

基于反向翻译的英语语法纠错应用研究

孙晓东,王丕坤,杨东强*

(山东建筑大学 计算机科学与技术学院,山东 济南 250101)

摘要:基于数据驱动和机器翻译模型的英语语法纠错是神经语言模型的主要应用之一。人工标注语料库的数量和质量是影响此类方法性能的重要因素。通过分析现有学习者语料的错误类型分布,对常见的错误类型如动词、名词、部分介词、拼写和标点建立混淆集。使用混淆集结合人工规则的方法对单语语料数据进行加噪处理,与学习者语料分别用于基于机器翻译的自动错误生成模型的预训练和微调;使用错误生成模型生成的合成数据与学习者语料共同训练语法纠错模型,模型性能在 CoNLL-2014 和 JFLEG 数据集上得到显著性提高。此外,通过使用语法纠正模型纠正学习者语料库源句,将产生的中间数据反馈输入到错误生成模型,并进行交替训练。纠错系统在标准数据集上的性能得到进一步提升。

关键词:数据增广;反向翻译;规则;语法纠错;交替训练

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2022)10-0143-08

doi:10.3969/j.issn.1673-629X.2022.10.024

Application Research of English Grammar Error Correction Based on Back-Translation

SUN Xiao-dong, WANG Pi-kun, YANG Dong-qiang*

(School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China)

Abstract: English grammar error correction method based on data-driven and machine translation models is one of the main applications of neural language models. The quantity and quality of artificially annotated corpora are important factors that affect the performance of such methods. By analyzing the distribution of error types in the existing learner corpus, a confusion set is established for common error types such as verbs, nouns, some prepositions, spelling and punctuation. Confusion sets is combined with artificial rules to add noise to the monolingual corpus data, and the learner corpus is used separately for the pre-training and fine-tuning of the automatic error generation model based on machine translation. The synthetic data generated by the error generation model and the learner's corpus are applied to train the grammatical error correction model, the performance of the model is significantly improved on the CoNLL-2014 and JFLEG data sets. In addition, by using the grammar correction model to correct the source sentences of the learner's corpus, the generated intermediate data is fed back into the error generation model, and alternate training is performed. The performance of the error correction system on the standard data set has been further improved.

Key words: data augmentation; back-translation; rule; grammatical error correction; alternating training

0 引言

语法纠错(Grammatical Error Correction, GEC)的主要目的是在不改变语句的基本语义前提下最大限度地纠正语法错误。早期面向英语学习者的GEC研究方法主要以基于规则和统计的模型为主^[1]。基于规则的方法人工设定语法规则数据库,只对有限的错误类型有效。而且对于相对复杂的语法错误,需要设计大量的检测和纠正规则,规则之间容易产生冲突。基于统计的GEC方法主要构建统计模型,一定程度上避免

了基于规则的GEC方法其纠错类型有限的弊端。在随后出现的基于分类的语法错误纠正方法^[2]中,不同错误类型的纠正被看作分类任务,通过结合上下文语境训练GEC分类器进行错误的预测纠正。而对于基于机器翻译的GEC方法主要将语法纠正看做语言翻译任务,利用噪声信道模型生成翻译规则,把错误的源语句“翻译”成正确的目标语句。

相比于上述早期的GEC方法,现阶段基于数据驱动和神经网络的机器翻译方法已成为英语语法纠错的

收稿日期:2021-10-28

修回日期:2022-03-02

基金项目:国家教育部人文社科基金资助项目(15YJA740054)

作者简介:孙晓东(1995-),男,硕士研究生,研究方向为自然语言处理;通讯作者:杨东强(1970-),男,副教授,博士,研究方向为自然语言处理。

主流^[3],其中应用神经网络编码器-解码器结构的模型已经取得显著成效。神经机器翻译(Neural Machine Translation, NMT)是基于神经网络以端到端的方式进行翻译的方法,与早期的机器翻译系统相比,采用注意力机制的 NMT 模型,解决了上下文中长距离语义依赖的问题,能够更好地实现语法纠正。典型的端到端 NMT 模型将束搜索(beam-search)得到的解码结果作为输出,同时还引入外部语言模型、编辑距离、编辑操作数等多个特征评价模型的解码输出。

基于 NMT 的 GEC 模型性能主要取决于训练数据的数量和质量。由于人工标注数据成本较高,目前研究主要集中在如何自动合成训练数据上。研究表明^[4]人工数据合成可以弥补学习者语料库不足的问题,提高 GEC 性能。对于 GEC 中数据增广方法,相对于基于错误规则或错误模式的无监督数据产生方法,监督式数据合成方法生成的数据更接近语法错误的实际分布状态,但是监督式数据合成方法需要足够数量的人工标注数据,并且有限的人工标注数据以及其中的错误类型分布和修改风格等会对生成模型的效果产生较大影响。因而通过结合有限的规则和反向翻译模型生成训练数据,不仅可以兼顾质量和数量的优势,还能够提高语法纠错模型的性能。

