Computer Science and Application 计算机科学与应用, 2022, 12(3), 565-571
Published Online March 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.123057

Hans汉斯

基于Transformer增强架构的中文文本纠错研究

杨靖翔，赵曙光

东华大学信息科学与技术学院，上海

收稿日期：2022年2月13日；录用日期：2022年3月8日；发布日期：2022年3月15日

摘　要

本文将Transformer模型应用于中文文本自动校正领域，并将Transformer模型中不同神经模块的输出动态结合，同时在模型训练时引入课程学习策略，以加快模型收敛速度。实验结果表明，本文所提出的增强模型及在训练中引入的课程学习策略对校正结果的准确率、召回率、纠错$F_{0.5}$值有较大提升。

关键词

中文文本校对，Transformer模型，动态残差，深度学习

Research on Chinese Text Error Correction Based on Transformer Enhanced Architecture

Jingxiang Yang, Shuguang Zhao

College of Information Science and Technology, Donghua University, Shanghai

Received: Feb. 13th, 2022; accepted: Mar. 8th, 2022; published: Mar. 15th, 2022

Abstract

In this paper, the transformer model is applied to the field of Chinese text automatic correction, and the outputs of different neural modules in the transformer model are dynamically combined. At the same time, the curriculum learning strategy is introduced in the model training to speed up the convergence speed of the model. The experimental results show that the proposed enhancement model and the curriculum learning strategy introduced in the training can greatly improve the accuracy, recall rate and error correction $F_{0.5}$ value of the correction results.

文章引用: 杨靖翔, 赵曙光. 基于 Transformer 增强架构的中文文本纠错研究[J]. 计算机科学与应用, 2022, 12(3): 565-571. DOI: 10.12677/csa.2022.123057

Keywords

Chinese Text Proofreading, Transformer Model, Dynamic Residual, Deep Learning

1. 引言

随着新媒体和大数据时代的到来，互联网每天都会产生海量的文本信息，其中的文本质量良莠不齐，随之而来的文本校对任务也越来越繁重，由于传统的人工校对方法效率低下，研究快速高效的自动校对方法逐渐成为热点问题。

目前国内在中文文本自动校正领域的研究方法主要有以下 3 种：1) 基于语言学和规则的方法，针对具体的错误类型，制定相应的规则进行文本纠错，该类方法的纠错准确性依赖于规则的质量且只能修改特定的错误种类，可扩展性较差[1]；2) 基于概率统计的方法，利用字词的文本特征，对语言进行 N-gram 建模，并选取合适的统计模型纠错，但该类方法对于未知词组预测能力低，存在数据稀疏性问题[2]；3) 基于深度学习的方法[3]，将文字通过编码转换为词向量，构建深度学习网络，无需考虑具体错误类型，完成端到端的文本纠错。常用的模型包括 LSTM [4]或 CNN 神经机器翻译模型[5]，但其存在共同的局限性，即认为语句中的每个字词具有相同的重要性，无法有选择性地进行关注。

因此，为了提高模型的并行计算能力以及选择性特征提取能力，本文采用基于多头注意力机制的 Transformer 模型作为纠错模型，并提出一种新的动态残差结构，增强模型语义特征提取能力。同时为了加快模型生成速度和收敛速度，在训练数据中找到更好的局部最小值，本文在训练纠错模型时引入课程学习策略。通过实验，结合本文提出的两种方法，模型在准确率、召回率、纠错 $F_{0.5}$ 值上均有更好的表现。

2. 基于动态残差结构的 Transformer 模型

2.1. Transformer 模型及实现

Transformer 是 Vaswani 等人[6]在 2017 年提出的一个新框架，作者采用多头注意力机制(Multi-Headed Attention)解决在提取长距离语义信息时，所存在的语义信息丢失问题，其核心结构如图 1 所示。

1) Word Embedding 文字可通过 Word2Vec、Glove 等词嵌入方法将词语投射到特定长度的向量空间，其中语义越接近的词语，词向量间的距离越近。

2) Positional Encoding 通过增加关于特征的相对、绝对位置提升模型的有序性。

Transformer 模型所采用的不同于 RNN，所以通过使用 Positional Encoding 确定单词在序列中的位置。

3) Multi-Head Attention 相当于将 n 个 self-attention 相结合，使得模型能够关注到不同子空间的语义信息，本文中 n 取 8。

4) Encoder and Decoder Transformer 模型沿用了 Encoder-Decoder 架构，本文中 Encoder 和 Decoder 模块均由 6 个相同神经元模块堆叠而成，每一层中包含 Multi-Head Attention 子层和 Feed Forward 子层，子层之间通过残差和归一化相连接。

