课程作业：中文文本纠错

自然语言处理2025

助教　陈宇飞

2025年4月1日

任务背景

中文文本纠错（Chinese Text Correction, CTC）是自然语言处理领域的一项重要任务，旨在自动检测并纠正中文文本中的拼写（Chinese spelling check，CTC）和语法错误（Chinese Grammatical Error Correction，CGEC）。

近年来，随着社交媒体、在线评论和电子商务等领域的快速发展，存在拼写错误、用词不当、语境不通的文本大量涌现。一个高效的文本纠错系统不仅能改善用户体验，也为下游任务（如机器翻译、情感分析）提供更干净的数据输入。

任务背景

中文文本纠错（Chinese Text Correction, CTC）是自然语言处理领域的一项重要任务，旨在自动检测并纠正中文文本中的拼写（Chinese spelling check，CTC）和语法错误（Chinese Grammatical Error Correction，CGEC）。

Error Type	Example sentence
Spelling Errors	进入大学，就是进入一个新的环境，结出（接触）新的人，你的所有过去对于他们来说是一张白纸。
Redundant Words	突然有一天，一个女人来看~~来看~~孩子。
Missing Words	今天要讲（的）是他在一年时间里面的教师生涯。
Word Ordering Errors	一般室内环境采用200系列材质即可，而室外~~需环境~~（环境需）使用304材质。

任务背景

中文的语言特点，如拼音、同音字和多义词等，使得文本处理面临诸多挑战。在自然语言处理领域，基于规则的方法与统计方法各有其优势和局限。

规则方法：依赖人工构建的详尽语言规则、词典以及正则表达式

- 对于识别固定格式和特定错误模式具有较高的精度和确定性
- 然而，这种方法在处理语言中的新词、隐晦表达及复杂语境时常常显得力不从心。

统计方法：利用语言模型和概率统计

- 能够更灵活地应对多变的错误模式和语境变化
- 但其表现也往往依赖于数据的丰富性和质量，且在语料不足的情况下可能会出现偏差。

数据集

数据集：CCTC

主要特点：

1. 数据集来源于真实场景中的中文文本，主要由母语者生成，具有较高的代表性和实用性。
2. 文本中包含了多种常见错误类型，如拼写错误、同音字混淆、词语搭配错误以及跨句子语境不连贯等。
3. 与传统单句纠错不同，存在跨句子（或跨语境）的错误。

数据示例：

{"source": "我的大脑在不断的思索。", "target": "我的大脑在不断地思索。", "label": 1}
{"source": "因此,午后股指仍可能继续。", "target": "因此,午后股指仍可能继续回落。", "label": 1}

作业要求

完成给定的代码TODO部分，主要包括数据分析，规则方法和统计方法，可以探索规则方法和统计方法的结合使用。

统计方法部分可以尝试你认识的任意方法和其组合，包括传统机器学习方法和神经网络方法，并在实验报告中比较不同方法的结果。

不允许使用LLM纠错，禁止使用现成的文本纠错工具和已经训练完成文本纠错模型，禁止结果造假和代码抄袭。

所提供的示例代码可以大幅修改，也可以选择你习惯的代码流程，但evaluation 文件不要改动。

评测方法

通过 Levenshtein Distance 计算源文本和目标文本之间的编辑操作，将纠错行为分类为插入，替换和删除，并以此评估纠错模型在两个关键维度上的性能：检测（Detection）和修正（Correction）。

示例：

- Source：我今天去学佼了
- Target：我今天去学校了
- Prediction：我今天去学效了

Source -> Target：在（5, 6）中替换为“校”

Source -> Prediction：在（5, 6）中替换为“效”

成功检测出错误但修正失败：Detection TP +1；Correction FP +1

如果检测失败：Detection FN +1， Detection FP +1 ； Correction FN +1

评测方法

对于 Detection 和 Correction 分别使用精确率（Precision）、召回率（Recall）、F1 分数和 F0.5 分数来量化检测效果。

Correction F0.5 分数被选为最终评估指标，因为它在精确率和召回率之间更倾向于精确率，能够更好地反映模型在实际应用中的纠错能力。

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$F_{0.5} = \frac{(1 + 0.5^2) \times Precision \times Recall}{0.5^2 \times Precision + Recall}$$

作业要求

DDL：4.23

提交格式要求：压缩包“学号_姓名. zip”，内含代码和实验报告，报告命名为“学号_姓名.pdf”。

实验报告需包含以下内容：实现了哪些方法并对自己设计的代码模块用简洁的语言描述；如何复现主要实现结果，包括执行命令和环境依赖；不同方法的实验结果如何；遇到的具体问题，如何解决；对该任务的思考。

主要依据报告质量、代码是否可复现、实验结果得分、对不同方法的探索和思考来综合评分。

推荐使用sklearn和pytorch等python库提升效率，使用深度学习方法，如Bert、LSTM等模型，来获得更高的得分。

Thanks!