该文使用 Transformer 模型^[5]分别作为语法错误产生和纠正的基础结构,主要探索结合规则的数据增广方法和反向翻译模型为 GEC 生成人工合成数据,从而提高语法纠错系统的性能。主要创新在于:

(1)与将人工规则方法产生的数据直接训练 GEC 模型不同,该文提出将基于规则的数据增广方法与 NMT 相结合,首先建立语法错误生成(Grammatical Error Generation, GEG)模型,然后将 GEG 模型产生的合成数据用于 GEC 模型的预训练。

(2)为了提高 GEC 模型的错误识别能力,该文提出一种由 GEG 到 GEC 再到 GEG 模型的交替训练结构。使用 GEC 系统的输出候选句对 GEG 模型再次训练,以提高其错误生成能力。

对比实验表明,提出的数据增广策略和交替训练方法能够产生更加接近学习者语料库质量的合成数据,并显著提高 GEC 系统的性能。基于 CoNLL-2014 和 JFLEG 测试集的实验结果表明:该方法的 GLEU 和 $F_{0.5}$ 值分别达到 61% 和 62%。

1 语法纠正领域数据增广方法相关工作

针对 GEC 的人工数据增广方法主要有基于人工规则的方法、基于机器翻译的方法以及基于维基百科等社交媒体的修订历史记录方法等。

1.1 基于人工规则

基于规则的数据增广方法是使用限制性规则完成对数据的加噪处理,已成为提高训练数据质量的一种重要手段。该方法简单有效,能够涵盖不同类型的语法错误。例如,在 CoNLL-2014 测试集上, Xu 等人^[6]引入五种错误规则合成训练数据,用此合成的数据训练基于 NMT 的 GEC 模型,取得优异的结果, $F_{0.5}$ 值达到 60.9%。在面向 GEC 的数据增广任务中,人工规则主要利用单词或字符级别上的增加、删除、取代、交换等四种编辑距离及其衍生变种。

1.2 基于机器翻译

基于机器翻译的数据增广方法主要包括两种:基于反向翻译和基于往返翻译的方法。基于反向翻译的数据增广方法是训练一个反向 NMT 错误生成模型,输入为语法正确的源语句,输出为含有语法错误的目标语句。该方法能够覆盖多种不同种类的错误,但是反向翻译模型往往需要大量带有标注的学习者语料库作为支撑,因而也面临着人工标注数据短缺的问题。基于往返翻译的数据增广方法^[7]是训练两个机器翻译模型,首先是由英语翻译到中间语言,再由中间语言翻译到英语。使用往返翻译合成的训练数据可以借助多种中间语言,不需要大量的人工标注数据,翻译模型性能的优劣通常会影响到合成数据的质量。该方法存在需要协调不同翻译模型的质量搭配问题。高质量往返翻译模型的合成数据可能不存在语法错误现象,而低质量的翻译模型产生的合成数据可能包含除语法错误之外的其他错误,如源语句的语义信息丢失等。

1.3 基于人工修订

维基百科包含其百科知识修订(wiki-edit)历史页面,可以从中提取相应数据作为“源-目标”语句对,其中源语句由较旧的修订历史提供,目标句是对应较新的连续修订历史提供。除此之外, Lang8 作为一个在线的语言学习网站,包含了 80 多种语言,主要利用社交媒介纠正语言学习中的错误,可以从中挖掘出大量的“源-目标”平行句对。

1.4 评价指标

尽管 GEC 研究取得了一定的成功,但仍然面临评价指标合理性的挑战。最大匹配分数(Max Match score, M^2)和 ERRANT 是评估 GEC 常用的评测方法,主要采用基于 Levenshtein 距离的对齐策略,计算将纠正语句转换成人工标注的参考语句所需要的编辑(插入、删除、替换)数量。最大匹配分数以精确度(P)、召回率(R)、 $F_{0.5}$ 值为主要评价指标。计算公式为:

$$P = \frac{\sum_{i=1}^n |e_i \cap g_i|}{\sum_{i=1}^n |e_i|} \quad (1)$$

$$R = \frac{\sum_{i=1}^n |e_i \cap g_i|}{\sum_{i=1}^n |g_i|} \quad (2)$$

$$F_{0.5} = \frac{P \times R(1 + 0.5^2)}{(0.5^2 \times P) + R} \quad (3)$$

式中, $e_i \cap g_i = \{e \in e_i \mid \exists g \in g_i, e == g\}$ 。其中 e_i 表示 GEC 系统预测的纠正编辑集合, g_i 表示人工标注、标准的纠正编辑的集合。通过计算 GEC 系统输出的纠正编辑集合和人工标注的纠正编辑集合之间的匹配程度来衡量系统的性能。

