Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Челябинский государственный университет»

На правах рукописи

НИКОЛАЕВ Иван Евгеньевич

МЕТОДЫ И АЛГОРИТМЫ ИНТЕЛЛЕКТУАЛЬНОЙ ПОДДЕРЖКИ ФОРМИРОВАНИЯ ТРЕБОВАНИЙ ВАКАНСИИ НА ОСНОВЕ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ ЯЗЫКА И АКТУАЛЬНЫХ ТРЕБОВАНИЙ РЫНКА ТРУДА

Специальность 2.3.4. Управление в организационных системах

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
доктор технических наук, профессор
Мельников Андрей Витальевич

Челябинск – 2024

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. Современный рынок труда характеризуется высокой динамикой и скоростью изменения требований к компетенциям работников. Формирование точного и полного списка требований к потенциальному работнику является одной из важнейших задач в процессе подбора персонала. Специалисты по подбору персонала сталкиваются со множеством проблем при составлении требований к кандидатам на вакантные должности. Во-первых, рынок труда очень динамичен и требования к соискателям быстро меняются, что затрудняет прогнозирование и определение актуальных требований к кандидатам на фоне быстрого устаревания знаний и навыков. Во-вторых, у самих рекрутеров часто недостает информации и опыта в конкретной сфере для понимания реальных потребностей компании. В-третьих, исследование и анализ требований требуют значительных временных и денежных затрат. В-четвертых, возникают сложности с формулировкой конкретных, объективных и адекватных требований, а также с их согласованием с заинтересованными лицами. Таким образом, процесс определения требований к кандидатам является одним из наиболее проблемных этапов в работе специалистов по подбору персонала.

Чтобы эффективно подбирать персонал, необходим постоянный мониторинг рынка труда и детальный анализ вакансий в разрезе отдельных компетенций. Однако возможности рекрутеров по сбору и обработке больших объемов соответствующих данных ограничены. Одним из решений может стать разработка интеллектуальных рекомендательных систем, которые позволят автоматизировать сбор и анализ информации о востребованных на рынке компетенциях. Такие системы способны выявлять важные закономерности и тенденции, помогать формулировать обоснованные требования к кандидатам с учетом специфики компании и отрасли. Разработка методов интеллектуальной поддержки формирования требований к вакансиям с использованием нейросетевых моделей языка позволит автоматизировать и оптимизировать процесс подбора персонала, сократит издержки

на поиск неподходящих кандидатов. Создание подобных интеллектуальных систем является актуальной задачей для всей сферы HR (англ. HR – human resources).

Тема исследования тесно связана с приоритетами государственной политики в области развития искусственного интеллекта и цифровых технологий в Российской Федерации. Разработка интеллектуальных методов анализа требований к вакансиям с использованием нейросетевого моделирования соответствует целям национальной стратегии развития искусственного интеллекта на период до 2030 года в части создания конкурентоспособных отечественных решений в сфере ИИ.

Результаты исследования могут быть использованы для решения задач цифровизации рынка труда и повышение эффективности подбора персонала. Эти задачи были выделены в рамках национальных инициатив, включая проекты «Цифровая экономика», «Кадры для цифровой экономики» в контексте национальной программы «Цифровая экономика России 2024», а также в контуре Национальной технологической инициативы (НТИ), сфокусированной на обеспечении кадрами для поддержки промышленного развития. Важность этих выводов особенно заметна в аспектах найма и обучения высококвалифицированных специалистов для ключевых и наиболее перспективных секторов.

Данное диссертационное исследование посвящено разработке методов и алгоритмов на основе нейросетевых моделей языка и кластерного анализа для автоматизации и повышения эффективности процесса формирования точного и актуального списка требований проекта вакансии, что позволит повысить качество и скорость процессов поиска и подбора кандидатов на вакансии, а также сократить затраты на подбор персонала. Таким образом, тема является актуальной и имеет большое практическое значение для всех участников рынка труда.

Степень разработанности темы. Методы анализа рынка труда представлены в работах А.П. Цыпина, М.М. Шайлиевой, А.С. Сорокина, И.Б. Хмелева, Е.В. Чекановой, О.П. Толкачевой, А.В. Сапунова, В.В. Гаврилюк, А.А. Ташпулатова, Ю.А. Шварца, А.И. Смирновой, А.В. Пахомова, Е.А. Пахомовой, О.В. Рожковой, И.А. Волошиной, Л.В. Козловой, П.Н. Новикова, П.А. Лебедева, Д.А. Шурыгиной,

А.Д. Волгина, В.Е. Гимпельсона, И.Н. Калиновской, О.А. Хохловой, А.Н. Хохловой, А.Ц. Чойжалсановой, Е.Н. Бавыкиной, С.С. Гущиной, Т.В. Корецкой, В.С. Половинко, Р.А. Долженко, С.Б. Долженко, А.О. Епанчинцева, Д.С. Ботова, В.А. Малых, R. Boselli, F. Mercorio, M. Mezzanzanica, A. Giabell, E. Colombo, I. Konstantinidis, M. Zhang, K.N. Jensen и других исследователей.

В существующих методах анализа рынка труда предложено множество подходов, но эти решения имеют ряд ограничений.

Большинство исследований рынка труда связаны с анализом макростатистики и макропоказателей, и направлены на комплексный анализ спроса и предложения рабочей силы для обеспечения эффективного функционирования рынка труда. Данные исследования не позволяют проанализировать спрос и предложение на уровне отдельных сущностей требований, знаний и навыков/компетенций.

В тех исследованиях, которые пытаются анализировать рынок труда на уровне анализа текстов вакансий, как правило, используются модели дистрибутивной семантики прошлого поколения, такие как TF-IDF, word2vec.

В последнее время наиболее востребованным направлением для анализа текстовой информации являются нейросетевые модели на архитектуре трансформеров с механизмом внимания. Отличительной особенностью таких моделей является контекстуальное понимание всех слов в тексте, что позволяет ей лучше распознавать долгосрочные зависимости между ними. В настоящее время такие модели достигли лучших оценок на большинстве задач обработки и понимания естественного языка, чем модели предыдущих поколений.

Данное диссертационное исследование посвящено решению актуальной задачи интеллектуальной поддержки процесса формирования точного и актуального списка требований на уровне отдельных сущностей знаний и навыков/компетенций при составлении вакансии на рынке труда. Проведенный анализ существующих публикаций и практических исследований показал, что предложенный набор методов и алгоритмов, основанных на интеллектуальном анализе текстов вакансий, ранее не упоминался и не использовалась в контексте задачи интеллектуальной поддержки

составления списка требований для вакансий. Таким образом, тема данной диссертационной работы, связанная с разработкой методов и алгоритмов для автоматизации анализа реальных (фактических) данных о требованиях рынка труда, является актуальной.

Цель и задачи диссертационной работы. Целью работы является разработка методов и алгоритмов интеллектуальной поддержки процедур (процесса) формирования требований к вакансиям, которые обеспечат повышение качества анализа современных тенденций рынка труда, повысят эффективность процессов подбора персонала и соответствие разрабатываемых требований в проектах вакансий реальным потребностям рынка труда.

Для достижения указанной цели необходимо было решить следующие задачи:

1. Разработать модель формализованного описания требований реального рынка труда на уровне отдельных сущностей знаний и навыков/компетенций, учитывающую структурные и семантические отношения между ними.
2. Разработать метод и алгоритм извлечения сущностей знаний и навыков/компетенций из текстов требований вакансий реального рынка труда на основе нейросетевых моделей языка и методов классификации.
3. Разработать интеллектуальный метод поддержки формирования списка требований вакансий на основе семантического сопоставления сущностей знаний и навыков/компетенций предложенной структурно-семантической модели и применения методов кластеризации.
4. Выполнить программную реализацию предложенных методов, моделей и алгоритмов в виде прототипа интеллектуальной рекомендательной системы информационной поддержки формирования требований вакансии.
5. Провести оценку эффективности реализованных методов и алгоритмов интеллектуальной поддержки формирования требований на текстовом корпусе проектов вакансий.

Объектом исследования является процесс формирования требований в проекте вакансий, включая формулирование и корректировку разрабатываемых требований в соответствии с потребностями реального рынка труда.

Предметом исследования являются методы и алгоритмы интеллектуальной поддержки формирования требований в проектах вакансий.

Научная новизна диссертационной работы заключается в следующем:

– Предложена модель формализованного описания требований рынка труда на уровне отдельных сущностей знаний и навыков/компетенций в виде структурно-семантической модели, которая позволяет учитывать структурные и семантические отношения между сущностями. Такое формализованное представление сущностей требований рынка труда может быть сформировано автоматически, в отличие от существующих работ, требующих ручных методов разработки онтологических моделей предметной области (п. 2 п.с.).

– Разработан метод извлечения сущностей знаний и навыков/компетенций из текстов требований вакансий реального рынка труда, на основе нейросетевых моделей языка и методов классификации, который в отличие от существующих методов не требует, чтобы искомые сущности были представлены в виде последовательности подряд идущих синтаксических или лексических конструкций, и позволяет извлекать существенно больше информации об отдельных сущностях требований из текстов вакансий реального рынка труда (п. 5 п.с.).

– Разработан интеллектуальный метод поддержки формирования требований вакансии на основе семантического сопоставления сущностей знаний и навыков/компетенций предложенной структурно-семантической модели и методов кластеризации, который обеспечивает соответствие разрабатываемых требований в проектах вакансий реальным потребностям рынка труда (п. 9 п.с.).

