基于字和词的中文拼写纠错

1.概要设计

1.1 问题引入

中文拼写纠错任务，就是检测出文本中的错误并且进行纠正。不仅能注明出错的位置，还会给出相应的修改策略。因此，一般的中文文本纠错任务包括两个部分：检测任务和纠正任务。中文拼写纠错任务一般都是基于字为单位的。如果是以词为基本单位，一方面是分词的过程有可能会引入别的错误，应一方面是文字的错误也会影响分词的效果。但是同时，对于中文，词所蕴含的语义信息往往更加丰富，所以有时候虽然纠错任务是以字为基本单位，仍可以将词的信息作为字的额外特征加入到模型中去，进一步丰富字的信息。

1.2 预期目标

整体而言，本设计的预期输入是一句带有错误的中文文本，预期输出是被修改完成的文本。同时，应满足下条件：

- 最大程度检测并修正句子的错误；
- 没有把握时，优先保证不提供错误的修改策略；
- 优先针对打字输入产生的音近字错误。

1.3 主要工作

在本设计中，个人采用了基于字的方法和基于词的方法这两种方法进行对比实验，以达成更好 的效果。首先，使用字典法和样例法结合的方法设计纠错算法，这是基于词的纠错方法； 借助ERNIE1.0模型的网络框架进行实验，实现了百度在ACL 2021上提出结合拼音特征的Softmask策略的中文错别字纠错的下游任务网络，以此设计对照深度学习算法，这是基于字的纠错方法；最终，对照效果，得出结论。

2.具体实现

2.1 基于词的实现

基于词的实现使用了下面的数据集和数据包作为支撑：

- jieba 中文分词库，pinyin 汉语拼音库；
- 中文有效词词典；
- 修改样例数据库。

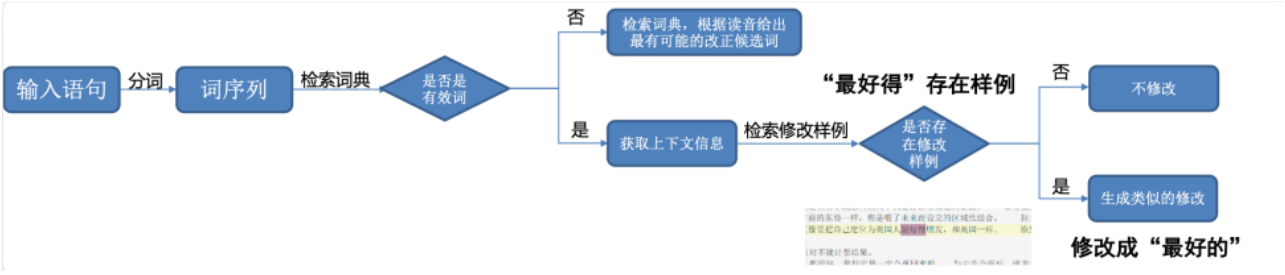
其中，修改样例数据库由下图展示的ASR（自动语音识别）技术大规模产生。

对于输入的文本，首先将其分词，并判断每个词是否是有效词。如果不是有效词，则按照读音的优先级给出可能的改正候选词；反之，获取上下文信息，并根据修改样例做出决策。

在判断读音的优先级时，按照如下的规则：

- 最高优先级：读音完全一致的词；
- 第二优先级：读音不完全一致，但是某一个字的声母或者韵母相同的词；
- 第三优先级：某一个字相同的其他词语；
- 最低优先级：原本的词语。

