Сопоставление векторных представлений вакансий и резюме с использованием больших языковых моделей

Л.А. Комарова, А.М. Колосов, В.И. Соловьев

***Аннотация*—В данной работе представлены современные подходы обработки естественного языка для анализа соответствия между текстами вакансий и резюме. Исследование сосредоточено на использовании больших языковых моделей (Large Language Models, LLM) для создания векторных представлений текстов вакансий, резюме и извлеченных из них навыков. Показано, что векторные представления текстов вакансий и резюме, а также навыков, извлеченных из них, образуют различные векторные пространства, в то время как стандартизированные навыки ESCO и слова находятся в едином пространстве. Для проверки гипотез о единстве векторных пространств использованы методы статистического анализа, такие как максимальная средняя дисперсия (MMD), а также алгоритмы понижения размерности (t-SNE и Ivis), позволяющие провести визуальный анализ распределения векторных представлений. Дополнительно в статье проведена аналитика по результатам экспериментов, и уделено особое внимание анализу свойств навыков ESCO, которые, благодаря стандартизации, образуют единое векторное пространство. Результаты исследования могут способствовать улучшению процессов подбора персонала, предлагая новые методы сопоставления навыков соискателей и требований работодателей. В работе также сделан вывод о важности стандартизации данных для их интерпретации и сопоставления.**

***Ключевые слова*—построение векторных представлений, LLM, Ivis, t-SNE, MMD**

I. Введение

На современном рынке труда соответствие сведений о соискателях, представленных в их резюме (CV), и ожиданий, изложенных в описании вакансии, играет ключевую роль в процессе найма. Статья посвящена области обработки естественного языка (NLP) для автоматического сопоставления сведений, представленных в резюме и ожиданий, изложенных в описании вакансии.

Статья получена 8 октября 2024. Статья представляет собой часть диссертационной работы аспиранта Комаровой Л.А.

Любовь Александровна Комарова Финансовый Университет при Правительстве РФ, Москва, Россия (e-mail: 229388@edu.fa.ru).

Алексей Михайлович Колосов, МГУ им. М.В. Ломоносова, Москва, Россия (e-mail: akolosov@cs.msu.ru).

Владимир Игоревич Соловьев, Финансовый Университет при Правительстве РФ, Москва, Россия (e-mail: vsoloviev@fa.ru).

Основные цели и задачи работы:

1. Построение векторных представлений вакансий, резюме и навыков с использованием LLM. Применяя большие языковые модели, решается задача построения векторных представлений вакансий, резюме, навыков из требований к вакансии и навыков из компетенций в резюме.

2. Анализ векторных представлений вакансий и резюме. Эта часть исследования позволяет продемонстрировать, что векторные представления полных текстов вакансий и резюме не занимают единое векторное пространство.

3. Анализ векторных представлений навыков, извлеченных из вакансий и резюме. Эта часть исследования позволяет продемонстрировать, что векторные представления навыков из вакансии и резюме занимают единое векторное пространство. Вывод важен для определения совместимости между навыками, которых ждут работодатели, и теми навыками, которые представлены соискателями.

II. Обзор подходов к построению векторных представлений вакансий, резюме и навыков

Применение технологий NLP в сфере управления персоналом началось с простых задач по обработке текста, таких как поиск, по ключевым словам, (rule-based), в резюме и объявлениях о вакансиях. Ранние исследования [1, 2] в значительной степени полагались на методы логического поиска, чтобы сопоставлять кандидатов с требованиями в вакансиях исключительно на основе наличия определенных слов или фраз. Эти подходы, хотя и просты, часто не учитывают нюансы требований к должностям или разнообразную квалификацию кандидатов.

По мере развития компьютерной лингвистики и машинного обучения росли и возможности систем NLP в сфере HR-технологий. Внедрение статистических методов в конце 1980-х и 1990-х годах привело к появлению более сложных методов классификации и

кластеризации текста [11], которые позволили лучше группировать похожие резюме и описания вакансий без явного сопоставления ключевых слов. В 2010-х для построения векторных представлений стали использовать Word2Vec и GloVe [18, 19]. Эти модели особенно эффективны для определения семантических отношений слов, которые имеют значение для задач в сфере HR-технологий. Базовой является работа [12, 13] по построению векторных представлений с помощью Word2Vec, использующей нейронные сети для изучения словесных ассоциаций из большого массива текста. Аналогично статья [17] дает представление об использовании для построения векторных представлений слов с помощью модели GloVe.

За последнее десятилетие произошел значительный сдвиг в сторону моделей глубокого обучения. Внедрение таких архитектур, как сверточные (CNN) и рекуррентные (RNN) нейронные сети, а затем и модель трансформеров (Bert), значительно улучшило способность систем понимать текст. Эти модели превосходно улавливают сложные закономерности в данных и устанавливают новые стандарты точности в таких задачах, как классификация текста, анализ настроений и распознавание объектов.

Например, исследование [3] показывает, что использование BERT для получения векторных представлений резюме и описаний вакансий значительно повышает точность сопоставления за счет понимания контекста.

Одним из наиболее значимых событий в новейшей истории NLP является появление предварительно обученных языковых моделей, таких как GPT и их производных [22, 23]. Эти модели обучаются на больших объемах текста, а затем адаптируются (дообучаются) для конкретных задач. В сфере управления персоналом эти модели предлагают возможности по извлечению навыков из резюме, пониманию требований к должностям и их сопоставлению. Также, есть ряд исследований, подтверждающих эффективность GPT моделей в задаче автоматизации HR процессов [5] и интерпретации сложных навыков и требований из вакансий и резюме [4].

Обзор дает представление о подходах, которые используются при построении систем искусственного интеллекта в технологиях управления персоналом. От элементарного поиска, по ключевым словам, до продвинутого семантического анализа на базе искусственного интеллекта — эта область развивалась, используя глубокое обучение, чтобы лучше подбирать соискателей подходящих вакансий. Таким образом, можно сделать вывод, что на текущий момент наиболее распространенным и изученным представляется использование семейств моделей BERT и GPT для построения векторных представлений и дальнейшего использования построенных векторных представлений в сфере подбора персонала.

