Многоязычный поиск ранее проверенных утверждений для верификации фактов

Николаев Е.Н.
nikolaev.en@phystech.edu

Центр когнитивного моделирования МФТИ

2025 – Научно-исследовательский семинар





Мотивация

Количество контента, публикуемое в медиа, существенно превосходит возможности специалистов по проверке фактов, особенно учитывая кросс-языковую природу данных. Автоматизированный подход на основе нейросетей может помочь им, извлекая уже существующие методы проверки фактов, относящиеся к исследуемому контенту.

Задача



Задача: Разработать систему для эффективного поиска проверенных утверждений (истинных или ложных) релевантных запросу

Формальная постановка задачи

Дано:

- Запрос Q, содержащий одно или несколько утверждений $\{Q_1, Q_2, \ldots, Q_k\}$.
- Корпус C состоящий из документов (проверенных утверждений) D_i, потенциально позволяющих подтвердить или опровергнуть утверждения, приведенные в запросе Q.

Формальная постановка задачи

Цель: Составить поисковую систему, выбирающая top-k релевантных документов, удовлетворяющих следующим требованиям:

1. Каждое утверждение Q_i из $\{Q_1, Q_2, \ldots, Q_k\}$ должно подтвержаться или опровергаться каким либо D_j из $\{D_1, D_2, \ldots, D_k\}$..
2. Каждое утверждение D_j из $\{D_1, D_2, \ldots, D_k\}$ должно подтвержать или опровергать какое либо утверждение Q_i из $\{Q_1, Q_2, \ldots, Q_k\}$.

Данные

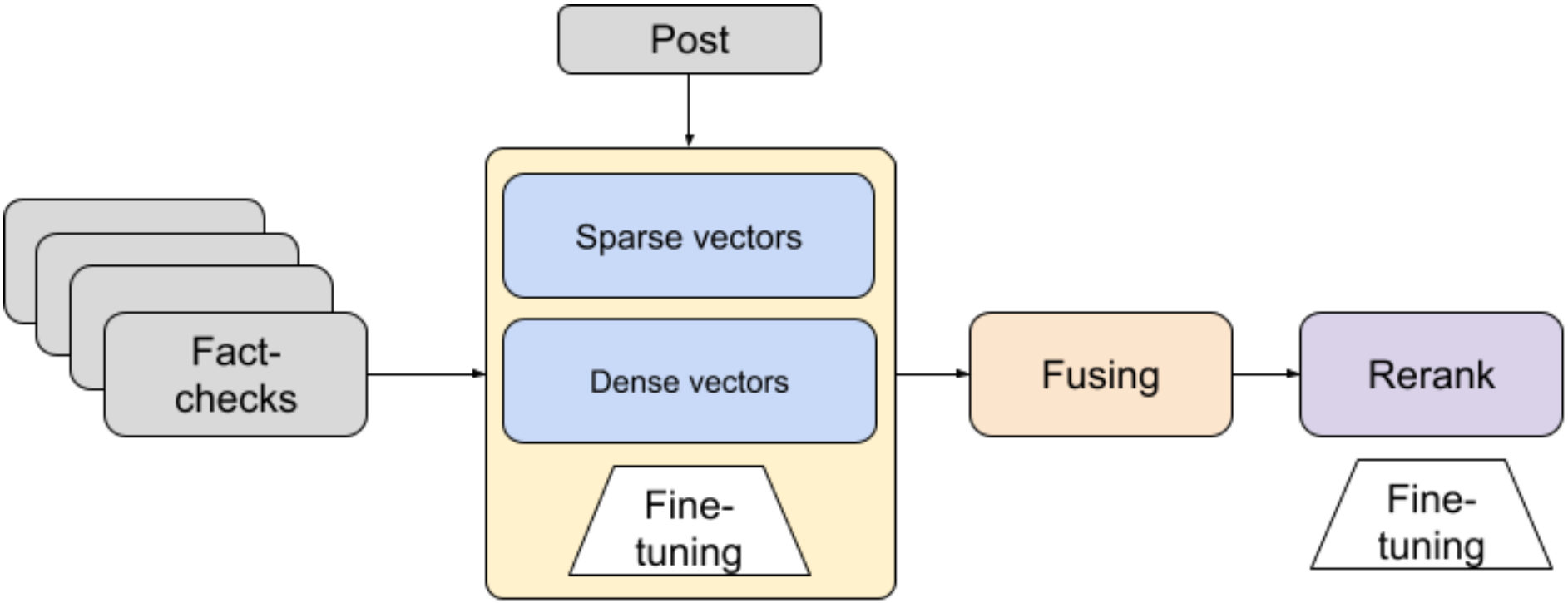
Для проведения экспериментов и оценки эффективности модели был выбран датасет MultiClaim

- 28 тысяч запросов на 27 языках из социальных сетей.
- 206 тысяч проверенных утверждений на 39 языках, написанных профессиональными специалистами по проверке фактов.
- 31 тысяча пара пост-утверждение.

Методы оценки

- Success-at-10 – вхождение всех необходимых проверенных утверждений в top-10 наиболее веротяных утверждений.

Базовая схема поиска



Базовый подход к поиску утверждений

Векторный поиск



Для опоры на лексическую составляющую текста были выбраны следующие подходы:

1. BM25
2. BGE-m3-lexical

Для опоры на семантическую составляющую текста были выбраны следующие модели:

1. stella_en_1.5B_v5
2. BGE-m3-dense
3. multilingual-e5-large-instruct (промт *“Given a post on a social network, retrieve the claims it contains”*)

Для объединения резултата был использован Reciprocal rank fusion (RRF).