Теоретическая и практическая значимость работы. Предложенные в работе модели, методы и алгоритмы дают научное обоснование механизмов для интеллектуальной поддержки формирования актуального и точного списка требований в процессе составления текста вакансии и сочетают формальные модели

представления отдельных сущностей требований и отношений между ними с методами семантического анализа текстов на основе нейросетевых моделей языка и кластерного анализа.

Созданный на основе предложенных моделей, методов и алгоритмов прототип интеллектуальной системы поддержки формирования актуального списка требований в процессе составления текста вакансии может быть применен в службах занятости и отделах кадров для формирования более полного и точного списка требований к соискателям. Это позволит решить проблему разрыва между содержанием текста вакансии и требованиями реального рынка труда, а также позволит сократить время и ресурсы на анализ и подготовку итогового списка требований при подготовке текста вакансии.

Методология и методы исследования. Диссертационное исследование базируется на различных методах, включая системный анализ, системное моделирование, теорию принятия решений, искусственный интеллект, инженерию знаний, методы машинного обучения, анализ данных, обработку естественного языка, теорию множеств и графов, разработку интеллектуальных систем, а также объектно-ориентированное проектирование и разработку информационных систем. Эти методы являются теоретической и методологической основой исследования.

Положения, выносимые на защиту:

1. Структурно-семантическая модель описания требований реального рынка труда на уровне отдельных сущностей знаний и навыков/компетенций, включающая формализованное описание сущностей в виде графа с учетом структурных и семантических отношения между сущностями.

2. Метод извлечения сущностей знаний и навыков/компетенций из текстов требований вакансий реального рынка труда, на основе нейросетевых моделей языка и методов классификации.

3. Интеллектуальный метод поддержки формирования требований вакансии на основе семантического сопоставления сущностей знаний и

навыков/компетенций предложенной структурно-семантической модели и методов кластеризации.

4. Оценка эффективности реализованных методов и алгоритмов интеллектуальной поддержки формирования требований на текстовом корпусе проектов вакансий.

Обоснованность и достоверность результатов. Полученные в диссертационном исследовании результаты обоснованы применением проверенных математических методов: теории графов, теории множеств, векторной алгебры, а также статистической обработки данных. Использованы известные теоретические положения в области анализа естественного языка и интеллектуальной поддержки принятия решений.

Адекватность предложенных методов и алгоритмов подтверждена экспериментально на значимых текстовых данных, а также внедрением и апробацией прототипа интеллектуальной рекомендательной системы на реальных предприятиях. Достоверность экспериментальных результатов обеспечена применением устоявшихся подходов к оценке качества информационного поиска, классификации и кластеризации текстов. Также проводилась корректная статистическая обработка данных и анализ высокой согласованности мнений экспертов.

Апробация результатов. Основные положения и результаты работы докладывались и обсуждались на следующих научных конференциях:

– 7-я всероссийская научная конференция с международным участием «Информационные технологии и системы» (ИТиС’2019) (12 – 16 марта 2019 г., Ханты-Мансийск);

– 9-я всероссийская научная конференция с международным участием «Информационные технологии интеллектуальной поддержки принятия решений» (ITIDS’2021) (7 – 9 декабря 2021 г., Уфа);

– Семинар «Методы искусственного интеллекта в решении прикладных задач» (23 – 27 августа 2023 г., Ханты-Мансийск).

Результаты диссертационной работы внедрены в следующих организациях: Челябинский государственный университет (г.Челябинск), Югорский научно-исследовательский институт (г.Ханты-Мансийск), компания ООО Фирма «Интерсвязь» (г.Челябинск), что подтверждается приведенными в приложении актами о внедрении научных положений и разработок диссертации в практику деятельности организации.

Публикации. По результатам диссертационного исследования опубликовано 9 печатных работ, в том числе 3 работы в рецензируемых печатных изданиях, рекомендованных ВАК [17, 15, 16], 5 работ в изданиях, индексируемых в Scopus и Web of Science [92, 91, 43, 112, 51], и 1 публикация в других научных журналах и сборниках трудов конференций [5]. Получено 1 свидетельство о государственной регистрации программы для ЭВМ [22].

Личный вклад автора. Модели и методы интеллектуальной поддержки принятия решений, метод извлечения отдельных сущностей знаний и навыков/компетенций, структура прототипа интеллектуальной системы были предложены, описаны и разработаны автором лично. Персональный вклад автора подробно отражают основные положения, выносимые на защиту, и публикации по теме диссертации.

Структура и объем диссертации. Диссертационная работа состоит из введения, четырех глав, заключения и списка литературы. Объем работы составляет 162 страница основного текста, включая 31 рисунка, 17 таблиц, 4 приложения. Список литературы содержит 118 наименование.

В главе 1 рассматриваются основные понятия и процессы, связанные с процессом подбора персонала в целом, и процессом формирования списка требований при подготовке проекта вакансии, в частности. Проводится обзор существующих подходов и методов анализа рынка труда, систем онлайн-рекрутмента, методов векторного представления текстов в задаче определения семантической близости, а также методов извлечения именованных сущностей, исследуются ключевые ограничения и проблемы. Определяются ключевые проблемы извлечения текстов

отдельных сущностей требований знаний и навыков/компетенций из текстов вакансий.

В главе 2 описывается концепция информационной поддержки формирования требований вакансии на основе семантического сопоставления сущностей структурно-семантической модели и методов кластеризации. Описывается модель формализованного описания требований рынка труда на уровне отдельных сущностей знаний и навыков/компетенций в виде структурно-семантической модели. Проводится подробный обзор нейросетевых моделей на архитектуре трансформеров, а также процесс их дообучения на текстах предметной области. Предлагается метод извлечения текстов отдельных сущностей требований знаний и навыков/компетенций.

В главе 3 описывается интеллектуальный метод поддержки формирования требований вакансии на основе семантического сопоставления сущностей знаний и навыков/компетенций предложенной структурно-семантической модели и методов кластеризации. Метод включает в себя три этапа. Описываются основные датасеты и поэтапное проведение экспериментов каждого этапа. Предлагается методика оценки качества результатов рекомендательной системы.

В главе 4 описывается процесс разработки и апробации прототипа интеллектуальной рекомендательной системы поддержки формирования списка требований к вакансиям на основе предложенных методов и алгоритмов. Описаны функциональные требования и определена общая структура прототипа и структура базы данных. Подробно рассмотрены особенности программной реализации различных модулей системы, а также используемые технологии и библиотеки. Приведены результаты апробации прототипа рекомендательной системы, дана комплексная оценка его эффективности в процессе формирования требований при подготовке вакансии.

ГЛАВА 1 АНАЛИЗ ПРОЦЕССА ПОДБОРА ПЕРСОНАЛА И СУЩЕСТВУЮЩИХ МЕТОДОВ АНАЛИЗА ТРЕБОВАНИЙ РЫНКА ТРУДА

В первой главе рассматривается основные понятия, процедуры и проблемы, связанные с процессом подбора персонала. Проводится обзор существующих подходов и методов анализа рынка труда, систем онлайн-рекрутмента, методов векторного представления текстов в задаче определения семантической близости, а также методов извлечения именованных сущностей, исследуются ключевые ограничения и проблемы. Определяются ключевые проблемы извлечения текстов отдельных сущностей требований знаний и навыков/компетенций из текстов вакансий.

1.1. Основные понятия и процессы подбора персонала

1.1.1. Понятия, термины и определения

Вакансия – это открытая должность или рабочее место в организации, для которой ищут подходящего кандидата.

Текст вакансии – это документ или объявление, опубликованное организацией, содержащее информацию о свободной должности, такую как название должности, обязанности, требования к кандидату, условия труда, информацию о компании и другие связанные с вакансией детали. Этот текст используется для предоставления информации потенциальным кандидатам о вакансии и привлечения подходящих кандидатов для заполнения этой должности. Ключевым разделом вакансии являются **список требований**, который включает в себя необходимые знания, навыки, компетенции и способности, а также профессиональный уровень, которым должен обладать соискатель для того, чтобы успешно справляться с задачами, предоставленными на данной должности.

Следует подчеркнуть, что существует множество трактовок терминов «знание», «навык» и «компетенция», каждая из которых имеет свои характеристики, зависящие от специфики области применения.

Давайте определим «знание», «навык» и «компетенцию» в контексте изучения требований на рынке труда.

Знание – это совокупность теоретических сведений, фактов и информации, которыми должен обладать кандидат для успешного выполнения работы на определенной должности. Знания приобретаются через обучение, чтение, наблюдение и опыт. В требованиях вакансий знания часто указываются как необходимые или желательные в определенных областях, например, знание конкретных законов, принципов, методологий, языков программирования и т.д.

Например, «знание принципов объектно-ориентированного программирования (ООП)». Кандидат должен понимать основные концепции ООП, такие как классы, объекты, наследование, полиморфизм, инкапсуляция. Он должен знать, как применять эти принципы при разработке программного обеспечения, как проектировать архитектуру приложения с использованием ООП и какие преимущества дает этот подход. Знание ООП является фундаментальным требованием для многих вакансий программистов и разработчиков.

