

ОБРАБОТКА ФАКТОЛОГИЧЕСКИХ ПРОТИВОРЕЧИЙ В НЕСТРУКТУРИРОВАННЫХ ТЕКСТАХ

Рассматривается подход обнаружения и обработки фактологических противоречий.

Введение

В современных системах обработки естественного языка одной из ключевых проблем остаётся обеспечение логической согласованности текстовых данных. При работе с большими корпусами документов, а также при построении баз знаний, возможны ситуации, когда различные утверждения противоречат друг другу, что снижает достоверность анализа и качество принимаемых решений. Целью данной работы является разработка системы обнаружения противоречий.

I. Алгоритм обработки данных

На первом этапе выполняется разбиение документа на фрагменты фиксированного размера. Далее с использованием LLM производится извлечение фактов.

Следующим этапом является деконтекстуализация, при котором неявные ссылки заменяются на конкретные сущности. После нормализации каждое утверждение преобразуется в вектор.

Полученные векторные представления сохраняются в векторной базе данных. Для каждого утверждения выполняется поиск наиболее близких по смыслу утверждений на основе косинусной близости, в результате чего формируется множество кандидатных пар.

II. Определение противоречий

Для каждой пары утверждений применяется модель Natural Language Inference. Модель определяет логическое отношение между утверждениями как противоречие, следование или нейтральное отношение. В итоговый набор включаются только те пары, для которых вероятность противоречия превышает заданный порог.

III. Реализация системы

Система реализована в виде модульного API-сервиса, обеспечивающего полный цикл обработки

данных от загрузки документа до формирования отчёта. Архитектура включает компоненты извлечения фактов, деконтекстуализации, векторизации, хранения и классификации. Взаимодействие между модулями организовано по конвейерной схеме, что обеспечивает масштабируемость и возможность независимой модификации отдельных компонентов.

Векторное хранилище реализовано с использованием специализированных баз данных, таких как Qdrant, что позволяет эффективно работать с большими объёмами данных. Для каждой входной выборки формируется ограниченное множество кандидатных пар, что снижает вычислительную сложность последующей классификации.

IV. Промежуточное тестирование системы

Было проведено тестирование NLI-модели, используемой в нашем проекте, получены результаты.

В результате тестирования было выявлено, что хуже всего модель обрабатывает аббревиатуры. При сравнении тех же утверждений, но уже с расшифровками аббревиатур, система работает корректно.

V. Выводы

Разработана система обнаружения противоречий в текстах, основанная на интеграции методов извлечения знаний, векторного поиска и моделей логического вывода. Предложенный подход обеспечивает автоматизацию анализа текстовой информации и может применяться при построении баз знаний, обработке документов и создании интеллектуальных обучающих систем. Полученные результаты подтверждают эффективность архитектуры и её применимость для дальнейшего развития.

Список литературы

1. He, P. DeBERTa: Decoding-enhanced BERT with Disentangled Attention / P. He, X. Liu, J. Gao, W. Chen // arXiv preprint arXiv:2006.03654. – 2020.

Сайковский Антон Валерьевич, студент кафедры интеллектуальных информационных технологий БГУИР, antonsaikovski@gmail.com.

Вашкевич Максим Викторович, студент кафедры интеллектуальных информационных технологий БГУИР, maxim.vashkevich2007@gmail.com.

Мяло Всеволод Вячеславович, студент кафедры интеллектуальных информационных технологий БГУИР, myalo.vsevolod@gmail.com.

Научный руководитель: Зотов Никита Владимирович, младший научный сотрудник НИЛ 3.3, ассистент кафедры интеллектуальных информационных технологий БГУИР, n.zotov@bsuir.by.