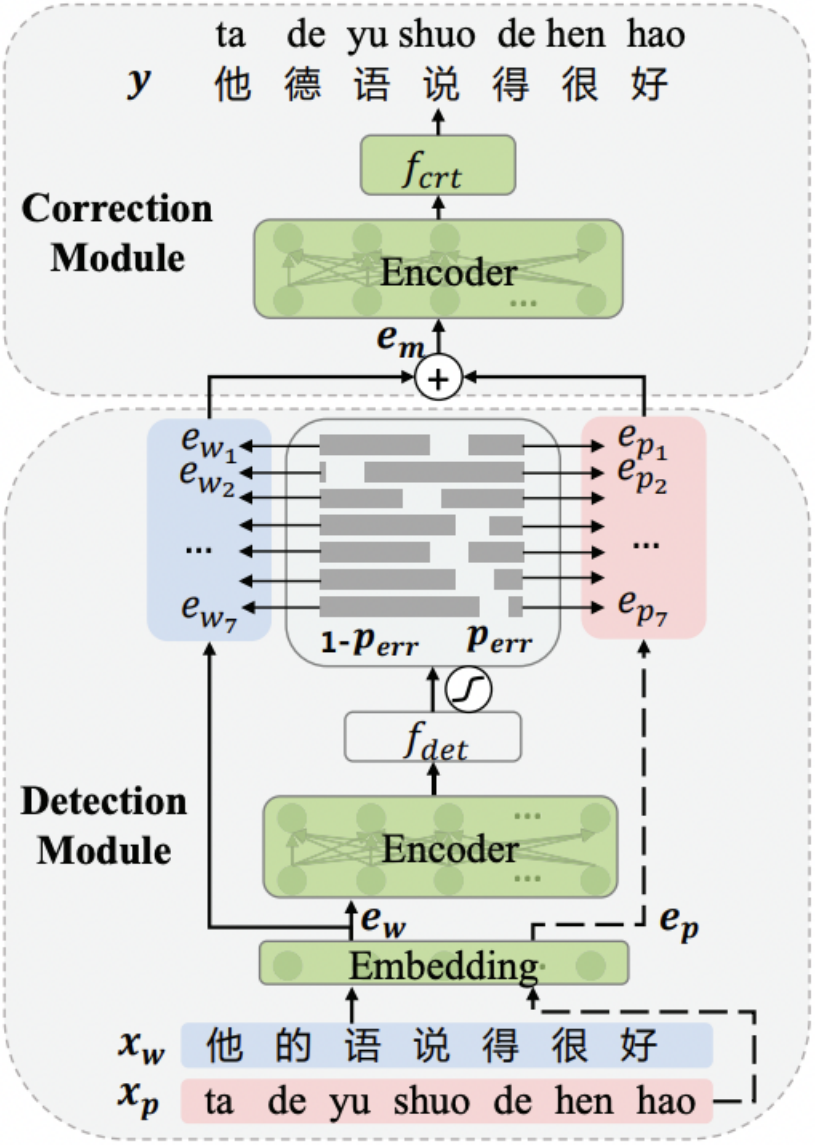
在获取上下文信息时，将词序列进行组合，形成短句。例如，对于 [最好；得；国家] 词序列而言， [最好得；得国家] 是其上下文信息。



2.2 基于字的实现

在这一部分的实现中借助百度paddlepaddle框架下的ERNIE模型，进行文本纠错实现。ERNIE是百度开创性提出的基于知识增强的持续学习语义理解框架，该框架将大数据预训练与 多源丰富知识相结合，通过持续学习技术，不断吸收海量文本数据中词汇、结构、语义等方面的知识，实现模型效果不断进化。在不断的迭代和学习之中，已经达到了较高的准确性，可以胜任文本纠错任务。

在本次实验中，借助ERNIE1.0模型的网络框架进行实验，实现了百度在ACL 2021上提出结合拼音特征的Softmask策略的中文错别字纠错的下游任务网络，并提供预训练模型。



论文中暂未开源融合字音特征的预训练模型参数(即MLM-phonetics)，所以本文提供的纠错模型是在ERNIE-1.0的参数上进行finetune，纠错模型结构与论文保持一致。由于其原始网络模型已经较好，个人并为对其网络模型进行修改，只是对其训练函数进行一些finetune，同时在原有数据的基础上，又加入271k条数据集进行模型训练。

3.效果测试

对于两种模型的效果我们做了一定的对比，这里选取比较有代表性的三个例子进行展示，分别说明了这两种方法之间存在的效果和准确度的区别。

A 日本首相指出，中国是最好得锅家。应是“日本首相指出，中国是最好的国家。”

B 世界杯决赛是阿根廷对对法国。应是“世界杯决赛是阿根廷队对法国。”

C 牛顿顿顿炖牛肉。句子无错误。

原句	基于词的方法	基于字的方法
A	无法改出“最好得”错误	完全改正
B	完全改正	完全改正
C	保持原句	错误地将“炖牛肉”改成“顿牛肉”

在综合分析后，得出了以下的结论：

- 基于词的方法可以完成基本的日常纠错任务，优点在于速度快（几乎不花时间）、修改细致、把握性高、可解释性强，但是需要依赖大量的修改语料数据作为支撑，且对于语义错误有时难以修改；
- 基于字的方法可以修改专业的、语义方面的错误，但是想要训练出一个较好的模型耗时长，需要大量算力投入，可解释性弱，同时把握度低，优化空间小。

4.小结

本设计尝试通过多种方法解决文本纠错问题，都基本达成目标标。目前新型的思路是借助LLM去完成文本纠错任务，接下来的优化重点是可以参考新型的思路去进行学习以及优化，寻找相关的测试数据来量化成果。

参考资料

[1] Correcting Chinese Spelling Errors with Phonetic Pre-training ACL2021

[2] DingminWang et al. A Hybrid Approach to Automatic Corpus Generation for Chinese Spelling Check, EMNLP, 2018

[3] Dynamic Connected Networks for Chinese Spelling Check ACL 2021

[4] Instantaneous Grammatical Error Correction with Shallow Aggressive Decoding ACL 2021

[5] PLOME: Pre-training with Misspelled Knowledge for Chinese Spelling Correction ACL2021