基于数据增强和多任务特征学习的中文语法错误检测方法

谢海华[1] ✉，陈志优[1]，程静[1]，吕肖庆[1,2]，汤帜[1,2]
1. 北大方正集团有限公司数字出版技术国家重点实验室，北京市海淀区，100871
2. 北京大学王选计算机研究所，北京市海淀区，100871
xiehh@founder.com.cn

摘要

由于中文语法的复杂性，中文语法错误检测（CGED）的难度较大，而训练语料和相关研究的缺乏，使得CGED的效果还远达不到能够实用的程度。本文提出一种CGED模型，采用数据增强、预训练语言模型和基于语言学特征多任务学习的方式，弥补训练语料稀缺的不足。数据增强能够有效地扩充训练集，预训练语言模型蕴含丰富的语义信息有助于语法分析，基于语言学特征多任务学习对语言模型进行优化则可以使语言模型学习到跟语法错误检测相关的语言学特征。本文提出的方法在NLPTEA的CGED数据集进行测试，取得了优于其他模型的结果。

关键词： 中文语法错误检测 ； CGED ； 数据增强 ； 多任务学习

Chinese Grammar Error Detection based on Data Enhancement and Multi-task Feature Learning

Haihua Xie[1] ✉, Zhiyou Chen[1], Jing Cheng[1], Xiaoqing Lyu[1,2], Zhi Tang[1,2]
1. State Key Laboratory of Digital Publishing Technology,
Peking University Founder Group Co. LTD., Beijing, China, 100871
2.Wangxuan Institute of Computer Technology,
Peking University, Beijing, China, 100871
xiehh@founder.com.cn

Abstract

Due to the complexity of Chinese grammars, Chinese grammar error diagnosis (CGED) is a challenging task, and the lack of training corpus and relevant study makes the current approaches of CGED still far from practical applications. In this paper, we propose a CGED model to compensate for the low-resource defect using data augmentation, pre-trained language model and multi-task learning of linguistic features. Data augmentation can effectively expand the training set, and pre-trained language models are rich in semantic information that is conducive to grammatical analysis. Meanwhile, enhancing language models based on multi-task learning of linguistic features enables the model to learn linguistic features useful for grammatical error diagnosis. The method proposed in this paper was tested on the CGED dataset of NLPTEA and obtained better results than other models.

Keywords: Chinese Grammar Error Detection , CGED , Data Enhancement , Multi-task Learning

第十九届中国计算语言学大会论文集，第761页-第770页，海南，中国，2020年10月30日至11月1日。

1 引言

中文语法错误检测（Chinese Grammatical Error Diagnosis，CGED）的目标是自动检测出中文自然语句中的语法错误，例如：成分缺失或多余，语序不当等。CGED的检测任务一般包含：是否存在错误、错误类型、错误发生位置。虽然不能给出纠正错误的建议，CGED对于辅助写作和文档审校等场景依然十分有意义。在辅助写作中，CGED给出语法错误类型和位置以让作者针对性地修改文章，可以提升写作的质量和效率。另外，在出版行业的审校环节，由于正式出版物的格式要求十分严格，CGED自动检测出一些基础的语法错误有助于节省审校人员大量的时间，而直接纠正语法错误则可能造成文章的内容和逻辑发生变化。

目前，有关语法错误检测的研究大多数是针对英文的。与英文相比，中文的语法更加复杂和灵活。中文不存在词语的单复数和时态等明确的语法规则，其语法错误经常涉及隐晦的语义解析而不能基于字词形态来判断 (Fu, et al., 2018)。因此，现有的英文语法错误检测方法不能很好地适用于CGED。另外，目前研究者倾向于运用生成式的方法直接进行语法改错，跳过了语法错误检测的步骤 (Chris, Dolan and Gamon, 2018; Zheng and Briscoe, 2018; Zhou, et al., 2018)。只有少量的研究采用序列标注方法进行中文语法错误检测。然而，由于缺乏大规模高质量的标注语料作为训练集，CGED的准确率往往不高，达不到实用水平。如何在训练数据有限的情况下提高语法错误检测的效果是该类研究的一个难点。

针对上述问题，本文提出一种基于数据增强和语言学特征多任务训练方法来提升中文语法错误检测的效果。针对训练语料不充足的问题，本研究使用大量无标签的正确中文语料，通过词性规则、句法规则以及语言模型概率统计等方法来生成接近真实语法错误用例的样本，以扩充训练语料。此外，本研究采用预训练语言模型对字词进行表征，以利用大规模语料蕴含的语义信息，并将词法学习、句法学习、语法错误检测等任务结合进行多任务学习，进一步获取中文语义和语法信息。本文提出的方法在NLPTEA CGED评测任务数据集进行测试，准确率和召回率分别为85.16%和72.53%（F1值为0.783）,性能优于其他中文语法检测模型。