除此之外, Naples 等人^[8]提出 GLEU 评价句子的流利度, 主要计算纠正语句相对于参考语句的 n -gram 重叠度。原始输入的错误句简称源句 S , 人工标注的标准句称为参考句 R 。英语语法纠正系统输出的纠正句子称为假设句 C , W_n 表示均匀分布的权重。计算假设句相对于参考句的 n -gram 精度 P_n , 计算公式为:

$$\text{GLEU}(C, R, S) = \text{BP} \times \exp\left(\sum_{n=1}^4 W_n \log P_n\right) \quad (4)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \quad (5)$$

在 BP(brevity penalty) 的计算公式中, c 为假设句的长度, r 为参考句的长度。GLEU 计算假设句与参考句之间的 n -gram 精度, 只适用 JFLEG 数据集。

2 方法与模型

该文首先利用基于规则的数据增广策略合成训练数据, 并与学习者语料共同训练 GEG 模型。之后利用 GEG 模型合成的训练数据与学习者语料训练 GEC 模型。受 Popel 等人^[9]交替训练英语到捷克语的翻译模型工作启发, 使用 GEC 模型纠正学习者语料中的源句, 将纠正的结果与学习者语料中的标准参考句重新构成平行语料加入到 GEG 模型的训练数据中, 再次训练 GEG 模型。模型重复多次训练直到满足系统需要为止, 其基本操作过程如图 1 所示。

2.1 基于规则的数据增广方法

人工标注语料库的规模与数量限制了基于数据驱动的神器机器翻译模型的性能, 因此, 如何通过数据增广扩展语料库规模是提高语法纠错性能的重要因素之一。对于基于规则的数据增广方法, 在单词级别上执行插入、替换、交换、删除四种操作, 该文四种规则的引入概率与 Abhijeet 等人^[10]保持一致, 分别为 0.3、0.25、0.25、0.2。其次, 通过对学习者语料库中错误类型的分布情况^[11]分析, 发现动词(约 7%)、名词(约 4.5%)、冠词(约 10.86%)、拼写(约 9.59%)、介词

(约 11.2%) 和标点错误(约 9.7%) 占比较高。针对这些错误类型, 分别建立动词、名词、冠词、介词、拼写和标点混淆集, 混淆集示例如表 1 所示。

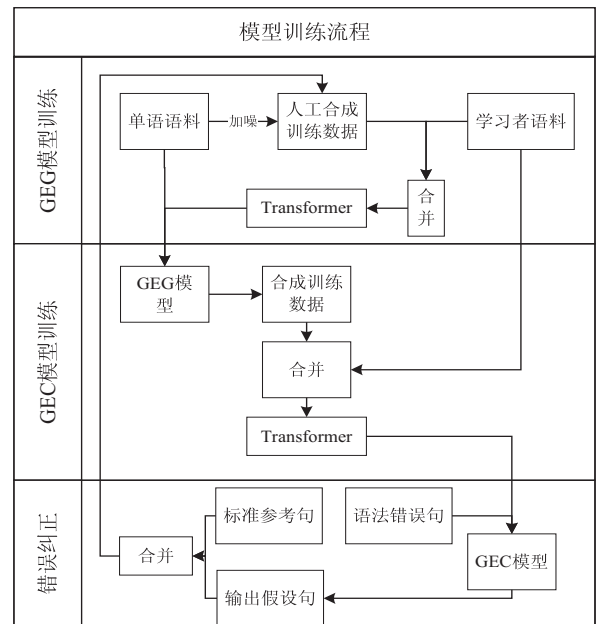


图 1 交替训练模型

表 1 不同词性候选集示例

错误类型	待替换词	候选集
动词	abandon	abandons、abandoned、abandoned、abandoning
名词	dictionary	dictionaries
冠词	an	a、the
介词	To	for、in、as、from、of、among、into、on、about、at、from、by、with
拼写错误	sucession	sucession、sucesion、succesion
标点符号	,	. ! : ? ;

合成数据的过程中, 对选中的待操作词随机执行插入、删除、替换、交换四种操作之一。对于替换规则, 从创建的不同词性候选集中随机选择一个单词进行替换。通过人工规则构建种类丰富的多词性错误可以为 GEG 和 GEC 模型提供更多的训练数据。

