Исследование подходов для оценки семантического сходства текстов в задаче сопоставления вакансий и резюме

Петрова Александра, 417 группа
Майсурадзе Арчил Ивериевич

Происхождение задачи

Задача **сопоставления вакансий и резюме** связана с необходимостью выбора работодателями кандидатов на основе большого количества резюме.

Актуальность задачи

Задача стала особенно актуальной с развитием интернет-платформ для поиска работы, где количество резюме может быть значительным.

Формальная постановка задачи

Пусть есть выборка вакансий V и резюме R, множество $\{\mathrm{R}_v\}_{v=1}^{l}$,
где R_v — множество релевантных резюме для вакансии v
Необходимо построить отображение $\mathrm{S}\colon V \to L^n$, L^n — множество упорядоченных списков резюме из R длины n

Такое что: $MAP@n = \frac{1}{l}\sum_{v=1}^{l} AP_v@n = \frac{1}{l}\sum_{v=1}^{l}\frac{1}{|R_v|}\sum_{k=1}^{n} I[L_v^n[k] \in R_v]\, P_v@k \to \max_S,$

где $P_v@k = \frac{\left|L_v^k \cap R_v\right|}{k}$, L_v^n — упорядоченный список резюме для вакансии v длины n

Обзор существующих методов

Job Recommendation Systems (JRS)

1. **Контентно-ориентированные системы (Content-Based JRS)** – анализируют соответствие описания вакансий и профилей кандидатов по ключевым параметрам (навыки, опыт, предпочтения). Этот подход предполагает, что вакансии, похожие на те, которые нравились пользователю в прошлом, также будут представлять интерес.
2. **Системы коллаборативной фильтрации (Collaborative Filtering JRS)** – рекомендуют вакансии на основе предпочтений и действий других пользователей с похожими интересами, не анализируя содержание.
3. **Гибридные системы (Hybrid JRS)** – объединяют подходы контентного анализа и коллаборативной фильтрации, компенсируя ограничения каждого метода.
4. **Системы на основе знаний (Knowledge-Based JRS)** – используют правила и базы знаний для подбора вакансий, что особенно эффективно для сложных или нестандартных требований.

В докладе будут рассматриваться методы, которые используют семантическое понимание и общий контекст, а не полагается на исторические данные или взаимодействия

Данные

Данные включают набор вакансий и резюме на русском языке, предоставленный HR-отделом компании ACD/Labs. Большинство вакансий и резюме относятся к сфере IT.

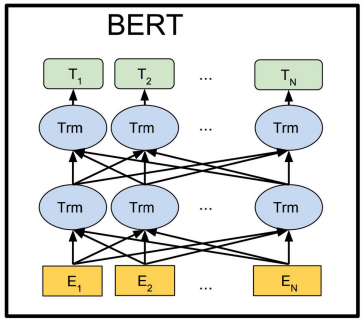
Для каждой вакансии известен перечень резюме кандидатов, приглашенных на собеседование. В выборке отсутствуют резюме, не связанные с конкретными вакансиями, и каждое резюме привязано к единственной вакансии.

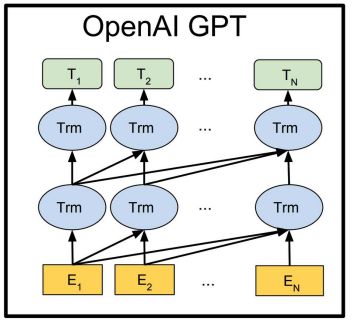
Количество резюме	Количество вакансий	Описание
800	800	1 вакансия - 1 резюме
2	2	1 вакансия - 2 резюме
3	1	1 вакансия - 3 резюме
4	2	1 вакансия - 4 резюме
5	1	1 вакансия - 5 резюме
8	1	1 вакансия - 8 резюме
9	1	1 вакансия - 9 резюме
11	2	1 вакансия - 11 резюме
13	2	1 вакансия - 13 резюме
17	1	1 вакансия - 17 резюме
890	813	Total

Таблица 1: Распределение вакансий по количеству резюме

Метод решения

1. Предобработка текста, токенизация
2. Получение векторных представлений текстов с помощью модели на архитектуре Transformer
3. Ранжирование на основе cosine similarity
4. Оценка качества





$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Выбор моделей

Model	Parameters	Layers	Languages
bert-base-multilingual-cased	110M	12	104
bert-base-ru-cased	110M	12	Russian
DeepPavlov rubert-base-cased	180M	12	Russian
rugpt2large	774M	48	Russian
rugpt3large	760M	96	Russian
paraphrase-MiniLM-L6-v2	22M	6	Multilingual

Результаты экспериментов

	Random	BERT multilingual	BERT Russian	RuGPT2 Large	RuGPT3 Large	MiniLM	text-embedding-ada-002
MAP@10	0.03	0.1	0.12	0.11	0.13	0.33	0.58
MAP@20	0.04	0.14	0.15	0.14	0.15	0.41	0.66

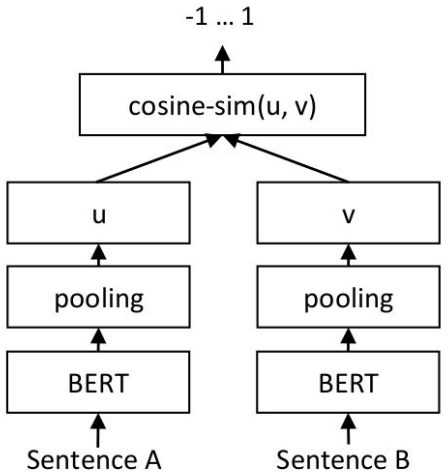
Fine-tuning

Классификация

1. Объединяем вакансию и резюме: [CLS] Вакансия [SEP] Резюме [SEP].
2. Пропускаем через BERT с классификационным слоем.
3. На выходе — вероятность соответствия пары.
4. Обучение через binary_crossentropy.

Сиамская сеть

1. Кодируем вакансию и резюме отдельными BERT.
2. Сравниваем эмбеддинги (косинусное сходство).
3. На выходе — вероятность соответствия.
4. Обучение через контрастивную loss или binary_crossentropy.



Выводы

- Модели на основе архитектур BERT и GPT продемонстрировали схожие показатели качества, однако BERT содержит меньше параметров и более эффективен для использования
- Sentence Transformer превзошел базовые трансформеры, обеспечив более высокие результаты. Это объясняется его адаптацией для задач сравнения текстов, таких как семантическое сходство, благодаря обучению на парах предложений.
- Для fine-tuning модели можно использовать два подхода: классификацию и сиамскую сеть.