Computer Science and Application 计算机科学与应用, 2023, 13(7), 1373-1381
Published Online July 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.137135

Hans汉斯

基于深度学习的语义级中文自动校对方法

邓晨曦*，蒋一锄*，李合军，彭姣丽，刘曜端，李凌云

湖南环境生物职业技术学院生态宜居学院，湖南 衡阳

收稿日期：2023年6月13日；录用日期：2023年7月11日；发布日期：2023年7月18日

摘　要

中文语法纠错任务是检查和纠正句子中的语法错误，相对于中文拼写错误纠正，中文语法错误纠正面对的错误不仅包括同音字和同形字的错误，还包括多字和少字的情况。本文通过大量的实验验证不同方法的优缺点，基于规则的方法需要消耗大量的人力来构造规则，而基于传统机器学习的方法面临特征提取能力不足的缺点，基于深度学习的方法是目前语法纠错的主要方法，因为语法纠错的文本存在不确定性，所以纠错的结果可能存在多种可能，因此Seq2Seq和预训练语言模型目前取得了较好的效果。

关键词

深度学习，中文语法纠错，Seq2Seq，预训练语言模型

A Semantic Level Chinese Automatic Proofreading Method Based on Deep Learning

Chenxi Deng*, Yichu Jiang*, Hejun Li, Jiaoli Peng, Yaoduan Liu, Lingyun Li

Ecolgical Livable College, Hunan Polytechnic of Environment and Biology, Hengyang Hunan

Received: Jun. 13th, 2023; accepted: Jul. 11th, 2023; published: Jul. 18th, 2023

Abstract

The task of Chinese grammar error correction is to check and correct grammatical errors in sentences. Compared with Chinese spelling error correction, Chinese grammar error correction not

*通讯作者。

文章引用: 邓晨曦, 蒋一锄, 李合军, 彭姣丽, 刘曜端, 李凌云. 基于深度学习的语义级中文自动校对方法[J]. 计算机科学与应用, 2023, 13(7): 1373-1381. DOI: 10.12677/csa.2023.137135

only includes homophone and homomorphic errors, but also includes redundant and missing characters. This paper verifies the advantages and disadvantages of different methods through a large number of experiments. Rule-based methods need to consume a lot of manpower to construct rules, while traditional machine learn-based methods face the disadvantage of insufficient feature extraction ability. Deep learn-based methods are the main methods for grammar error correction at present. Because there is uncertainty in the text of syntax correction, the result of error correction may have a variety of possible results, so Seq2Seq and the pre-trained language model have achieved good results.

Keywords

Deep Learning, Chinese Grammatical Error Correction, Seq2Seq, Pre-Trained Language Models

1. 引言

近年来，由于学习汉语的人数正在逐渐增加，汉语成为了世界上最受欢迎的语言之一，其中有 70 多个国家颁布法令，将汉语作为母语之外的第二外语。在汉语在全是广为流传的大背景下，汉语学习者的新挑战在于如何高效的学习汉语，经过调查，汉语学习者遇到了学习瓶颈，主要原因是汉语语法知识的学习不系统，再加上母语环境的影响，很难准确地识别和纠正语法错误，而语法错误是汉语中最常见的错误，因此汉语语法纠错系统显得尤为重要。不仅如此，在出版行业，语法错误的纠正也同样重要，目前，网络信息技术发展非常迅速，电子文档的数量显著增加并变得越来越丰富，人工纠错方法需要更多的时间和精力，难以适应当前大规模文本纠错的环境，也难以保证人工纠错的质量。中文语法纠错系统的使用可以大大降低人工成本，提高文本质量和纠错速度。在大数据时代，广电行业也依然采用人工校对的方式，而中国载体形式的多样化也增加了人工校对的工作量。现有的文本校对软件需要针对不同的格式和形式，不同的文本载体进行语法校正，这在具体的实践中是难以承受的，因此有必要实现一个高效的语法纠错系统来解放人力资源[1] [2] [3]。