В работе производится построение векторных представлений вакансий и резюме для каждого из двух выбранных подходов. Для построенных с помощью каждого из подходов набора векторных представлений производится сравнение порядков близостей, полученных векторных представлений резюме и представлений вакансий, с известным ответом [20]. Известным ответом в исследовании является экспертная разметка, полученная от рекрутеров. Эксперты сопоставили 2 набора из 10 и 25 резюме для одной вакансии и поставили ранг, как уровень соответствия данного резюме вакансии.

	bert_prob	gpt_prob	expert_rank	bert_rank	gpt_rank	reordered_expert_rank
0	0.048261	0.399899	24	22	2	2
1	0.027958	0.622020	1	3	25	25
2	0.041865	0.521966	19	17	15	7
3	0.037314	0.497133	2	12	11	24
4	0.026935	0.610132	3	2	24	23
5	0.044169	0.492613	4	20	10	22
6	0.029167	0.551077	5	8	20	21
7	0.032766	0.601093	6	10	23	20
8	0.043085	0.512576	20	19	13	6
9	0.014509	0.092344	25	1	1	1
10	0.050996	0.508569	22	24	12	4
11	0.045935	0.462213	21	21	6	5
12	0.028600	0.402617	14	6	3	12
13	0.038698	0.463134	17	14	7	9
14	0.049330	0.423319	18	23	4	8
15	0.037962	0.550765	15	13	19	11
16	0.039167	0.486303	16	15	8	10
17	0.042019	0.596027	23	18	22	3
18	0.040685	0.447574	13	16	5	13
19	0.055044	0.515727	12	25	14	14
20	0.028443	0.490716	11	5	9	15
21	0.029302	0.530094	10	9	17	16
22	0.028602	0.523688	9	7	16	17
23	0.028398	0.554639	8	4	21	18
24	0.036674	0.534648	7	11	18	19

Рис. 1. Результаты ранжирования моделями Bert и GPT на основании косинусного расстояния

Далее для оценки результатов были рассчитаны коэффициент корреляции Спирмена (Таблица 1).

Таблица 1. Корреляция Спирмена для близостей полученных с помощью моделей с матрицы известным ответом

model	ρ	p-value
BERT	-0.48	0.015
GPT	0.56	0.0036

По результатам проведенной оценки (Рис. 1) стоит отметить, что у BERT косинусные расстояния сконцентрированы в районе 0.04, тогда как GPT рассчитывает косинусные расстояния на интервале [0.09, 0.6].

Можем сделать вывод о более высоком качестве векторных представлений, полученных с помощью модели GPT и выше способности распознавать семантические особенности текста по сравнению с моделью BERT.

Вывод

На основании проведенного анализа работ по построению векторных представлений в области подбора персонала и проведенном эксперименте по сравнение близостей полученных векторных представлений резюме и представлений вакансий с известным ответом делается вывод, что для построения векторных представлений вакансий, резюме максимальное качество векторных представлений достигается с использованием модели GPT (text-embedding-3-small). Выбранная модель также используется для построения векторных представлений навыков, извлеченных из вакансий и резюме.

III. Анализ построенных векторных представлений вакансий, резюме и навыков

Раздел посвящён анализу векторных представлений вакансий, резюме и навыков. Используя методы статистического анализа и методы понижения размерности, проверяется гипотеза о том, что векторные представления вакансий и резюме занимают единое векторное пространство. Также проверяется гипотеза о том, что векторные представления навыков, извлечённых из вакансий, и векторные представления навыков, извлечённых из резюме, занимают единое векторное пространство. Проверка обеих гипотез важна для установления соответствия между вакансиями и резюме и совместимости между навыками, которых ждут работодатели, и теми, которые представлены соискателями.

A. Гипотеза единого векторного пространства для вакансий и резюме

Гипотеза единого векторного пространства для вакансий и резюме заключается в том, что векторные представления вакансий и резюме находятся в едином векторном пространстве. Для подтверждения гипотезы необходимо проверить, находятся ли векторные представления полных текстов резюме и вакансий в едином векторном пространстве. В эксперименте 1 для проверки гипотезы применяется метод MMD, а в эксперименте 2 методы Ivis и t-SNE для уменьшения размерности векторных представлений полных текстов резюме и вакансий для визуального подтверждения гипотезы. Векторные представления получены с использованием модели GPT (text-embedding-3-small), которая выбрана в результате обзора подходов к построению векторных представлений вакансий и резюме.

1) Описание данных

Были собраны наборы вакансий и резюме посредством API портала hh.ru. Всего было собрано 31000 резюме и 18000 вакансий по рабочим специальностям.

Из собранных наборов вакансий и резюме были извлечены описания навыков при помощи LLM модели от GPT и библиотеки Langchain [24, 25].

2) Методы анализа данных

Экспериментальное исследование включает в себя восемь экспериментов для проверки единства векторных пространств: два эксперимента для векторных представлений навыков из резюме и векторных представлений навыков из вакансий, два эксперимента для векторных представлений полных текстов резюме и вакансий, два для проверки векторных представлений слов и два для проверки навыков ESCO.

В исследовании применено два метода подтверждения гипотезы единства векторных пространств: статистический тест MMD и методы понижения размерности t-NSE и Ivis. Далее приведено краткое описание рассматриваемых методов.

Maximum Mean Discrepancy (MMD)

Максимальная средняя дисперсия (MMD) — статистический тест, используемый для определения, являются ли данные два набора выборок, принадлежащих одному распределению [10].

