

0.1. Мурашкина А.В. Повышение точности алгоритма распознавания старопечатных тибетских документов

Старопечатные документы на тибетском языке — ценное историко-культурное наследие, веками передаваемое по наследству людьми Тибета. Однако с течением времени и под воздействием различных факторов, в том числе антропогенных, бумажные носители подвержены физическому разрушению [1]. Оцифровка исторических документов является одним из наиболее надёжных способов их сохранения и систематизации [2].

В Тибетском фонде рукописей и ксилографов Института монголоведения, буддологии и тибетологии СО РАН хранится около 70 000 единиц, требующих оцифровки. В рамках сотрудничества ИВМиМГ СО РАН и ИМБТ СО РАН был разработан программный комплекс по распознаванию и транслитерации текста со старопечатных тибетских документов.

Так как алгоритмы OCR работают с недостаточно высокой точностью, стандартным этапом является коррекция результирующего текста [3]. Настоящая работа посвящена разработке и практической реализации алгоритмов выявления ошибок, возникающих в процессе функционирования программного комплекса, для повышения точности распознавания текста.

Научный руководитель – доктор техн. наук Бараннин В. Б.

Список литературы

- [1] LAMU Y. Protection status of domestic ancient Tibetan manuscripts and literatures // J. Ethnol. 2012. Т. 3. № 6. С. 54–58.
- [2] MA L. ET AL. Segmentation and Recognition for Historical Tibetan Document Images // IEEE Access. 2020. Т. 8. С. 52641–52651.
- [3] MOLINA A., TERRADES O. R., LLAD?S J. Fetch-A-Set: A Large-Scale OCR-Free Benchmark for Historical Document Retrieval [Электронный ресурс]. URL: <https://arxiv.org/abs/2406.07315v2> (дата обращения: 01.10.2024).