

CCL23-Eval 任务7系统报告：基于序列标注和指针生成网络的语法纠错方法

于右任, 张仰森, 畅冠光, 高贝贝, 姜雨杉, 肖拓
北京信息科技大学, 智能信息处理研究所
{a154377713,zys,cgg,gbg,jys,xt}@bistu.edu.cn

摘要

针对当前大多数中文语法纠错模型存在错误边界识别不准确以及过度纠正的问题, 我们提出了一种基于序列标注与指针生成网络的中文语法纠错模型。首先, 在数据方面, 我们使用了官方提供的lang8数据集和历年的CGED数据集, 并对该数据集进行了繁体转简体、数据清洗等操作。其次, 在模型方面, 我们采用了基于ERNIE+Global Pointer的序列标注模型、基于ERNIE+CRF的序列标注模型、基于BART+指针生成网络的纠错模型以及基于GECToR的纠错模型。最后, 在模型集成方面, 我们使用了投票和基于ERNIE模型计算困惑度的方法, 来生成最终预测结果。根据测试集的结果, 我们的COM指标达到了48.68, 位居第二名。

关键词: 序列标注; 指针生成; 困惑度

System Report for CCL23-Eval Task 7: A Syntactic Error Correction Approach Based on Sequence Labeling and Pointer Generation Networks

Youren Yu, Yangsen Zhang, Guanguang Chang, Beibei Gao, Yushan Jiang, Tuo Xiao
Institute of Intelligent Information Processing, Beijing Information Science and Technology University
{a154377713,zys,cgg,gbg,jys,xt}@bistu.edu.cn

Abstract

Aiming at the problems of inaccurate recognition of error boundaries and overcorrection in most current Chinese grammar error correction models, we propose a Chinese grammar error correction model based on sequence annotation and pointer generation network. First, in terms of data, we use the officially provided lang8 dataset and the CGED dataset of past years, which has been subjected to the operations of traditional to simplified Chinese, data cleaning and so on. Secondly, in terms of model, we used the sequence annotation model based on ERNIE+Global Pointer, the sequence annotation model based on ERNIE+CRF, the error correction model based on BART+Pointer Generation Network, and the error correction model based on GECToR. Finally, in terms of model integration, we used voting and ERNIE-based models to compute the perplexity to generate the final predictions. According to the results of the test set, we achieved a COM metric of 48.68, which placed us in second place.

Keywords: Sequence Labeling, Pointer generation, Perplexity

1 引言

中文语法错误检测 (Chinese Grammatical Error Diagnosis, CGED) 是一项旨在识别中文文本中的语法错误并确定其位置与类型的任务。这些语法错误可以分为四类：赘余 (Redundant Words, R)、遗漏 (Missing Words, M)、误用 (Word Selection, S) 和错序 (Word Ordering Errors, W)。准确地检测和纠正中文文本中的语法错误对于提高写作质量、加深语言理解以及有效沟通至关重要。中文语法检测有着广泛的应用，包括自动化写作、机器翻译、语音识别和智能助手等。

为了实现中文语法错误的检测与纠正，我们使用了多个模型进行集成。首先，为了提高所训练出的模型的鲁棒性和泛化能力，我们对数据集进行了繁体转简体、数据清洗等操作。其次，为了更好地捕捉错误边界并缓解过度纠正问题，我们采用了基于ERNIE+Global Pointer和基于ERNIE+CRF的序列标注模型，以利用上下文信息对语法错误进行标注。此外，我们还采用了基于BART+指针生成网络和基于GECToR的纠错模型，从不同角度提升语法纠错性能。

2 背景

2.1 任务设置

如表1所示，此次任务涉及四种类型的错误，包括缺失(M)、赘余(R)、误用(S)和乱序(W)。对于S类型和M类型的错误，需要提供相应的纠正结果。

错误类型	输入文本	输出文本
缺失(M)	有的国家解决吸烟造成的影响而采取了跟吸烟有关的政策。	有的国家为了解决吸烟造成的影响而采取了跟吸烟有关的政策。
赘余(R)	这表示你的肺部不是正常。	这表示你的肺部不正常。
误用(S)	所以我少就不理他了。	所以我早就不理他了。
乱序(W)	那就意味着有一有位亲人要去世！	那就意味着有一位亲人要去世！