深度学习相关方法，是目前人工智能的核心技术，广泛应用于计算机视觉和自然语言处理任务，文本纠错任务作为自然语义处理任务的子任务，也越来越多的使用到了深度学习技术。深度学习的主要优点是摒弃了传统机器学习中复杂的人工特征提取，而在庞大的语料库中的自动特征提取，是深度学习的主要优势，因此深度学习逐渐流行起来。因此，本文对当前主流的深度学习进行实验，分析了它们在中文语法纠错上的效果。

中文语法纠错领域主流的方法主要包括传统的机器学习方法和深度学习方法[4] [5] [6] [7] [8]，例如 Zhang 等人提出了一种中文拼写纠错纠正方法，将中文的拼写纠错任务分为三个主要的步骤，即错误检测、候选集生成以及纠错纠正，其中错误纠正使用机器学习的方法来对候选词进行选择[5]。面对与更加复杂的语法错误，Wang 等人提出了一种 Seq2Seq 方法，为了避免过多地对原始句子修改，提出了基于复制机制的方法，将原始文本变换为正确的文本，以此来达到纠错的目的[9]。

由于传统的机器学习方法需要人工构造文本特征，来实现查错和纠错，随着时代的发展，不断有新的词汇新的句式表达被提出，因此传统的机器学习方法慢慢被深度学习方法代替。下面介绍几个主要的

深度学习方法在文本纠错的贡献。Wu K 提出了一种深度学习模型用于解决中国古典自然语言处理问题：文本窗口去噪自编码器，以及完整的预训练解决方案[10]。Jun Wei Chen 提出了一种混合序列模型，该方法适用于汉语学习，降低学习成本和反馈时间，并帮助写作者检查错字[11]。Rui Zhang 等人构建了一个两层语义知识库来辅助错误检测和纠错[12]。

如今预训练语言模型(Pre-trained Language Models, PTM)在自然语言处理(Natural language processing, NLP)相关任务上取得较好的效果，NLP 相关的任务都会优先考虑使用 BERT 来对文本编码[13]，PTM 相关模型也是一种深度学习模型，只不过他们会对大量的数据进行预训练，然后应用到下游任务。另一方面，因为 BERT 的预训练的过程与文本纠错的过程十分相似，所以 BERT 也比较适合做文本纠错任务，主要体现在以下两点，第一点 BERT 的架构是基于双向的 Transformer 架构[14]，使用了自注意力机制，能够很好地获取上下文的特征，第二点体现在 Bert 自监督学习过程，会对文本进行掩码处理。

2. 基于深度学习的文本纠错技术

深度学习是人工智能的一个子领域，可以认为它是一种具有多层次表示的表示学习方法，它可以将抽象的概念或模式一层一层地表达出来，类似于将多个简单函数组合在一起的过程，从而使深度学习模型可以表示复杂的变换。早期的词袋模型虽然忽略了上下文语义信息的关联，将每个词作为一个独立的个体，导致大量重要信息的丢失，但同时由于计算能力资源的限制，手工特征提取仍然不能产生良好的结果。自动特征提取可以更快地从中找到最优特征，深度学习的出现大大节省了人工特征提取的人力支出，如利用词嵌入技术将词映射到向量空间，通过端到端训练可以快速完成如：“莫斯科 – 俄罗斯 + 东京 = 日本”的推理。深度学习并不神秘，它是基于是一种神经网络模型，以及数据编程产品的开发思路，深度学习作为一种被大众认可的通用工具，在实际应用场景中，利用深度学习做以下步骤。

1) 明确待解决的问题，找到问题背后隐藏的数学模型。
2) 根据需要将各种不同结构特征的神经网络进行排列组合，找出神经网络背后的数学意义。
3) 根据给定的数据找出优化算法。
4) 如何有效利用硬件性能，训练模型，并指定合理的损失函数，防止函数的过拟合问题。
5) 在选择合适的超参数，防止数值计算的陷阱，选择好的特征方面积累经验。

