	验证集正确率/%	测试集正确率/%
γ_2 的选择	1.0	1.0	99.02	**99.08**
	1.0	1.2	99.07	99.02
	1.0	1.4	99.06	99.06
γ_1 的选择	1.0	1.0	99.02	**99.08**
	1.2	1.0	99.03	99.02
	1.4	1.0	99.07	99.03

将 γ_2 确定后，对 γ_1 进行调整，依次验证 γ_1 为 1.0，1.2，1.4 时的模型性能，如表 6 所列，当 $\gamma_1=1.0$ 时，模型取得最好效果 99.08%。最终确定样本权重因子 $\gamma_1=1.0$，$\gamma_2=1.0$。

4.3 消融实验

为了深入分析本文模型各个模块的有效性，本文在 CPP 数据集上对模型进行了消融实验。为了避免其他因素的影响，所有消融实验的参数均相同。具体实验如下：

1）RoBERTa：仅使用 RoBERTa 预训练模型。

2）RoBERTa＋跨语言互译：在 RoBERTa 的基础上引入跨语言互译模块，目的是证明跨语言互译模块的有效性。

3）RoBERTa＋跨语言互译＋层级学习率策略：在 2）的基

础上引入层级学习率策略，目的是证明层级学习率策略的有效性。

4)RoBERTa+跨语言互译+层级学习率策略+样本权重：本文模型 CLTRoBERTa。在 3)的基础上引入样本权重模块，目的是证明样本权重模块的有效性。

如表 7 所列，每一个模块的加入都会使模型在测试集上的效果得到提升，这充分展示了本文模型每个模块的有效性。

表 7 消融实验

Table 7 Ablation experimental results

(%)

模型组合	验证集正确率	测试集正确率
RoBERTa	98.99	99.01
RoBERTa+跨语言互译	99.03	99.02
RoBERTa+跨语言互译+层级学习率策略	99.04	99.07
RoBERTa+跨语言互译+层级学习率策略+样本权重(CLTRoBERTa)	99.02	**99.08**

4.4 实验结果和分析

针对 CPP 数据集分布不均衡的特点，如表 2 所列，本文按照多音字含有的读音数量对其进行分类，分为类别 1(含有 2 个读音的多音字)、类别 2(含有 3 个读音的多音字)和类别 3(含有 4 个及以上读音的多音字)，对不同类别进行实验。

CPP 数据集存在一定的不均衡性，在样本数量上，类别 3 仅占整个数据集的 1.6%，类别 2 占 10.2%，类别 1 最多，占 88.2%。如图 4 所示，样本数量过少在训练过程中存在很大劣势，不采用样本权重策略，三者的测试集准确率相差较大；而使用采用了样本权重的 CLTRoBERTa 可以将占比为 1.6%的类别 3 测试正确率提升至和占比为 10.2%的类别 2 基本相当的水平，且这两类样本的测试准确率又均比不采用样本权重时有所提升；同时，因为模型兼顾少数类别，所以占比最大的类别 1 测试准确率稍有下降。总的来说，从图 4 可以看出，本文缩小了模型在少数类别和多数类别上的性能差异。

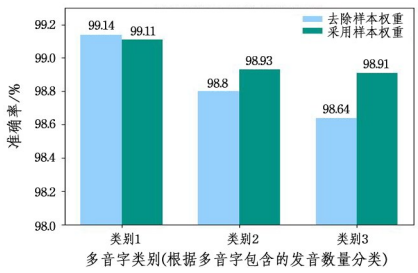

图 4 针对 CPP 数据集中不同类别的多音字测试结果

Fig. 4 Test results of different categories of polyphonic characters in CPP dataset