2.2 模型

GEG 和 GEC 模型均使用基于注意力(attention)机制的 Transformer 结构, 将 GEC 和 GEG 作为基于编码器-解码器的机器翻译任务, 反向传播最小化系统输出与真实输出之间的交叉熵损失函数。该模型采用开源的 FAIR Sequence-to-Sequence 工具包实现, GEG 系统解码使用 beam search(束搜索)解码。词向量的

维度和目标端的维度为 512, 编码器和解码器包含的网络层为 6 层, 设置前向神经网络子层的隐含层 (FFN) 维度为 4 096, dropout 率设置为 0.2, 优化算法使用 Adam, 8 个注意力头, 初始化学习速率为 0.002, 标签平滑率设置为 0.2, warmup 步长为 16 000。GEC 系统使用 C-Copy-Transformer 结构, 模型参数与 Zhao 等人^[12]的设置保持一致。

2.2.1 Encoder

Transformer 中 attention 机制让模型关注到上下文中有价值的信息。当对词进行编码时, 不仅仅考虑当前词, 还考虑当前词的上下文语境。把整个上下文语境融入到当前的词向量当中。Transformer 模型没有循环神经网络的迭代操作, 位置信息的融合使模型更容易识别出语言中的顺序关系。编码器由若干层组成, 每层包括两个子层, 分别是自注意力层和前馈神经网络层。多注意力层拼接成多头注意力层, 如公式 (6) ~ 公式 (9) 所示。

$$\text{attention_output} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \quad (6)$$

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_n) W^0 \quad (7)$$

$$\text{head}_i = \text{Attention}(\mathbf{Q} W_i^Q, \mathbf{K} W_i^K, \mathbf{V} W_i^V) \quad (8)$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V} \quad (9)$$

其中, \mathbf{Q} 表示查询向量, \mathbf{K} 表示相关性向量, \mathbf{V} 表示被查询信息的向量, 即实值矩阵。 \sqrt{d} 是固定因子, d 是隐藏层维度。

2.2.2 Decoder

解码器与编码器层数一致, 包括自注意力层、编码-解码注意力层和前馈神经网络层, 编码-解码注意力层将帮助解码器将注意力集中在输入句子的相关部

分。编码器和解码器中的每一个子层采用残差连接和归一化, 归一化把神经网络中的隐藏层归一到标准正态分布, 加快训练速度和收敛速度。

2.3 交替训练机制

不同于以往工作将基于规则的数据增广方法合成的训练数据直接训练 GEC 模型, 该文首先将其用于 GEG 模型的预训练, 然后再使用学习者语料库微调 GEG 模型, 期望通过不同数据增广方法的融合提高合成数据的质量。与 Sennrich 等人^[13]的反向翻译模型近似, 使用的目标端句子均具有真实性。为了进一步提高语法纠错的性能, 使用 GEC 模型纠正学习者语料库中的训练数据, 并将其纠正结果与训练集中的标准参考句构成平行语料, 作为重复训练 GEG 模型的扩充数据, 因而 GEG 模型将产生更多接近学习者语料库的训练数据。

3 实验与结果分析

3.1 实验数据

在现有 NUCLE、FCE、W&I+LOCNESS 三种学习者语料库^[11]及单语语料库 One Billion Word^[14]上进行验证实验。除此之外, 为了方便与现有工作做对比, 在最后使用 Lang-8 语料微调 GEC 模型。表 2 给出了使用的训练语料及相应的数据规模。通过统计学习者语料库中的句长分布, 约束单语语料库中句子的长度。其中单语语料库中训练数据长度均被规约到 5 至 100 的标记 (token) 长度大小。

实验将使用常用的 CoNLL-2014 测试集和 M²作为评测指标, 除此之外, 还使用 JFLEG 测试集和 GLEU 值对 GEC 的纠正结果进行流利程度分析。

表 2 训练语料规模

语料库	语料名称	句子数	标记 (tokens)
学习者语料库	NUCLE	57 151 (约 57 K)	约 1.2 M
	FCE	28 350 (约 28 K)	约 455 K
	Wi & locness	34 308 (约 34 K)	约 628 K
社交媒体数据	Lang-8	104 万 (约 1.04 M)	约 11.86 M
通用语料库	One Billion word	170 万 (约 1.7 M)	约 19 M

3.2 数据增广过程

3.2.1 基于规则的数据增广方法

步骤 1: 使用现有的三种学习者语料 (NUCLE、FCE、W&I+LOCNESS) 训练 GEC 模型, 在 CoNLL-2014 测试集上验证结果。

步骤 2: 使用基于规则的数据增广方法分别生成文件规模为 20 M (句子数量约 190 K)、80 M、140 M、200 M 的训练数据, 用不同规模的合成数据分别训练