Для оценки значения MMD [6,7] для двух выборок, используется уравнение (1), также вычисляется значение p-value, соответствующее расчетному значению MMD. Если значение p-value ниже заданного порога α, то две выборки относятся к одному и тому же распределению.

$$MMD(P, Q, F) = \sup_{f \in F} \left| E[f(X)] - E[f(Y)] \right| \quad (1)$$

, где P, Q - два проверяемых распределения, F — множество функций отображения признаков

Стоит отметить, что в последнее десятилетие MMD широко применяется для обнаружения сходства распределений в реальных данных, включая данные физики высоких энергий, амплитудно-модулированные сигналы [14] и наборы данных изображений, например, MNIST и CIFAR-10.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE) — нелинейный алгоритм уменьшения размерности, используемый для исследования многомерных данных [8, 9]. Он отображает многомерные данные в двух- или

трехмерное пространство, которое подходит для визуального анализа. t-SNE преобразует евклидовы расстояния между точками в условные вероятности, которые представляют меру сходства между распределениями. Затем алгоритм моделирует похожие многомерные объекты по близлежащим точкам и отличные объекты по удаленным точкам. Расположение этих точек итеративно оптимизируется посредством градиентного спуска, чтобы лучше отражать сходство в пространстве более низкой размерности, сводя к минимуму расхождение между распределением вероятностей в многомерном пространстве и аналогичным распределением на плоскости более низкой размерности.

Ivis

Ivis — это алгоритм снижения размерности, применяемый в основном для визуализации многомерных данных в пространстве низшей размерности с сохранением внутренней структуры данных. Он особенно эффективен для визуализации данных с нелинейными взаимосвязями.

Алгоритм IVIS работает путем итеративной оптимизации функции отображения (f), которая проецирует точки многомерного пространства данных на пространство низшей размерности. В основе лежит архитектура сиамской нейронной сети, которая обучается таким образом, чтобы максимально сохранить локальное сходство между точками данных путем оптимизации триплет функции потерь.

IVIS обычно использует стохастические методы оптимизации, такие как стохастический градиентный спуск (SGD) или Adam, для итеративного обновления параметров функции отображения f с целью минимизации расстояния Кульбака-Лейблера

Подход позволяет Ivis [15, 16] сохранять внутреннюю геометрию исходных данных в пространстве меньшей размерности, что особенно полезно для визуализации и дальнейших задач машинного обучения.

Эксперимент 1: MMD

Для проверки гипотезы единого векторного пространства для полных текстов вакансий и резюме, проведем статистический тест MMD. В данном и последующих экспериментах используется библиотека mmdagg[1], написанная на языке Python для проведения теста MMD. Тест возвращает словарь значений содержащий бинарный флаг принадлежат ли выборки одному распределению или нет и значение p-value. В статистическом тесте флаг 1 означает, что нулевая гипотеза отвергается, флаг 0 — что нулевая гипотеза не отвергается. В данном случае нулевая гипотеза утверждает, что две выборки происходят из одного и того же распределения. Тест возвращает флаг 1, если обнаруживает статистически значимые различия между выборками, и флаг 0, если таких различий не обнаружено. Результаты подтверждались вариацией параметра seed, использовались три значения в каждом эксперименте [42, 420. 4200].

Результаты теста MMD для полных текстов вакансий и резюме представлены в таблице 2.

Таблица 2. Результаты теста MMD для гипотезы единого векторного пространства для вакансий и резюме

Данные	Язык	В одном распределении?	p-value
Полные тексты	Русский	Нет	0.0005
Полные тексты, понижена размерность (t-SNE)	Русский	Нет	0.00005
Полные тексты	Английский	Нет	0.0005
Полные тексты, понижена размерность (t-SNE)	Английский	Нет	0.0005

Результаты эксперимента 1

Тест MMD показал, что расстояние между распределениями в вакансиях и резюме на порядок больше, чем расстояние внутри каждого из распределений. Таким образом, актуален переход от вакансий и резюме к рассмотрению навыков от них.

Эксперимент 2: Ivis и t-SNE

После статистического теста проведена визуализация векторных представлений для визуальной оценки кластеров, на которые они делятся и оценки есть ли пересечение у двух совокупностей или они образуют отдельные группы. Визуализация проведена двумя методами для получения 2d и 3d представлений (Рис.2, 3). Если визуализации показывают значительное перекрытие между векторными представлениями текстов вакансий и резюме, это может дополнительно подтвердить или же опровергнуть достоверность рассматриваемой гипотезы.

[1] https://github.com/antoninschrab/mmdagg

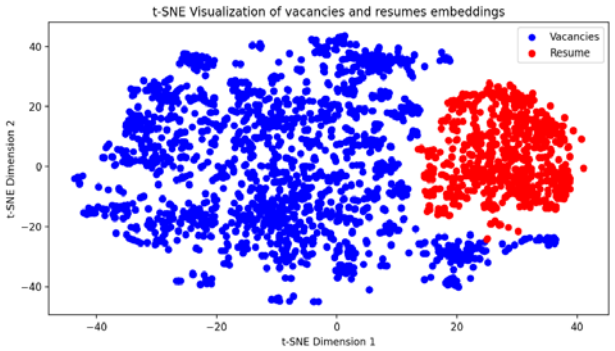

Рис. 2. Результаты визуализации t-SNE для гипотезы единого векторного пространства для вакансий и резюме

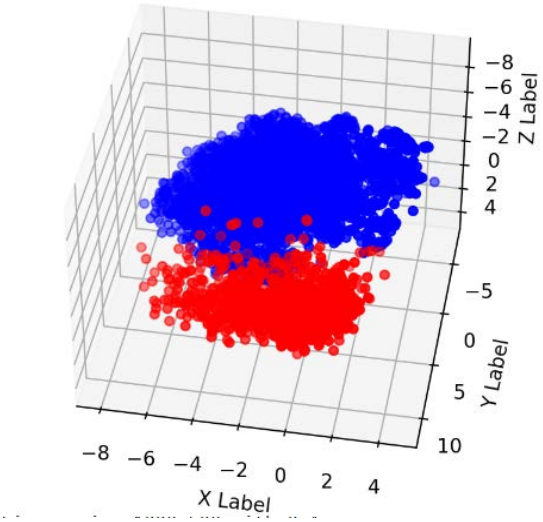

Рис. 3. Результаты визуализации Ivis для гипотезы единого векторного пространства для вакансий и резюме

Результаты эксперимента 2

По результатам визуализации эксперимента 2 делается вывод о наличии разделимости между полными текстами вакансий и резюме на основании визуального анализа 2d и 3d отображений векторных представлений и наличия плоскости разделения между двумя кластерами векторов. Делается вывод, что векторные представления полных текстов резюме и вакансий находятся в разных распределениях. Однако, было замечено, что полные тексты вакансий и резюме имеют разделимость, которая требует дополнительного исследования.