另外，为了证明本文模型 CLTRoBERTa 的有效性，本文使用以下方法作为基线模型进行比较。

1)MASK-BASED[10]：提出了掩码模型并进行多音字消歧任务，采用经过修正的焦点损失函数以应对样本不平衡问题。

2)g2pM(BiLSTM)[22]：将多音字消歧视为分类任务，并且使用 BiLSTM 进行训练。

3)g2pM(BERT)[22]：将多音字消歧视为分类任务，采用预训练模型 BERT 进行训练。

4)Distant supervision[25]：将多音字消歧视为序列-序列问题，并利用声学对齐模块产生大量的序列对。采用远程监督的方式训练核心模型。

5)PDF(with BERT)[12]：将多音字消歧视为分类任务，融合预训练模型和命名实体识别领域提出的 FLAT 方法进行多音字消歧，使模型既能避免分词错误，又利用了拼音特征。

6)BERT with LSTM[12]：采用 BERT 与 LSTM 的组合进行多音字消歧。

7)ELECTRA with MARC[23]：将图像分类任务中的长尾算法 MARC 应用到多音字消歧任务中，并与预训练模型 ELECTRA 相结合。

如表 8 所列，本文提出的多音字消歧模型在 CPP 基准数据集上相比其他基线模型有更好的消歧能力。本次研究对多音字消歧任务进行了广泛的实验评估，比较了多种不同模型的性能。下面将对实验结果进行详细分析和解释。

表 8 各模型实验结果

Table 8 Experimental results of different models

序号	模型	测试集正确率/%
1	MASK-BASED	97.68
2	g2pM(BiLSTM)	97.31
3	g2pM(BERT)	97.85
4	Distant supervision	97.51
5	PDF(with BERT)	98.83
6	BERT with LSTM	98.04
7	ELECTRA with MARC	98.81
8	CLTRoBERTa	**99.08**

首先，一些基于 BiLSTM 和 CNN 的模型(如 g2pM(BiLSTM)，MASK-BASED)在多音字消歧任务中取得了一定的效果，但性能相对较差，原因在于这些方法在捕捉多音字和词语的复杂语义方面存在局限性。此外，基于预训练模型的方法(如 g2pM(BERT))在多音字消歧任务中展现出了强大的性能(如 PDF with BERT，BERT with LSTM)，在充分利用预训练模型的语义表示能力的同时，还结合了其他方法来进一步提升性能，然而这些方法没有针对多音字数据分布不均衡的措施。最后，MASK-BASED，ELECTRA with MARC，Distant supervision 虽然针对性地解决了多音字的不均衡分布问题，但是其未充分利用模型的层间特性以及多音字的语义信息。

总的来说，本文方法在解决多音字消歧问题上具有显著优势，相对于其他方法获得了更高的准确率。

结束语 本文提出了一种基于预训练模型 RoBERTa 的多音字消歧方法 CLTRoBERTa，该方法融合跨语言互译模块、层级学习率优化策略和样本权重模块，解决了以往模型没有充分利用多音字所在词语的语义信息和模型的层间特性以及多音字数据分布不均衡的问题。本文在 CPP 基准数据集上进行了实验，与以往的方法相比，CLTRoBERTa 取得了更好的效果。在将来的工作中，将考虑如何更全面地整合外部知识，以便模型能够处理更多真实世界的多音字问题。

参 考 文 献

[1] BRUGUIER A，BAKHTIN A，SHARMA D. Dictionary Augmented Sequence-to-Sequence Neural Network for Grapheme to Phoneme Prediction[C]// INTERS-PEECH. 2018：3733-3737.

[2] HE M，YANG J，HE L，et al. Neural lexicon reader：Reduce pronunciation errors in end-to-end tts by leveraging external textual knowledge[J]. arXiv：2110. 09698，2021.

[3] GOU D，LUO W. Processing of polyphone character in chinese tts system[J]. Chinese Information，1991，1：33-36.

[4] DONG H，TAO J，XU B. Grapheme-to-phoneme conversion in Chinese TTS system[C]// 2004 International Symposium on Chinese Spoken Language Processing. IEEE，2004：165-168.