2 相关工作

中文语法错误自动检测模型采取的方法从最初的统计学习方法 (Chang, Wu and Prasetyo, 2012)和基于规则的分析 (Lee, et al., 2013)，到现在主流的深度学习算法 (Fu, et al., 2018; Yang, et al., 2017)，以及多种模型混合的方法 (Li, et al., 2018)。大多数研究采用序列标注模型来进行语法错误检测，并使用LSTM和CRF来实现 (Fu, et al., 2018; Yang, et al., 2017; Zhao, Li and Lin, 2018)。使用LSTM模型实现语法错误检测时，特征的选择十分重要，除了通常使用的字向量特征、词向量特征、词性POS特征，很多研究提出了许多新的特征 (Fu, et al., 2018; Li, et al., 2018; Zhao, Li and Lin, 2018)。例如：高斯互信息（ePMI）、向量词的共现(AWC)、依赖关系词语的共现（DWC）、基于语境的词表达等。也有一些研究针对LSTM模型结构进行改进，比如在LSTM模型中加入策略梯度 (Li and Qi, 2018)。这些研究的重点在于学习中文语法规律，基于无标注语料统计词语规律和词语用法，并提出相应的特征来提高检测效果。然而，统计特征不能捕获深层的语法和语义信息，因此一些隐晦的语法错误无法被发现。

针对训练语料不足的问题，一些研究者使用未标注的中文语料来构造错误用例，例如：通过随机增加、删除、替换字词和打乱字词顺序来生成错误样本 (Wang, et al., 2019)；统计已有训练语料中的语法错误分布，并构造相应的错误样本 (Zhang, et al., 2018)。前者采用随机方式构造的语法错误样本，往往显得不够真实，其语法错误分布与正常写作者所犯错误的分布相差较大。而后者构造的错误数据过于拟合已有的训练样本，不利于模型的泛化。

近年来，一些学者利用基于大规模语料预训练的语言模型来获取文本的语言学特征，以弥补训练语料的不足。基于预训练语言模型的语法错误检测模型，其效果优于通过融合多种特征构建的模型 (Bell, Yannakoudakis and Rei, 2019; Kaneko and Komachi, 2019)。不过这些方法都以英文为研究对象，它们尚未在中文数据集上进行试验或者测试性能。

大多数情况下，语法错误检测的目的是为了对语法错误进行纠正。在检测出语法错误的类型和发生位置之后，可以根据错误类型，采用相应的方法来修改语法错误。例如：错误提示为“成分冗余”，则直接删除该成分；错误提示为“用词不当”，则基于词语统计信息（例

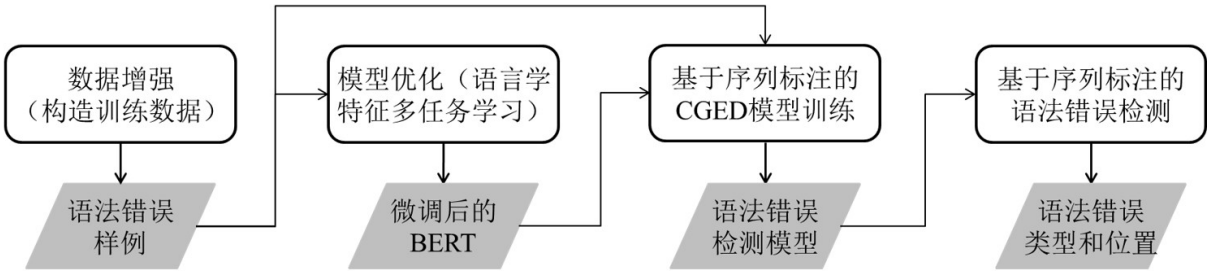

Figure 1: 基于数据增强和语言学特征多任务学习的中文语法错误检测系统框架

如：PMI）推荐候选词语以替换错误词语 (Fu, et al., 2018; Zhang, et al., 2018)。不过目前中文语法错误纠正的研究大多采用端到端的生成式方法，使用统计翻译模型 (Chris, Dolan and Gamon, 2018)、深度学习模型 (Zheng and Briscoe, 2018)、融合规则和统计的算法 (Zhou, et al., 2018)等，由错误句子直接生成正确的句子。但是生成的结果有时会改变原文的表达方式甚至语义和逻辑，在很多情况下不能产生令人满意的结果。