Реранкер

Для улучшения качества выдачи также был произведено повторное ранжирование с помощью модели BAAI/bge-reranker-v2-m3.

Эта модель имеет кросс-энкодерную архитектуру т.е. оценивает соотвествие запроса и ответа в виде query[SEP]answer

Дообучение

Обучение проводилось в 2 этапа с опорой на результат sparse-составляющей векторного поиска:

1. Позитивные примеры берутся из top-5. Негативные примеры выбираются случайно по всей выборке. Отношение pos:neg = 1:5
2. Позитивные примеры берутся входят в диапозон 4-15 наиболее вероятных при sparse-поиске. Негативные примеры берутся из top-25. Отношение pos:neg = 1:8

Дообучение

Для дообучения dense-составляющей была использована модель BAAI/BGE-m3 и использованием contrastive loss

Contrastive Loss

$$L(x_1, x_2, y) = \frac{1}{2}\left(y \cdot d(x_1, x_2)^2 + (1-y) \cdot \max(0, m - d(x_1, x_2))^2\right)$$

где:

- x_1, x_2 — векторы признаков для двух объектов,
- y — метка: $y = 1$, если объекты похожи, и $y = 0$, если непохожи,
- $d(x_1, x_2)$ — косинусное расстояние между x_1 и x_2,
- m — зазор (margin), который определяет минимальное желаемое расстояние между непохожими объектами.

Дообучение

Для обучения реранкера была взята модель BAAI/bge-reranker-v2-m3.

В силу существенного дисбаланса позитивного и негативного классов при обучении использовалась функция потерь Focal Loss с коэфициентами $\alpha = 0.9$ и $\gamma = 4.0$

Focal Loss

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

где:

- p_t — вероятность правильной классификации,
- α_t — весовой коэффициент для класса,
- γ — фокусирующий параметр (гамма).

Результат

Sparse Retrieval	Dense Retrieval	Fusion	Rerank	Mono	Cross
BGE-m3-lexical	–	–	No	79.18	57.87
–	BGE-m3-dense	–	No	78.91	65.9
BGE-m3-lexical	BGE-m3-dense	–	No	85.20	67.02
–	BGE-m3-dense-finetuned	–	No	79.04	65.4
BGE-m3-lexical	BGE-m3-dense-finetuned	1:1	No	82.14	66.0
		1:1	Yes	<u>85.90</u>	**67.2**
		8:2	No	85.43	<u>66.95</u>
		8:2	Yes	**86.01**	66.57
BM25	–	–	No	62.96	56.15

Таблица: Результат межъязыковых (Кросс) и одноязычных (моно) задач. Измеряется как Success@10, %. Слияние $k : n$ означает, что sparce вес равен k, а dence вес равен n при RRF.

Виртуальные документы



Идея: вместо поиска по векторизированным проверкам фактов генерировать документы, которые могли бы быть подходящим запросом для этой проверки

Проверяемый факт: Земля плоская

Виртуальные документы:

- А вы слышали, что земля квадратная?
- Общеизвестный факт: Земля круглая
- Популярное мнение: Земля треугольная

Для создания виртуальных документов использовалась LLama3-8B-Instruct-int4

Нормализация текстов



Проблема: Из-за домена данных (соцсети) тексты имеют специфическую лексику и стиль написания
Идея: Нормализовать текст с помощью LLM (убрать специфическую лексику, сокращения, жаргонизмы, повторы)

В силу большого количества документов была дополнительна обучена google/flan-t5-small

Нормализация текстов



Пример: OMG, literally SO excited for the weekend! Weekend vibes incoming! FR FR! Can't wait. Like, REALLY can't wait. Weekend! Weekend! Weekend! Gonna be LIT!

Преобразование: I am feeling very excited about the upcoming weekend. I am eagerly anticipating its arrival and have high expectations for enjoyable activities.

GraphRAG



- **Идея:** Графовый подход для ответов на вопросы по частным текстовым корпусам, который масштабируется как по общности вопросов, так и по объему исходного текста.
- **Ключевые особенности:**
  - Использование LLM для построения графового индекса в два этапа.
  - Сочетание сильных сторон RAG и QFS.

GraphRAG: Построение индекса



Двухэтапное построение графового индекса с помощью LLM:

1. **Извлечение графа знаний сущностей:**
   - LLM анализирует исходные документы.
   - Идентифицирует ключевые сущности и связи между ними.
   - Формируется граф знаний (entity knowledge graph).

GraphRAG: Построение индекса



Двухэтапное построение графового индекса с помощью LLM:

1. **Извлечение графа знаний сущностей:**
   - LLM анализирует исходные документы.
   - Идентифицирует ключевые сущности и связи между ними.
   - Формируется граф знаний (entity knowledge graph).
2. **Предварительная генерация обобщений для сообществ (Community Summaries):**
   - В графе знаний выявляются группы тесно связанных сущностей (сообщества).
   - Для каждого такого сообщества LLM заранее генерирует обобщающее описание (community summary).

Дальнейшие шаги



1. Продолжение работы над GraphRAG
2. Продолжение дообучения моделей
3. Использование дополнительных моделей для поиска
4. Улучшение механизма агрегации поиска

Спасибо за внимание!

nikolaev.en@phystech.edu