Навык определяется как способность использовать знания в реальных условиях для достижения определённых целей. Это умение, развиваемое через практический опыт и обучение. В описании должностей навыки часто указываются как специфические действия, которыми должен владеть соискатель, включая, например, умение программировать на конкретном языке, владение определёнными инструментами и технологиями, умения в области анализа данных, навыки общения и другие.

Например, «навык работы с базами данных SQL». Кандидат должен уметь писать запросы на языке SQL для извлечения, фильтрации, сортировки и агрегации данных из реляционных баз данных. Он должен быть способен создавать и изменять структуру базы данных, определять таблицы, индексы, ключи. Также

важно умение оптимизировать запросы для повышения производительности, работать с транзакциями, обрабатывать ошибки. Навык работы с SQL необходим для вакансий, связанных с разработкой баз данных, анализом данных, созданием отчетов.

Компетенция, более широкое понятие, и определяется как набор знаний, умений, способностей, черт личности и поведенческих паттернов, необходимых для успешного осуществления специфической профессиональной работы. Это понятие включает в себя как умения и знания, так и личностные особенности человека. В объявлениях о работе компетенции зачастую указывают на ожидаемые результаты и профессиональное умение в различных сферах, таких как умение управлять проектами, принимать решения, работать в команде, обладать лидерскими качествами и другое.

Например «компетенция в управлении проектами». Кандидат должен обладать знаниями методологий управления проектами (например, Agile, Scrum, Waterfall), понимать процессы планирования, организации, контроля и завершения проектов. Он должен уметь определять цели и требования проекта, составлять план работ, распределять ресурсы, отслеживать прогресс, управлять рисками и изменениями. Важны навыки использования инструментов управления проектами, таких как Jira, Trello, MS Project. Кроме того, необходимы развитые коммуникативные и лидерские навыки, умение работать в команде, решать конфликты, мотивировать участников проекта. Компетенция в управлении проектами востребована для руководящих позиций, менеджеров проектов, скрам-мастеров.

Таким образом, при анализе требований вакансий:

– знания указывают на необходимую теоретическую базу и информированность кандидата в определенных сферах;

– навыки описывают конкретные практические умения, которые требуются для выполнения должностных обязанностей;

– компетенции представляют собой комбинацию знаний, навыков и личностных характеристик, необходимых для успешной работы на данной позиции.

Четкое понимание этих трех понятий помогает лучше структурировать и анализировать требования вакансий, оценивать соответствие кандидатов ожиданиям работодателей и принимать обоснованные решения в процессе подбора персонала.

Далее в работе термин *знание* будет использоваться как самостоятельная сущность, а термины *навык,* и *компетенция* будут использоваться как одна сущность *навык/компетенция*, т.к. с точки зрения анализа текстовой информации эти понятия практически невозможно отделить друг от друга.

1.1.2. Анализ процесса подбора персонала

Этапы процесса подбора персонала могут варьироваться в зависимости от конкретной организации и ее потребностей.

В общем случае процесс подбора персонала включает в себя следующие этапы:

1. Анализ потребностей и определение требований к кандидатам

На этом этапе руководитель подразделения самостоятельно или вместе с со специалистом по подбору персонала проводят комплексный анализ, включающий в себя следующие шаги.

Анализ текущих HR-показателей:

– численность персонала по подразделениям;
– текучесть кадров;
– показатели производительности;
– анализ затрат на персонал.

Аудит укомплектованности штата:

– сравнение фактической численности с штатным расписанием;
– выявление вакантных позиций.

Анализ планов и целей бизнеса:

– определение потребности в персонале исходя из планов компании;

– учет открытия новых направлений, проектов.

Анализ текущего состава персонала:

– аудит компетенций сотрудников;

– выявление пробелов для решения стоящих задач;

– собеседование с другими сотрудниками;

– анализ должностных обязанностей будущих кандидатов.

Анализ рынка труда:

– исследование аналогичных вакансий и резюме, связанных с выбранной должностью, чтобы выявить, какие требования и квалификации наиболее востребованы на данный момент.

После проведенного анализа определяются первичный список требований и необходимые качества для вакансии, формируется перечень требующихся должностей и необходимого числа сотрудников.

2. Разработка вакансии

Определение требований к соискателям. На этом этапе специалист по подбору сотрудничает с руководителем подразделения, от которого он получил первоначальный список требований, чтобы получить подробную информацию о требованиях и ответственностях, связанных с вакансией. Он выясняет, какую квалификацию, навыки и опыт требуется от кандидатов, а также какие задачи они будут выполнять в рамках должности, оценивает приоритет требований, проводит работу по уточнению формулировок требований, и редактирует список, если это необходимо в зависимости от конкретной должности или ситуации на рынке труда, расширяет получившийся список требований требованиями к образованию, опыту, знаниям, навыкам и компетенциям соискателей.

Создание привлекательного объявления. Используя полученную информацию, специалист по подбору персонала составляет объявление о вакансии. Очень важно, чтобы в объявлении о вакансии были понятно и полно описаны

информация о компании, название должности, требования к знаниям и навыкам кандидата, его должностные обязанности, квалификационные требования и условия работы и бенефиты.

3. Поиск кандидатов

Размещение объявления. После того как объявление о работе готово, рекрутер определяет самые эффективные пути для привлечения потенциальных сотрудников к этой позиции: публикация на сайтах для поиска работы, через агентства по найму персонала, на веб-сайте компании, а также с помощью профессиональных соцсетей, тематических групп и форумов, использование рекомендаций от текущих работников и профессиональных связей в сфере, не говоря уже об использовании внутренних ресурсов. Основная задача – привлечь лучших кандидатов и обеспечить широкий охват потенциальных претендентов.

Уточнение деталей. Если у кандидатов возникают вопросы относительно вакансии или работы в компании, специалист по подбору персонала отвечает на них, уточняет детали и предоставляет необходимую информацию. Он также может проводить информационные сессии для потенциальных претендентов, расширяя их понимание о компании и вакансии.

Рекрутинговый маркетинг. Важным аспектом на этом этапе является привлечение наиболее подходящих кандидатов. Для этого специалист по подбору персонала может использовать методы рекрутингового маркетинга, такие как активное привлечение, SEO-оптимизацию объявления на сайтах поиска работы, продвижение вакансий в социальных сетях и т. д.

4. Предварительный отбор

По результатам поиска специалист по подбору персонала проводит первичную оценку предоставленных резюме и соответствие кандидатов требованиям вакансии. На этом этапе, в зависимости от того, какой уровень знания и опыта требуется на вакансию, может быть привлечен руководитель структурного подразделения или эксперт для более подробного рассмотрения кандидатов.

5. Собеседование и оценка

Кандидаты, прошедшие предварительный отбор, приглашаются на собеседование. Цель собеседования – более подробно изучить кандидатов, их навыки, опыт работы и возможность приспособиться к организационной культуре. На этом этапе специалиста по подбору персонала могут применять различные методики и форматы интервью, такие как структурированные интервью, случайные собеседования, групповые интервью и пр. Некоторые организации могут также проводить тестирование и оценку компетенций кандидатов на данном этапе. Для их проведения могут быть привлечены руководители, психологи, технические специалисты или внешние агентства.

После собеседования и оценки компетенций кандидата может проводиться проверка рекомендаций и предыдущего опыта работы.

6. Принятие решения о найме

На последнем этапе руководитель подразделения и специалист по подбору персонала обсуждают претендентов и принимают решение о приеме или отклонении кандидатуры. Руководитель подразделения играет важную роль в этом процессе, так как обладает более глубоким пониманием требований и специфики должности.

Руководитель подразделения участвует на этапах анализа потребностей и определения требований к кандидатам, а также на этапе собеседования, оценки компетенций и принятия решения. Специалист по подбору персонала, в свою очередь, преимущественно занимается поиском кандидатов, предварительным отбором, организацией собеседований, проведением оценки и проверкой референсов.

Важно отметить, что процесс подбора персонала может быть длительным и должен быть осуществлен с соблюдением законодательства и принципов равного руководства. Кроме того, в различных организациях могут применяться дополнительные этапы или подходы в зависимости от их особенностей и потребностей.

Можно выделить несколько групп проблем, с которыми могут столкнуться специалисты при формировании требований к вакансиям:

1. Недостаток информации и опыта

– Отсутствие технического опыта в данной сфере работы. Специалисты могут не иметь достаточного технического знания и опыта в определенной области, что затрудняет определение конкретных требований к соискателям.

– Недостаток информации о конкретной должности / профессии / отрасли. Специалистам может быть трудно собрать достаточно информации о конкретной должности / профессии / отрасли, чтобы определить точные требования.

– Непонимание специфики работы и требований в данной отрасли. Отсутствие знаний о специфике работы в конкретной отрасли может затруднить определение соответствующих требований.

– Отсутствие данных о вакансиях и компетенциях на рынке.

2. Скорость и масштабность изменения на рынке труда

– Быстрое изменение требований на рынке труда. Технологии, методы работы и требования в различных отраслях постоянно меняются. Специалистам может быть трудно следить за этими изменениями и корректно обновлять требования к вакансиям.

– Сложность прогнозирования будущих изменений и потребностей компании. Специалистам может быть сложно определить будущие потребности компании и адаптировать требования к вакансии в соответствии с этими потребностями.

– Различия между требованиями на внутренний и внешний рынок труда. Требования к кандидатам могут различаться в зависимости от того, обращается ли компания на внутренний или внешний рынок труда.

3. Ограничения и проблемы взаимодействия

– Недостаток времени для детального анализа и определения требований к вакансиям.