3 基于数据增强和语言学特征多任务学习的CGED模型

这一节将详细介绍我们提出的CGED模型。为了解决训练语料缺乏的问题，我们采用数据增强方法来扩充训练数据集，采用预训练语言模型BERT (Devlin, et al., 2019)作为基础的文本表征提取工具，并运用多任务训练数据来调整BERT参数以使它学习到更多的语言学特征。我们的语法错误检测系统框架见Figure 1。

我们的主要贡献是提出了基于句法分析与预训练语言模型采样的数据增强方法和基于语言学特征多任务学习的模型优化方法。以下章节将对Figure 1所示流程和上述两个贡献进行详细阐述。

3.1 基于句法分析与预训练语言模型采样的数据增强（构造训练数据）

中文语法错误检测研究的主要问题之一是训练语料的缺乏。我们使用大量未经标注的正确语句构造含有语法错误的训练样例，以弥补训练数据不足的问题。中文维基百科覆盖面广且表达方式丰富，人民日报表达方式规整规范，所以我们以维基百科和人民日报中文数据集为基础，抽取其中正确的语句，并对数据进行处理后构造训练样本。主要步骤的介绍如下。

1. 数据集预处理，主要的处理手段如下。
   - 增加数据的一致性和减少噪音，例如：将中文维基百科的繁体中文转化成简体中文，把全角字符转化为半角字符。
   - 运用中文处理工具对文本进行分词、词性标注、命名实体识别和依存句法分析。
   - 选择质量较高的句子，例如：去除过长（词数超过100个）和过短（词数小于3个）的句子。
2. 错误样例构造。本步骤将一些正确的语句改造为含有语法错误的语句。在语句经过分词、词性标注和依存句法分析之后，我们采用以下措施，构建不同类型的语法错误的训练样本。
   - 成分冗余构造：在语句的词语之间随机插入没有实际意义的词语。候选的插入词语选自停用词表。
   - 成分缺失构造：从主谓结构片段中删除主语或者谓语，从动宾结构片段中删除谓语或者宾语，从状中结构或者定中结构片段中删除被修饰成分。
   - 语序不当构造：修改动宾结构、状中结构、定中结构等结构片段中成分的顺序。
   - 用词不当构造：随机选取一个词语并将其遮盖（用MASK将其替换），然后用BERT的Masked LM预测出的候选字替换原来的字符。

