基于细粒度强化学习增强噪声数据的低资源跨语言摘要

黄于欣[1,2]，顾怀领[1,2]，余正涛[1,2]，高玉梦[1,2]，潘通[1,2]，徐佳龙[1,2]
1昆明理工大学信息工程与自动化学院，中国昆明市，650504
2昆明理工大学云南省人工智能重点实验室，中国昆明市，650504

摘要：跨语言摘要是从源语言文档生成目标语言摘要的任务。最近，端到端跨语言摘要模型通过使用大规模、高质量数据集取得令人瞩目的结果，这些数据集通常是通过将单语摘要语料库翻译成跨语言摘要语料库而构建的。然而，由于低资源语言翻译模型性能有限，翻译噪声会严重降低模型性能。提出一种细粒度强化学习方法解决基于噪声数据的低资源跨语言摘要问题。引入源语言摘要作为黄金信号，减轻翻译后噪声目标摘要的影响。具体来说，通过计算源语言摘要和生成目标语言摘要之间的词相关性和词缺失度设计强化奖励，并将其与交叉熵损失相结合优化跨语言摘要模型。为验证所提出模型性能，构建汉语-越南语和越南语-汉语跨语言摘要数据集。实验结果表明，所提出模型在ROUGE分数和BERTScore方面优于其他基线。

关键词：跨语言摘要；低资源语言；噪声数据；细粒度强化学习；词相关性；词缺失度
https://doi.org/10.1631/FITEE.2300296