– Ограничения и недостатки в бюджете и ресурсах для исследования рынка и анализа требований к вакансии. Отсутствие финансовых и временных ресурсов может затруднить сбор достаточной информации о рынке труда и требованиях к вакансии.

– Разнообразие работы по привлечению кандидатов и необходимость определения разных требований для каждой должности. Специалисты могут столкнуться с проблемой подготовки различных требований для различных должностей и секторов рынка труда. Разные вакансии требуют разных преследуемых целей и профессионального бэкграунда, поэтому специалистам приходится тщательно анализировать каждую должность и формулировать требования, чтобы соответствовать потребностям компании.

4. Проблемы в формулировке и согласовании требований

– Сложность оценки важности и приоритетности определенных навыков и опыт. Специалист должен определиться, какие квалификации крайне необходимы для выполнения работы на этой должности, а какие навыки могут быть развиты в процессе работы и являются дополнительными. Слишком жесткий список требований может стать барьером для потенциальных кандидатов и уменьшить количество откликов на вакансию, тогда как слишком мягкий список может привести к привлечению кандидатов, которые не соответствуют критериям, необходимым для выполнения работы.

– Проблемы с балансированием между желаемыми требованиями и реальными возможностями соискателей.

– Проблемы с составлением объективных и конкретных требований, чтобы привлечь наиболее подходящих кандидатов.

– Сложность согласования требований с руководителями отделов и другими заинтересованными сторонами. Процесс согласования требований к вакансии может затянуться из-за разногласий или разного представления о наборе требований в вакансии, поскольку разные стороны могут иметь разные представления о требованиях и важности определенных навыков или опыта.

Формулировка требований к вакансиям требует определенного уровня экспертизы и понимания требований компании, отрасли и должности.

Понимание перечисленных проблем поможет специалистам по подбору персонала лучше организовать работу и поможет идентифицировать основные области, в которых требуется уделить больше внимания для успешного формирования требований к вакансиям.

1.2. Обзор существующих методов анализа рынка труда

Существует несколько основных методов, которые используются для различных видов анализа рынка труда: изучения спроса и предложения рабочей силы, динамики занятости, заработной платы и других ключевых показателей. Далее приводится обзор некоторых из наиболее распространенных подходов.

Статистический анализ [32, 33, 26, 21] предполагает использование официальных статистических данных о занятости, безработице, заработной плате, вакансиях и т.д., предоставляемых государственными органами и службами статистики, а также анализ динамики и структуры занятости по отраслям, профессиям, регионам, демографическим характеристикам и расчет ключевых показателей рынка труда, таких как уровень безработицы, коэффициент напряженности, индекс заработной платы и др.

Статистический анализ имеет ряд преимуществ – он позволяет получить объективные количественные данные, отслеживать тенденции и динамику. Однако он требует наличия достоверной и актуальной статистической информации, что не всегда доступно. Кроме того, статистический анализ сам по себе не объясняет причины наблюдаемых явлений. Трудоемкость данного подхода умеренная, так как он в основном базируется на обработке вторичных данных.

Опросы и обследования [8, 23] включают проведение опросов работодателей, работников, соискателей для сбора первичных данных о ситуации на рынке труда, изучение потребностей работодателей в кадрах, требований к квалификации и компетенциям, условий труда и заработной платы, а также анализ трудоустройства

выпускников, их удовлетворенности работой, соответствия полученного образования требованиям рынка.

Опросы и обследования дают возможность глубже понять мотивы, ожидания и проблемы различных участников рынка труда. Они позволяют получить первичную информацию, которая не всегда отражается в статистике. Вместе с тем, опросы требуют больших временных и финансовых затрат на организацию и проведение. Кроме того, их результаты могут быть субъективными и не всегда репрезентативными. Трудоемкость данного подхода высокая.

Эконометрическое моделирование [34, 18, 7] направлено на создание экономико-математических моделей, которые используются для прогнозирования спроса и предложения на рынке труда, а также для анализа воздействия различных факторов (экономических, демографических, технологических) на динамику занятости и заработной платы, разработку сценариев развития рынка труда при различных условиях.

Эконометрическое моделирование дает возможность прогнозировать развитие рынка труда, оценивать влияние различных факторов. Однако построение адекватных моделей требует высокой квалификации исследователей и наличия качественных исходных данных. Кроме того, модели могут не учитывать всех нюансов и взаимосвязей на рынке труда. Трудоемкость данного подхода также высокая.

Анализ вакансий и резюме [12, 6, 10, 31] включает сбор и анализ данных о вакансиях и резюме из открытых источников, анализ требований, предъявляемых работодателями к потенциальным сотрудникам, анализ заработной платы и условия труда, определение наиболее востребованных профессиональных навыков и компетенций, а также анализ специфики спроса на рабочую силу в различных регионах.

Анализ вакансий и резюме позволяет оперативно отслеживать изменения в спросе на рабочую силу и предложении со стороны соискателей. Это относительно недорогой и быстрый способ получения информации. Однако он не дает полной

картины, так как охватывает лишь открытые вакансии и активных соискателей. Трудоемкость данного подхода умеренная.

Качественные исследования [2, 19, 13] предполагают проведение интервью, фокус-групп, экспертных опросов для глубокого изучения проблем и тенденций на рынке труда, анализ кейсов успешного трудоустройства, построения карьеры, адаптации к изменениям на рынке, изучение лучших практик в области управления человеческими ресурсами, развития персонала, социального партнерства.

Качественные исследования дают глубокое понимание проблем и тенденций на рынке труда, выявляют скрытые мотивы и причинно-следственные связи. Они помогают интерпретировать количественные данные. Вместе с тем, качественные методы субъективны по своей природе и требуют высокой квалификации исследователей. Трудоемкость данного подхода высокая.

Сетевой анализ [27] включает изучение структуры и динамики социальных связей между участниками рынка труда, анализ потоков рабочей силы между отраслями, профессиями, регионами, выявление ключевых факторов и их влияния на процессы на рынке труда.

Сетевой анализ позволяет изучить структуру и динамику социальных связей между участниками рынка труда, что важно для понимания процессов перетока кадров. Однако он требует специфических методик и инструментов, а также наличия данных о социальных связях. Трудоемкость данного подхода также высокая.

Перечисленные методы часто используются в комбинации для получения более полной и достоверной картины ситуации на рынке труда. Выбор конкретных подходов зависит от целей и задач исследования, доступности данных и ресурсов, специфики изучаемого сегмента рынка труда.

Сравнительный анализ перечисленных методов представлен в таблице 1.

Таблица 1 – Сравнительный анализ методов анализа рынка труда

Характеристика	**Статистический анализ**	**Опросы и обследования**	**Экономическое моделирование**	**Анализ вакансий и резюме**	**Качественные исследования**	**Сетевой анализ**
Источники данных	Государственная статистика, отчетность компаний	Прямые опросы работодателей и соискателей	Статистические данные, экономические показатели	Онлайн-платформы, сайты по поиску работы	Фокус-группы, глубинные интервью	Социальные сети, профессиональные сообщества
Глубина анализа	Общие тенденции, макроэкономические показатели	Более детальная информация о предпочтениях	Прогнозирование и выявление закономерностей	Актуальная информация о спросе и предложении	Углубленное понимание мотивов и поведения	Выявление связей и взаимодействий на рынке труда
Скорость получения результатов	Относительно быстрый	Более длительный процесс	Требует времени на построение моделей	Оперативный доступ к актуальным данным	Наиболее длительный процесс	Зависит от доступности данных в социальных сетях
Возможность количественной оценки	Высокая	Средняя	Высокая	Средняя	Низкая	Средняя
Охват рынка	Широкий охват на макроуровне	Зависит от выборки респондентов	Широкий охват, но может упускать локальные особенности	Охватывает активную часть рынка	Глубокий, но ограниченный охват	Зависит от активности в социальных сетях
Уровень детализации	Обобщенные показатели	Более детальная информация	Может учитывать множество факторов	Детальная информация о спросе и предложении	Максимальная детализация	Зависит от доступности данных в социальных сетях
Объективность	Высокая объективность	Может быть смещена субъективными факторами	Высокая объективность при корректном построении модели	Относительно объективный, но может быть смещен	Наиболее субъективный подход	Зависит от достоверности данных в социальных сетях
Стоимость проведения	Относительно низкая	Более высокая	Высокая, требует специальных знаний	Относительно низкая	Наиболее высокая	Может быть низкой при наличии доступа к данным

В последние годы значительно вырос интерес в части применении алгоритмов и структур искусственного интеллекта и методов машинного обучения для систематического анализа данных о рынке труда с целью поддержки стратегического планирования и принятия решений [52, 36, 108]. Данный вид исследований получил название «интеллектуального анализа рынка труда» (англ. LMI – Labor Market Intelligence).

Особенно важным направлением в рамках LMI стало использование методов интеллектуального анализа текстов на естественном языке для анализа изменений на рынке труда на основе данных из текстов вакансий из открытых источников.

Современные интеллектуальные методы анализа текстов вакансий представляют собой мощный инструмент, позволяющий:

– определять востребованность отдельных профессий, должностей, квалификаций, знаний, навыков и компетенций.

– определять изменений спроса на определенные знания, навыки, компетенции и технологии в рамках конкретной профессии или отрасли;

– сравнивать рынки труда в разных отраслях и регионах;

– сегментировать вакансий и отдельные требования по различным группам на основе различных критериев.