Table 1: 数据集示例

2.2 评测标准

下述是对系统性能的评价，从以下五个方面以精确率、召回率和F1值进行评估：

- (1) 假阳性 (False Positive)：正确的段落单元被错误地判断为包含错误的比例。
- (2) 侦测层 (Detective-level)：对段落单元是否包含错误进行二分判断。
- (3) 识别层 (Identification-level)：该层的子任务是多分类问题，即确定错误点的错误类型。

(4) 定位层 (Position-level)：对错误点的位置和覆盖范围进行判断。错误的边界以词边界为界限，分词颗粒度参考jieba缺省模式。

(5) 修正层 (Correction-level)：提供了针对错误字符串 (S) 和字符串缺失 (M) 两种错误类型的修正答案。对于每个赛题的S型和M型错误，均提供1-3个正确答案。

2.3 相关工作

中文语法纠错是自然语言处理 (NLP) 领域中一个重要的研究方向。随着中文在全球范围内的广泛应用，正确的语法表达对于有效沟通和交流变得尤为重要。然而，中文的语法结构复杂，包含大量的规则和特殊情况，给语法纠错任务带来了挑战。目前，中文语法纠错的研究主要集中在以下几个方面：基于规则的方法、基于统计和机器学习的方法以及基于深度学习的方法。基于规则的方法主要依赖于手工编写的规则和语法知识库，但受限于规则的完备性和准确性。基于统计和机器学习的方法通过训练模型来捕捉中文语法的模式和规律，但由于中文语法的复杂性，这些方法在中文语法纠错上的效果有限。

最近，基于深度学习的方法在中文语法纠错中展现出了巨大的潜力。(zhang et al., 2020)提出了一种软掩码 (soft masking) 机制，利用Bi-GRU网络对文本进行检错，并预测每个位置上字符的错别字概率。然后利用这些概率对该位置上的字符进行软掩码处理，并借助BERT模型进行纠错，从而有效地降低了模型的误纠率。(Hong et al., 2019)提出了一种基于BERT的

深度降噪编码器(DAE)和以置信度-字音字形相似度为基础的解码器(CSD)的文本纠错模型。在DAE阶段,该模型利用BERT模型动态生成候选集,代替传统的混淆集。而在CSD阶段,通过计算置信度和字音字形相似度两个维度来选择候选集,取代传统单一阈值的方法,从而提高纠错效果。(Cheng et al., 2020)提出了一种利用图卷积网络(Graph convolutional network,GCN)进行文本纠错的模型。该模型通过GCN学习文本中字音和字形之间的结构关系,并将这些信息融入字的嵌入向量中。在纠错分类时,模型更倾向于预测字在混淆集中的可能性。(Lai et al., 2022)提出了一种类型驱动的多轮修正方法,用于语法错误的纠正。通过按照不同错误类型对原始训练数据进行分解,使得模型能够逐步学习如何纠正错误,并理解不同类型错误之间的内在关联,从而获得更好的效果。(Li et al., 2022)提出了一种适用于中文拼写检查的错误驱动对比概率优化模型。该模型利用对比学习的思想,将正确字符作为正样本,将模型容易预测错误的字符作为负样本,并设计了额外的对比损失函数,从而进一步提高正确字符的预测概率,同时降低错误字符的预测概率。

3 参赛系统

本节介绍了我们参赛系统所采用的方法和策略。该系统主要由四个部分组成:(1)数据预处理:包括繁体转简体和数据清洗等操作;(2)语病检测模块:采用基于ERNIE+Global Pointer和基于ERNIE+CRF的序列标注模型,用于检测文本中的错误信息;(3)语病纠错模块:采用基于BART+指针生成网络和基于GECToR的纠错模型,用于对文本中的错误信息进行纠正;(4)模型集成模块:利用投票和基于ERNIE模型的文本困惑度方法,对模型生成的结果进行集成。

3.1 数据预处理

由于lang8数据集和CGED历年数据集有简体和繁体而测试集均为简体,并且数据集中存在噪音,因此要对数据集进行清洗。具体流程如下:

- (1) 去除特殊字符和重复标点。
- (2) 使用opence包统一将繁体转为简体。
- (3) 去除输入长度与输出长度差异过大的文本。
- (4) 去除长度过短文本。
- (5) 去除中文字符占比小于20

3.2 语病检测模块

3.2.1 ERNIE

在语病检测模块,我们使用基于ernie-3.0-xbase-zh模型对文本进行编码。ERNIE 3.0模型(Sun et al., 2021)是一种由百度公司研发的基于BERT模型改进的预训练语言模型。它首次引入了大规模知识图谱,并采用了海量无监督文本与大规模知识图谱的平行预训练方法。这种方法通过联合掩码训练,将大规模知识图谱的实体关系与大规模文本数据输入到预训练模型中,从而促进了结构化知识和无结构文本之间的信息共享,提升了模型对知识的记忆和推理能力。

ERNIE 3.0模型采用了通用语义表示网络和任务语义表示网络的两层框架。通用语义表示网络学习基础和通用的知识,而任务语义表示网络基于通用语义表示学习任务相关的知识。任务语义表示网络可以通过自编码结构或自回归结构实现,并通过底层共享来实现交互和增强。在学习过程中,任务语义表示网络只学习对应类别的预训练任务,而通用语义表示网络会学习所有的预训练任务,从而提升了模型效果。

3.2.2 Global Pointer

Global Pointer是一种利用全局归一化的思路,可以无差别识别非嵌套实体(flat-ner)和嵌套实体(nested-ner)的解码器。设长度为 n 的文本经过ERNIE模型编码后的向量序列为 $[h_1, h_2, \dots, h_n]$,首先,如公式(1)(2)所示,使用线性变换得到向量序列 $[q_{1,\alpha}, q_{2,\alpha}, \dots, q_{n,\alpha}]$ 和 $[k_{1,\alpha}, k_{2,\alpha}, \dots, k_{n,\alpha}]$ 。

$$q_{i,\alpha} = \mathbf{W}_{q,\alpha} h_i + b_{q,\alpha} \quad (1)$$

$$k_{i,\alpha} = \mathbf{W}_{k,\alpha} h_i + b_{k,\alpha} \quad (2)$$

其中, $W_{q,\alpha}, W_{k,\alpha}$ 为权重矩阵, $b_{q,\alpha}, b_{k,\alpha}$ 为偏置项, h_i, h_j 代表下标为*i*、*j*的输入序列。

为了增加Global Pointer解码器对输入文本中实体的长度与跨度的敏感度, 使用相对位置编码RoPE (Su et al., 2021)对向量序列添加相对位置信息, 并计算出下标从*i*到*j*的连续序列为类型实体的得分函数, 具体如公式(3)所示。

$$\begin{aligned} s_{\alpha}(i, j) &= (R_i q_{i,\alpha})^T (R_j k_{j,\alpha}) \\ &= q_{i,\alpha}^T R_i^T R_j k_{j,\alpha} \\ &= q_{i,\alpha}^T R_{j-i} k_{j,\alpha} \end{aligned} \tag{3}$$

其中, $R_i R_j$ 为RoPE中的变换矩阵, 满足关系 $R_i^T R_j = R_{j-i}$ 。

3.2.3 语病检测模型

ERNIE+Global Pointer 如图1所示, 该模型的结构如下: 首先, 将输入序列通过ernie-3.0-xbase-zh编码以获取每个字符的表征信息。然后, 使用Global Pointer解码器来输出预测的标签序列。同时, 为了增强模型的效果, 我们对模型采用了差分学习率。

