基于知识及流利度提升的中文语法纠错模型

王 岩[1] 梁椰玲[1]
WANG Yan LIANG Yeling

摘 要 语法错误纠正（grammatical error correction，GEC）旨在将包含语法错误的句子纠正为正确的句子。目前语法错误纠正研究主要基于 Transformer 模型，但由于模型参数规模大，中文 GEC 任务语料不足，Transformer 无法得到充分训练来学习文本中足够的语义信息。提出了基于知识及流利度提升策略的中文 GEC 模型，将 MacBERT 预训练模型作为外部知识来源，并利用流利度提升策略缓解 GEC 模型单轮推理纠错不完全的局限。为了验证所提出的 GEC 模型的有效性，在 NLPCC 2018 中文 GEC 共享任务数据集上进行了大量实验，其性能优于 NLPCC 2018 GEC 共享任务中开发的最佳模型。

关键词 中文语法纠错；Transformer 模型；知识增强学习；流利度提升策略；预训练语言模型

doi：10.3969/j.issn.1672-9528.2024.05.023

0 引言

中文语法纠错的目的是自动检测并纠正句子中包含的语法错误，包括句子中的选词（word selection，S）、冗余词（redundant words，R）、缺词（missing words，M）和词序（word order，W）错误。基于编码器 - 解码器架构的 Transformer[1] 模型已成为 GEC 任务中的主流方法，该模型通过注意力机制进一步提高 GEC 的性能。目前，该方法存在两个局限性：泛化能力差以及无法纠正句子中的多个错误 [2]。

预训练语言模型（pre-training language models，PLMs）在自然语言处理任务中取得了较大的成功。BERT 作为最成功的 PLMs 模型之一，已被应用于 GEC 任务，并在性能上取得了实质性的改进 [3]。掩码语言模型（MLM）BERT 建立在 Transformer 模型的编码器架构上。MLM 采用 [MASK] 符号随机遮盖输入序列，并利用上下文预测被遮盖的词，促使模型提取训练语料库中的各种语言特征。然而，在 GEC 模型的训练过程中不会出现 [MASK] 符号，并且 BERT 的预训练语料库与 GEC 任务的训练语料差异很大，因此 MLM 既不能适应 GEC 任务，也不能适应 GEC 语料库中特殊的语法错误部分。BERT 和 GEC 任务之间的差距使得基于 BERT 的 GEC 模型无法充分利用 BERT 中的语言知识。

MacBERT[4] 采用 Mac（MLM as correction）任务进行预训练，使用相似的单词来遮盖单词，取代 BERT 原始的 MLM 任务，相较于 MLM 任务采用 [MASK] 符号遮盖单词，Mac 任务更接近于 GEC 任务。Mac 预训练任务填补了预训练任务与下游 GEC 任务之间的空白，促使 GEC 任务尽可能多地利用 MacBERT 中的知识信息，因此基于 MacBERT 的方法有望提高 GEC 任务的性能。

为了解决 GEC 模型中存在的局限性，本文提出了一个基于知识及流利度提升策略的 GEC 模型（KFPGECM）。该模型通过采用预训练语言模型 MacBERT 提供的语言知识以及流利度提升策略，在 NLPCC2018 GEC 共享任务测试集上 $F_{0.5}$ 值达到 31.88，优于共享任务中的 Top-3 模型。为了进一步改善 KFPGEC 模型的性能，使用构造数据扩充训练语料对模型进行训练，并进行模型集成，纠错性能进一步得到提升。

1 KFPGEC 模型原理

在本研究中，提出了一个基于编码器 - 解码器架构的中文 GEC 模型，该模型将预训练语言模型 MacBERT 与流利度提升策略相结合，并将语言生成过程分为两个阶段：知识增强学习阶段和流利度促进推理阶段。

知识增强学习采用 MacBERT 作为 GEC 任务的知识来源。MacBERT 采用 Mac 作为预训练任务，使用相似词替换部分词，并根据上下文预测被覆盖的词，该任务更接近于中文 GEC 任务。因此，在中文 GEC 任务中引入 MacBERT，可以减少预训练任务与 GEC 任务之间的差距所带来的负面影响，最大限度地提高 GEC 任务对语言知识的有效利用。在神经机器翻译研究的启发下，流利度促进推理通过对语句进行多次推理，执行多轮纠正，直至句子流利度不再提升，进而改善 GEC 性能。