Интеллектуальный анализ текстов вакансий становится ключевым инструментом для исследования и понимания динамики рынка труда, что имеет важное значение для стратегического планирования и принятия управленческих решений.

Рассмотрим каждое из этих направлений подробнее.

Определение востребованности отдельных профессий, должностей, квалификаций, знаний, навыков и компетенций. Это направление предполагает детальный анализ перечня требований, указанных в вакансиях. Выявление наиболее часто встречающихся навыков, компетенций и квалификаций позволяет понять, какие именно умения и знания наиболее ценятся работодателями. Аналогично, определение наиболее популярных должностей и профессий дает

представление о структуре спроса на рынке труда. Это знание помогает соискателям сфокусировать свое профессиональное развитие на наиболее востребованных направлениях.

Определение изменений спроса на определенные знания, навыки, компетенции и технологии в рамках конкретной профессии или отрасли. Отслеживание динамики изменений в требованиях работодателей, востребованности различных навыков и профессий позволяет выявлять ключевые тренды на рынке труда. Это включает анализ влияния внешних факторов, таких как технологические, экономические и социальные изменения. Понимание этих тенденций помогает прогнозировать будущие потребности работодателей и адаптировать стратегии трудоустройства.

Сравнение рынков труда в разных отраслях и регионах. Сопоставление вакансий в различных отраслях и географических локациях дает возможность выявить отраслевые и региональные особенности рынка труда. Это позволяет определить наиболее перспективные направления для трудоустройства с учетом специфики конкретных рынков. Такой анализ помогает соискателям сделать более обоснованный выбор в отношении профессионального развития и поиска работы.

Сегментировать вакансий и отдельные требования по различным группам на основе различных критериев. Данный метод предполагает группировку и классификацию вакансий и/или отдельных требований по различным параметрам, таким как отрасль, регион, уровень должности, требуемые навыки и т.д. Это позволяет выявить особенности и закономерности в различных сегментах рынка труда, которые могут быть упущены при общем анализе. Более глубокое понимание специфики отдельных сегментов помогает определить наиболее подходящие возможности для трудоустройства.

Комплексное использование перечисленных подходов позволяет получить наиболее полную картину состояния рынка труда, что помогает специалистам по управлению персоналом и подбору персонала принимать обоснованные решения. Данное направление исследований незаменимо для HR-специалистов, аналитиков

рынка труда и всех, кто заинтересован в эффективном трудоустройстве и развитии человеческого капитала.

Далее приводится обзор исследований с использованием методов интеллектуального анализа в контексте задачи анализа рынка труда.

В исследованиях [48, 42], представлен анализ квалификаций, востребованных на итальянском рынке труда в онлайн-объявлениях о работе. Использовались методики анализа естественного языка, включая технологии векторного представления слов и кластеризации, чтобы выявить как специализированные, так и универсальные навыки, необходимые для различных специальностей. Исследователи также рассмотрели, как автоматизация влияет на потребность в социальных и цифровых компетенциях, и предложили инструментарий для отслеживания эволюции терминологии и появления новых профессиональных навыков через анализ вакансий в интернете.

В публикации [69] изложен метод оценки степени цифровизации профессий в Германии, основанный на данных из онлайн-объявлений за 2014-2018 годы. Авторы создали индекс для анализа спроса на цифровые умения и их динамики, отметив рост необходимости хотя бы одного цифрового навыка с 38,1% до 47,5% за указанный период. Результаты показывают, что цифровые компетенции востребованы в разной степени в различных секторах, особенно актуальны в сферах ИТ, связи, финансов и образования.

В статьях [112, 92] авторы сравнили несколько подходов к моделированию текстовых данных, в частности, требований к кандидатам в объявлениях о вакансиях, используя разные модели векторного представления слов, включая TF-IDF, pLSA, word2vec, fastext, ELMo и многоязычный BERT, а также исследовали эффективность различных стратегий кластеризации. Особое внимание было уделено возможности улучшения результатов за счет дообучения моделей на специализированном корпусе текстов вакансий.

На сегодняшний день наиболее перспективным и востребованным подходом для анализа текстов на естественном языке является использование нейронных

сетей построенных на архитектуре трансформеров, использующих механизм внимания [111] В частности, в 2018 году был представлен BERT [50], инновационная модель, которая учитывает контекст слов при формировании их векторных представлений, показывая тем самым лучшие результаты в различных аспектах обработки естественного языка по сравнению с предшественниками, например, word2vec и LSTM [54]. Это достигается за счет более эффективного учета взаимосвязей в тексте и сохранения контекстуальной значимости слов. Сейчас активно ведется работа над новыми версиями BERT, такими как RoBERTa [78] и GPT-3 [47], которые в некоторых областях показывают еще большую эффективность. Особый интерес представляют исследования, посвящённые расширению возможностей модели BERT с использованием структур из баз данных [93, 79].

Тексты требований из вакансий, как правило, представляют собой короткие тексты (3-15 слов). Длина текста влияет на качество решаемых NLP задач [77]. Особенностью коротких текстов – сам текст: в силу своего размера, хранит достаточно ограниченное количество информации. Поэтому задача разработки методов расширения (обогащения) контекста коротких текстов для задач компьютерной обработки является актуальной. В этом направления предлагаются следующие подходы: на основе открытой базы DBpedia Ontology [57], с использованием сверточных нейронных сетей [1], сверточных сетей с механизмом внимания [64], графовых сверточных сетей [106], с использование длинных текстов [37], с использованием таксономий [103], и на базе векторных моделей представления слов [118].

Так, кроме методов анализа текстов, особенно коротких текстов, отдельным актуальным вопросом для исследования остаются методы формализации знаний предметной области. Одной из самых эффективных структур представления знаний, позволяющей с одной стороны хранить огромное количество объектов и связей между ними, а с другой предоставлять высокоскоростной доступ к хранящимся данным, является граф знаний (knowledge graph). В последние годы

появилось большое количество работ, в которых рассматриваются различные подходы (в т.ч. автоматического) построения графов знаний из текстов для различных предметных областей [89, 102, 76, 71, 70].

В статье [61] описывается реализация системы для анализа РТ (GraphLMI) для Европейского проекта ESCO. Описан процесс формализации и проектирования модели данных GraphLMI, а затем процесс ее реализации ее в виде графа знаний, созданной путем обработки более 5,3 миллионов текстов онлайн вакансий, составленных из произвольного текста и собранных в период с 2018 по 2019 год для Франции, Германии и Великобритании. Предложенный подход позволяет оценить актуальность и схожесть занятий и навыков на европейском рынке труда, а также оценить его динамику и тенденции на уровне отдельных занятий и навыков. Результаты работы позволили обогатить европейскую стандартную таксономию занятий и навыков (ESCO), чтобы она лучше соответствовала ожиданиям рынка труда. Авторами также предложен декларативный язык запросов, позволяющий понять, сравнить и оценить динамику рынка труда в странах для поддержки политики и принятия решений на европейском уровне.

Графовые структуры могут также выступать как инструмент расширения контекста коротких текстов. Для извлечения именованных сущностей из коротких текстов предлагается использовать подходы NER (от англ. Named-Entity Recognition), а для их связывания между собой, подходы к решению задачи связывания именованных сущностей (от англ. NEL – Named Entity Linking) [80].

В последние годы универсальные языковые модели достигли значительных успехов, требуя разработки методов и инструментов для их всесторонней проверки и оценки обобщающих интеллектуальных способностей. В статье «RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark» представлена новая методика оценки сильного ИИ для русскоязычных текстов, которая проверяет его способность к обобщению, логическому мышлению, здравому смыслу и умению целеполагания в контексте текстовых данных [101].

Существующие методы анализа требований рынка труда основаны либо на статистической информации, лишенной возможности подробно оценить изменения в каждой профессии или отрасли, либо на неэффективных подходах к анализу текстовой информации на основе векторных представлений слов и методов машинного обучения прошлого поколения. Для эффективного мониторинга изменений требований и потребностей рынка труда можно воспользоваться методами анализа и обработки естественного языка с применением новых методов и технологий на основе современных нейронных сетей с механизмом внимания, а также технологий представления данных в виде графов знаний. Переход от статистического анализа на уровень анализа отдельных сущностей требований поможет повысить эффективность подбора персонала и оптимизировать процесс поиска кандидатов. Хотя анализ требований на уровне конкретных компетенций и знаний может быть трудоемким, разработка методов извлечения и анализа данных способна автоматизировать процесс выделения ключевых элементов из текстов вакансий и резюме, что позволит рекрутерам получать более точную информацию о кандидатах и требованиях к ним уже на начальном этапе подбора персонала.

В обзоре существующих методов и подходов интеллектуального анализа стоит особенно отметить проект европейской классификации навыков, компетенций, квалификаций и профессий ESCO (англ. European Skills, Competences, Qualifications and Occupations).

Эффективность извлечения ценных знаний из больших объемов данных в системах онлайн-рекрутмента по подбору персонала напрямую зависит от наличия обновленных баз данных, классификаторов и таксономий. Эти инструменты играют ключевую роль для эффективного использования алгоритмов машинного обучения и большинства задач по обработке и пониманию текстов на естественном языке. Современные исследования рынка труда часто опираются на проект ESCO - многоязычную европейскую классификацию навыков, компетенций, квалификаций и профессий [53].

