**MuCGEC & NaSGEC文本纠错数据集共享协议（2025年2月版本）**

甲方：苏州大学“语析LAGroup”

乙方：[研究机构或公司的名字]

MuCGEC（Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction）和NaSGEC（Multi-Domain Chinese Grammatical Error Correction Dataset from Native Speaker Texts）数据集分别来自汉语学习者和母语者文本，旨在标注高质量、多领域多来源和多参考的中文文本纠错并行语料库，以促进文本纠错技术的广泛应用。为了促进文本纠错技术的进展，甲方将 MuCGEC和NaSGEC数据集 免费共享给乙方，用于科学研究。

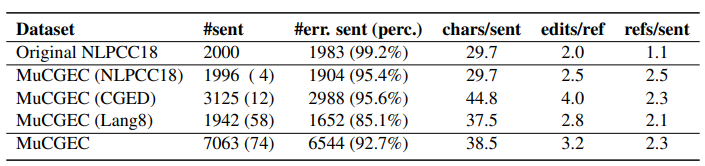
本开源数据由苏州大学“语析LAGroup”整理、汇集及公开分享。该数据的来源广泛，本开源数据部分来源于公开网络渠道、第三方公开数据集或经授权的其他信息源，部分由语析LAGroup自主注释生成。

本开源数据主页：<https://github.com/SUDA-LA/CGECData>

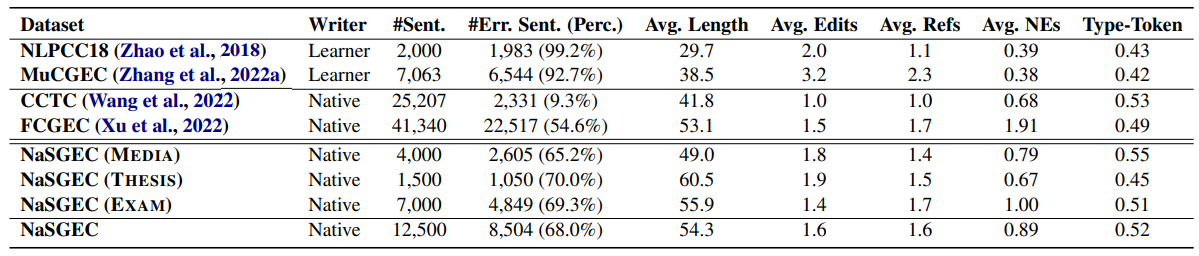
**特别说明**

1. 甲方直接开源MuCGEC和NaSGEC的Exam子集。如需NaSGEC的Media和Thesis子集，请您所在课题组、实验室、科研团队负责人打印本协议，填写相关信息并签字后，使用官方邮箱将扫描版发送给李正华([zhli13@suda.edu.cn](mailto:zhli13@suda.edu.cn))。
2. MuCGEC和NaSGEC只免费提供给“高校、科研院所”用于科学研究使用。如果用于商业产品、或者以公司名义使用，请联系李正华购买。
3. 未来我们还会继续共享更多文本纠错数据，旧的数据也可能发生更新（修改错误等）。如果需要新的共享数据，通常不需要重新签署协议（除非协议内容有所更新），只需使用官方邮箱再次发邮件索要新的数据。对于新的共享数据，即使无需重新签署协议，乙方仍需遵守以下协议条款。
4. 如果将来发生一些变化，如MuCGEC和NaSGEC改名、主页网址变化、甲方名称改变等，不影响树库共享的持续性，乙方仍需遵守以下协议条款。