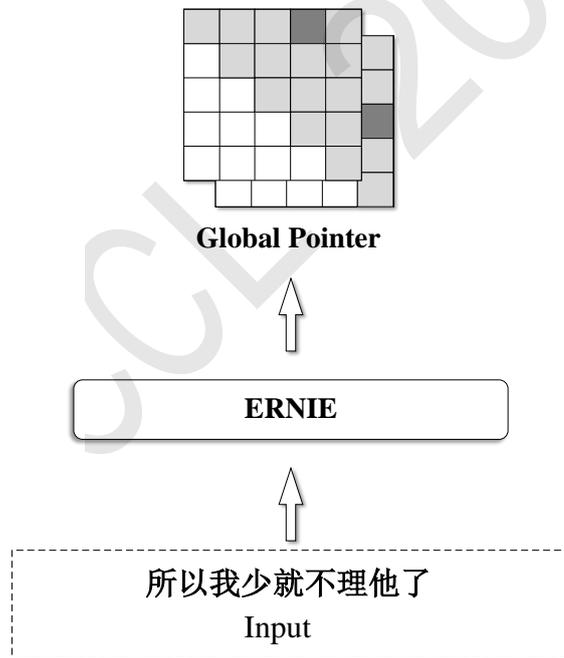


Figure 1: ERNIE+Global Pointer模型结构图

ERNIE+CRF 模型结构如图2所示, 首先, 与上述模型相同, 将输入序列通过ernie-3.0-xbase-zh编码以获取每个字符的表征信息。然后, 将表征向量通过一层分类器获取每个字符在语病标签空间的分布, 将其作为CRF层的发射矩阵。最后, 得到输入序列的语病标签输出序列。

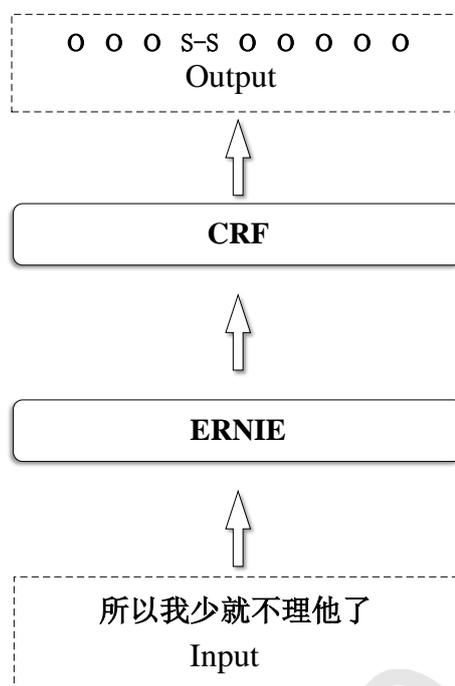


Figure 2: ERNIE+CRF模型结构图

3.3 语病纠错模块

3.3.1 BART

BART (Bidirectional and AutoRegressive Transformer) (Lewis et al., 2020)是一种结合了BERT的双向编码器和GPT的从左到右解码器的模型。它是建立在标准的seq2seq Transformer模型的基础上，相较于BERT更适用于文本生成任务，并且相较于GPT，BART还包含了双向上下文语境信息。BART模型在生成任务上取得了显著的进步，同时在一些文本理解类任务上也达到了最先进的水平。

BART模型被设计为用于预训练序列到序列模型的降噪自动编码器。它的训练包括两个关键步骤：首先，利用任意一种噪声函数对文本进行处理；其次，学习一个模型来将处理后的文本重构回原始文本。BART采用了更多样的噪声方法，旨在破坏序列结构信息，防止模型过度依赖这些信息，具体包括：

Token Masking: 类似于BERT的方法，将随机的单词替换为[MASK]。

Token Deletion: 随机删除输入中的字符。与字符屏蔽相反，模型需要确定缺少输入的哪些位置。

Text Infilling: 随机将一段连续的单词（称为span）替换为一个[MASK]，其中span的长度服从一定的泊松分布。特别需要注意的是，当span长度为0时，相当于在该位置插入一个[MASK]。

Sentence Permutation: 根据句号将文档分割为多个句子，然后随机排列这些句子的顺序。

Document Rotation: 从文档序列中随机选择一个单词，并将该单词作为文档的开头。

这些噪声方法的引入增加了BART模型的鲁棒性和泛化能力，使其能够更好地处理文本生成任务和文本理解类任务。

3.3.2 指针生成网络

指针生成网络 (Pointer-Generator Networks) 是一种结合了传统的序列到序列 (seq2seq)

模型和指针机制的模型，用于处理涉及生成和复制两种操作的任务。指针生成网络的核心思想是在模型的输出中引入指针机制，使模型能够决定是生成一个词汇表中的词语，还是从输入序列中复制一个片段。具体来说，指针生成网络在词汇表中为每个词分配一个概率，同时为输入序列中的每个位置分配一个概率。然后，它根据这些概率来决定是生成一个词，还是从输入序列中复制一个片段。