ESCO направлена на облегчение информационного обмена о компетенциях и квалификациях между различными странами, способствуя тем самым мобильности рабочих в рамках Европейского союза. Классификация предоставляет поддержку в сферах занятости, образования и переквалификации рабочих. ESCO состоит из трех взаимосвязанных частей, образуя базу данных, которая доступна на 28 языках. Она включает в себя профессиональные профили, описания навыков и компетенций, а также квалификации, соответствующие стандартам EQF и ISCED. Структура данных в модели ESCO представлена на рисунке 1.

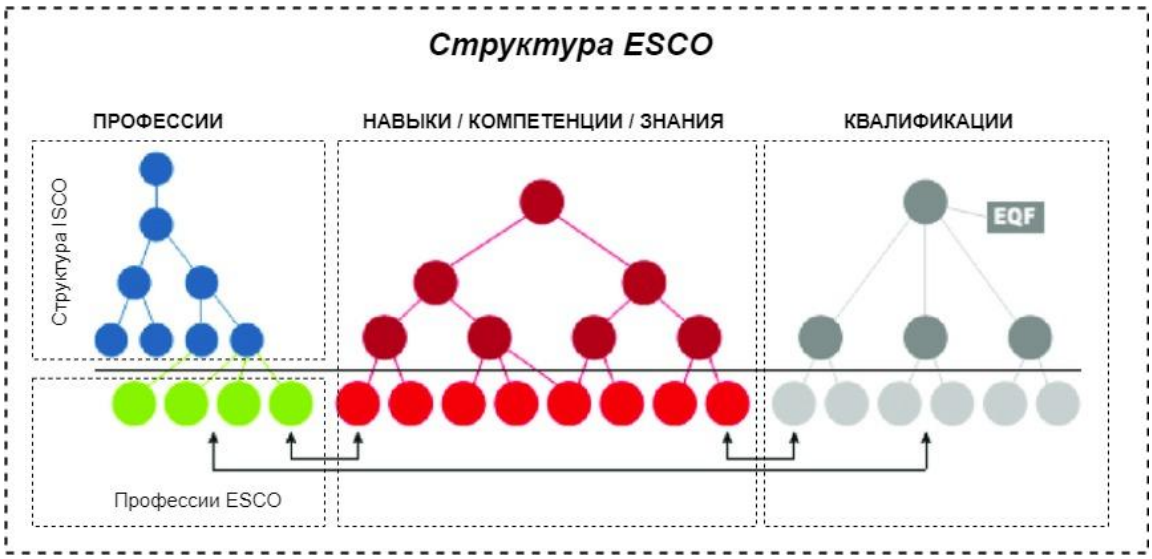

Рисунок 1 – Структура данных в модели ESCO

В качестве примера, профессия «инженер по компьютерному оборудованию» с кодом ESCO 2152.1.1 в таксономии включает 22 вариативных наименования, часть из которых указана в таблице 2. Для этой же профессии определены необходимые навыки, компетенции и знания, каждый из которых имеет свой список альтернативных наименований. Для профессии «инженер по компьютерному оборудованию» экспертами было определено 47 основных и 25 дополнительных навыков или компетенций, а также 16 основных и 20 дополнительных знаний (см. таблицу 3).

Таблица 2 – Примеры названий для профессии «инженер по компьютерному оборудованию» в классификации ESCO

Название профессии	Альтернативные названия
инженер по компьютерному оборудованию	специалист по компьютерному оборудованию
	инженер по компьютерной технике
	инженер по аппаратной части ПК
	специалист по ИТ-оборудованию

Таблица 3 – Примеры приоритетного и альтернативных названий знаний и навыков/ компетенций для профессии «инженер по компьютерному оборудованию» из классификации ESCO

Приоритетное название	Альтернативные названия	Тип
собирать аппаратные компоненты	сборка компьютерной техники, установка оборудования, сборка компьютерных комплектующих, сборка компонентов компьютера	навык / компетенция
установка программного обеспечения	установка компьютерного программного обеспечения, загрузка программного обеспечения, загрузка компьютерного программного обеспечения, установка компьютерного программного обеспечения, установка программного обеспечения, загрузка программного обеспечения, …	навык / компетенция
создавать технические планы	создавать планы относительно технических деталей, создавать промышленные планы, создавать технические чертежи	навык / компетенция
принципы электричества	электрический ток, напряжение, физика электричества, наука об электричестве, теория электричества, сопротивление, напряжение	знание
аппаратные компоненты	аппаратные компоненты системы, типы аппаратных компонентов, компоненты оборудования, компоненты для аппаратных систем, части для аппаратных систем, компоненты аппаратных систем, аппаратные части системы, типология аппаратных компонентов	знания

ESCO основана на онтологической модели, облегчающей связывание знаний, умений и квалификаций с подходящими должностями и профессиональными областями. Это способствует общему пониманию профессиональных ролей и упрощает процесс сравнения квалификаций между разными государствами и компаниями.

Главное достоинство данной системы классификации - использование ясных и понятных терминов для обозначения профессий, навыков и знаний на общеупотребительном языке. Это значительно облегчает и расширяет ее применение для анализа текстов требований к работникам из объявлений о работе, написанных на общеупотребительном языке, с применением современных методов машинного обучения.

Проектные команды применяют классификацию ESCO как основу для создания интеллектуальных технологий анализа рынка труда, что позволяет автоматизированно и систематически изучать спрос на определенные знания, умения и компетенции. Благодаря тому, что ESCO предлагает подробную и структурированную информацию о компетенциях, эти методы могут эффективно проводить сравнительный анализ требований рынка труда в разнообразных отраслях и регионах. Некоторые исследовательские группы разрабатывают методы и алгоритмы для того, чтобы автоматически расширять таксономию компетенций ESCO, используя данные из объявлений о вакансиях в интернете [62, 81].

Проект ESCO занимает ключевую позицию в разработке и использовании интеллектуальных способов анализа рынка труда, обеспечивая надежную и единообразную основу для тщательного изучения и оценки необходимых навыков и компетенций. Это, в свою очередь, может быть полезно для создания образовательных программ, планирования профессиональной траектории и поддержки в принятии решений по вопросам занятости.

1.3. Анализ моделей векторного представления текстов в задаче оценки семантической близости

Семантическая близость текстов (англ. semantic similarity) — это мера, показывающая степень схожести смысла, идеи или темы между двумя или более текстовыми фрагментами. Это понятие широко используется в области обработки естественного языка (англ. NLP – Natural Language Processing), информационного поиска и искусственного интеллекта для различных задач, таких как автоматическая суммаризация, поиск по смыслу, системы рекомендаций, анализ тональности текста и многих других.

Отношения семантической близости могут быть разделены на несколько категорий [11]:

1. Семантическая синонимия. Два слова считаются семантически близкими, если они имеют схожие или синонимичные значения. Например, слова «автомобиль» и «машина» считаются семантически близкими, так как они описывают одно и то же понятие.

2. Семантическая антонимия. Два слова считаются семантически близкими, если у них противоположные значения. Например, слова «холодный» и «горячий» считаются семантически близкими, так как они представляют противоположные состояния температуры.

3. Гиперонимия и гипонимия. Гиперонимия и гипонимия обозначают отношения между словами, где одно из слов служит более общим термином, а другое - его более узким или специфическим эквивалентом. В качестве примера, «овощ» выступает как гипероним для «помидор», подчеркивая их семантическую связь через структуру иерархии понятий.

4. Ассоциации. Возникают, когда два слова часто встречаются вместе или ассоциируются друг с другом. Например, слова «кофе» и «бодрость» часто ассоциируются в связи с влиянием кофе на состояние бодрствования.

5. Когипонимы. Слова, относящиеся к одному родовому понятию. Например, «береза», «клен», «дуб» – когипонимы, к ним можно добавить общее название «лесные деревья».

6. Меронимы и холонимы. Мероним – часть целого, холоним – целое. Например, «ветка» – мероним, «дерево» – холоним.

7. Семантические поля. Представляют собой группы слов, относящихся к одной тематической или понятийной области. Например, «образование», «наука», «учеба» относятся к семантическому полю «учение».

Это только некоторые категории отношений семантической близости, в реальности существует много других форм и связей между словами и понятиями. Анализ и представление этих отношений является важной задачей в области обработки естественного языка и позволяет более точно понять и интерпретировать смысл текстов и слов, помогает определить систему значений в языке, строить семантические сети понятий и решать многие NLP задачи, такие как автоматическое понимание текстов или машинный перевод.

В области компьютерной лингвистики существует множество подходов, которые позволяют преобразовать тексты и слова в числовые векторы, кодирующие их семантические свойства и отношения. Затем эти векторные представления можно использовать для решения различных задач NLP, таких как семантическое сопоставление текстов или извлечение сущностей.

Термины «эмбеддинг», или «векторное представление», а также «плотный вектор», часто используются взаимозаменяемо для обозначения одного и того же концепта. Эти векторы, имеющие небольшое количество нулевых значений при произвольной размерности, генерируются с использованием алгоритмов машинного обучения.

Далее по тексту описываются модели векторного представления текстов, которые используются в задачах семантического сопоставления и извлечения отдельных сущностей требований рынка труда из текстов вакансий.

Существуют различные классы моделей векторного представления текстов.

Модели на основе частоты слов

Bag of words (BOW): в этой модели текст представляется в виде набора слов без учета их порядка. Векторное представление строится путем подсчета частоты каждого слова в тексте.