2.1. 多层感知机

多层感知器(Multi Layer Perceptron, MLP)是神经网络的基础模型，如图 1 所示，其主要结构是由全连接层和 Sigmoid 函数构成。隐层神经元的数量和层数可以看作是超参数，这些超参数往往会影响多层感知机的学习性能和效率。

图 1 展示了多层感知器的结构，可以发现，多层感知器只是一个在单层感知器基础上改进的分层结构，主要是在单层感知器中加入一个或多个隐藏层，数据传输从下往上单向传播，实现数据集输入端数据到数据集输出端的映射关系。由于在任何一层输入中都没有计算，因此上面图 1 所示的 MLP 是一个两层结构，全连接层计算过程如式(1) (2)所示：

$$a_n = \theta\left(W_1 X + b_1\right) \tag{1}$$

$$a_o = a_b W_2 + b_2 \tag{2}$$

如公式(1)所示，其中输入向量记为 X，θ 为激活函数，隐层的输出权值记为 W_1，W_2，偏置因子记为 b_1，b_2。

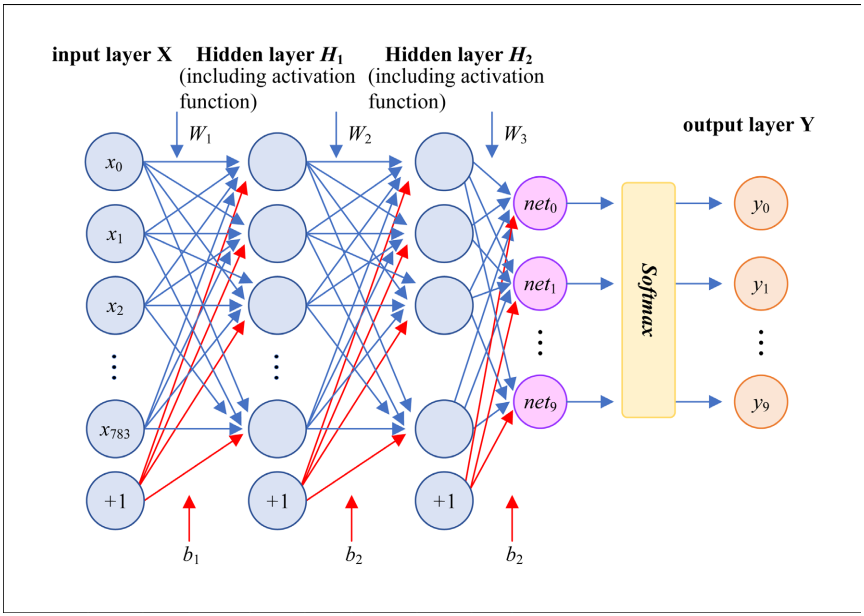

Figure 1. Structure of multi-layer perceptron
图 1. 多层感知机的结构

2.2. 循环神经网络

循环神经网络(Recursive Neural Network, RNN)能从输入的时间维度上，抓取随着时间变化的特征，一段文本序列是基于时间序列的一段数据，RNN 主要用于处理时间序列类型的数据，因此 RNN 更擅长处理文本数据，通过设置状态变量来存储历史信息，此时历史信息和当前状态的输入数据共同决定了此时的输出结果。目前仍使用 RNN 结构和其他类型的深度学习模型来模拟语言模型，因此语言模型是自然语言处理技术发展的基础。前面讨论的 n-gram 基于当前词，只能考虑有限数量的前词的历史信息，无法避免历史信息的缺失。如果 n 出现，它必然会增加 n-gram 的参数数量，其复杂性呈指数级增长，因此 RNN 放弃了这种刚性记忆的方式，增加了隐藏状态来存储历史信息。