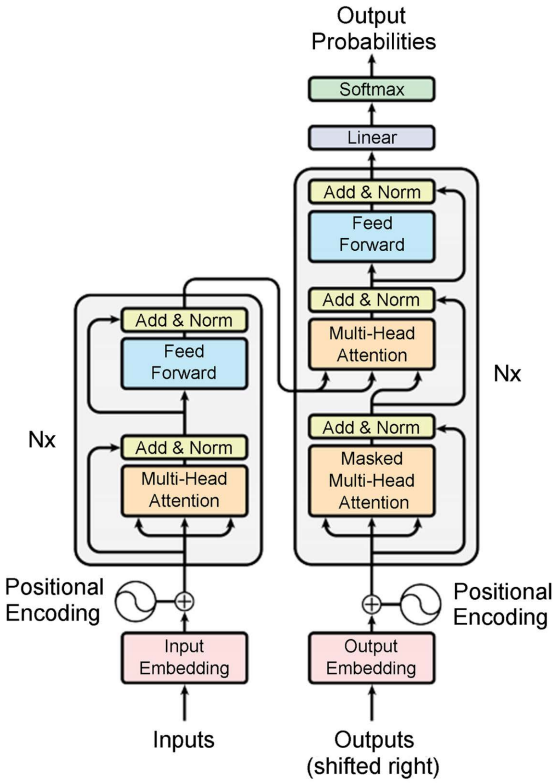

Figure 1. Transformer model structure
图 1. Transformer 模型结构图

2.2. 动态残差结构

传统 Transformer 模型中的神经模块可简化为图 2，每个模块的输出如式(1)：

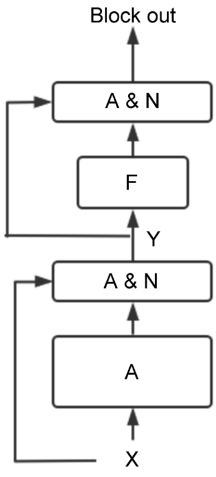

Figure 2. Simplified neural module
图 2. 简化后的神经模块

$$y = BN\left(A\left(x\right)+x\right),\quad out = BN\left(F\left(y\right)+y\right) \tag{1}$$

其中，A 为多头注意力层，$A\&N$ 为归一化函数，F 为前馈层的线性变换，对

式(1)求导得式(2)。

$$\frac{\partial out}{\partial x}=\frac{\partial out}{\partial y}*\frac{\partial y}{\partial x},\quad \frac{\partial y}{\partial x}=\frac{\partial BN}{\partial x}*\left(\frac{\partial A}{\partial x}+1\right),\quad \frac{\partial out}{\partial y}=\frac{\partial BN}{\partial y}*\left(\frac{\partial F}{\partial y}+1\right) \tag{2}$$

通过分析可得，求导后的$\frac{\partial out}{\partial y}$中由于乘法因子$\frac{\partial BN}{\partial y}*\frac{\partial BN}{\partial x}$的作用，在模块不断叠加的过程中，存在累乘效应，存在梯度消失的风险。

为了解决上述问题，本文对 Transformer 模型进行改进，在编码器和解码器端分别采用动态残差结构，如图 3 所示，可以更充分地结合高层和低层神经模块的语义信息，具体计算如式(3)。

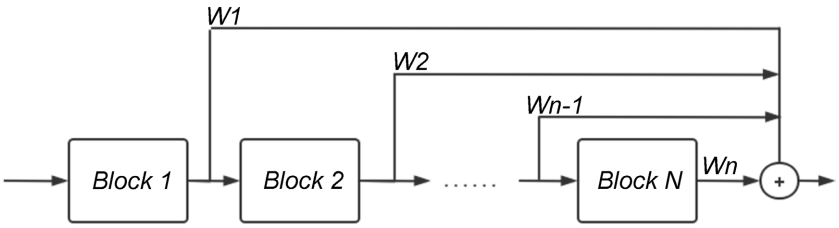

Figure 3. Dynamic residual structure
图 3. 动态残差结构

$$out=\sum_{i=1}^{N}Block_i\cdot W_i \tag{3}$$

3. 课程学习策略

课程学习的概念在 2009 年由 Yoshua Bengio [7]等人提出，与人类学习机制类似，即先学习简单的技能，再学习困难的技能。如果训练数据以特定的顺序输入，先从简单的数据开始学，等到模型有一定的能力后再去学习难的数据，这样即符合人类的直觉；同时，从机器学习的角度去看，这种方法也可以避免过早陷入不好的局部最优解。