TF-IDF (Term Frequency-Inverse Document Frequency) [105]. Индивидуальное весовое значение присваивается каждому слову в тексте на основе его частоты в этом тексте и обратной частоте его появления во всех текстах корпуса. TF-IDF учитывает, что некоторые слова могут быть более информативными для различных текстов.

Тематические модели

Вероятностный латентно-семантический анализ (PLSA, Probabilistic latent semantic analysis) [66]. Эта модель использует вероятностные методы для поиска скрытых тематик в тексте. Каждый текст представляется в виде комбинации тем с определенными вероятностями.

Латентное размещение Дирихле (LDA, Latent Dirichlet Allocation) [40]. Похоже на PLSA, LDA также моделирует текст как комбинацию тематик, но с использованием распределения Дирихле для генерации вероятностей.

Модели на основе нейронных сетей

Модели на основе распределенных представлений слов: Word2vec [85] -> Doc2Vec [73], GloVe [95], Fasttext [41]. Эти модели используют нейронные сети для обучения векторного представления слов или документов. Word2vec обычно используется для представления слов, а Doc2Vec – для представления документов. GloVe и Fasttext также строят векторы слов, учитывая контекст информации.

Вторая группа на основе предобучения языковых моделей включает ELMo [95] и BERT [50]. ELMo использует контекстуализацию для создания эмбеддингов слов. Она учитывает все предыдущие слова в предложении при создании эмбеддинга, что позволяет уловить контекстуальную информацию. BERT – это еще более продвинутая модель, которая предобучается на больших объемах текста и позволяет учитывать двунаправленный контекст. Она обычно используется для

различных задач NLP, таких как классификация текста, вопросно-ответная система или заполнение пропущенных слов.

Обе эти модели на основе предобученных языковых моделей, ELMo и BERT, позволяют получить контекстуализированные эмбеддинги слов, что помогает улучшить обработку языка и понимание текста. Эти модели обучаются на больших объемах текстовых данных, чтобы понять семантику языка и синтаксические особенности. Затем они могут быть использованы для решения различных NLP-задач и обработки текстовых данных.

Это основные классы моделей, которые позволяют представить текст в виде векторов, удобных для дальнейшей обработки и анализа с использованием методов машинного обучения. Выбор конкретной модели зависит от решаемой задачи и имеющихся данных.

1.3.1. Меры семантической близости для векторных представлений текстов

Мера семантической близости – это способ количественной оценки смыслового сходства между единицами языка (словами, текстами, предложениями). Она позволяет определить, насколько близки значения сравниваемых единиц.

Для вычисления меры семантической близости существует различные подходы [94]:

1. Косинусная мера близости – вычисляется как косинус угла между векторными представлениями сравниваемых единиц. Чем меньше угол, тем выше близость. Широко используется для векторов, полученных методами Word2vec, BERT и др.

2. Евклидово расстояние – геометрическое расстояние между векторами. Чем меньше расстояние, тем выше близость. Также применяется для векторных представлений.

3. Количество общих слов и элементов – чем больше общего в сравниваемых текстах, тем выше близость. Используются меры типа Jaccard index и TF-IDF.

4. Сетевая близость – основана на анализе связей в семантической сети. Узлы ближе по смыслу, если между ними меньше «расстояние» (количество дуг).

5. Вероятностные меры – основаны на вероятности перехода от одной единицы к другой. Чем выше вероятность, тем ближе смысл. Например, модели типа Skip-gram.

6. Комбинированные – объединяют различные вышеупомянутые меры для повышения точности. Например, суммирование косинусной близости и евклидова расстояния.

Выбор конкретной меры семантической близости зависит от нескольких факторов:

Тип входных данных

Если единицы языка представлены в виде векторов фиксированной размерности (например, полученных с помощью Word2Vec или BERT), то естественным выбором будут косинусная мера или евклидово расстояние. Они хорошо подходят для работы с плотными векторными представлениями.

Если единицы языка представлены в виде текстов или наборов слов, то лучше использовать меры, основанные на подсчете общих элементов, такие как коэффициент Жаккара или TF-IDF.

Если данные имеют структуру графа или сети, то применимы меры сетевой близости, учитывающие связи между узлами.

Размерность и разреженность данных

Для данных высокой размерности (тысячи и более признаков) косинусная мера обычно работает лучше, чем евклидово расстояние, так как менее чувствительна к абсолютным значениям векторов.

Для разреженных данных (большинство признаков равны нулю) лучше подходят меры, учитывающие только ненулевые значения, такие как коэффициент Жаккара.

Интерпретируемость результатов

Косинусная мера и евклидово расстояние дают значения в абсолютной шкале, которые могут быть сложны для интерпретации.

Меры на основе общих слов (Жаккар, TF-IDF) и вероятностные меры дают более интуитивно понятные оценки в процентах или вероятностях.

Вычислительная сложность

Косинусная мера и евклидово расстояние вычисляются достаточно быстро даже для векторов высокой размерности.

Меры на основе общих слов требуют предварительного построения словарей и подсчета частот, что может быть затратно по времени и памяти для больших текстовых коллекций.

Вероятностные меры и меры сетевой близости могут требовать обучения сложных моделей на больших объемах данных.

Специфика задачи

Выбор меры близости должен учитывать особенности конкретной задачи и данных. Например, для коротких текстов (заголовков, запросов) хорошо работают косинусная мера и Жаккар, а для длинных документов – TF-IDF. Для сравнения текстов на разных языках лучше подходят меры на основе многоязычных векторных представлений (например, LASER, LaBSE).

Выбор меры семантической близости – это компромисс между эффективностью, интерпретируемостью и адекватностью для конкретной задачи и типа данных. На практике полезно экспериментировать с разными мерами и оценивать их качество на размеченных данных или с помощью экспертных оценок.

В области компьютерной лингвистики или задачах обработки естественного языка для оценки семантической схожести двух текстов чаще всего используется

косинусная мера близости [44] между векторными представлениями этих текстов (1.1):

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}, \quad (1.1)$$

где:

$A \cdot B$ – скалярное произведение векторов A и B

$\|A\|$ и $\|B\|$ – евклидовы нормы (длины) векторов A и B

A_i и B_i – i-е компоненты векторов A и B

n – размерность векторного пространства

Косинусная мера близости принимает значения от -1 до 1: 1 означает, что векторы сонаправлены (угол 0°); 0 означает, что векторы ортогональны (угол 90°); -1 означает, что векторы противоположно направлены (угол 180°). Чем ближе значение косинусной меры к 1, тем более похожи векторы.

Косинусная мера широко используется в задачах информационного поиска, анализа текстов и рекомендательных системах для оценки сходства между векторными представлениями объектов (документов, запросов, предпочтений пользователей).

Для всех дальнейших экспериментов также будет использоваться косинусная мера для оценки семантической близости между векторами текстов.

1.3.2. Векторные индексы

Векторные индексы представляют собой специализированные структуры данных и алгоритмы, нацеленные на ускорение поиска векторов, наиболее близких или похожих, в обширных коллекциях векторной информации. Использование этих индексов позволяет более быстро находить ближайших соседей в векторных пространствах большой объемности и размерности, значительно увеличивая скорость и эффективность работы разнообразных систем и алгоритмов. Это делает векторные индексы эффективным инструментом в различных сферах, таких как машинное обучение, анализ данных и информационный поиск.

Основная проблема в том, что полный перебор и сравнение запроса со всеми векторами из набора данных требует огромного числа операций. Это не масштабируется на большие объемы данных и не позволяет достичь приемлемого быстродействия.

Векторные индексы решают эту проблему следующими способами:

– сокращают число сравниваемых векторов за счёт разбиения пространства на ячейки и выбора кандидатов только из релевантных ячеек;

– аппроксимируют вектора с помощью квантизации для ускорения вычисления расстояний между ними;

– используют эвристические алгоритмы поиска в графах или иерархических структурах данных.

Это позволяет ускорить поиск в сотни и тысячи раз без существенной потери точности. Благодаря векторным индексам становится возможным масштабирование семантического поиска на большие наборы текстов, изображений и других данных.

Векторные индексы основаны на идее, что:

– данные в векторном пространстве организуются в отдельные ячейки;

– каждая ячейка содержит индексированные вектора, которые могут быть упорядочены разными способами для более быстрого поиска;

– при выполнении поискового запроса, сначала находится подходящая ячейка, после чего поиск осуществляется исключительно внутри неё, используя её уникальную систему индексации.

Такой подход позволяет многократно сократить количество операций сравнения векторов по сравнению с линейным перебором.

Существует множество разновидностей векторных индексов, использующих разные методы кластеризации и индексации для оптимизации скорости и точности поиска.

Например, векторные индексы могут быть использованы для:

– поиска семантически близких фраз или документов в задачах обработки естественного языка;

– поиска похожих изображений или областей изображений в задачах компьютерного зрения;

– поиска похожих паттернов или объектов в задачах анализа данных;

– поиска похожих товаров или пользователей на основе их предпочтений в рекомендательных системах.

В настоящее время одной из самых популярных и востребованных библиотек, реализующей векторные индексы для быстрого поиска в больших массивах векторных данных является библиотека FAISS.

FAISS (англ. Facebook AI Similarity Search) [55] – это библиотека для эффективного поиска сходных векторов в больших наборах данных. Она предоставляет инструменты для быстрого поиска ближайших соседей в пространствах большой размерности, и используется для реализации семантического поиска и кластеризации в векторных пространствах.