[5] ZHANG Z R，CHU M，CHANG E. An efficient way to learn rules for grapheme-to-phoneme conversion in Chinese[C]// International Symposium on Chinese Spoken Language Processing. 2002.

[6] LIU F，ZHOU Y. Polyphone disambiguation based on tree-guided tbl[J]. Computer Engineering and Applications，2011，47(12)：137-140.

[7] LIU F，SHI Q，TAO J. Maximum entropy based homograph disambiguation[C]// NCMMSC2007. 2007：41-46.

[8] SHAN C，XIE L，YAO K. A bi-directional lstm approach for polyphone disambiguation in mandarin chinese[C]// 2016 10th International Symposium on Chinese Spoken Language Processing(ISCSLP). IEEE，2016：1-5.

[9] CAI Z，YANG Y，ZHANG C，et al. Polyphone disambiguation for mandarin chinese using conditional neural network with multi-level embedding features[J]. arXiv：1907. 01749，2019.

[10] ZHANG H，PAN H，LI X. A Mask-Based Model for Mandarin Chinese Polyphone Disambiguation[C]// INTERSPEECH. 2020：1728-1732.

[11] GEHRING J，AULI M，GRANGIER D，et al. Convolutional sequence to sequence learning[C]// International Conference on Machine Learning. PMLR，2017：1243-1252.

[12] ZHANG H T. Polyphone Disambiguation in Chinese by Using FLAT[C]// INTERSPEECH. 2021.

[13] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv：1810. 04805，2018.

[14] LIU Y，OTT M，GOYAL N，et al. Roberta：A robustly optimized bert pretraining approach[J]. arXiv：1907. 11692，2019.

[15] LAN Z，CHEN M，GOODMAN S，et al. Albert：A lite bert for self-supervised learning of language representations[J]. arXiv：1909. 11942，2019.

[16] CLARK K，LUONG M T，LE Q V，et al. Electra：Pre-training text encoders as discriminators rather than generators[J]. arXiv：2003. 10555，2020.

[17] DAI D，WU Z，KANG S，et al. Disambiguation of Chinese Polyphones in an End-to-End Framework with Semantic Features Extracted by Pre-Trained BERT[C]// INTERSPEECH. 2019：2090-2094.

[18] ZHANG S，ZHENG K，ZHU X，et al. A Poly-phone BERT for Polyphone Disambiguation in Mandarin Chinese[J]. arXiv：2207. 12089，2022.

[19] SHI Y，WANG C，CHEN Y，et al. Polyphone disambiguation in mandarin chinese with semi-supervised learning[J]. arXiv：2102. 00621，2021.

[20] BROWN P F，DELLA PIETRA S A，DELLA PIETRA V J，et al. The mathematics of statistical machine translation：Parameter estimation[J]. Computational linguistics，1993，19(2)：263-311.

[21] HOWARD J，RUDER S. Universal language model fine-tuning for text classification[J]. arXiv：1801. 06146，2018.

[22] PARK K，LEE S. g2pm：A neural grapheme-to-phoneme conversion package for mandarin chinese based on a new open benchmark dataset[J]. arXiv：2004. 03136，2020.

[23] GAO Y，XIONG Y J，YE J C. Double-Weighted Disambiguation Algorithm for Long-tail Polyphone Problem[J]. Journal of Chinese Information Processing，2022，36(11)：169-176.

[24] ZHANG S，LI Z，YAN S，et al. Distribution alignment：A unified framework for long-tail visual recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：2361-2370.

[25] ZHANG J，ZHAO Y，ZHU J，et al. Distant Supervision for Polyphone Disambiguation in Mandarin Chinese[C]// INTERSPEECH. 2020：1753-1757.

GAO Beibei，born in 2000，postgraduate. Her main research interests include natural language processing and machine learning.

ZHANG Yangsen，born in 1962，postdoctor，professor，Ph. D supervisor，is a member of CCF(No. 16640S). His main research interests include natural language processing and artificial intelligence.

（责任编辑：何杨）