指针生成网络通常由两个子模型组成：生成模型（Generator）和复制模型（Pointer）。生成模型使用传统的seq2seq模型结构，通过学习输入序列和输出序列之间的映射关系来生成序列。复制模型使用注意力机制（Attention）来计算输入序列中每个位置的重要性，并为每个位置分配一个概率。这些概率表示模型应该从输入序列中复制哪个片段。在训练过程中，指针生成网络通过最大化目标序列的对数似然来学习参数。这包括生成模型和复制模型的参数。在推理过程中，模型可以根据生成模型的概率分布或者复制模型的概率分布来决定生成或复制。

3.3.3 语病纠错模型

GECToR GECToR (Omelianchuk et al., 2020)是一种序列标注的语病纠错模型，采用了seq2edit方法来解决语法错误修改的问题。模型结构如图3所示，首先，使用StructBert编码器对源句子序列进行编码，然后，在每个字符位置上使用分类器来预测最可能的编辑标签。每个字符都被映射到一个编辑操作，包括“KEEP”（保留当前字符）、“DELETE”（删除当前字符）、“APPEND”（在当前字符后添加特定字符）和“REPLACE”（将当前字符替换为特定字符）。其中，“KEEP”和“DELETE”是单标签操作，而“APPEND”和“REPLACE”则需要添加或替换相应的字符。

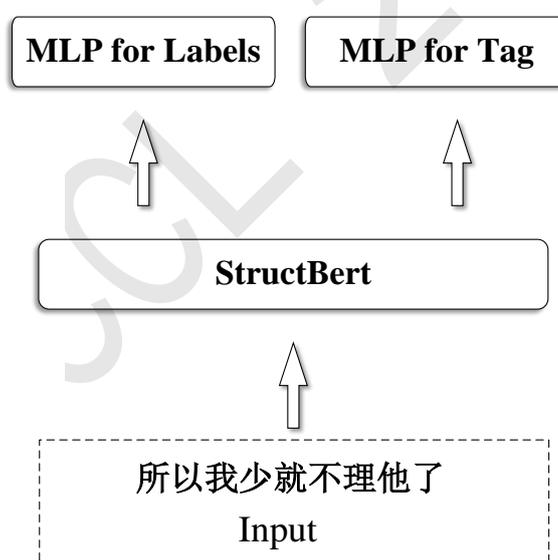


Figure 3: GECToR模型结构图

BART+指针生成网络 如图4所示，该模型的结构如下：首先，将输入序列通过bart-large-chinese编码以获取每个字符的表征信息。然后，使用指针生成网络来更好地解决纠错模型的过度纠正问题。同时，使用label smooth缓解错误标签的现象。

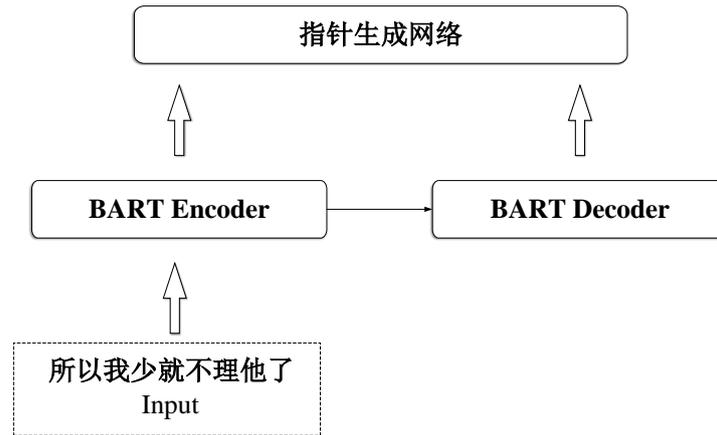


Figure 4: BART+指针生成网络模型结构图

3.3.4 模型融合模块

参与最终融合的模型共有5个，它们包括：1个GECToR模型、2个结合了BART+指针生成网络的模型、1个基于ERNIE+Global Pointer序列标注模型，以及1个基于ERNIE+CRF的序列标注模型。模型融合的主要流程为：