1.1 知识增强学习

分别探索了通过融合 [5]（MacBERT-fuse）及初始化 [6-7]（MacBERT-init）两种知识引入策略，将 MacBERT 预训练

1. 郑州科技学院信息工程学院 河南郑州 450064

模型引入纠错任务中。

1.1.1 MacBERT-init

为了利用 MacBERT 学习到的已有知识，采用初始化策略将 MacBERT 中的语言知识引入 Transformer 模型中。首先对其解码器采用随机初始化的方法，因为 MacBERT 预训练模型基于 Transformer 模型编码器进行构建，架构相同，所以使用 MacBERT 学习到的权重对 Transformer 编码器进行初始化。MacBERT-init 结构如图 1 所示。

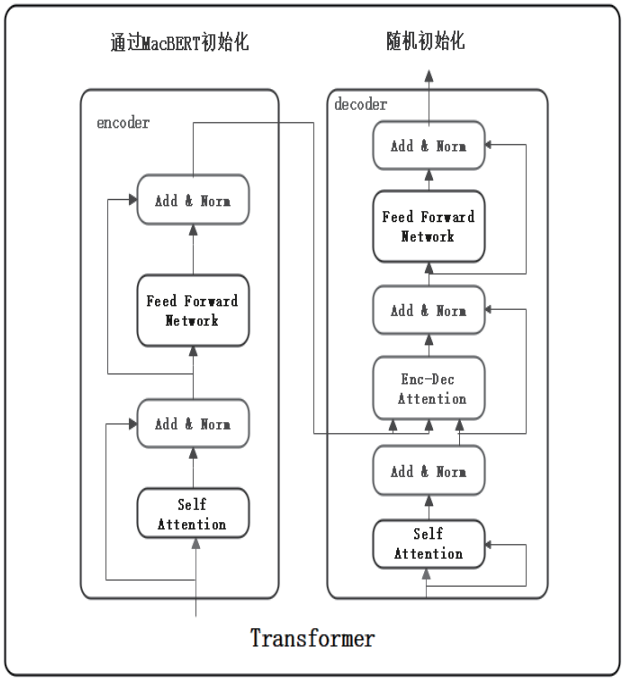

图 1 MacBERT-init 策略

1.1.2 MacBERT-fuse

通过融合策略，将 MacBERT 引入 Transformer 的编码器 - 解码器架构中。首先获得 MacBERT 最后一层的特征表示，然后通过注意力机制将其与编码器 - 解码器每一层进行融合，以确保充分利用 MacBERT 中预训练的知识信息，如图 2 所示。

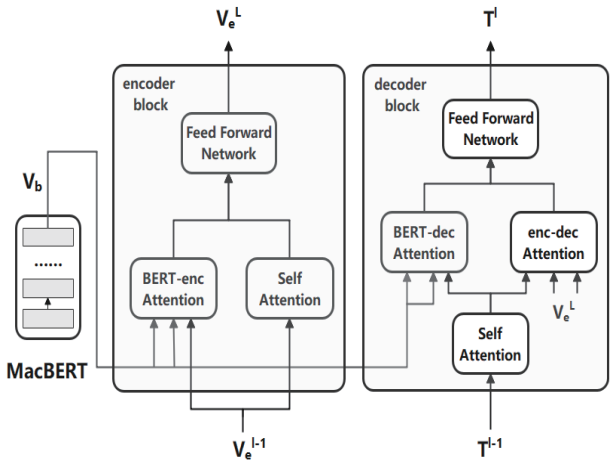

图 2 MacBERT-fuse 策略

X 和 Y 分别表示输入序列和输出序列。输入序列 X 通过 MacBERT 编码得到 V_b: $V_b = MacBERT(X)$，V_b 是 MacBERT 最后一层的输出。V_e^l 表示编码器第 l 层的隐藏表示，V_i^l 表示 V_e^l 中的第 i 个元素，其中 $i \in [l_x]$，ATT_{self} 表示编码器中的自注意力机制，ATT_{enc} 表示编码器注意力机制，则 V_i^l 的计算公式为：

$$V_i^l = \frac{1}{2}(ATT_{self}(V_i^{l-1}, V_e^{l-1}, V_e^{l-1}) + ATT_{enc}(V_i^{l-1}, V_b, V_b)) \quad (1)$$