为了保证改造后的句子在含有语法错误的同时，保持语句的基本语义和结构，以免发生意思改变，我们设计了以下规则。

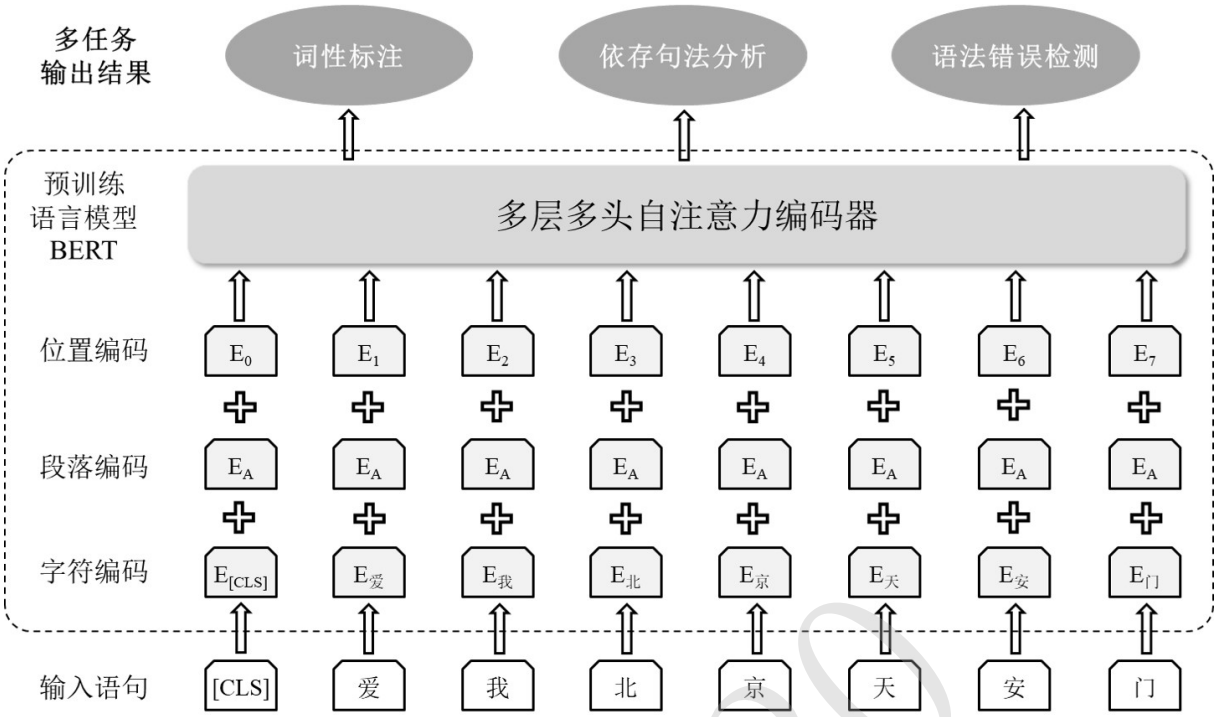

Figure 2: 基于语言学特征的多任务学习进行BERT模型优化

1. 不对命名实体进行修改。命名实体在句子中往往是主体成分，修改命名实体会改变句子的意思，例如：句子“协和医院是中国最好的医院之一，专治各种疑难杂症”，如果对“协和医院”进行修改，语句的意思就会发生变化。
2. 对于短句子，我们构造的样例中只含有一个错误。对于15个词语以上的句子，我们会随机增加错误。
3. 在成分缺失和语序不当构造时，规避修改依赖距离很远的结构成分，防止破坏语句结构。
4. 关于用词不当构造，除了构造“的地得”之间的误用情况，不对虚词、语气词之类无意义的词语进行修改以构造该类错误。实际样例中，虚词的使用错误主要是成分缺失和冗余。

以下是两个构造的错误样例。

1. 样例一：语序不当构造
   - 原句：**加速推广**菌草技术，将其列入国家开发计划。
   - 构造句：**推广加速**菌草技术，将其列入国家开发计划。
2. 样例二：用词不当构造
   - 原句：我跟朋友们经常用手机**打**电话聊天。
   - 构造句：我跟朋友们经常用手机**找**电话聊天。

3.2 基于语言学特征多任务学习的模型优化

在以往的CGED研究中，研究者使用的主流模型是BiLSTM-CRF结构。由于中文语法错误的复杂性和多样性，语法的正确使用与语言学特征高度相关，因此使用少量的训练数据很难训练出一个鲁棒性好的CGED模型，人们会在模型中加入词性、N-gram、PMI等语言学特征。但是，大量特征的使用使得模型结构繁琐，而且提取这些特征信息也大大降低了模型的运行速度。

我们使用BERT之类的预训练语言模型作为基础来构建CGED模型，以利用它们在预训练阶段学习到的深层语义信息。然后，我们采取多任务学习方式对BERT的参数进行调整，使模

	我	爱	北	京	天	安	门
我	0	SBV	0	0	0	0	0
爱	SBV	HEAD	0	0	VOB	VOB	VOB
北	0	0	0	0	ATT	ATT	ATT
京	0	0	0	0	ATT	ATT	ATT
天	0	VOB	ATT	ATT	0	0	0
安	0	VOB	ATT	ATT	0	0	0
门	0	VOB	ATT	ATT	0	0	0

HEAD： 主干词
SBV： 主谓关系
VOB： 动宾关系
ATT： 定中关系

Figure 3: 依存句法结构矩阵示例

型学习到各种语言学知识，并在预测阶段不必进行语言学特征提取，以提高模型的性能和效率。

多任务学习是指为模型设置多个训练目标，这些任务之间具有一定关联，并在训练阶段可以互相促进以达到更好的训练效果。多任务学习可以通过在模型上设置一些共享参数来实现。本文提出的方法使用BERT作为模型的共享部分，并使用不同结构来实现词性标注、句法分析和语法错误分类三个具体任务。基于语言学特征的多任务学习进行BERT模型优化的结构如Figure 2。

Figure 2所示模型的输出目标包括：词性标注，依存句法分析和语法错误检测。基于这三项任务的训练，可以对BERT的参数进行优化，以使BERT能学到更多的语言学知识。我们认为，这三个任务之间有互相促进的作用，词性和句法分析的结果能辅助判断语句是否有语法错误，例如：图2示例句是一个语法错误句，它的词性标注的结果是：动词-代词-名词，这个词性序列在中文语句中不常见，因此该句很可能含有语法错误。同样地，判断出语句含有语法错误，也有益于更准确地分析语句的词性和句法。这三个任务的详细描述如下。