MuCGEC数据的详情信息如下表：



NaSGEC数据的详情信息如下表：



本着互相促进及尊重知识产权的原则，乙方必须遵守下述条款：

1. 免费共享MuCGEC和NaSGEC后，仅用于学术研究。如果用于商业产品、或者以公司名义使用，请联系李正华购买。
2. 该资源仅在签署协议的组织内部使用，不以任何方式将资源扩散给第三方。如果申请共享数据的老师没有课题组或其他组织，通常签署协议的组织应该为从事相关研究的课题组，而不应该是学院或学校级别。
3. 高校或研究所获得MuCGEC和NaSGEC后，在和公司进行非商业性质的、以发表文章为目的的科研合作中，允许使用我们的数据。如果科研合作终止，除非取得我们的许可，并重新签署协议，否则公司不可以单独继续使用我们的数据。
4. 代码公开和实验结果可重现条款：我们希望通过共享数据，推动文本纠错技术的研究。代码公开和实验结果可重现非常重要，可以保证研究的良性发展。如果公开发表文本纠错相关论文（包括 Arxiv 文章），且实验中使用了我们的数据，则需要：
   1. 需发布完整的实验代码，包括实验配置、实验日志等所有信息，保证可以重现出论文中的结果。如果由于特殊原因确实无法发布代码或无法发布完整代码，也请积极帮助后续研究者尽快重现方法和结果。
   2. 如果在实验中使用其他人无法获得的额外资源，研究人员应同时报告未使用额外资源情况下的详细结果，使后续研究人员能够清楚地了解额外资源的影响。
   3. 不能随意改变数据的划分（train/dev/test）和评价指标（ChERRANT计算的字片段级别的F0.5分数），以便不同工作之间的实验比较。如果由于特殊原因需要改变，请在论文中详细说明原因，并尽可能汇报正常设置的结果。
   4. 需尽可能与前人的结果进行适当、公平比较，促进研究圈子良性发展。我们计划在CGEC数据集主页上及时整理和更新在MuCGEC和NaSGEC上的结果。
   5. 如果文章或其他成果中使用到MuCGEC和NaSGEC，请引用相关论文。乙方负责人需完整、清晰填写如下信息。

甲方在共享数据时，发布以下免责声明：

1. MuCGEC和NaSGEC的整理、汇编以及公开分享过程遵循了合法合规的原则，但SudaLAGroup无法完全确保数据来源方原始的版权状况。对于数据中所涉及的第三方版权内容，您有责任自行核实其合法性，并且应严格遵循相关版权法规开展后续使用行为。
2. MuCGEC和NaSGEC按“原样”提供，不附带任何形式的明示或暗示的保证，包括但不限于适销性、特定用途适用性、准确性、完整性、无侵权以及数据的时效性保证。在任何情况下，语析LAGroup对于因使用本数据而引发的任何直接、间接、偶然、特殊、惩罚性或后果性的损害赔偿（包括但不限于业务中断、利润损失、程序错误、数据丢失、系统故障等），不论基于合同、侵权行为还是其他法律理论，均不承担责任。
3. 用户明确知晓并同意，在使用MuCGEC和NaSGEC数据时，有义务采取合理的预防措施，以保护自身系统、业务及第三方权益不受潜在风险影响。若用户因违反适用法律法规、第三方协议或本免责声明，而引发任何争议、索赔或法律责任，均由用户自行承担全部法律后果，语析LAGroup概不负责。
4. 若乙方基于本开源数据创作了衍生作品，该衍生作品的版权归属于您，但需保证该衍生作品不会侵犯原数据中的第三方版权，同时，您不得以任何方式暗示您的衍生作品与语析LAGroup存在官方关联，除非获得我们的书面许可。
5. 根据使用协议，语析LAGroup明确不保证相关内容的正确性/权利清洁性，并且明确用户不得出于商业目的进行复制、使用等行为。也就是说，这些数据本身可能有其他的权利方，存在侵权的可能性；不过仅用作开源的话可以认为不是商业目的，避免在开源说明页面/其他途径透露集团模型训练用到了这些数据即可。
6. 本免责声明受中华人民共和国法律管辖，与本数据使用相关的任何争议，均应提交至指定管辖法院进行裁决。再次提醒，您使用本开源数据即视为您已充分阅读、理解并同意本免责声明的全部条款。如有任何疑问或异议，请立即停止使用本数据，并与我们联系

姓名：

职务或职称：

电话：

电子邮件：

通讯地址（邮编）：

乙方负责人签名：

年 月 日