- (1) 将所有模型的预测结果合并，形成一个包含所有模型预测结果的集合。
- (2) 利用模型的预测结果进行投票，如果有至少3个模型预测该样本为正确，那么将该样本判定为正确。
- (3) 删除起始和结束位置相同的W类型错误点。
- (4) 由于BART+指针网络模型在处理乱序类型的错误方面表现良好，并在验证集中的W类型指标精确率较高，因此在融合过程中，将该模型输出的W类型错误判定为无误，并将与之重合的错误或之前被判定为正确的样本剔除。
- (5) 将每个样本的错误点按照位置排序，依次遍历每个错误点。将位置有重合的错误点放入一个列表中，对该重合列表中的错误点进行筛选。选择被多个模型判定的错误位置作为最终判定结果，即判定数目越高，置信度越大。
- (6) 对于融合结果中同一个位置可能存在多个错误类型，并且参与判定的模型数量差异不大的情况，采用局部困惑度进行筛选。通过初始化基于ERNIE的语言模型来计算困惑度。如果修改前后的困惑度降低值（delta）越大（>0），说明修改前句子的困惑度较高，修改后困惑度降低，句子变得更准确。因此，可以倾向于将具有较大降低值的错误类型判定为正确的结果。

4 实验与结果

4.1 实验流程与参数设置

在训练过程中，我们使用两阶段训练的方法，第一阶段使用lang8数据集+CGED历年数据集，第二阶段只使用CGED历年数据集。

ERNIE+Global Pointer 在第一阶段中，我们使用差分学习率进行训练，ERNIE模型学习率设置为 $2e-5$ ，Global Pointer解码器学习率设置为 $6e-5$ ，epoch为10，dropout为0.1，全局随机种子为42，batch size为32，输入序列的最大长度为200，优化器为AdamW，并使用Cosine Warmup学习率调整器。在第二阶段中，模型的学习率设置为 $5e-6$ 。

ERNIE+CRF 在第一阶段中，我们设置学习率设置为 $1e-5$ ，epoch为10，dropout为0.1，全局随机种子为99，batch size为32，输入序列的最大长度为200，优化器为AdamW，并使用Cosine学习率调整器。在第二阶段中，模型的学习率设置为 $5e-6$ ，dropout为0.3。

GECToR 在第一阶段中，我们只对MLP for Labels层和MLP for Tag层进行训练，学习率设置为 $1e-3$ ，batch size为16，epoch为2。在第二阶段，模型的学习率设置为 $1e-5$ ，batch size为64，梯度累积为4，epoch为20，patience为3。

BART+指针生成网络 在第一阶段中，我们设置学习率设置为 $1e-5$ ，epoch为8，dropout为0.1，全局随机种子为42，batch size为128，梯度累积为16，输入序

列的最大长度为180，beam search的k为3，优化器为AdamW，并使用Cosine Warmup学习率调整器。在第二阶段中，模型的学习率设置为5e-6，batch size为64，梯度累积为4，epoch为10。

4.2 实验结果与分析

模型	COM	FPR	DET	IDE	POS	COR
GECToR	37.94	40.11	82.04	54.64	33.28	21.94
BART+指针生成网络	44.48	21.39	82.30	56.17	36.05	24.78
最终融合结果	48.68	19.47	83.25	63.45	40.15	27.34

Table 2: 模型对比结果

根据实验结果（见表2），GECToR模型的COM值为37.94，BART+指针生成网络模型的COM值为44.48。然而，我们发现在使用ERNIE+Global Pointer模型、ERNIE+CRF模型、GECToR模型和BART+指针生成网络模型融合后，达到了最佳效果，COM值提升至48.68。

这些结果表明，在文本校对任务中，模型融合可以显著提升性能。ERNIE+Global Pointer模型和ERNIE+CRF模型的融合，以及GECToR模型和BART+指针生成网络模型的融合，都能够在COM值方面取得明显的改进。通过结合不同模型的优势，能够克服单一模型的局限性，从而实现更好的文本校对效果。此外，BART+指针生成网络模型在单独使用时也取得了较高的COM值。这表明该模型在处理文本校对任务上具有潜力，但与其他模型相比，在模型融合方面仍存在改进空间。