1. 词性标注

   我们采用序列标注方法来实现词性标注任务，在BERT之后增加一个全连接层直接输出词性结果。由于BERT采用字符嵌入方式，对于多字符词语，我们采用“BI”的标注方式（‘B’表示词语开始位置，‘I’表示词语中间或结束位置）进行词性标注。在准备训练数据时，词性标注的标签可以由中文处理工具（例如pyltp (pyltp, 2020)）直接生成，标注示例如Table 1。

Character	爱	我	北	京	天	安	门
POS tag	B-v	B-r	B-n	I-n	B-n	I-n	I-n

Table 1: 词性标注示例

2. 依存句法分析

   依存句法分析的目的是确定语句的句法结构，通常以句法树的形式，用有向弧表示词语之间的修饰及指向关系（即依存关系）。在本文中，我们将句法结构（或词语之间的依存

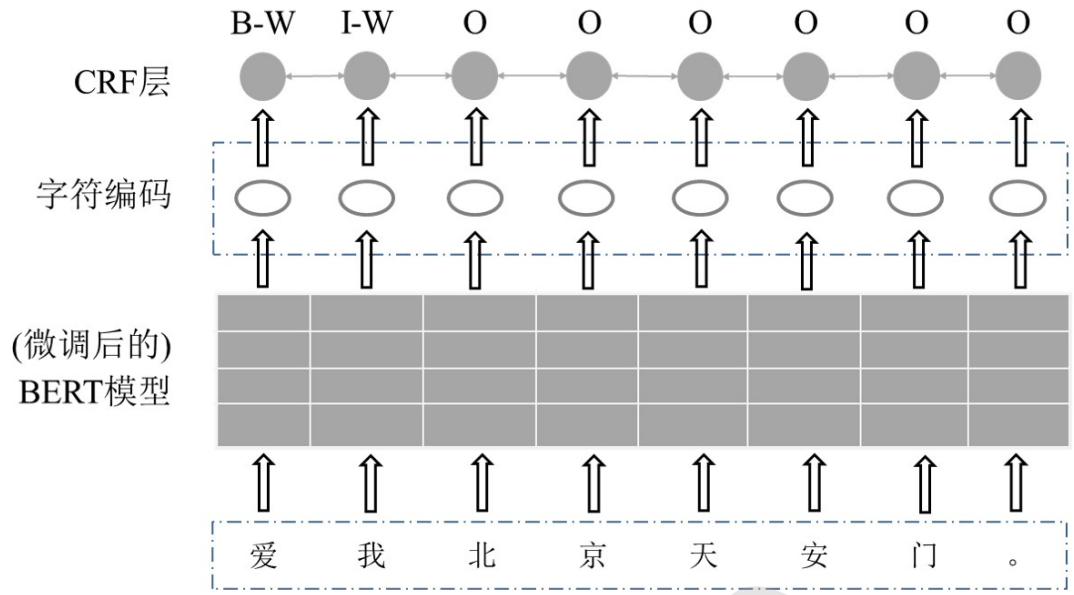

Figure 4: 基于BERT-CRF架构的中文语法错误检测模型

关系）用矩阵形式来表示。对于一个含有n个字的句子，用一个$n \times n$的矩阵表达词语之间的依存关系。为了避免关系矩阵（记为M）过于稀疏，我们将依存关系进行简化，取消修饰词和被修饰词之间的指向关系，所以M是一个对称矩阵。假设语句的第i个词（含有一个字符，在句子中的序号设为i^c）与第j个词（含有三个字符，在句子中的序号分别为j_1^c，j_2^c，j_3^c）之间的关系为动宾关系（VOB），则有$M_{i^c j_1^c} = M_{i^c j_2^c} = M_{i^c j_3^c} = VOB$，而且$M_{j_1^c i^c} = M_{j_2^c i^c} = M_{j_3^c i^c} = VOB$。我们将语句的主干词对应的对角线位置的值设置为Head，而对角线上其他位置的值设为0。以矩阵表示的句法结构示例如Figure 3。