Вывод.

По результатам экспериментов делается вывод о нахождении векторных представлений полных текстов вакансий и резюме в разных распределениях, что подтверждается, как статистически, так и визуально при кластеризации представлений. Соответственно, гипотеза о единстве векторного пространства между векторными представлениями полных текстов вакансий и резюме отвергается.

B. Гипотеза единого векторного пространства для навыков

Основываясь на результатах предыдущих экспериментов, было решено проверить находятся ли в едином векторном пространстве навыки, извлеченные из полных текстов вакансий и резюме, как более короткие и стандартизированные сущности. Для проверки предположения о единстве векторных пространств навыков из резюме и вакансий, необходимо проверить находятся ли векторные представления в едином векторном пространстве для этого проведены эксперименты 3 и 4, аналогичные экспериментам 1 и 2, но для извлеченных из текстов навыков.

Эксперимент 3: MMD

Для второй гипотезы, которая предлагает единое векторное пространство для навыков, проведен статистический тест MMD, результаты представлены в Таблице 3.

Таблица 3. Результаты теста MMD для гипотезы единого векторного пространства для навыков

Данные	Язык	В одном распределении?	p-value
Навыки	Русский	Нет	0.0005
Навыки, понижена размерность (t-SNE)	Русский	Нет	0.00005
Навыки	Английский	Нет	0.0005
Навыки, понижена размерность (t-SNE)	Английский	Нет	0.0005

Результаты эксперимента 3

Тест MMD показал, что расстояние между распределениями извлеченных из текстов вакансий и резюме навыков на порядок больше, чем расстояние внутри каждого из распределений, это следует из результата теста на нахождение в одном распределении. Таким образом, актуален переход от навыков из текстов вакансий и резюме к рассмотрению атомарных структур - слов.

Эксперимент 4: Ivis и t-SNE

На рис. 4 представлена визуализация векторных представлений навыков из вакансий и резюме, выполнено с t-SNE, а на рис. 5 3d представление при помощи Ivis.

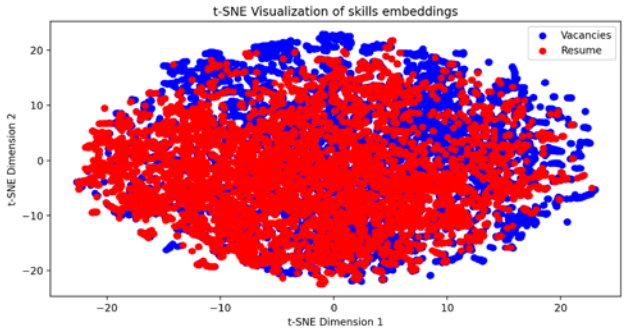

Рис. 4. Результаты визуализации t-SNE для гипотезы единого векторного пространства для навыков

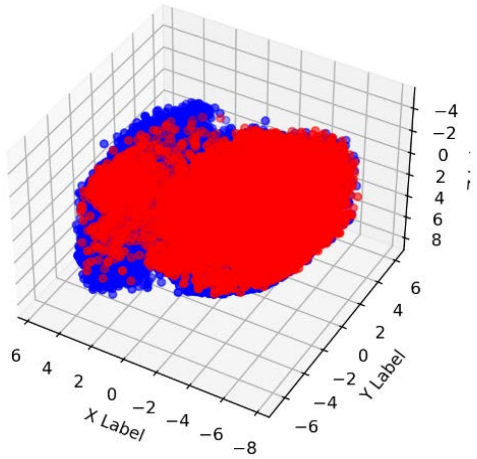

Рис. 5. Результаты визуализации Ivis для гипотезы единого векторного пространства для навыков

Результаты эксперимента 4

По результатам эксперимента с визуализацией векторных представлений делается вывод о наличии перекрытия двух облаков векторов между навыками из полных текстов вакансий и резюме на основании визуального анализа 2d и 3d отображений векторных представлений. Однако, стоит отметить смещение центра плотности одного облака относительно другого.

Выводы

По результатам экспериментов делается вывод о нахождении векторных представлений навыков, извлеченных из текстов вакансий и резюме в разных распределениях. Несмотря на то, что визуально навыки составляют два облака, накладывающихся друг на друга облака, статистический тест не подтверждает единство пространства.

C. Гипотеза единого векторного пространства для слов

Третья гипотеза предполагает, что слова могут быть представлены в едином векторном пространстве.

Эксперимент 5: MMD

Использован также тест MMD для проверки гипотезы, измеряя среднее расстояние между векторными представлениями наборов слов. Единое векторное пространство будет обозначено низким значением MMD, что предполагает, что векторные представления слов близки друг к другу в многомерном пространстве. Результаты представлены в Таблице 4.

Таблица 4. Результаты теста MMD для гипотезы единого векторного пространства для слов

Данные	Язык	В одном распределении?	p-value
Слова	Русский	Да	0.36
Слова, понижена размерность (t-SNE)	Русский	Нет	0.00005
Слова	Английский	Да	0.26
Слова, понижена размерность (t-SNE)	Английский	Нет	0.0005

Результаты эксперимента 5

По результатам эксперимента 5 делается вывод о единстве пространств векторных представлений слов.

Эксперимент 6: Ivis и t-SNE

Эксперимент направлен на визуализацию векторных представлений слов и показывает, группируются ли векторные представления слов вместе или имеют границу разделений.

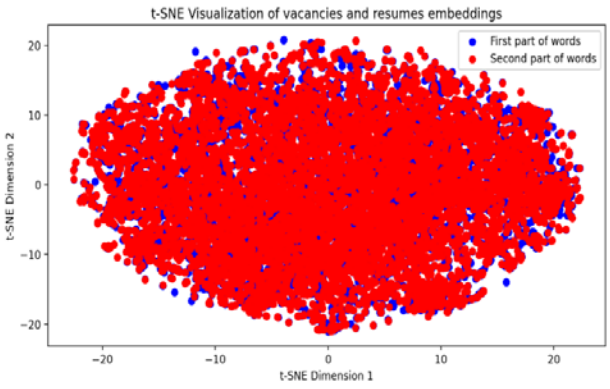

Рис. 6. Результаты визуализации t-SNE для гипотезы единого векторного пространства для слов.