4.3 案例分析

原始文本	BART+指针生成网络	GECToR	最终融合结果
美国的诺贝尔奖获得者的人数远远 轟 其他国家。	美国的诺贝尔奖获得者的人数远远 高于 其他国家。	美国的诺贝尔奖获得者的人数远远 比 其他国家。	美国的诺贝尔奖获得者的人数远远 高于 其他国家。
他到以后，他们聊天一 会 。	他到以后，他们聊一 会天 。	他到了以后，他们聊天一 会 。	他到了以后，他们聊一 会天 。
我 拿出去 饼干和啤酒去客厅	我 拿着 饼干和啤酒去公司	我 拿 饼干和啤酒去公司	我 拿着 饼干和啤酒去公司

Table 3: 模型预测结果案例

根据表3所示的结果，在原始文本中，句子中的一些表达存在一些语法错误或不通顺的问题。在使用BART+指针生成网络模型进行文本校对后，第一行的句子被修正为“美国的诺贝尔奖获得者的人数远远高于其他国家”，明显改善了表达的准确性。使用GECToR模型进行文本校对后，第二行的句子被修正为“他到了以后，他们聊天一会”，修正了原始文本中的一个语法错误。最终的融合结果是将多个模型的输出进行综合，第二行的句子被修正为“他到了以后，他们聊一会天。”，结合了BART+指针生成网络模型和GECToR模型的优点，使得表达更加准确和自然。

在进行文本校对任务时，采用投票和基于ERNIE模型的困惑度作为融合方法，将ERNIE+Global Pointer模型、ERNIE+CRF模型、GECToR模型和BART+指针生成网络模型进行融合，相较于仅使用BART+指针生成网络模型或GECToR模型，能够综合各模型的优点，从而显著提升效果。

5 总结

在这次评测任务中，我们采用了ERNIE+Global Pointer和ERNIE+CRF序列标注模型来检测错误地址和错误类型。同时还应用了BART+指针生成网络模型和GECToR模型对文本进行纠错。通过将多个模型进行融合，我们的COM指标达到了48.68%的成绩。

然而，在接下来的中文语病诊断研究中，还面临着一些挑战。其中之一是如何更好地解决过度纠正问题。虽然我们的系统在纠错过程中表现出较高的准确性，但有时可能会过度修改文本，导致意思的扭曲或不必要的改动。因此需要进一步探索新的方法和策略，以确保在纠错过程中能够保持语义的一致性和原始文本的风格。

此外，我们还致力于进一步提升纠错效果。这可以通过优化模型的结构和参数，改进数据预处理方法以及引入更多的上下文信息等方式来实现，通过不断地研究和创新，我们能够取得更好的成果，提高中文语病诊断和纠错的准确性和实用性。

参考文献

- Zhang S, Huang H, Liu J, et al. 2020. *Spelling error correction with soft-masked BERT*. In proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
- Hong Y, Yu X, He N, et al.. 2019. *FASPELL: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm*. In Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pages 160-169.
- Cheng X, Xu W, Chen K, et al. . 2020. *Spellgen: Incorporating phonological and visual similarities into language models for chinese spelling check*. In proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
- Lai S, Zhou Q, Zeng J, et al.. 2022. *Type-Driven Multi-Turn Corrections for Grammatical Error Correction*. In proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pages 3225-3236.
- Li Y, Zhou Q, Li Y, et al.. 2022. *The past mistake is the future wisdom: Error-driven contrastive probability optimization for Chinese spell checking*. In proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pages 3202-3213.
- Sun Y, Wang S, Feng S, et al. . 2021. *Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation*. arXiv preprint arXiv:2107.02137.
- Su J, Lu Y, Pan S, et al.. 2021. *Roformer: Enhanced transformer with rotary position embedding*. arXiv preprint arXiv:2104.09864, 2021.
- Lewis M, Liu Y, Goyal N, et al.. 2020. *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. In proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871-7880.
- Omelianchuk K, Atrasevych V, Chernodub A, et al. . 2020. *GECToR-grammatical error correction: tag, not rewrite*. In proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 163-170.