在准备训练数据时，语句的句法结构矩阵可以由中文处理工具生成的句法树修改而成。在参数优化阶段，假设输入语句为S，其文本序列长度为t，经过BERT之后的语义表征为S_{BERT}，它的维度为$t \times 768$。然后采用以下公式产生两个中间变量H_1和H_2。

$$H_i = f(W_i S_{BERT} + b_i) \tag{1}$$

f表示对矩阵进行形变操作的函数，W_i和b_i是随机初始化并在训练中更新的参数。产生的H_1和H_2的维度都是$64 \times t \times 12$。然后基于以下公式产生句法结构分析结果。

$$M = Softmax[W(H_1 \cdot H_2^T)] \tag{2}$$

M的维度$64 \times t \times t$，对应$t \times t$矩阵的每个元素的数值（维度是1×64），即句法结构矩阵的结果。

3. 语法错误检测

我们采用多标签分类的方法完成语法错误检测任务，在BERT之后增加一个全连接层直接输出分类结果。分类的结果是句子含有的语法错误的类型。如果语句不含语法错误则输出“没有错误”，如果它含有多个语法错误则输出多个语法错误标签。语法错误检测的训练数据是由前文所述方法构造出来或者实际写作中产生。

上述三个任务模型的损失函数都用交叉熵来计算。多任务学习模型的损失函数是这三个模型的损失函数之和，模型训练的目标是最小化该损失函数。

Model	FPR	Detection level				Identification level				Position level			
		Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
NLPTEA-18-HSK													
*B0	0.2138	0.7153	0.8395	0.6759	0.7489	0.6253	0.6518	0.4305	0.5185	0.4935	0.4395	0.248	0.3171
*B0+MTL	0.2438	0.7416	0.8354	0.733	0.7809	0.6293	0.6309	0.4854	0.5486	0.472	0.4062	0.2785	0.3304
*B0+MTL+DA	0.2106	0.743	0.8516	0.7253	0.7833	0.6462	0.6619	0.4827	0.5583	0.4843	0.4234	0.2769	0.3348
NLPTEA-16-HSK													
*B0	0.2182	0.7828	0.7771	0.7792	0.7782	0.7371	0.6993	0.582	0.6353	0.6492	0.5703	0.4173	0.482
*B0+MTL	0.2619	0.7768	0.7491	0.8173	0.7817	0.7176	0.663	0.6225	0.6421	0.617	0.5313	0.452	0.4884
*B0+MTL+DA	0.1975	0.795	0.7922	0.8074	0.7997	0.7457	0.7063	0.5996	0.6485	0.6543	0.5713	0.4341	0.4933
NLPTEA-16-TOCFL													
*B0	0.2132	0.6905	0.7512	0.6005	0.6676	0.618	0.5858	0.3753	0.4575	0.5024	0.3671	0.1986	0.2576
*B0+MTL	0.2666	0.7044	0.7514	0.6773	0.7124	0.608	0.5514	0.4322	0.4863	0.5066	0.3698	0.2301	0.2836
*B0+MTL+DA	0.2495	0.706	0.7498	0.6742	0.71	0.617	0.5761	0.4244	0.4886	0.5123	0.3718	0.2283	0.2828

Table 2: 中文语法错误检测模型的对比实验结果

3.3 基于序列标注的CGED模型训练和应用

我们把CGED视为序列标注问题，并选用BERT-CRF结构作为模型的基本架构，其中BERT的参数经过经过2.2节所述方法进行调整，见Figure 4。由于我们处理的对象是中文数据，我们使用中文BERT模型，它是基于大量中文维基百科语料预训练而成。在BERT之后使用CRF模型 (Sutton and McCallum, 2012)，一种经典的序列标注方法，直接生成语法错误检测的结果。语法错误标签使用“BIO”方式编码，“B”代表错误的开始位置，“I”表示中间或者结束位置，“O”表示当前字符没有语法问题。例如对于错误X，“B-X”代表“X”错误的第一个位置，“I-X”表示其他位置。

在训练阶段，训练数据集的部分数据来自人们在实际写作中出现的语法错误，而另一部分则来自前文所述方法构造出的数据。训练模型和预测模型的结构是一样的，输出的结果包含：是否存在错误，错误类型以及错误发生的位置。

4 中文语法错误检测实验

我们采用NLPTEA中文语法错误检测评测数据集[18]试验了我们的方法。NLPTEA提供一份标注过的语法错误数据集，语料来源是汉语非母语的汉语学习者在中文写作当中产生的错误样例。该数据集将语法错误分为四种类型：redundant errors（记为‘R’，即成分冗余），missing words（记为‘M’，即成分缺失），word selection errors（记为‘S’，即用词不当）和word ordering errors（记为‘W’，即词序不当）。数据集里的语句可能没有语法错误，也可能含有一个或多个语法错误。语法错误检测系统需要从以下三个方面对语句进行检测：