GEC 模型, 对比使用不同规模数据对训练 GEC 的影响。

3.2.2 基于规则与反向翻译的数据增广融合使用过程

步骤 1 (GEG₁ → GEC): 使用三种学习者语料训练基于反向翻译的错误生成模型 (GEG₁), 再使用 GEG₂ 模型生成训练数据, 最后用生成的数据训练 GEC 模型。

步骤 2($GEG_2 \rightarrow GEC$):为了进一步优化 GEG_1 模型,尝试将两种数据增广方法结合使用。首先使用基于规则的合成数据预训练 GEG,再用学习者语料对其微调,得到优化后的错误生成模型 GEG_2 。使用 GEG_2 模型处理单语语料,分别生成文件规模为 20 M、80 M、140 M、200 M 的合成数据,并将合成的数据用于 GEC 模型的训练,以此来验证不同数据增广方法的融合使用效果。

3.3 交替训练

步骤 1:选取 3.2.2 节中训练得到的性能最佳的 GEC 模型处理三种学习者语料库中的源句,将模型输出的候选句与学习者语料库中的标准参考句合成训练集。与基于规则的数据增广方法合成的训练数据混合后再次训练 GEG,并使用三种学习者语料库微调得到错误生成模型 GEG_3 。

步骤 2($GEG_3 \rightarrow GEC$):使用 GEG_3 模型处理单语语料库,再次生成 20 M、80 M、140 M、200 M 不同规模的合成数据,训练 GEC 模型。

步骤 3:在现有合成数据的基础上,为了进一步提高 GEC 模型的性能,尝试扩大预训练数据规模。

3.4 实验类型与结果

在单语数据充足的情况下,将研究如何使用较少的标注数据最大化模型性能。相对于有限的标注数据,尝试使用较多的单语数据自动生成学习者语料以便提高系统性能。共设置了三组实验:

第一组实验:验证基于规则的数据增广方法效果。

第二组实验:验证基于规则和反向翻译的数据增广方法融合使用的有效性。

第三组实验:验证交替训练对提升 GEC 模型合成数据的质量和提提高 GEC 纠错性能的有效性。

3.4.1 基于规则的数据增广方法结果

分别使用基于规则的数据增广方法生成的不同规模训练数据,以及学习者语料训练 GEC 模型。其在 CoNLL-2014 测试集上的结果如图 2 所示。随着合成数据规模的不断增加,GEC 模型的 $F_{0.5}$ 值达到 37.9%,超出仅使用学习者语料库训练 GEC 模型得到的 $F_{0.5}$ 值 11.4% 左右。

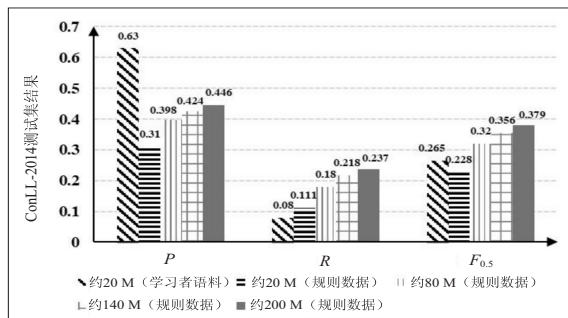


图 2 不同规模数据训练的 GEC 模型性能

3.4.2 基于规则和反向翻译的数据增广方法融合使用结果

在不同数据增广方法融合使用的效果验证过程中, GEG_1 表示仅使用学习者语料库训练错误生成模型, GEG_2 表示融合使用基于规则和反向翻译的数据增广方法训练的错误生成模型。使用 GEC 模型生成不同规模合成数据训练 GEC,其结果如表 3 所示。使用 GEG_1 得到最佳 GEC 模型的 $F_{0.5}$ 值达到 0.286,使用 GEG_2 合成的 200 M 训练数据得到的 GEC 模型性能比前者的 P 、 R 、 $F_{0.5}$ 值分别提高约 5.2%、3.9%、4.9%。

表 3 利用合成数据训练 GEC 性能对比

模型	训练 GEC 的合成数据规模											
	20 M			80 M			140 M			200 M		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$
$GEG_1 \rightarrow GEC$	0.275	0.135	0.228	0.291	0.184	0.261	0.304	0.193	0.273	0.319	0.201	0.286
$GEG_2 \rightarrow GEC$	0.359	0.153	0.282	0.36	0.232	0.343	0.372	0.243	0.335	0.371	0.240	0.335
$GEG_3 \rightarrow GEC$	0.445	0.123	0.292	0.443	0.212	0.364	0.417	0.272	0.377	0.427	0.273	0.384