课程学习策略的特点包括：1) 提高模型生成速度和加快收敛速度；2) 在非凸训练数据中找到更好的局部极小值。利用课程学习策略的方法可大致分为两类：一是控制训练数据，如每 k 个训练步骤就增加 p%的数据、将数据分为 m 个批次，每 k 个训练步骤就增加一个批次等；二是控制训练数据被采样的概率，如定义一个与训练步骤相关的平滑函数等。本文采用类型一的课程学习策略应用于纠错任务中，框架概念如图 4 所示：

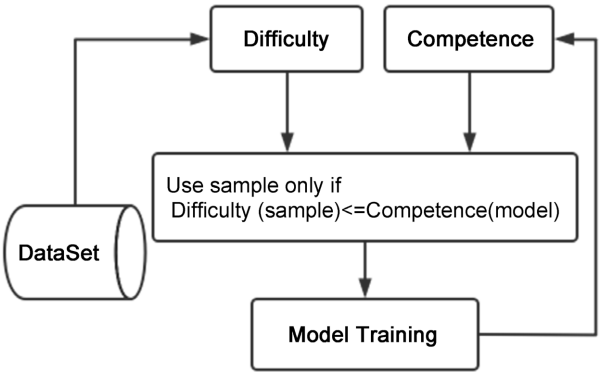

Figure 4. Curriculum learning framework
图 4. 课程学习框架

本文所应用的课程学习策略方法有两个重要概念：Difficulty 和 Competence。其中 Difficulty 代表某个训练样本的难度值，由样本句子长度和单词相对词频共同决定，计算公式如式(4)：

$$d(s_i) = -\sum_{k=1}^{N_i} \log p\left(w_k^i\right) \tag{4}$$

其中，s_i 为第 i 个样本句子，N_i 为样本句子长度，$p\left(w_k^i\right)$ 为样本第 k 个词语的相对词频，词频计算公式如式(5)：

$$p\left(w_j\right) = \frac{1}{N_{total}} \sum_{i=1}^{M} \sum_{k=1}^{N_i} a_{w_k^i = w_j} \tag{5}$$

其中，M 为样本总数，a 为条件函数，满足条件为 1 反之为 0。

另一个概念 Competence 表示模型的训练进度，定义模型训练状态的一个函数，介于 0 和 1 之间。该方法将模型在 t 时刻的能力 $c(t)$定义为 t 时刻允许使用的训练数据比例。训练样本根据难度进行排序，模型只允许在时刻 t 使用相应顶部部分数据。 Linear 形式的 $c(t)$计算公式如式(6)：

$$c(t) = \min\left(1, t\frac{1-c_0}{T} + c_0\right) \tag{6}$$

其中 c_0 为初始值，T 为时间步阈值，当超过该阈值时，认为模型具有能力，t 为时间步。

本论文的 Competence 除了上述提到的 Linear 形式，还提出了一种依据 loss 选择训练数据的方法。因为模型的能力强弱除了可以间接地使用训练步骤彰显，最直观的反映模型能力强弱的便是模型的 loss，其 $c(t)$计算如式(7)：

$$c_{loss}(t) = \min\left(1, \sqrt{1 - \frac{loss}{T}} + c_0\right) \tag{7}$$

基于以上两个概念，本文将课程学习策略应用于中文纠错任务中，以达到提升性能的目的。首先对语料中的语句分别计算 Difficulty 值，并计算难度分数的累计密度函数，使其分布在 0 到 1 之间，在模型训练时，根据不同的训练阶段计算获取模型当前的 Competence 分数，根据当前分数从数据集中均匀采样符合 Difficulty 值条件的数据进行训练，最终输出模型。本文设计的课程学习策略算法如表 1 所示：

Table 1. Curriculum learning strategy algorithm
表 1. 课程学习策略算法

输入：数据集 $D = \{s_i\}_{i=1}^{M}$，训练模型 T，Difficulty 值 d，Competence 值 c
输出：训练好的模型

1) 计算句子难度 $d(s_i)$
2) 计算难度分数的累计密度函数，使难度分数在 0 到 1 之间
3) **for** training step $t = 1, 2, \cdots, T$ **do**
计算当前模型 Competence 值 $c(t)$
从数据集中均匀采样相应批次数据 B_t，其中 $s_i < c(t)$
数据 B_t 作为输入，调用模型 T 进行训练
4) 输出模型

4. 实验与分析

4.1. 数据预处理

本文所使用的数据为 NLPCC 2018 GEC 数据，其中包含 717,302 条平行语料，本文选取 700,000 条

作为训练集，17,302 条作为测试集。预处理过程为：

加载语料数据，使用 opencc 工具包将繁体转化为简体；去除数据中的空格，将语料数据格式转化为平行句子对；删除源端重复和目标端重复，删除长度小于 5 和大于 80 的平行句子对；最后，使用 jieba 分词工具对平行句子进行分词。

