УДК 004.912

Объяснимая классификация логических ошибок в русскоязычных текстах с помощью больших языковых моделей

Николаева А. К.[1]
Научный руководитель – к.ф.-м.н, доцент Каф. Информационно-аналитических систем СПБГУ, Михайлова Е. Г. [1]
Консультант – к.ф.-м.н., старший преподаватель Каф. Информационно-аналитических систем СПБГУ, Азимов Р. Ш. [1]
[1]СПбГУ
anna.nikolaevakons@gmail.com

Введение

Задача классификации логических ошибок важна для анализа текстов на предмет качества аргументации и дезинформации: многие утверждения могут звучать убедительно, но при этом являться ошибочными с точки зрения логики. В настоящее время существуют подходы, основанные на применении БЯМ (больших языковых моделей), для подобного анализа текстовых данных на английском языке, для русскоязычных же данных нет таких аналогов. Кроме того, важным аспектом является интерпретация результатов, полученных с помощью БЯМ, иначе не понятно на какие паттерны в утверждениях языковая модель обращает больше всего внимания и как эти паттерны зависят от типа логической ошибки и от контекста. Для объяснимости классификации текстовых данных также существует ряд англоязычных методов. Применение существующих англоязычных решеёний к русскоязычным данным затруднительно, так как это может отразиться на качестве классификации и объяснений из-за дефектов перевода. Дообучение же русскоязычных моделей для задачи объяснимой классификации логических ошибок затруднительно, так как для этого необходимо наличие текстовых датасетов на русском языке, которых в открытом доступе не имеется.

Основная часть

Для проведения экспериментов создан обучающий и тестовый набор русскоязычных данных. Набор состоит из утверждений, аннотированных меткой класса ошибки и маской слов, входящих в объяснение. Для выявления оптимального подхода рассматриваются три/четыре стратегии решения задачи объяснимой классификации логических ошибок в русскоязычных утверждениях:

1. Обучение русскоязычных моделей на русскоязычном корпусе.
2. Обучение англоязычной модели на англоязычном корпусе, применение обученной модели на переведенном на английский язык русскоязычном тексте.
3. Обучение англоязычной модели на переведенном русскоязычном корпусе, применение обученной модели на переведенном на английский язык русскоязычном тексте.
4. Применение необученных под целевую задачу больших языковых моделей с использованием промпта с определениями классов логических ошибок и примерами неверных высказываний для каждого класса. На основе данного промта модель классифицирует целевые данные.

Выводы

Выявлена наиболее оптимальная стратегия для решения задачи объяснимой классификации логических ошибок в русскоязычных данных.

Список использованных источников:

1. Explain Yourself! Leveraging Language Models for Commonsense Reasoning / Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, Richard Socher // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics / Ed. by Anna Korhonen, David Traum, Lluís Màrquez. — Florence, Italy : Association for Computational Linguistics, 2019. — . — P. 4932–4942. — URL: https://aclanthology.org/P19-1487/.

2. Ghasemi Madani Mohammad Reza, Minervini Pasquale. REFER: An End-to-end Rationale Extraction Framework for Explanation Regularization // Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL) / Ed. by Jing Jiang, David Reitter, Shumin Deng. — Singapore : Association for Computational Linguistics, 2023. — . — P. 587–602. — URL: https://aclanthology.org/2023.conll-1.40/.

3. Nguyen Thi Huyen, Fisichella Marco, Rudra Koustav. A trustworthy approach to classify and analyze epidemic-related information from microblogs // IEEE Transactions on Computational Social Systems. — 2024. — Vol. 11, no. 5. — P. 6229–624

4. Sahai Saumya Yashmohini, Balalau Oana, Horincar Roxana. Breaking down the invisible wall of informal fallacies in online discussions // ACL-IJCNLP 2021-Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. — 2021

5. e-snli: Natural language inference with natural language explanations / Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, Phil Blunsom // Advances in Neural Information Processing Systems. — 2018. — Vol. 31