接着，前馈神经网络 (FFN) 对 V_i^l 进行处理，得到第 l 层的输出 V_e^l，编码器每一层迭代之后从最后一层输出 V_e^L。

令 $T_{<t}^l$ 表示解码器第 l 层前 t 个时间步的隐藏表示，T_t^l 表示 l 层在第 t 个时间步的隐藏表示。则：

$$T_t^{l_1} = ATT_{self}(T_t^{l-1}, T_{<t+1}^{l-1}, T_{<t+1}^{l-1}) \quad (2)$$

$$T_t^{l_2} = \frac{1}{2}(ATT_{dec}(T_t^{l_1}, V_b, V_b) + ATT_{enc-dec}(T_t^{l_1}, V_e^L, V_e^L)) \quad (3)$$

$$T_t^l = FFN(T_t^{l_2}) \quad (4)$$

式中：ATT_{dec}、$ATT_{enc\text{-}dec}$ 分别表示解码器注意力机制和编码器 - 解码器注意力机制。可以在多次迭代后得到 T_t^L，然后通过线性变换和 softmax 函数对第 t 个词进行预测。将所有预测得到的单词组合起来，可以获得最终的输出序列。

1.2 流利度促进推理

在流利度促进推理阶段采用流利度提升策略，对存在多个语法错误的语句，该算法可以先对其中的一些错误进行修正，修正后的语句可以使得上下文更加清楚，从而帮助模型修正其它错误。如图 3 所示，流利度提升策略可以使模型通过多轮编辑对错误句子进行修改。通常纠错模型将源语句作为模型输入，第一轮推理结果（$Round_1$）作为模型输出。然而，流利度提升策略并不会将其作为最终预测结果，它会对 $Round_1$ 进行打分，分数越高，说明句子流利度越高。如果句子流利度得到提升，流利度提升策略将以 $Round_1$ 作为模型输入，生成下一轮推理结果 $Round_2$。推理过程将持续进行，直到句子流利度不再提升为止。

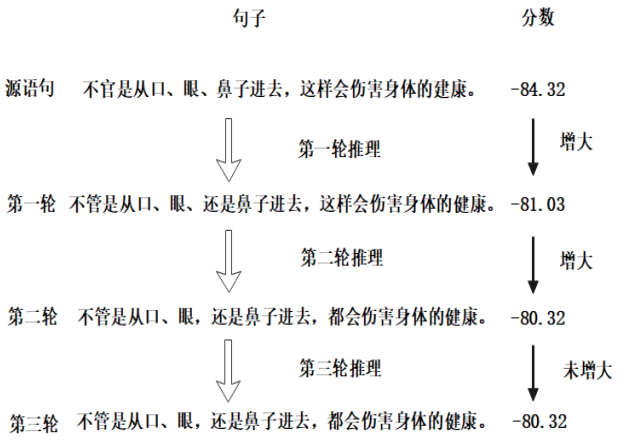

图 3 流利度促进推理阶段

2 实验设置

本研究中使用NLPCC2018 GEC共享任务[8]提供的中文GEC数据集，整个数据集包含1 157 708个中文句子对和2000个标记测试语句。去除数据集中文本长度大于64以及源语句和目标语句完全相同的部分句子对，获得1 025 550个句子对。最后，从中随机选取5000个句子对作为验证集，剩余1 020 550个句子对为训练集，2000条标记语句作为测试集。

训练过程中的模型参数设置如下：对于融合策略，选择Adam优化器训练模型，最大令牌为4096，损失函数使用标签平滑交叉熵，dropout概率为0.3，设置初始学习率为3×10^{-5}，并采用线性衰减的学习率衰减方法。对于初始化策略，学习率设置为5×10^{-4}，采用Adam优化器训练模型，批次大小为32，设置dropout概率为0.1，损失函数采用cross entry。