3.4.3 交替训练结果

如表 3 模型 $GEG_3 \rightarrow GEC$ 所示,通过交替训练得到的语法错误生成模型 GEG_3 ,其合成的训练数据在

GEC 上的 $F_{0.5}$ 值达到 0.384,高出利用 GEG_2 模型训练 GEC 得到的 $F_{0.5}$ 值约 4.1%。

表 4 交替训练及 GEC 预训练模型数据扩充实验结果对比

GEC 模型训练	CoNLL-2014		
	P	R	$F_{0.5}$
Pre-train(200 M(GEG_1)+fine-tune(三种学习者语料)→GEC	0.626	0.275	0.499
Pre-train(200 M(GEG_2)+fine-tune(三种学习者语料)→GEC	0.617	0.319	0.52
Pre-train(200 M(GEG_3)+fine-tune(三种学习者语料)→GEC	0.659	0.32	0.543

续表 4

GEC 模型训练	CoNLL-2014		
	<i>P</i>	<i>R</i>	<i>F</i> _{0.5}
Pre-train(200 M(GEG ₃)+200 M(规则))→GEC	0.472	0.327	0.434
Pre-train(200 M(GEG ₃)+200 M(规则))+fine-tune(三种学习者语料)→GEC	0.677	0.338	0.564
Pre-train(200 M(GEG ₃)+200 M(规则))+fine-tune(三种学习者语料+Lang-8)→GEC	0.728	0.389	0.620

如表 4 所示,如果在 GEC 训练过程中加入学习者语料进行微调,对于 200 M 数据预训练 GEC 而言,利用 GEG₃ 合成数据训练 GEC 模型(GEG₃→GEC)的 *P*、*R*、*F*_{0.5} 值分别达到 0.659、0.32、0.543。进一步扩大预训练数据规模,将由 GEG₃ 合成的 200 M 训练数据与基于规则的数据增广方法合成的 200 M 数据混合后预训练 GEC 模型,然后使用三种学习者语料微调。由表 4 的实验结果可知,数据规模扩大后,*P*、*R*、*F*_{0.5} 值分别达到 0.677、0.338、0.564,超出不使用学习者语料微调 GEC 模型 20%、2%、13% 左右。

3.5 现有研究成果

为了更好地与现有工作对比,在三种学习者语料的基础上加入 Lang-8 语料,仅用于 GEC 模型的微调,模型性能进一步提高。在没有使用多模型集成和重排序等情况下, GEG₃ 的 *F*_{0.5} 达到 0.62,如表 5 所示。其中,在 JFLEG 测试集上的 GLEU 值达到 0.61,超出 Xie 等人^[15]在 JFLEG 测试集结果约 5%。该文 GEC 模型精确度达到 0.73,在纠正英语错误的准确度上具有明显优势。

表 5 现有工作实验结果对比

现有相关工作	数据增广方法				学习者语料				测试集			
	基于规则	基于翻译		纠错系统	FCE	NUCLE	W&+ LOCNESS	Lang-8	JFLEG GLEU	CoNLL-2014(M ²)		
		反向翻译	往返翻译							<i>P</i>	<i>R</i>	<i>F</i> _{0.5}
Awasthi 等 ^[10] (2019)	✓			bert	✓	✓		✓	0.60	0.66	0.43	0.59
Xie 等人 ^[15] (2018)		✓		NMT				✓	0.56	0.54	0.35	0.49
Grundkiewicz ^[16] (2019)	✓			NMT	✓		✓	✓	0.61	-	-	0.61
Choe ^[17] (2019)	✓			NMT	✓	✓	✓	✓	-	0.75	0.34	0.60
Lichtarge ^[18] (2020)			✓	NMT	✓	✓	✓	✓	0.64	0.69	0.44	0.62
Kanekol ^[19] (2020)	✓			bert	✓	✓	✓	✓	0.61	0.69	0.46	0.63
文中方法	✓	✓		NMT	✓	✓	✓	✓	0.61	0.73	0.39	0.62

3.6 测试实例结果

该文使用 GEG₁、GEG₂、GEG₃ 模型合成的 200 M 训练数据与三种学习者语料共同训练三种语法错误纠正模型,相应语法错误纠正模型的纠正句分别为纠正句 1、纠正句 2、纠正句 3。源句是包含语法错误句,参考句是标准的纠正句,CoNLL-2014 测试集中的部分实例纠正结果如表 6 所示。结果表明,使用 GEG₃ 模型合成的数据训练英语语法纠错模型的性能优于 GEG₁、GEG₂ 模型。

4 讨论与分析

4.1 不同 GEC 模型纠正结果对比分析

由表 6 示例 1 所示,通过纠正句 1,发现模型可以对冠词错误较好的初始纠正,这得益于构建词性候选集融入的冠词错误。纠正句 2 所有错误均被纠正,源于规则与反向翻译的数据增广方法融合使用,使生成的训练数据覆盖更加丰富的错误类型,合成的数据不

