УДК 004.89

РАЗРАБОТКА МОДЕЛИ КЛАССИФИКАЦИИ ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ ПРОЦЕССНО-СМЫСЛОВОГО АНАЛИЗА ТЕКСТОВ

Никифорова А.Д. (федеральное государственное автономное образовательное учреждение высшего образования Университет ИТМО)

Роговой В. (федеральное государственное автономное образовательное учреждение высшего образования Университет ИТМО)

Научный руководитель – доцент, кандидат физико-математических наук Суров И.А. (федеральное государственное автономное образовательное учреждение высшего образования Университет ИТМО)

В данной работе разрабатывается классификатор для смыслового анализа текста в рамках процессного подхода к моделированию семантики.

Введение. Моделирование семантики – важное и развивающееся направление в сфере обработки и анализа естественного языка. Одним из подходов к моделированию семантики является процессный. В нём предполагается, что смысл информационных блоков порождается процессом, в котором они участвуют с точки зрения субъекта. В качестве универсальной структуры процесса используется жизненный цикл. Любой цикл представлен следующими последовательными этапами: восприятие (получение от окружающего мир информации с помощью органов чувств), анализ (обработка полученной информации), планирование (определение цели и шагов по её достижению), действие (выполнение плана), прогресс (доведение предварительного результата до финала) и оценка (подведение итогов, рефлексия) [1]. Циклы могут накладываться и пересекаться. Смысловой анализ текста в рамках процессного подхода к моделированию семантики предполагает поиск и выделение подобных циклов. В этой работе рассматривается метод процессно-смыслового анализа, основанный на применении машинного обучения.

Основная часть. Для процессно-смыслового анализа текстов был разработан шестиклассовый классификатор текстовых фрагментов, работающий с данными на английском языке. За основу была взята предобученная большая языковая модель RoBERTa [2] с оптимизатором AdamW и планировщиком cosine scheduler, так как в сравнении со своими аналогами (BERT [3], DistilBERT [4]) при одинаковой настройке гиперпараметров она показала наилучший результат.

Модель была обучена на вручную размеченных текстовых данных, разбитых на шесть классов в соответствии с этапами жизненного цикла. Часть предложений имела оценки одного эксперта, часть – двух или трёх. В качестве обучающих данных были использованы 1288 предложений, имеющих более одной метки. В размеченных данных наблюдался дисбаланс классов, для устранения данной проблемы были опробованы два подхода: проведение балансировки путем отбрасывания части примеров из более объемных классов и применение специальной функции потерь DiceLoss [5]. Второй подход показал лучший результат.

Обученный шестиклассовый классификатор был применен к датасету ROCStories [6], содержащему небольшие логически завершенные тексты, разделенные на пять предложений. В силу того, что данные тексты логически целостны, было сделано предположение, что классификатору удастся выделить в них структуру, приблизительно советующую полному смысловому циклу (восприятие-анализ-планирование-действие-прогресс-оценка). Предположение подтвердилось, была обнаружена корреляция между номером предложения в тексте и последовательными этапами жизненного цикла. Так, большая доля первых предложений была отнесена классификатором к классам «восприятие» и «анализ», а большая доля пятых, т. е. последних, - к классу «оценка».

На текущий момент ведется работа над улучшением качества модели и устранением проблемы пересекающихся смысловых циклов.

Также дополнительно в рамках работы над задачей была проверена гипотеза о том, что стандартно выделяемые типы речи (повествование, описание, рассуждение) коррелируют с рассматриваемой шестиклассовой классификацией. Для решения задачи классификации текстов на три класса в соответствии с типами речи была использована модель логистической регрессии, обученная на данных 23-го задания из ЕГЭ по русскому языку [7]. И хотя некоторые закономерности между типами речи и этапами жизненного цикла действительно удалось обнаружить, они оказались недостаточными для использования типа речи в качестве значимого предиктора.

Выводы. Для выделения в текстовых фрагментах классов в соответствии с моделью жизненного цикла был разработан и обучен классификатор, работающий с текстами на английском языке. Усредненная по шести классам f-мера составила 63%. Применение модели к небольшим текстам, представляющим собой логически завершенные истории, показало её работоспособность. В дальнейшем классификатор возможно использовать в задачах реферирования и аннотирования текстов.

Список использованных источников:

1. *Суров И.А.* Жизненный цикл: смысловая матрица процессного моделирования // Онтология проектирования. 2022. Т.12, №4(46). С.430-453. DOI:10.18287/2223-9537-2022-12-4-430-453.
2. Liu Y. et al. Roberta: A robustly optimized bert pretraining approach // arXiv Prepr. arXiv1907.11692. 2019.
3. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv Prepr. arXiv1810.04805. 2018.
4. Sanh V. et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter // arXiv Prepr. arXiv1910.01108. 2019.
5. Li X. et al. Dice loss for data-imbalanced NLP tasks //arXiv preprint arXiv:1911.02855. – 2019.
6. Nasrin Mostafazadeh et al. A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories // arXiv:1604.01696. 2016
7. РешуЕГЭ [Электронный ресурс]. – Режим доступа: https://rus-ege.sdamgia.ru/ – Дата доступа: 13.11.2022.

Никифорова А.Д. (автор) Подпись

Роговой В. (автор) Подпись

Суров И.А. (научный руководитель) Подпись