多层感知器模型只考虑了输入信息到输出信息的映射，而忽略了对输出结果起关键作用的历史信息，这正是 RNN 擅长的地方，即对包含历史单词消息的文本序列进行建模。RNN 家族最擅长处理基于时间维度的文本数据，而递归神经网络更擅长通过添加隐藏状态来改善 n 元记忆问题。

首先，考虑这样一种情况：序列化数据的输入具有时间顺序，如图 2 所示，如果输入的输入序列表示为 X_t，而 H_t 是 t 时刻的隐藏状态。通过与图 1 的对比，可以发现多层感知器的缺点是没有对历史信息的记录。在图 2 所示的 RNN 结构中，RNN 表示前一时刻的历史输出为 H_{t-1}，此时的权值记 w_2，计算范式如(3)和(4)所示。

$$H_t = \phi\left(X_t w_1, + H_{t-1} w_2 + b_1\right) \tag{3}$$

$$O_t = H_t w_3 + b_3 \tag{4}$$

隐藏状态的历史记录好输入序列数据到目前为止，这些计算循环，RNN 的记忆，RNN 结构丰富多样，可以根据需要建立，上部结构的 2 是更常见的经典结构，重量 w_1、w_2、w_3，偏差因子 b_1，b_2 是 RNN 参数，随着时间的推移，RNN 参数保持不变，没有几何参数的数量随着时间的增长，就像语法模型。RNN 的参数保持不变，参数数量不像 n-gram 模型那样随时间呈几何增长。在 RNN 的训练过程中，可以使用反向传播算法来实现 RNN 参数的整定，然而，这带来了一个棘手的梯度消失和梯度爆炸问题。

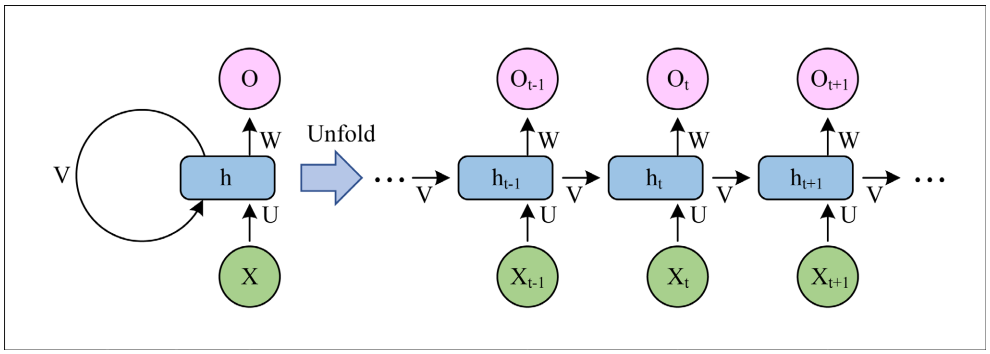

Figure 2. RNN structure diagram
图 2. RNN 结构图

2.3. 基于深度学习的文本纠错关键技术

传统的机器学习在中文语法纠错上，需要手工创建特征，很难包含所有的规则，而深度学习不需要任何复杂的特征工程。首先，对于文本的预处理，深度学习与传统的机器学习没有区别，仍然是将文本转换为计算机可以理解的矩阵表示，深度学习通过人工神经网络自主学习相关特征。在训练神经网络模型阶段，将文本转换为相应的特征向量作为输入端，并通过训练深度模型，使用高级语义信息代替传统的人工特征工程，采用深度学习的中文语义纠错系统可以获得更高的准确率。另一方面，机器翻译领域端到端模型的爆发也带动了语义纠错技术的发展，使得主流机器翻译模型广泛应用于语义纠错领域，如基于 RNN 的 Seq2seq 模型、结合注意机制的 Transformer 模型、基于卷积神经网络的 Seq2seq 模型等。以下包含了目前深度学习用于文本纠错的关键技术：