评估工具选择MaxMatch（M_2）工具包。M_2方法[9]是一种广泛应用于GEC任务的评价工具，可以计算系统假设与标准纠正之间的匹配程度。首先，M_2计分器计算源语句与系统假设之间可能的编辑序列，并找到与标准纠正集g_i重叠度最高的编辑序列e_i，然后依据最优序列计算精确率（P）、召回率（R）和$F_{0.5}$，计算公式为：

$$P=\frac{\sum_{i=1}^{n}|e_i\cap g_i|}{\sum_{i=1}^{n}|e_i|} \tag{5}$$

$$R=\frac{\sum_{i=1}^{n}|e_i\cap g_i|}{\sum_{i=1}^{n}|g_i|} \tag{6}$$

$$F_{0.5}=5\times\frac{P\times R}{P+4\times R} \tag{7}$$

3 结果及分析

实验结果如表1所示。实验9为Transformer模型，BERT-fuse和BERT-init模型结果相较于实验9均有所提升，说明BERT中已有的语义知识有利于提高中文语法纠错性能。采用BERT-fuse和BERT-init作为基线模型。设置实验3和实验4，通过融合和初始化策略将MacBERT预训练模型引入到纠错任务中，纠错性能均得到改善，验证了知识增强学习阶段的有效性，并且GEC模型可以更有效利用MacBERT中的知识。实验5和实验6为本文提出的KFPGEC模型，纠错性能相较于基线模型提升了+2.14和+2.03，融合策略对纠错性能的影响大于初始化策略，最佳结果达到31.88。此外，与实验3和实验4相比，结果进一步增加，表明流利度提升策略有利于GEC任务。

表1 不同改进对模型的影响

ID	Model	P/%	R/%	$F_{0.5}$/%
1	BERT-fuse	31.90	23.40	29.74
2	BERT-init	31.81	22.41	29.35
3	MacBERT-fuse	36.56	20.34	31.53
4	MacBERT-init	32.79	24.59	30.74
5	KFPGEC-fuse	36.78	20.80	31.88
6	KFPGEC-init	33.21	25.71	31.38
7	数据扩充	38.63	24.82	34.76
8	模型集成	41.04	27.23	37.26
9	Transformer	36.57	14.27	27.86
10	YouDao	35.24	18.64	29.91
11	BERT-encoder	41.94	22.02	35.51
12	BN-CGECM	51.57	17.43	37.05

实验7采用基于规则和基于反向翻译的方法构造数据集，和NLPCC2018原始纠错数据共同用于KFPGEC-fuse模型的训练。实验8同时采用数据扩充及模型集成方法来提升KFPGEC-fuse模型纠错性能，实验结果最终达到37.26。与前人研究进行比较，如实验10至实验12。实验10为NLPCC2018中文语法纠错共享任务中有道团队[10]开发的性能最好的纠错系统。实验11将BERT预训练模型引入Transformer编码器中构建BERT-encoder模型，并基于该模型训练了一个4-ensemble模型。实验12构建BART噪声器[11]向输入语句中生成噪声，将其用于模型训练，并对模型进行集成。本文提出的集成模型相较于NLPCC2018中文语法纠错共享任务性能最好的模型，$F_{0.5}$值高出7.35个百分点。并且和实验11及实验12相比，$F_{0.5}$值分别提升了1.75和0.21个百分点，验证了本文工作的有效性。

表2 各错误类型校正的$F_{0.5}$值

Error Type	BERT-fuse	MacBERT-fuse	BERT-init	MacBERT-init
S	34.27	37.16	35.12	38.35
M	32.64	33.92	33.20	35.46
R	35.73	38.31	33.85	31.79
W	17.18	13.62	11.26	9.59

从测试结果中随机抽取400条句子对，分析模型在知识增强学习阶段对四种类型错误的纠正效果，实验结果如表2所示。模型对词序类型的错误纠正效果最差，并且MacBERT预训练模型的引入对该类错误没有帮助，$F_{0.5}$值明显降低。MacBERT预训练模型可以提升模型对其他三种类型错误的