Результат кластеризации с помощью t-SNE (Рис. 7) показывает, что слова образуют два накладывающихся друг на друга облака. Аналогично подтверждает гипотезу визуализация Ivis (Рис. 8).

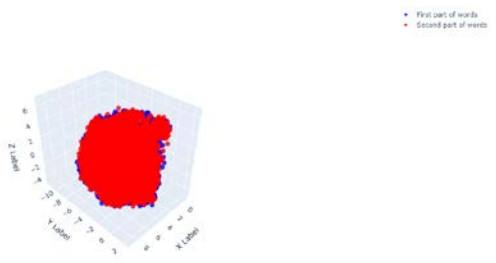

Рис. 7. Результаты визуализации Ivis для гипотезы единого векторного пространства для слов

Результаты эксперимента 6

По результатам эксперимента с визуализацией векторных представлений делается вывод о наличии перекрытия двух облаков векторов слов на основании визуального анализа 2d и 3d отображений векторных представлений.

Вывод

По результатам экспериментов делается вывод о нахождении векторных представлений слов в одном распределении, основываясь на статистически значимом результате теста MMD и результатов визуализации.

D. Гипотеза о единстве пространства навыков ESCO

Дополнительно протестировано нахождение навыков ESCO (Европейская база навыков, компетенций и квалификаций) в едином векторном пространстве.

Эксперимент 7: MMD

Результат статистического теста MMD представлен в таблице 5.

Таблица 5. Результаты теста MMD для гипотезы единого векторного пространства для навыков ESCO

Данные	Язык	В одном распределении?	p-value
Навыки ESCO	Русский	Да	0.9
Навыки ESCO, понижена размерность (t-SNE)	Русский	Нет	0.00005
Слова	Английский	Да	0.9
Слова, понижена размерность (t-SNE)	Английский	Нет	0.0005

По результатам делается вывод о том, что навыки ESCO находятся в едином векторном пространстве.

Результаты эксперимента 7

Тест MMD показал, что расстояние между распределениями навыков ESCO близко и сравнимо с расстоянием внутри каждого из распределений. По результатам эксперимента 7 можно сделать вывод о единстве пространств векторных представлений навыков ESCO.

Эксперимент 8: Ivis и t-SNE

Для подтверждения вывода проведены визуальные эксперименты t-SNE (Рис. 8) и Ivis (Рис. 9).

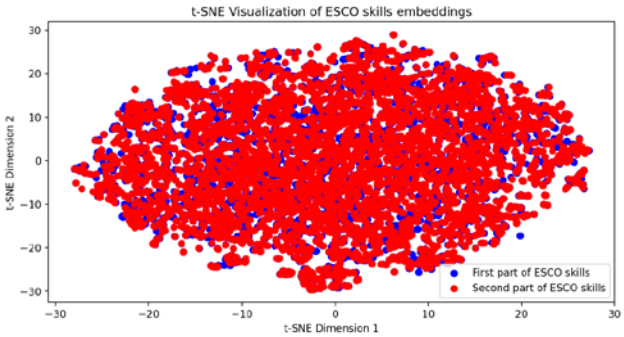

Рис. 8. Результаты визуализации t-SNE для гипотезы единого векторного пространства для навыков ESCO

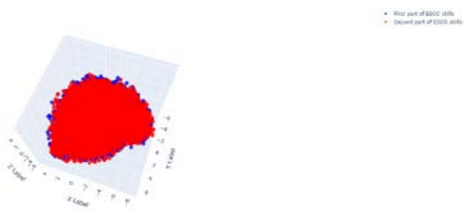

Рис. 9. Результаты визуализации Ivis для гипотезы единого векторного пространства для навыков ESCO

Результаты эксперимента 8

По результатам эксперимента с визуализацией векторных представлений делается вывод о наличии перекрытия двух облаков векторов навыков ESCO на основании визуального анализа 2d и 3d отображений векторных представлений.

Вывод

По результатам экспериментов делается вывод о наличии перекрытия двух облаков векторов навыков ESCO на основании визуального анализа 2d и 3d отображений векторных представлений.

IV. Анализ результатов экспериментов про вакансии и резюме

На основании экспериментов 1 и 2 вынесено предположение, что полные тексты вакансий и резюме

имеют плоскость разделимости и разделимы на основании семантического несходства. Для подтверждения гипотезы необходимо проверить несколько предположений:

1. Если есть разделимость по каждой из компонент вектора, то есть разделимость по вектору целиком
2. Выполнение требования близости по расстоянию Левенштейна
3. Если выполняется требование семантической близости, то есть разделимость по смыслу

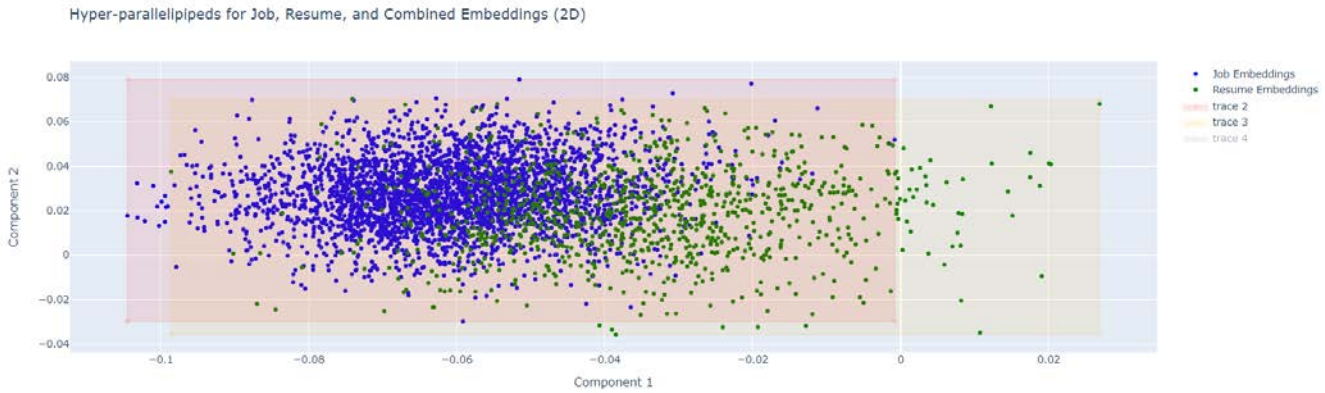

Рис. 10. 2d визуализация гиперпараллелепипеда макс. и мин. компонент векторов текстов вакансий и резюме