1) 基于 Seq2seq 的文本纠错模型。将语法纠错任务转换成类似机器翻译的任务，即将错误的句子翻译成正确的句子。

2) 基于 RNN 和注意力机制的文本纠错模型(RNN Attention)。RNN 序列模型适合文本任务，而加入注意机制的 RNN 对较长的文本具有更强的纠错效果。因为加入注意力机制，RNN 可以考虑远程文本信息。

3) 基于 Seq2Seq 和注意力机制的文本纠错模型(Seq2Seq Attention)。在 Seq2seq 的文本纠错模型基础上，加入注意力机制，能够更好地抓取上下文的语义信息。

4) 基于 Transformer 的文本纠错模型。该模型用自注意力机制结构代替 LSTM 来解决序列对序列问题，语义特征提取效果较好。

5) 基于 BERT 的文本纠错模型。该模型是中文微调模型，利用 MASK 特征进行拼写纠错。

6) 基于 Kenelm 的文本纠错模型。Kenelm 是统计语言建模工具，利用 Kenelm 构建相应的规则，可以实现文本纠错，缺点是构建规则的过程中需要消耗大量的人力。

7) 基于 Conv Seq2Seq 的文本纠错模型。一种基于卷积神经网络的 Seq2Seq 模型，融合了注意机制，由于语义错误与相邻词的关系更密切，使用 CNNs 比 LSTMs 能更好地捕捉相邻词的关系，多层 CNNs 还能捕捉更远距离上的单词交互信息。

2.4. 中文纠错数据集的构建

以 NLPCC2018 GEC 官方数据集作为训练数据的基础，不对其进行分词处理。加入 lang8 平行语料库，对其进行分词处理。在上述语料库的基础上，加入 CGED16、CGED17、CGED18 的数据，经过分词、将繁体字转换为简体字、打乱数据顺序等预处理，生成一个熟食语料库用于纠错，共计 130 万句对。

2.4.1. 训练集

训练集来自 http://lang-8.com，这是一个语言学习网站，母语为英语的人可以自由选择修改学习者的文本。通过探索“语言交换”社交网络服务(SNS)，我们收集了大量汉语普通话学习者的语料库。在这个社交网站上大约有 68,500 名汉语学习者。通过收集他们的中文论文和从中国当地人那里得到的修订本，初步建立了一个语料库，从 135,754 篇论文中收集了 1,108,907 个句子。通过以上整理工作，获得了 61 个不同类别 717,241 句的语料库。

2.4.2. 测试集

测试集从北京大学汉语学习者语料库中提取。北京大学汉语学习者语料库是北京大学中文系为促进汉语媒介语言的国际教育和研究而建立的，它是由外国大学生写的论文组成的。在实验中，从语料库中收集了 2000 句句子，本实验对其进行标注以纠正语义错误，标注准则遵循编辑距离最小的一般原则。该原则指定如何重建包含错误的句子的正确形式，以及如何选择使编辑距离最小化的句子。错误分为四种类型：冗余词(用大写字母“R”表示)、缺词(M)、选词错误(S)和排序错误(W)。

3. 实验

3.1. 传统机器学习语法纠错方法

基于规则的方法难以包含所有错误的情况，面对不同的领域需要构造不同的规则集合，需要对错误类型进行分类并形成混淆集来提高基于规则的语义错误纠正的准确性。在分类方法中，文本语义纠错被认为是一个多分类问题，并使用词汇标签和依赖句法等特征，为给定的错误类型指定一个混淆集，在本实验中，混淆集指的是单词和词汇模式的组合。对于给定的错误类型，纠错任务被认为是一个分类任务，它可以从大量的本地文本数据中学习语义表示。

如表 1 所示，将常用的文本分类算法 TF-IDF、朴素贝叶斯算法和 SVM 的算法性能进行了比较，并通过实验结果进行了验证。通过实验发现，TF-IDF 和朴素贝叶斯算法这两种算法在本实验的应用场景中分类效果都较差，图 3 所示，当混淆集扩展时，ONE-STEP-FORWARD 分类效果更好。