纠错性能，其中对于选词类型纠错性能的提升最高。

为展示模型推理过程，对采用流利度提升策略后的推理过程进行分析，部分样本如表 3 所示，其中 Source 为源语句，Target 为目标语句，$Round_i$ 为模型第 i 次推理后输出的语句。例如，在样本 1 中，经过两次推理预测，模型对包含多个错误的源语句进行逐步纠正。模型首先将“今大”改为“今天”，使得模型流利度提升，更易于理解，促使模型在后续的推理中对其他位置的错误进行进一步的纠正。然而，流利度提升策略在某些情况下也会出现错误。例如，在样本 2 中，KFPGEC 在语句中添加了不必要的单词“里”。

表 3 纠错样本示例

样本编号	句子
1	Source：今大我跟我朋友去餐厅吃饭。 Target：今天我跟朋友去餐厅吃饭。 Round 1：今天我跟我朋友去餐厅吃饭。 Round 2：今天我跟朋友去餐厅吃饭。
2	Source：我在公圆散步。 Target：我在公园散步。 Round 1：我在公园散步。 Round 2：我在公园里散步。

4 结论

本文提出了基于 MacBERT 预训练模型和流利度提升策略的中文语法纠错模型 KFPGEC。在 NLPCC 2018 中文 GEC 共享任务数据集上进行大量实验验证，结果表明，MacBERT 预训练模型已有的知识信息和流利度提升策略可以有效改善 GEC 性能。此外，采用数据构造及集成方法进一步改善纠错性能。结果分析表明，词序纠错仍有改进的空间，将在今后的工作中继续探索这一问题。

参考文献：

[1]VASWANI A, SHAZEER N, PARMAR N, et al.Attention is all you need[J].Advances in neural information processing systems,2017,30:6000-6010.

[2]GE T, WEI F, ZHOU M.Fluency boost learning and inference for neural grammatical error correction[C]//56th Annual Meeting of the Association for Computational Linguistics: Long Papers,vol.2.Stroudsburg,PA:Association for Computational Linguistics,2018: 1055-1065.

[3]WANG H, KUROSAWA M, KATSUMATA S, et al.Chinese grammatical correction using BERT-based pre-trained model[EB/OL].(2020-11-04)[2024-03-01].https://doi.org/10.48550/arXiv.2011.02093.

[4]CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[C]//Findings of ACL: EMNLP 2020, Part 1.Stroudsburg, PA: Association for Computational Linguistics,2020: 657-668.

[5]ZHU J, XIA Y, WU L, et al.Incorporating bert into neural machine translationd[EB/OL].(2020-02-17)[2024-03-01].https://doi.org/10.48550/arXiv.2002.06823.

[6]LAMPLE G, CONNEAU A.Cross-lingual language model pretraining[EB/OL].(2009-01-22)[2024-03-01].https://doi.org/10.48550/arXiv.1901.07291.

[7]ROTHE S, NARAYAN S, SEVERYN A.Leveraging pre-trained checkpoints for sequence generation tasks[J].Transactions of the association for computational linguistics,2020, 8: 264-280.

[8]ZHAO Y, JIANG N, SUN W, et al.Overview of the nlpcc 2018 shared task: Grammatical error correction[C]//Natural Language Processing and Chinese Computing,Part II.Berlin: Springer, 2018:439-445.

[9]DAHLMEIER D, NG H T.Better evaluation for grammatical error correction[C]//2012 Conference of the North American Chapter of the Association for Computational Linguistics: human language technologies.Stroudsburg,PA:Association for Computational Linguistics,2012:568-572.

[10]FU K, HUANG J, DUAN Y.Youdao ' s winning solution to the nlpcc-2018 task 2 challenge: a neural machine translation approach to chinese grammatical error correction[C]//Natural Language Processing and Chinese Computing: 7th CCF International Conference.Cham:Springer,2018:341-350.

[11] 孙邱杰，梁景贵，李思 . 基于 BART 噪声器的中文语法纠错模型 [J]. 计算机应用 ,2022, 42(3):860-866.

【作者简介】

王岩（1994—），男，山西吕梁人，硕士，助教，研究方向：自然语言处理、数字图像处理。

（收稿日期：2024-03-12）