УДК 004.8

СРАВНЕНИЕ МОДЕЛЕЙ МАШИННОГО ПЕРЕВОДА И ОПРЕДЕЛЕНИЕ МЕСТА НЕЙРОННЫХ СЕТЕЙ В ПЕРЕВОДЧЕСКОМ ПРОЦЕССЕ.

Дутов Д.А. (ГУАП)
Научный руководитель – Светова С.Ю. (директор компании «Т-Сервис»)

Введение. В настоящее время активно развиваются нейронные архитектуры encoder-decoder с использованием блоков Transformer. Так как данная технология вышла на новый уровень с появлением больших языковых моделей (Large Language Models, LLM), снова возник вопрос о возможности использования этих методов для задач перевода. Следовательно, требуется исследование качества работы представленных на рынке моделей, чтобы предположить вектор развития переводческой деятельности.

Основная часть. Для исследования качества моделей мы разработали малый корпус текстов в формате язык перевода – гипотеза модели – постредактированный вариант. В качестве текстов на английском языке мы взяли тексты из открытого источника arXiv[1]. Корпус подразумевает перевод с английского на русский язык. После его разработки мы провели лингвистический анализ ошибок перевода, составили классификацию. Также оценили метрики (WER, BLEU, LEPOR [2], COMET [3]) на постредактированных текстах, что позволило провести оценку без погрешностей из-за языковой синонимии. На этом этапе мы проанализировали возможности той или иной модели в сфере технического перевода исходя из которых предположили, что вероятно создание модели, которая сможет распознавать частотные сложные паттерны для переводческих моделей. На основании малого корпуса текстов проверили распределение метрик при оценке непостредактированного референсного текста и гипотезы. Исходя из этих данных мы подготовили датасет на основании мультиязычного параллельного корпуса Tatoeba в формате язык перевода – гипотеза модели – целевой параметр: верная гипотеза или нет. После этого приступили к создании модели на основе энкодера bert-base-multilingual-cased.

Выводы. В данной работе приведен анализ ошибок моделей машинного перевода с учётом сравнения с постредактированным вариантом, а не референсным, что позволяет избежать искажения значений метрик. Исходя из результатов было выявлено, что модели ошибаются по частотным паттернам, а следовательно гипотетически возможно создание модели, которая сможет предсказать возможную ошибку в машинном переводе. Проведено исследование разных архитектур и подходов бинарной классификации предложений. В соответствии с этим были разработаны и обучены модели, решающие поставленную задачу, а также оценено их качество.

Список использованных источников:

1. ArXiv a free distribution service and an open-access archive for nearly 2.4 million scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics // Cornell University – URL: https://arxiv.org/ (дата обращения: 30.01.2024).
2. Lifeng Han (Aaron). LEPOR: An Augmented Machine Translation Evaluation Metric. Master of Science in Software Engineering 2014, abs/1703.08748
3. Ricardo Rei, Ana C Farinha, Jose G. C. de Souza, Pedro G. Ramos, Andre F. T. Martins, Luisa Coheur, Alon Lavie. Searching for COMETINHO: The Little Metric That Could. URL: https://aclanthology.org/2022.eamt-1.9.pdf (дата обращения: 30.01.2024)

Автор _____ Дутов Д.А.
Научный руководитель _____ Светова С.Ю.