На Рис. 10 построена некоторая выпуклая оболочка (которую задают неравенства) компонент векторов. Области резюме и вакансий пересекаются, однако не полностью и имеют долю непересекающихся точек. На основании Рис. 10 делается вывод о нестрогой разделимости векторных представлений вакансий и резюме.

Требование близости по расстоянию Левенштейна

Для проверки второго предположения взят один текст эталонного резюме, его векторное представление и список из произвольных слов. В цикле производилась замена в исходном тексте по одному из произвольных слов и измерялось косинусное расстояние между модифицированным текстом и измененном.

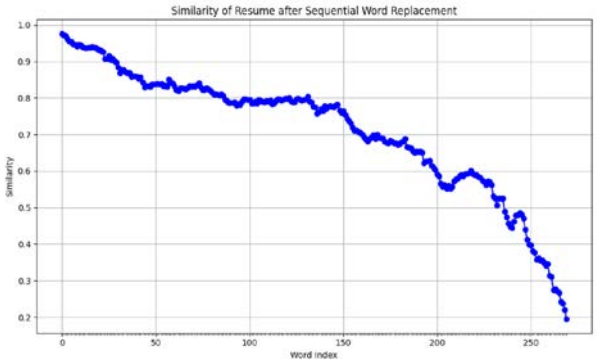

Рис. 11. Результаты измерения косинусного сходства векторных представлений резюме с постепенной заменой слов на произвольные

На основании графика (Рис. 11) можно сделать вывод о гладкости векторного пространства, чем на большее число слов отличается текст от исходного, тем меньше косинус угла между ним и векторным представлением эталонного текста.

Требование семантической близости

Для проверки третьего предположения о семантической близости вакансий и резюме рассчитано попарное косинусное расстояние между всем корпусом векторных представлений вакансий и резюме и каждого резюме со всем корпусом резюме. Далее рассчитана разница между максимальным расстоянием для пар вакансия-резюме и резюме-резюме (Таблица 6).

Таблица 6. Результаты min/max косинусного расстояния при добавлении текста к корпусу

	Вакансии		Резюме	
	Min	Max	Min	Max
Добавление вакансии	0.53	0.87	0.31	0.66
Добавление резюме	0.33	0.64	0.34	0.76

Проверив для всех резюме, делается вывод, что резюме больше похожи друг на друга, чем на вакансии. Такая ситуация называется разделимостью по модальностям, а модальность - класс объектов, для которого выполнено требование семантической близости. Исходя из этого можно определить единую модальность – “найм” и две подмодальности – “вакансии” и “резюме”.

V. Анализ результатов экспериментов про навыки, слова и навыки ESCO

По результатам экспериментов 7 и 8 проведена аналитика по особенностям навыков ESCO и причинах их

нахождении в одном распределении.

Распределение длин навыков по количеству слов

Первым шагом решено было проверить длину фраз в навыках ESCO и навыках извлеченных из текстов вакансий и резюме.

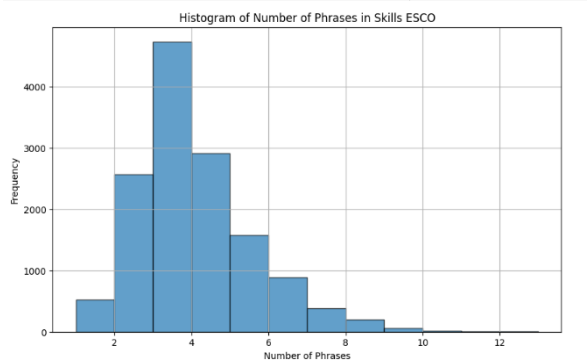

Рис. 12. Гистограмма распределения количества фраз в навыках ESCO

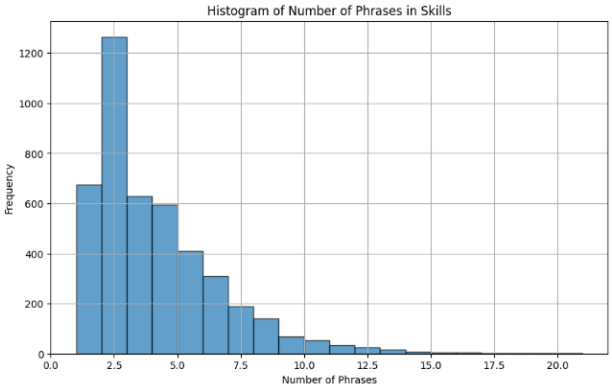

Рис. 13. Гистограмма распределения количества фраз в навыках из текстов вакансий и резюме

По гистограммам (Рис. 12, 13) видно, что как в навыках ESCO, так и навыках из резюме/вакансий количество слов в фразе примерно равно 3. Соответственно можно сделать вывод о соответствии длины фраз извлеченных навыков и ESCO.

Отношение средних длин по количеству слов для извлеченных навыков и ESCO

Дополнительно для подтверждения результатов гистограммы были рассчитаны отношения длин извлеченных навыков и ESCO.

Таблица 7. Отношение средних длин по количеству слов для извлеченных навыков и ESCO.

Англ.	ESCO	Навыки	Слова
ESCO	x	3,6/3,6	3,6/1
Навыки	x	x	3,6/3,6
Слова	x	x	x
Рус.			
ESCO	x	3,65/3,18	3,65/1
Навыки	x	x	3,18/1
Слова	x	x	x

Результаты также подтверждаются (Таблица 7) и также свидетельствуют о том, что навыки ESCO и извлеченные навыки имеют примерно одинаковую длину. Поэтому схожесть по средней длине не является причиной того, что ESCO находятся в едином распределении, а извлеченные навыки не находятся в едином распределении.