4.2. 模型参数设置

本文使用模型的超参数具体为：词嵌入矩阵选用 512 维词表，模型结构中采用各含 6 个神经模块的编解码器 ，注意力头数量采用 8 个。在训练时选用 Adam 优化器，学习率初始时设置为 1×10^{-7}，经过 5000 个 batch 的训练之后增长为 1×10^{-5}，后续逐步下降。

4.3. 评价指标

本此实验采用 M^2 算法评估校对模型，M^2 算法通过对比模型输出和源句子之间短语级别的编辑和标准输出之间的差别完成评估。评价指标包括准确率 P、召回率 R、纠错值 $F_{0.5}$。

设标准句子对原句子的编辑集为$\{g_1, g_2, \cdots, g_n\}$和模型输出对原句子的编辑集为$\{e_1, e_2, \cdots, e_n\}$，相关指标如式(8)：

$$P = \frac{\sum_{i=1}^{N}|e_i \cap g_i|}{\sum_{i=1}^{N}|e_i|}, \quad R = \frac{\sum_{i=1}^{N}|e_i \cap g_i|}{\sum_{i=1}^{N}|g_i|}, \quad F_{0.5} = 5 \times \frac{P \times R}{P + 4 \times R} \tag{8}$$

其中，$e_i \cap g_i = \{e \in e_i \mid \exists g \in (match(e, g))\}$。

4.4. 实验设置及结果

本文实验基线模型使用传统的 Transformer 模型进行文本纠错；第二组实验将提出的动态残差结构加在模型的编码器端；第三组实验将动态残差结构加在模型的解码器端；第四组在编码器和解码器端均加入动态残差结构；第五组是在第一组实验的基础上，在模型训练时采用课程学习策略；第六组是在第三组实验的基础上，在模型训练时采用课程学习策略。实验结果如表 2 所示：

Table 2. Effects of different improvements on model performance
表 2. 不同改进对模型性能的影响

模型类	准确率(P)/%	召回率(R)/%	$F_{0.5}$/%
Transformer	39.01	13.87	28.63
+动态残差结构(编码器端)	32.85	9.16	21.65
+动态残差结构(解码器端)	41.43	15.87	31.34
+动态残差结构(编码器端和编码器端)	38.03	15.74	29.64
+课程学习策略	40.56	14.43	29.77
+动态残差结构(解码器端) &课程学习策略	42.98	18.11	33.72

前四组实验结果显示，在 Transformer 基线模型的基础上，在解码器端添加动态残差结构可增强模型性能。第五组实验反映出在训练传统 Transformer 模型的基础上使用课程学习策略可提高纠错性能，第六组实验将二者综合，使模型性能达到最优。

5. 总结

本文将 Transformer 模型应用于中文文本校正领域，并提出一种动态残差结构有效应用于解码器端，用以增强捕获语句信息的能力。同时在训练模型时采用课程学习策略，加快模型收敛速度。最终在 NLPCC 2018 GEC 公开数据集上完成实验，在纠错准确率、召回率、$F_{0.5}$ 值方面得到提升。

参考文献

[1] 段娜, 杨妍, 赵军民. 基于规则的短文本识别算法[J]. 计算机产品与流通, 2019(2): 173-174.

[2] 王琼, 旷文珍, 许丽. 基于改进的 N-gram 模型和知识库的文本查错算法[J]. 计算机应用与软件, 2021, 38(10): 310-315.

[3] 高印权. 基于深度学习的文本语法自动纠错模型研究与实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2020. https://doi.org/10.27005/d.cnki.gdzku.2020.001563

[4] Zhao, W. (2021) Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community. Lynne Bowker and Jairo Buitrago Ciro. *Digital Scholarship in the Humanities*, **36**, 256-260. https://doi.org/10.1093/llc/fqaa074

[5] Fu, K., Huang, J. and Duan, Y. (2018) Youdao's Winning Solution to the NLPCC-2018 Task 2 Challenge: A Neural Machine Translation Approach to Chinese Grammatical Error Correction. Springer, Cham, 879-884. https://doi.org/10.1007/978-3-319-99495-6_29

[6] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[7] Bengio, Y., Louradour, J., Collobert, R., *et al*. (2009) Curriculum Learning. *Proceedings of the* 26*th Annual International Conference on Machine Learning*, *ICML* 2009, Montreal, 14-18 June 2009, 41-48. https://doi.org/10.1145/1553374.1553380