受固定规则、固定候选集的限制,进一步提高 GEC 模型纠正动词等错误的性能。

如示例 2 所示,纠正句 1 与源句相比,不存在标点错误,GEC 模型对标点错误有一定识别纠正能力。但是纠正句 2 并没有进一步纠正,说明在某一上下文语境中,GEC 模型对特定语法错误的纠正仍存在缺陷。交替训练的使用,部分程度弥补了此缺陷,如纠正句 3 所示。GEC 模型对错误的检测与纠正力度进一步提高,但是仍然存在部分错误未被纠正。

如示例 3,通过该文建立的基于规则的数据增广方法,合成的训练数据存在固定词语搭配失误、介词搭配失误、拼写错误等类型,因此训练得到的 GEC 模型一定程度强化了对此类型的识别与纠正。

4.2 基于规则的数据增广方法结果分析

使用基于规则的数据增广方法,随着数据规模的不断增加,训练得到的语法纠错模型性能不断提高。当使用 200 M 合成数据训练 GEC 模型时,其在召回率

上的表现超过仅使用学习者语料训练 GEC 模型的表现,这得益于该文使用的数据增广策略,即在生成训练数据的过程中,融入不同单词的词性错误。与简单的从词表中随机选词替换相比,错误类型更加丰富且具有针对性,召回率得到显著提高。

表 6 GEC 模型在不同错误语句中的表现

示例 1	
源句	First of all , people saves money by using internet to contact other people and reading news .
纠正句 1	First of all , peoplesaves money by using the internet to contact other people and reading news .
纠正句 2	First of all , peoplesave money by using the internet to contact other people and read news .
纠正句 3	First of all , peoplesave money by using the internet to contact other people and read news .
参考句 1	First of all , peoplesave money by using the internet to contact other people and to read news .
参考句 2	First of all , peoplesave money by using the Internet to contact other people and read the news .
示例 2	
源句	If a gene runs in the family , one of the family member test positive , whom does he need to tell .
纠正句 1	If a gene runs in the family , one of the family members test positive , whom does he need to tell?
纠正句 2	If a gene runs in the family , one of the family members test positive , whom does he need to tell?
纠正句 3	If a gene runs in the family , one of the family memberstests positive , whom does he need to tell ?
参考句	If a gene runs in the familyand one of the family members tests positive , whom does he need to tell ?
示例 3	
源句	In conclude , socia media benefits people in several ways but in the same time harms people .
纠正句 1	Inconclusion , social media benefits people in several ways but at the same time harms people .
纠正句 2	Inconclusion , social media benefits people in several ways but at the same time harms people .
纠正句 3	Inconclusion , social media benefits people in several ways but at the same time harms people .
参考句 1	Inconclusion , social media benefit people in several ways , but at the same time harm people .
参考句 2	To conclude, social media benefits people in several ways , but at the same time harms people .

4.3 基于规则、翻译的数据增广方法融合使用结果分析

该文将规则方法合成的数据与学习者语料共同训练错误生成模型。与直接将规则数据训练 GEC 模型相比,使用前者错误生成模型合成的数据训练 GEC 的效果更佳,表明不同数据增广方法融合使用的有效性。

如表 3 所示,与 GEG₁ 模型相比,使用 GEG₂ 模型产生的 80 M 合成数据训练 GEC 模型,提高了纠错模型的召回率和 $F_{0.5}$ 值。原因是相比于基于规则的数据增广方法,基于反向翻译的数据增广方法为模型提供更加丰富的训练信息。同时,规则合成数据的使用加强了模型对特定规则错误的检测与纠正,最终提高模型的召回率。

4.4 交替训练及现有工作对比分析

由表 3 可以看出,通过 GEG 模型和 GEC 模型的交替训练,再次提高了 GEG₃ 的性能。由于交替训练不断强化模型识别不易检测的错误,明显改进错误生成模型合成数据的质量,进而提高 GEC 模型性能。GEC 模型的提高还得益于学习者语料的多次使用: GEG₁ 仅使用规则数据训练,未使用学习者语料; GEG₂ 模型在使用规则数据预训练的基础上使用学习者语料微调; GEG₃ 与 GEG₂ 模型相比,训练数据再次增加 GEC 模型纠正学习者语料后的数据。

如表 4 所示,200 M GEG₃ 模型合成数据与学习者语料对 GEC 模型预训练、微调的方法比使用同规模合成数据、训练方式的 GEG₁、GEG₂ 模型取得更优的结果,原因是学习者语料的重复利用,提高模型的性能。将 GEC 模型预训练数据扩增:使用规模 400 M 的混合数据显著高于仅使用 200 M 合成训练的 GEC 模型,这表明预训练数据的增加一定程度上提高了模型的性能。

