中文文本纠错数据集汇总

任务简介

中文文本纠错任务（Chinese Text Correction，CTC）旨在对中文文本中所包含的所有类型错误进行纠正，主要包含两类子任务：中文拼写纠错（Chinese Spelling Check，CSC）和中文语法纠错（Chinese Grammatical Error Correction，CGEC）。其中，中文拼写纠错任务仅需要对文本中因为音近/形近混淆而产生的拼写错误进行纠正，而中文语法纠错任务则需要对文本中所有类型（替换、冗余、缺失、词序四类）的错误进行纠正。

数据集

中文拼写纠错（Chinese Spelling Check，CSC）

SIGHAN系列

来自于SIGHAN2013、SIGHAN2014、SIGHAN2015评测任务，分别包括350/6526/3174句训练集和974/526/550句测试集。

论文链接：

> Hih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee (2013). Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013. *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing (SIGHAN'13)*, Nagoya, Japan, 14 October, 2013, pp. 35-42.
> Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen (2015). Introduction to SIGHAN 2015 Bake-off for Chinese Spelling Check. *Proceedings of the 8th SIGHAN Workshop on Chinese Language Processing* (*SIGHAN'15*), Beijing, China, 30-31 July, 2015, pp. 32-37.

下载地址：

https://github.com/onebula/sighan_raw

OCR Dataset

爱奇艺基于OCR技术生成的伪CSC训练集，4575句。

论文链接：

> Yuzhong Hong, Xianguo Yu, Neng He, Nan Liu, and Junhui Liu. 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, Hong Kong, China. Association for Computational Linguistics.

下载地址：

https://github.com/iqiyi/FASPell

Hybrid Dataset

基于OCR和ASR技术伪造的CSC训练集，约27万条。

论文链接：

> Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A Hybrid Approach to Automatic Corpus Generation for Chinese Spelling Check. In *Proceedings of the 2018 Conference*

> *on Empirical Methods in Natural Language Processing*, pages 2517—2527, Brussels, Belgium. Association for Computational Linguistics.

下载地址：
https://github.com/wdimmy/Automatic-Corpus-Generation

ECSpell

苏州大学开放的多领域拼写纠错数据集，包括金融、医药等领域。

论文链接：

> Lv, Qi, Ziqiang Cao, Lei Geng, Chunhui Ai, Xu Yan, and Guohong Fu. "General and Domain Adaptive Chinese Spelling Check with Error Consistent Pretraining." *arXiv preprint arXiv:2203.10929* (2022).

下载地址：
https://github.com/wdimmy/Automatic-Corpus-Generation

中文语法纠错（Chinese Grammatical Error Correction, CGEC）

CGED

北京语言大学团队开放的CGED系列数据集，面向二语者文本，早期仅包含语法错误检测任务，后期包含语法错误纠正任务。
论文链接：

> Rao, Gaoqi, Erhong Yang, and Baolin Zhang. "Overview of NLPTEA-2020 shared task for Chinese grammatical error diagnosis." In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pp. 25-35. 2020.

下载地址：
https://github.com/wdimmy/Automatic-Corpus-Generation

NLPCC18

北京大学团队于NLPCC2018上开放的语法纠错评测任务数据集，面向二语者文本，同期开放的还有Lang8训练数据集。
论文链接：

> Zhao, Yuanyuan, Nan Jiang, Weiwei Sun, and Xiaojun Wan. "Overview of the nlpcc 2018 shared task: Grammatical error correction." In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 439–445. Springer, Cham, 2018.

下载地址：
http://tcci.ccf.org.cn/conference/2018/dldoc/trainingdata02.tar.gz

MuCGEC

苏州大学和阿里巴巴达摩院开放的CGEC数据集，面向二语者文本，包含3个领域和多答案。
论文链接：

> Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. MuCGEC: a Multi–Reference Multi–Source Evaluation Dataset for Chinese Grammatical Error Correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118—3130, Seattle, United States. Association for Computational Linguistics.

下载地址：
https://github.com/HillZhang1999/MuCGEC

YACLC

北京语言大学团队开放的CGEC数据集，面向二语者文本，包含多答案。
论文链接：

> Wang, Yingying, Cunliang Kong, Liner Yang, Yijun Wang, Xiaorong Lu, Renfen Hu, Shan He et al. "YACLC: A Chinese Learner Corpus with Multidimensional Annotation." *arXiv preprint arXiv:2112.15043* (2021).

下载地址：
http://tcci.ccf.org.cn/conference/2018/dldoc/trainingdata02.tar.gz

CTC

科大讯飞和哈工大开放的CGEC数据集，面向母语者文本。
下载地址：
https://github.com/destwang/CTC2021

Midu

蜜度公司2022–WAIC文本智能校对大赛数据集，面向母语者文本。

下载地址：
https://aistudio.baidu.com/aistudio/competition/detail/404/0/introduction

CAIL–2022

哈工大法研杯2022开放的数据集，面向母语法律文本。

下载地址：
http://cail.cipsc.org.cn/task_summit.html?raceID=2&cail_tag=2022

致谢

本列表由阿里云天池数据科学团队和阿里达摩院提供长期维护。