УДК 004.8

РАЗРАБОТКА МУЛЬТИЯЗЫЧНОГО ПУНКТУАТОРА С ИСПОЛЬЗОВАНИЕМ СОВРЕМЕННЫХ АРХИТЕКТУР

Дутов Д.А. (ИТМО), **Митрофанов А.А.** (ИТМО)
Научный руководитель – Митрофанов А.А. (ИТМО)

Введение. В настоящее время особой популярностью пользуются модели автоматического распознавания речи. Однако для повышения читабельности текста требуется не только точное его распознавание, а также восстановление пунктуации и капитализации. Ввиду важности масштабирования моделей необходима архитектура, которая может быть использована для ряда языков.

Основная часть. В данной работе в качестве базовой архитектуры нейросетевой модели использовался предобученный трансформерный кодер BERT (Bidirectional Encoder Representations from Transformers), который модифицирован путем замены классификационной головы. Основная цель экспериментов – исследование различных вариантов архитектур классификационного блока для повышения эффективности модели в задаче расстановки пунктуации и капитализации.

В рамках исследования тестировались различные предобученные версии BERT, включая модели с разным числом параметров и размером обучающего корпуса. Кроме того, рассматривались альтернативные варианты классификационной головы, среди которых полносвязная нейронная сеть (Fully Connected Network, FCN), рекуррентные архитектуры LSTM (Long Short-Term Memory) и BLSTM (Bidirectional LSTM), а также более современные модели, такие как xLSTM [1], Mamba [2] и другие.

Для обучения и валидации модели использовались открытые корпусные данные, включающие разнообразные текстовые датасеты. Среди наиболее значимых источников можно выделить русскоязычный корпус Taiga и мультиязычный параллельный корпус Tatoeba [3]. Помимо этих ресурсов, в экспериментах использовались и другие датасеты, содержащие тексты с пунктуационной разметкой.

В мультиязычных моделях было предложено деление языков по родственным группам. Например: английский - немецкий, португальский - испанский - французский.

Выводы. В ходе работы была реализована и протестирована серия нейросетевых моделей, отличающихся архитектурой классификационной головы. Также был проведён сравнительный анализ их эффективности. Это позволило определить оптимальные конфигурации для решения задачи пунктуации и капитализации текста. Среднее качество лучшей русскоязычной модели достигло F1=0.8, а мультиязычной модели – F1=0.74.

Список использованных источников:

1. Beck M. et al. xLSTM: Extended Long Short-Term Memory //arXiv preprint arXiv:2405.04517. – 2024.
2. Gu A., Dao T. Mamba: Linear-time sequence modeling with selective state spaces //arXiv preprint arXiv:2312.00752. – 2023.
3. Tatoeba // https://tatoeba.org/ru/ (дата обращения: 30.01.2024).