Вывод

Одним из возможных объяснений полученных результатов является то, что слова составляют собственное единое подпространство, а навыки являются составными объектами в этом подпространстве. Поэтому слова образуют единое подпространство, а навыки не образуют. В то же время ESCO составляют иное единое подпространство, в котором являются элементами, а не составными объектами. Об этом свидетельствует тот факт, что слова навыки и ESCO попарно не лежат в едином распределении.

Другим возможным объяснением может быть то, что навыки ESCO, будучи стандартизированными, образуют однородное векторное пространство. В то время как навыки, извлечённые из реальных текстов резюме и вакансий, демонстрируют большую вариативность и контекстуальные различия, что приводит к их распределению в различные векторные пространства.

VI. Заключение

В ходе исследования, применяя большие языковые модели, была решена задача построения векторных представлений вакансий, резюме и навыков с использованием LLM (GPT, BERT). Далее по результатам эксперимента была выявлена наибольшее качество векторных представлений, полученных с помощью модели GPT.

Проведенный анализ векторных представлений вакансий и резюме позволяет сделать вывод, что векторные представления полных текстов вакансий и резюме не находятся в едином векторном пространстве. Только атомарные объекты находятся в одном распределении. Представление текста - функция от слов и само представление не является элементом векторного пространства, а агрегатом от векторов слов.

Анализ векторных представлений навыков ESCO позволяет сделать вывод о том, что векторные представления занимают единое векторное пространство. Вывод важен для определения совместимости между навыками, которых ждут работодатели, и теми, которые представлены соискателями.

Дополнительно в ходе исследования был сделан вывод о том, что понижение размерности не сохраняет принадлежность к одному распределению.

Выводы, полученные в результате исследования, могут служить основой для дальнейшей работы по решению задачи сопоставления вакансий и резюме.

Библиография

[1] J. Malinowski, T. Keim, and O. Wendt, "Matching People and Jobs: A Bilateral Recommendation Approach," 2006.

[2] R. Kessler, F. Béchet, M. Roche, and J.-M. Torres-Moreno, "Automatic extraction of skills from resumes: The ESCO experience," in *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 2013.

[3] P. Nikolaev, V. Rangarajan Sridhar, and P. B. Bogen, "BERT for Matching Resumes and Job Descriptions."

[4] T. Brown *et al*., "Language Models are Few-Shot Learners," 2020.

[5] H. Aguinis, J. R. Beltran, A. Cope, "How to use generative AI as a human resource management assistant", *Organizational Dynamics*, Volume 53, Issue 1, 2024, doi: 10.1016/j.orgdyn.2024.101029

[6] S. Panda, C. Shen, R. Perry, J. Zorn, A. Lutz, C. E. Priebe, and J. T. Vogelstein, "Nonpar MANOVA via Independence Testing," *arXiv preprint* arXiv:1910.08883 [cs, stat], April 2021.

[7] C. Shen and J. T. Vogelstein, "The exact equivalence of distance and kernel methods in hypothesis testing," *AStA Advances in Statistical Analysis*, September 2020. doi:10.1007/s10182-020-00378-1.

[8] A. C. Belkina *et al.*, "Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets," *Nature Communications*, vol. 10, no. 1, pp. 1–12, 2019.

[9] L. J. P. van der Maaten and G. E. Hinton, "Visualizing High-Dimensional Data Using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[10] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A Kernel Two-Sample Test," *Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.

[11] L. Van-Duyet, V. Minh-Quan, and D. Quang-An, "Skill2vec: Machine Learning Approaches for Determining the Relevant Skill from Job Description," 2019.

[12] S. Pudasaini, "Scoring of Resume and Job Description Using Word2vec and Matching Them Using Gale–Shapley Algorithm," 2021.

[13] C. M. Jaramillo, "Word embedding for job market spatial representation: tracking changes and predicting skills demand," 2020.

[14] W. Jitkrittum, Z. Szabo, K. Chwialkowski, and A. Gretton, "Interpretable Distribution Features with Maximum Testing Power," 2016.

[15] L. Bulej, "IVIS: Highly customizable framework for visualization and processing of IoT data," 2020. DOI: 10.1109/SEAA51224.2020.00095.

[16] H. S. Lee and C. Wallraven, "Visualizing the embedding space to explain the effect of knowledge distillation," *arXiv preprint* arXiv:2110.04483, 2021. https://doi.org/10.48550/arXiv.2110.04483.

[17] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. DOI: 10.3115/v1/D14-1162.

[18] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: a review," *Artificial Intelligence Review*, vol. 56, pp. 10345–10425, 2023. https://doi.org/10.1007/s10462-023-10419-1.

[19] S. J. Johnson, M. R. Murty, and I. Navakanth, "A detailed review on word embedding techniques with emphasis on word2vec," *Multimedia Tools and Applications*, vol. 83, pp. 37979–38007, 2024. https://doi.org/10.1007/s11042-023-17007-z.

[20] L. Shamir, "Automatic identification of rank correlation between image sequences," *International Journal of Data Science and Analytics*, vol. 17, pp. 1–11, 2024. https://doi.org/10.1007/s41060-023-00450-4.

[21] R. Chattamvelli, "Rank Correlation," in *Correlation in Engineering and the Applied Sciences*, Synthesis Lectures on Mathematics & Statistics. Springer, Cham, 2024. https://doi.org/10.1007/978-3-031-51015-1_3.

[22] G. Briganti, "How ChatGPT works: a mini review," *European Archives of Oto-Rhino-Laryngology*, vol. 281, pp. 1565–1569, 2024. https://doi.org/10.1007/s00405-023-08337-7.

[23] F. D. Souza and J. B. de O. e S. Filho, "Embedding generation for text classification of Brazilian Portuguese user reviews: from bag-of-words to transformers," *Neural Computing and Applications*, vol. 35, pp. 9393–9406, 2023. https://doi.org/10.1007/s00521-022-08068-6.