- Detection-level：检测语句是否含有语法错误。
- Identification-level：语句含有的语法错误的类型。
- Position-level：语句含有的语法错误的发生位置。

4.1 数据收集和处理

我们使用pyltp中文处理工具对语句进行分词、词性标注和依存句法分析，同时采用pyltp的标注体系。在多任务学习优化BERT时，我们使用了一些公开数据集来提升分词的准确性，以提高词性标注和依存句法分析的准确度。

我们收集了NLPTEA 2016，IJCNLP 2017和NLPTEA 2018的CGED任务的评测数据集，共有语句数量为20,451，按照句号、问号和感叹号拆分之后的语句数量为104,141。选择其中的80%数据作为训练数据，其余数据为校验数据。同时，我们收集和整理了中文维基百科数据集和人民日报数据集，使用2.1节介绍的数据构造方法生成训练数据（语句总数为138,825）并加入到训练集。为了维持正确语句和错误语句的比例，我们在数据集中加入了同等数量的不含语法错误的语句。

Model	FPR	Detection level				Identification level			Position level		
		Acc.	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
B0+MFF+DA	0.2106	0.743	0.8516	0.7253	0.7833	0.6619	0.4827	0.5583	0.4234	0.2769	0.3348
HFL											
*run1	0.1613	0.7101	0.8276	0.609	0.7017	0.7107	0.4173	0.5259	0.5341	0.2729	0.3612
*run2	0.7554	0.6436	0.6171	0.9572	0.7504	0.3931	0.7331	0.5118	0.1441	0.3886	0.2102
*run3	0.1754	0.7278	0.8254	0.6517	0.7283	0.6874	0.4588	0.5503	0.4752	0.2906	0.3606
CMMC-BDRC											
*run1	0.5314	0.6889	0.6736	0.8621	0.7563	0.4834	0.5952	0.5335	0.2741	0.3177	0.2943
*run2	0.3574	0.6988	0.7266	0.7408	0.7336	0.5831	0.4955	0.5357	0.3839	0.2966	0.3346
*run3	0.347	0.663	0.7109	0.6709	0.6903	0.4853	0.4096	0.4442	0.2482	0.1814	0.2096
NCYU											
*run1	0.9987	0.5596	0.5598	0.9985	0.7174	0.2381	0.9749	0.3828	0.0030	0.0390	0.0056
*run2	0.9994	0.5599	0.5599	0.9995	0.7177	0.2382	0.9752	0.3828	0.0030	0.0384	0.0056
*run3	0.9994	0.5599	0.5599	0.9995	0.7177	0.2382	0.9752	0.3828	0.0030	0.0380	0.0055

Table 3: BERT+MFF+DA与NLPTEA 2018 CGED评测模型的对比

4.2 实验结果

我们按照2.2节介绍的方法，运用训练数据对BERT的参数进行调整。然后使用训练数据对语法错误检测的BERT+CRF模型进行训练，使用校验数据进行测试。我们同时使用不同的模型进行了对比实验，Table 2显示了对比实验的结果。其中，B0表示未经过优化的BERT模型，MTL表示多任务学习方法，DA表示数据增强，B0+MTL+DA则表示文本采用的方法。不同的模型分别在NLPTEA 2018 CGED任务的HSK测试集（NLPTEA-18-HSK）、NLPTEA 2016 CGED任务的HSK测试集（NLPTEA-16-HSK）和TOCFL（NLPTEA-16-TOCFL）测试集上进行了实验。

对比实验结果表明，使用语言学特征对BERT进行优化之后，语法错误检测的效果在各方面都有明显的提升，特别是检测的召回率得到很大提高。但是随着召回率的上升，检测精确率有一定程度的下降，不过数据增强的使用很好地弥补了这个问题，使得模型能够同时提高检测得召回率和精确率，并使F1指标提升。

我们与NLPTEA 2018 CGED评测结果进行了横向对比。我们没有采用模型融合以进一步提高检测效果，只用单一模型来与NLPTEA 2018评测效果较好的模型进行对比，结果见Table 3。HFL，CMMC-BDRC和NCYU是NLPTEA 2018评测结果里面准确率，召回率或者F1值较高的模型。在Detection Level和Identification Level这两个测试指标上，我们的单模型都取得了最优的F1值。但是在Position Level指标上，我们方法的效果不如HFL。经过分析，我们认为这可能是因为构造的错误案例与实际测试的错误案例在错误分布的不一致而造成的。