3.2. 深度学习语法纠错方法

实验中使用的纠错数据集来自 2018NLPCC 发布的训练语料库，共使用了 717,241 对汉语句子。以 30,000 条错误到正确的数据对作为测试集，其余为训练语料库，将划分方法随机化，如下图 4 所示。

实验使用了 2018 年 NLPCC 基准测试集，通过设置文本相似度阈值来判断纠错是否有效，即语法纠错结果与正确的结果文本相似度阈值大于 0.75，误差修正被认为是正确的，否则，错误的判断是错误的。这个实验运行在 centos-7 上，测试集包括错误的句子和正确的句子，一共 30,000 条数据，首先将文本相似度阈值设置为 0.75，然后将错误的句子放入深度模型中，深度模型将反馈一个修正结果，然后，修改后的句子只要文本相似度大于 0.75 的阈值，就判断修正结果是正确的，反之则判断修正结果是错误的。表 2 展示了各语法纠错模型的误差修正结果。

Table 1. Experimental results of traditional machine learning methods
表 1. 传统机器学习方法实验结果

	ACC	P	R	F1
TF-IDF	0.75	0.80	0.85	0.82
NAÏVE-BAYESIAN	0.90	0.81	0.91	0.86
ONE-STEP-FORWARD	0.92	0.89	0.89	0.89

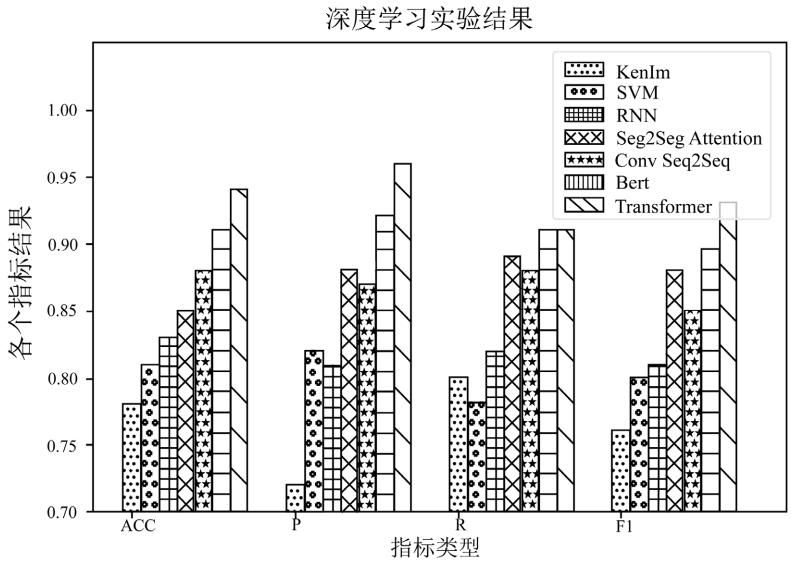

Figure 3. Traditional machine learning experiment results
图 3. 传统机器学习实验结果

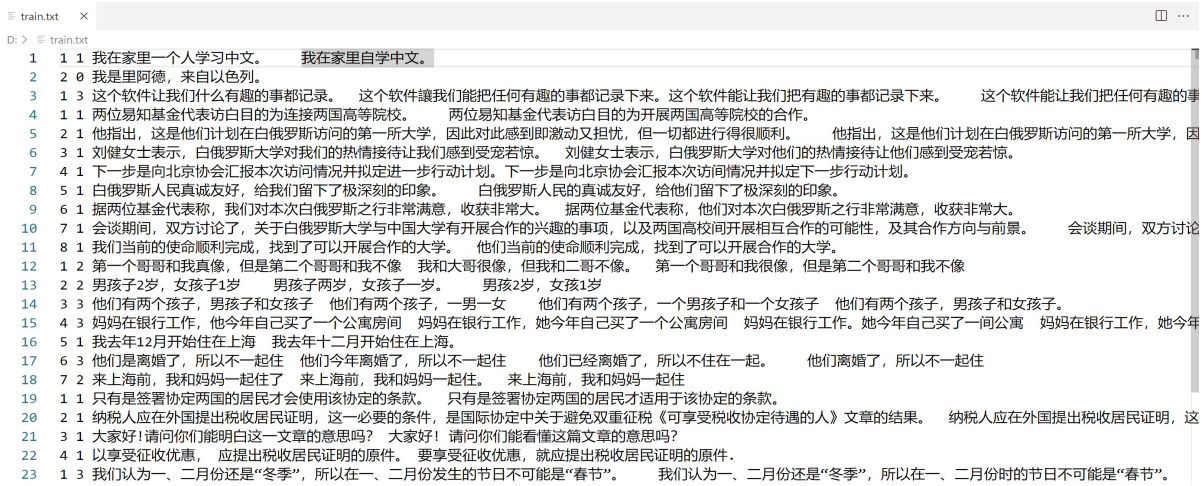

Figure 4. Data display