如表 5 所示,文中模型在精确度上优于 Xie 等人提出的模型。由于重复使用较多的学习者语料,模型对错误的纠正精确度随着交替训练不断被强化。其次,模型在纠正过程中更好地利用语句上下文信息,模型对于语句流利程度的纠正控制表现较好,在 GLEU 值上表现超过 Xie 等人的 8% 左右。Grundkiewicz 等人使用与文中相近规模的数据集,在精确度上低于文中模型。该文使用规则与反向翻译融合的数据增广方法,利用不同数据增广方法之间的优势可以合成质量较高的训练数据,并且交替训练的加入,强化模型学习到更多原来未学习到的语法信息,为模型的训练带来较大增益,加强了 GEC 模型对错误类型的识别与纠正。

5 结束语

在训练数据短缺的情况下,如何充分利用丰富的单语数据以提高 GEC 模型的性能变得至关重要。在大量的单语数据下,优异的数据增广方法可以使预训练模型获得好的初始化参数。提出利用不同数据增广方法之间的优势来提高 GEG、GEC 模型的性能。同时,模型交替训练的加入,明显改进了模型的性能。未

来工作中,将继续探索 GEG、GEC 模型的优化方法以及探索通过单一错误纠正、多模型集成、重排序等策略进行错误纠正工作。

参考文献:

- [1] 谭咏梅,杨一泉,杨 林,等. 基于 LSTM 和 N-gram 的 ESL 文章的语法错误自动纠正方法[J]. 中文信息学报,2018,32(6):19-27.
- [2] 诸凯丽. 基于分类模型的英语语法纠错算法研究[D]. 杭州:浙江大学,2019.
- [3] 邓俊锋. 基于神经机器翻译方法的英语语法错误纠正研究[D]. 哈尔滨:哈尔滨工业大学,2019.
- [4] 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,2021,41(12):3540-3545.
- [5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. Long Beach: ACM,2017:6000-6010.
- [6] XU S, ZHANG J, CHEN J, et al. Erroneous data generation for grammatical error correction [C]//Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications. Florence: ACL,2019:149-158.
- [7] LICHTARGE J, ALBERTI C, KUMAR S, et al. Corpora generation for grammatical error correction[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies. Stroudsburg: ACL,2019:3291-3301.
- [8] NAPOLES C, SAKAGUCHI K, POST M, et al. Ground truth for grammaticality correction metrics [C]//Association for computational linguistics. Australia: ACL,2015:588-593.
- [9] POPEL M, TOMKOVA M, TOMEK J, et al. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals [J]. Nature Communications,2020,11(1):1-15.
- [10] AWASTHI A, SARAWAGI S, GOYAL R, et al. Parallel iterative edit models for local sequence transduction [C]//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing. Australia: ACL,2019:4260-4270.
- [11] BRYANT C, FELICE M, ANDERSEN E, et al. The BEA-2019 shared task on grammatical error correction [C]//Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications. Florence: ACL,2019:52-75.
- [12] ZHAO W, WANG L, SHEN K, et al. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data [C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies. Stroudsburg: ACL,2019.
- [13] SENNRICH R, HADDOW B, BIRCH A J A. Improving neural machine translation models with monolingual data [C]//54th annual meeting of the association for computational linguistics. association for computational linguistics. Berlin: ACL,2016:86-96.
- [14] CHELBA C, MIKOLOV T, SCHUSTER M, et al. One billion word benchmark for measuring progress in statistical language modeling [C]//15th annual conference of the international speech communication association. Singapore: ISCA,2014:2635-2639.
- [15] XIE Z, GENTHIAL G, XIE S, et al. Noising and denoising natural language: diverse backtranslation for grammar correction [C]//Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies. Stroudsburg: ACL,2018:619-628.
- [16] GRUNDKIEWICZ R, JUNCZYS-DOWMUNT M, HEAFIELD K. Neural grammatical error correction systems with unsupervised pre-training on synthetic data [C]//Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications. Stroudsburg: ACL,2019:252-263.
- [17] CHOE Y J, HAM J, PARK K, et al. A neural grammatical error correction system built on better pre-training and sequential transfer learning [C]//Proceedings of the fourteenth workshop on innovative use of nlp for building educational applications. Stroudsburg: ACL,2019:213-227.
- [18] LICHTARGE J, ALBERTI C, KUMAR S. Data weighted training strategies for grammatical error correction [J]. Transactions of the Association for Computational Linguistics,2020,8:634-646.
- [19] KANEKO M, MITA M, KIYONO S, et al. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction [C]//Proceedings of the 58th annual meeting of the association for computational linguistics. [s. l.]: ACL,2020:4248-4254.