5 总结与展望

我们针对中文语法错误检测研究存在的主要问题之一，训练语料的缺乏，采用数据增强、预训练语言模型和语言学特征多任务学习的方式，有效地提高了语法错误检测的效果。使用语言学特征对语言模型进行优化能够使它学习到显式的语言学特征以及隐藏的语义信息，而语言学特征和语法使用是十分相关的，所以它对语法错误检测效果有明显的改善作用。

由于中文语法的复杂性，我们目前的工作依然存在很多不足，错误类型和位置的检测效果不好。在下一步工作计划中，我们将进一步提高数据构造的合理性，使构造的错误样本更符合人们实际所犯的语法错误。另外，我们会对语言学特征的多任务学习的结构进行改善，以进一步提高CGED任务的检测效果。

致谢

本项研究以下项目的支持：国家重点研发计划（No. 2019YFB1406302），国家自然科学基金项目（No. 61472014, No. 61573028, No. 61432020），北京市自然科学基金项目（No.

4142023），北京新星计划项目（XX2015B010）。感谢所有审稿人给出的宝贵意见和建议。

参考文献

Ruiji Fu, Zhengqi Pei, Jiefu Gong, et al. 2018. Chinese grammatical error diagnosis using statistical and prior knowledge driven features with probabilistic ensemble enhancement. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, NLP-TEA@ACL 2018*: 52-59.

Ru-Yng Chang, Chung-Hsien Wu, and Philips Kokoh Prasetyo. 2012. Error Diagnosis of Chinese Sentences Using Inductive Learning Algorithm and Decomposition-Based Testing Mechanism. *ACM Transactions on Asian Language Information Processing*, 3:1-3:24

Lung-Hao Lee, Li-Ping Chang, Kuei-Ching Lee, et al. 2013. Linguistic rules based Chinese error detection for second language learning. *Proceedings of the 21st International Conference on Computers in Education, 2013*: 27-29.

Yi Yang, Pengjun Xie, Jun Tao, et al. 2017. Embedding grammatical features into LSTMs for Chinese grammatical error diagnosis task. *Proceedings of the IJCNLP 2017, Shared Tasks, 2017*: 41-46.

Chen Li, Junpei Zhou, Zuyi Bao, et al. 2018. A hybrid system for Chinese grammatical error diagnosis and correction. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, NLP-TEA@ACL 2018*: 60-69.

Jianbo Zhao, Si Li, Zhiqing Lin. 2018. Contextualized character representation for Chinese grammatical error diagnosis. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, NLP-TEA@ACL 2018*: 172-179.

Changliang Li, Ji Qi. 2018. Chinese grammatical error diagnosis based on policy gradient LSTM model. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, NLP-TEA@ACL 2018*: 77-82.

Brockett Chris, William B. Dolan, and Michael Gamon. 2018. Correcting ESL errors using phrasal SMT techniques. *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, ACL 2006*: 249-256.

Yuan Zheng, and Ted Briscoe. 2018. Grammatical error correction using neural machine translation. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016*: 380-386.

Junpei Zhou, Chen Li, Hengyou Liu, et al. 2018. Chinese grammatical error correction using statistical and neural models. *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018*: 117-128.

Yongwei Zhang, Qinan Hu, Fang Liu, et al. 2018. CMMC-BDRC solution to the NLP-TEA-2018 Chinese grammatical error diagnosis task. *Proceedings of the 5th Workshop on Natural Language Proceeding Techniques for Educational Applications, NLP-TEA@ACL 2018*: 180-187.

王辰成, 杨麟儿, 王莹莹, 杜永萍, 杨尔弘. 2019. 基于Transformer增强架构的中文语法纠错方法. *The Eighteenth China National Conference on Computational Linguistics, CCL 2019*.

Samuel Bell, Helen Yannakoudakis, and Marek Rei. 2019. Context is key: grammatical error detection with contextual word representations. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2019*: 103-115.

Masahiro Kaneko, and Mamoru Komachi. 2019. Multi-head multi-layer attention to deep language representations for grammatical error detection. *Computacion y Sistemas Vol. 23, No. 3*: 883–891.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*: 4171-4186.

pyltp: the python extension for LTP. 2020. https://github.com/HIT-SCIR/pyltp. *Last accessed 21 May 2020*.

Charles A. Sutton, and Andrew McCallum. 2012. An Introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning 4(4)*: 267-373.

Gaoqi Rao, Qi Gong, Baolin Zhang, et al. 2018. Overview of NLPTEA-2018 Share Task Chinese Grammatical Error Diagnosis. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, NLP-TEA@ACL 2018*: 42-51.