图 4. 数据展示

Table 2. Experimental results of deep learning methods
表 2. 深度学习方法实验结果

	ACC	P	R	F1
Kenlm	0.78	0.72	0.80	0.76
SVM	0.81	0.82	0.78	0.80
RNN	0.83	0.81	0.82	0.81
Seq2Seq Attention	0.85	0.88	0.89	0.88
Conv Seq2Seq	0.88	0.87	0.84	0.85
Bert	0.91	0.92	0.87	0.89
Transformer	0.94	0.96	0.91	0.93

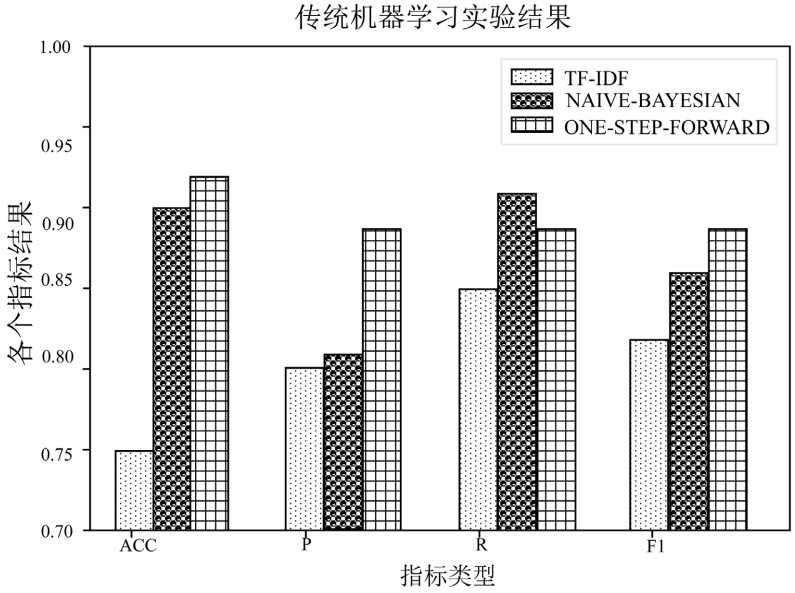

Figure 5. Results of deep learning experiment
图 5. 深度学习实验结果

实验结果表明，基于规则的纠错模型的纠错效果最稳定，并且取得了不错的效果，但是代价是需要人工构造完备的规则集合；基于传统的机器学习方法，由于提取特征能力不足，所以效果不太理想；而基于深度学习的方法，在训练数据充足的情况下，明显好于传统的机器学习方法。图 5 显示了各种类型的深度模型的实验验证。

4. 结论

汉语语法纠错是自然语言处理领域的一个重要课题。这项工作的主要目的是检查和纠正句子中的语法错误，语法纠错已被用于中文文本自动校对和中文学习辅助领域。近年来，随着汉语在全球范围内的影响力越来越大，语法语义纠错的任务也有了很大的突破。本文的主要贡献在于通过实验分析了目前主流方法的优缺点，并且分析了语法纠错的发展方向：由于语法纠错的文本存在不确定性，纠错的结果可能存在多种可能，因此 Seq2Seq 和预训练的方式将会成为语法纠错的发展方向。