[24] S. Ott, K. Hebenstreit, V. Liévin *et al.*, "ThoughtSource: A central hub for large language model reasoning data," *Scientific Data*, vol. 10, p. 528, 2023. https://doi.org/10.1038/s41597-023-02433-3.

[25] L. Wang, C. Ma, X. Feng *et al.*, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, p. 186345, 2024. https://doi.org/10.1007/s11704-024-40231-1.

Alignment of Job and Resume Vector Representations with LLM

Lyubov Komarova, Alexey Kolosov, Vladimir Soloviev

***Abstract*—This study presents modern natural language processing (NLP) approaches for analyzing the alignment between job descriptions and resumes. The research focuses on using Large Language Models (LLMs) to create vector representations of job descriptions, resumes, and the skills extracted from them. It demonstrates that vector representations of job descriptions, resumes, and their extracted skills occupy distinct vector spaces, while standardized ESCO skills and words exist in a unified vector space. To test the hypotheses regarding the unity of vector spaces were utilized, statistical method - Maximum Mean Discrepancy (MMD) and dimensionality reduction algorithms (t-SNE and Ivis), to visualize of vector distribution and analyze. The study also provides an in-depth analysis of experimental results, with special attention to the properties of ESCO skills, which form a cohesive vector space due to their standardization. The findings of this research can improve recruitment processes by offering innovative methods for matching candidate skills with employer requirements. Additionally, the study highlights the importance of data standardization in facilitating accurate interpretation and alignment.**

***Keywords*— Large Language Models, vector representations, MMD, t-SNE, Ivis**

REFERENCES

[1] J. Malinowski, T. Keim, and O. Wendt, "Matching People and Jobs: A Bilateral Recommendation Approach," 2006.

[2] R. Kessler, F. Béchet, M. Roche, and J.-M. Torres-Moreno, "Automatic extraction of skills from resumes: The ESCO experience," in *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 2013.

[3] P. Nikolaev, V. Rangarajan Sridhar, and P. B. Bogen, "BERT for Matching Resumes and Job Descriptions."

[4] T. Brown *et al.*, "Language Models are Few-Shot Learners," 2020.

[5] H. Aguinis, J. R. Beltran, A. Cope, "How to use generative AI as a human resource management assistant", *Organizational Dynamics*, Volume 53, Issue 1, 2024, doi: 10.1016/j.orgdyn.2024.101029

[6] S. Panda, C. Shen, R. Perry, J. Zorn, A. Lutz, C. E. Priebe, and J. T. Vogelstein, "Nonpar MANOVA via Independence Testing," *arXiv preprint* arXiv:1910.08883 [cs, stat], April 2021.

[7] C. Shen and J. T. Vogelstein, "The exact equivalence of distance and kernel methods in hypothesis testing," *AStA Advances in Statistical Analysis*, September 2020. doi:10.1007/s10182-020-00378-1.

[8] A. C. Belkina *et al.*, "Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets," *Nature Communications*, vol. 10, no. 1, pp. 1–12, 2019.

[9] L. J. P. van der Maaten and G. E. Hinton, "Visualizing High-Dimensional Data Using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[10] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A Kernel Two-Sample Test," *Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.

[11] L. Van-Duyet, V. Minh-Quan, and D. Quang-An, "Skill2vec: Machine Learning Approaches for Determining the Relevant Skill from Job Description," 2019.

[12] S. Pudasaini, "Scoring of Resume and Job Description Using Word2vec and Matching Them Using Gale–Shapley Algorithm," 2021.

[13] C. M. Jaramillo, "Word embedding for job market spatial representation: tracking changes and predicting skills demand," 2020.

[14] W. Jitkrittum, Z. Szabo, K. Chwialkowski, and A. Gretton, "Interpretable Distribution Features with Maximum Testing Power," 2016.

[15] L. Bulej, "IVIS: Highly customizable framework for visualization and processing of IoT data," 2020. DOI: 10.1109/SEAA51224.2020.00095.

[16] H. S. Lee and C. Wallraven, "Visualizing the embedding space to explain the effect of knowledge distillation," *arXiv preprint* arXiv:2110.04483, 2021. https://doi.org/10.48550/arXiv.2110.04483.

[17] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. DOI: 10.3115/v1/D14-1162.

[18] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: a review," *Artificial Intelligence Review*, vol. 56, pp. 10345–10425, 2023. https://doi.org/10.1007/s10462-023-10419-1.

[19] S. J. Johnson, M. R. Murty, and I. Navakanth, "A detailed review on word embedding techniques with emphasis on word2vec," *Multimedia Tools and Applications*, vol. 83, pp. 37979–38007, 2024. https://doi.org/10.1007/s11042-023-17007-z.

[20] L. Shamir, "Automatic identification of rank correlation between image sequences," *International Journal of Data Science and Analytics*, vol. 17, pp. 1–11, 2024. https://doi.org/10.1007/s41060-023-00450-4.

[21] R. Chattamvelli, "Rank Correlation," in *Correlation in Engineering and the Applied Sciences*, Synthesis Lectures on Mathematics & Statistics. Springer, Cham, 2024. https://doi.org/10.1007/978-3-031-51015-1_3.

[22] G. Briganti, "How ChatGPT works: a mini review," *European Archives of Oto-Rhino-Laryngology*, vol. 281, pp. 1565–1569, 2024. https://doi.org/10.1007/s00405-023-08337-7.

[23] F. D. Souza and J. B. de O. e S. Filho, "Embedding generation for text classification of Brazilian Portuguese user reviews: from bag-of-words to transformers," *Neural Computing and Applications*, vol. 35, pp. 9393–9406, 2023. https://doi.org/10.1007/s00521-022-08068-6.

[24] S. Ott, K. Hebenstreit, V. Liévin *et al.*, "ThoughtSource: A central hub for large language model reasoning data," *Scientific Data*, vol. 10, p. 528, 2023. https://doi.org/10.1038/s41597-023-02433-3.

[25] L. Wang, C. Ma, X. Feng *et al.*, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, p. 186345, 2024. https://doi.org/10.1007/s11704-024-40231-1