基金项目

湖南省教育厅资助科研项目(21C1118、22B0977)；湖南省教育科学研究工作者协会“十四五”规划重点课题(XJKX22A066)。

参考文献

[1] 冯雅. 基于深度学习的中文语法纠错研究[D]: [硕士学位论文]. 上海: 上海师范大学, 2022. https://doi.org/10.27312/d.cnki.gshsu.2022.000464

[2] 赵国红. 中文语法纠错方法的研究综述[J]. 现代计算机, 2021, 27(28): 65-69.

[3] 郭琰, 张矛. 基于深度学习的语法纠错算法建模研究[J]. 信息技术, 2021(4): 148-152, 158. https://doi.org/10.13274/j.cnki.hdzj.2021.04.027

[4] Yu, J.J. and Li, Z.H. (2014) Chinese Spelling Error Detection and Correction Based on Language Model, Pronunciation, and Shape. *Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, Wuhan, 20-21 October 2014, 220-223.

[5] Zhang, S.Y., Xiong, J.H., Hou, J.P., Zhang, Q. and Cheng, X.Q. (2015) Hanspeller++: Aunified Framework for Chi-

nese Spelling Correction. *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, Beijing, 30-31 July 2015, 38-45. https://doi.org/10.18653/v1/W15-3107

[6] Wang, D.M., Song, Y., Li, J., Han, J.L. and Zhang, H.S. (2018) A Hybrid Approach to Auto-Matic Corpus Generation for Chinese Spelling Check. *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*, *Brussels*, 31 October-4 November 2018, 2517-2527. https://doi.org/10.18653/v1/D18-1273

[7] Zhang, L., Zhou, M. and Pan, H.H. (2018) Automatic Detecting/Correcting Errors in Chinese Text by an Approximate Word-Matching Algorithm. *Proceedings of the* 38*th Annual Meeting on Association for Computational Linguistics*, October 2000, 248-254. https://doi.org/10.3115/1075218.1075250

[8] Zhao, J.B., Li, M.Z., Liu, W.J., Li, S. and Lin, Z.Q. (2018) Detection of Chinese Grammatical Errors with Context Representation. 2018 *International Conference on Network Infrastructure and Digital Content* (*IC-NIDC*), Guiyang, 22-24 August 2018, 25-29. https://doi.org/10.1109/ICNIDC.2018.8525629

[9] Wang, D.M., Tay, Y. and Zhong, L. (2019) Confusionset-Guided Pointer Networks for Chinese Spelling Check. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, July 2019, 5780-5785. https://doi.org/10.18653/v1/P19-1578

[10] Wu, K., Gao, Z., Peng, C. and Wen, X. (2013) Text Window Denoising Autoencoder: Building Deep Architecture for Chinese Word Segmentation. In: Zhou, G., Li, J., Zhao, D. and Feng, Y., Eds., *NLPCC* 2013: *Natural Language Processing and Chinese Computing*, Springer, Berlin, 1-12. https://doi.org/10.1007/978-3-642-41644-6_1

[11] Chen, J.W., Sigalingging, X.K., Leu, J.S. and Takada, J.I. (2020) Applying a Hybrid Sequential Model to Chinese Sentence Correction. *Symmetry*, **12**, Article 1939. https://doi.org/10.3390/sym12121939

[12] Zhang, R., Zhang, Y., Huang, G. and Chen, R. (2021) Research on Proofreading Method of Semantic Collocation Error in Chinese. In: Sun, X., Zhang, X., Xia, Z. and Bertino, E., Eds., *ICAIS* 2021: *Advances in Artificial Intelligence and Security*, Springer, Cham, 709-722. https://doi.org/10.1007/978-3-030-78615-1_62

[13] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805.

[14] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**, 5998-6008.