В FAISS существует несколько типов индексов, которые могут использоваться для различных типов данных и задач.

Плоский индекс (IndexFlat). Простой метод, который вычисляет расстояния между всеми векторами в наборе данных полным перебором. Хотя этот метод требует больше времени для построения индекса, он обеспечивает быстрый поиск ближайших соседей.

Индекс LSH (англ. Locality-Sensitive Hashing) [67, 104]. Использует хэширование для поиска ближайших соседей, применяя несколько хэш-функций для разбиения пространства на более мелкие подпространства.

Индексы IVF (англ. Inverted File index) [27]. Используют алгоритмы кластеризации для разделения данных на кластеры, что ускоряет поиск ближайших соседей.

Индекс HNSW (англ. Hierarchical Navigable Small World) [98, 83, 82]. Использует графовую структуру для быстрого поиска ближайших соседей в пространствах большой размерности.

Каждый из этих индексов имеет свои особенности и применим в зависимости от конкретной задачи и характеристик данных.

Выбор подходящего индекса в библиотеке FAISS зависит от нескольких факторов.

Размерность векторов. Если размерность векторов невелика (до нескольких сотен), то можно использовать плоский индекс (IndexFlat). Он обеспечивает быстрый и точный поиск, но требует больше памяти и времени на построение. Для векторов высокой размерности (тысячи и более) лучше подойдут индексы, использующие приближенный поиск, такие как IVF или HNSW. Они позволяют ускорить поиск за счет некоторой потери точности.

Объем данных. Для небольших наборов данных (до миллиона векторов) можно использовать плоский индекс. Для больших объемов данных (миллионы и миллиарды векторов) нужны индексы, масштабируемые на большие коллекции, такие как IVF или HNSW.

Требования к скорости и точности поиска. Если нужен максимально быстрый поиск и можно пожертвовать точностью, то хорошим выбором будут индексы LSH или IVF с небольшим количеством кластеров. Если важна высокая точность поиска, то лучше использовать плоский индекс или HNSW с большим количеством связей.

Характер данных. Если данные имеют кластерную структуру (естественным образом разбиваются на группы схожих векторов), то эффективны будут индексы IVF. Если данные распределены более равномерно, то можно использовать LSH или HNSW.

Доступные вычислительные ресурсы. Индексы IVF и HNSW требуют больше памяти и вычислительных ресурсов на этапе построения, чем плоский индекс или LSH. На практике часто пробуют несколько типов индексов с разными

параметрами и выбирают лучший по соотношению скорости и точности поиска на конкретных данных.

FAISS предоставляет удобные инструменты для сравнения качества разных индексов. Также можно комбинировать индексы между собой, используя, например, IVF поверх HNSW для больших коллекций векторов высокой размерности.

1.4. Анализ методов извлечения именованных сущностей

Извлечение именованных сущностей является важной задачей в области обработки естественного языка и информационного поиска (иногда в литературе можно встретить аналогичное название для этой группы методов «Классификация токенов» или англ. «Token Classification»). Существуют различные методы для извлечения именованных сущностей: методы на основе словарей, методы на основе правил, методы на основе машинного обучения или их комбинации.

Первый подход основан на словарях, который является наиболее простым и фундаментальным для NER. Он использует словарь с разными словами, синонимами и словарными формами. Алгоритм проверяет, есть ли определенная сущность из текста в словаре, в результате чего выполняется кросс-проверка сущностей. Один из недостатков этого подхода заключается в необходимости постоянного обновления словарей для эффективной работы модели NER.

Второй подход использует системы на основе правил. Информация извлекается на основе заранее установленных правил, определяющих, какие слова могут быть именованными сущностями. Существуют правила на основе шаблонов и контекстные правила, которые зависят от значения или контекста слова в документе. Эти методы могут быть эффективными в определенных областях, однако они также требуют обновления правил для новых данных и неспособны обрабатывать нестандартные ситуации.

Третий подход основан на машинном обучении и использует статистическое моделирование для обнаружения сущностей. В этом подходе текстовые документы

представляются в виде признаков, что позволяет модели распознавать различные типы сущностей, даже при небольших различиях в написании, используя контекстную информацию.

Методы машинного обучения для выявления именованных сущностей применяют разнообразные алгоритмы машинного обучения, включая скрытые марковские модели [86], условные случайные поля [1] и нейронные сети [74, 72], для тренировки моделей на большом количестве аннотированных данных. Эти подходы могут обеспечивать гибкость и способность обрабатывать комплексные ситуации, однако они могут потребовать значительное количество аннотированных данных и времени для обучения.

Смешанные (комбинированные) методы сочетают в себе подходы, основанные на правилах, и машинное обучение, чтобы повысить точность и полноту извлечения сущностей. Эти методы способны объединять преимущества обеих подходов, хотя их разработка и настройка может представлять сложность.

Для русского языка ярким представителем систем на основе правил является YARGY-парсер [88] – это инструмент на основе контекстно-свободных грамматик для извлечения структурированной информации из текстов на естественном языке. Данный парсер использует морфологический и синтаксический анализ для извлечения сущностей, отношений и событий из текстов.

YARGY-парсер существует как элемент библиотеки Natasha. Он предоставляет возможность создавать пользовательские правила для извлечения информации из текстов и поддерживает различные типы сущностей, таких как именованные сущности, даты, числа и т.д.

YARGY-парсер может применяться для анализа текстов на естественном языке в различных областях, таких как обработка новостей, медицинская документация, юридические материалы, обработка естественного языка и другие сферы. Он может быть полезен исследователям, разработчикам и специалистам по обработке естественного языка для извлечения структурированной информации из текстов.

Для русского языка также доступны решения от проектной группы DeepPavlov, основанные на нейронных сетях [74, 72]. В своих работах авторы исследовали различные архитектуры глубоких моделей нейронных сетей, начиная с базового двунаправленного Bi-LSTM, затем добавляя условные случайные поля (CRF), сети магистралей и, в конечном итоге, внешние вложения слов. Применение Bi-LSTM с CRF значительно улучшило качество предсказаний. Кодирование входных токенов при помощи внешних вложений слов сократило время обучения и позволило достигнуть современного уровня развития для задачи NER.

Можно выделить несколько типов проблем, возникающие в задаче извлечения именованных сущностей.

Дефрагментация сущностей (разрывность сущностей). Существующие методы извлечение именованных сущностей (NER) требует, чтобы искомая сущность в тексте была представлена как непрерывная последовательность слов для достижения высокой точности в распознавании и извлечении таких сущностей. Это условие существенно снижает эффективность NER при работе с реальными текстами требований, особенно для русского языка, где из-за человеческого фактора тексты требований часто представляют собой сложные структуры (см. разновидности сложных предложений на русском языке [29]).

На практике многие именованные сущности требований представляют собой не одно слово, а последовательность слов, которые могут быть разделены, другими словами. Например, текст требования из вакансии «Описание и моделирование бизнес-процессов». Текст сущности «описание бизнес-процессов» разделен здесь несколькими словами. Традиционные методы извлечения сущностей часто не могут обработать такие разрывные конструкции и не объединяют отдельные слова в одну сущность.

Это означает, что если именованная сущность состоит из нескольких слов или токенов, то методы извлечения именованных сущностей могут столкнуться с проблемами их корректного обнаружения.

Неоднозначность интерпретации. Данная проблема возникает в следствии: синонимии, многозначности отдельных слов, нестандартного написания или опечаток.

Например, «Замок» может быть и строением, и устройством. Традиционные методы не умеют корректно разрешать такую неоднозначность, определяя единственный тип сущности без учета контекста. Одна и та же последовательность слов может интерпретироваться как разные типы сущностей в зависимости от контекста. В текстах могут встречаются опечатки, разного рода искажения правильных написаний имен, фамилий, названий. Это тоже затрудняет извлечение стандартных сущностей.

Размытые границы сущностей. Бывает сложно определить точные границы сущности в потоке текста, отделить её от контекстных слов. Это важно для корректного извлечения.

Зависимость от языка и предметной области. Правила и особенности извлечения сущностей могут сильно меняться для разных языков и предметных областей (например, новости, медицина, бизнес). Это осложняет создание универсальных методов.

Ограничения существующих методов извлечения именованных сущностей включают в себя необходимость в большом количестве размеченных данных для методов машинного обучения, сложности при обработке нестандартных случаев и недостаточную гибкость правил. Также, многие из методов могут быть зависимы от конкретного языка.

1.5. Анализ существующих в России систем онлайн-рекрутмента

Системы онлайн-рекрутмента – это интернет-платформы, предназначенные для поиска работы и подбора персонала онлайн. Они позволяют соискателям размещать резюме и откликаться на интересные вакансии, а работодателям – публиковать вакансии и искать подходящих кандидатов. Эти

платформы позволяют быстро найти интересующую должность или сотрудника, имеют развитый функционал для подбора персонала и поиска работы.

Популярные российские системы онлайн-рекрутмента:

– headhunter (hh.ru) – крупнейшая база вакансий и резюме в России;
– superjob (superjob.ru) – вторая по величине база вакансий после HH;
– работа.ру (rabota.ru) – более 300 тысяч вакансий по всей России;
– job.ru – база вакансий от ведущих российских компаний;
– career.ru – вакансии крупных международных компаний в России.

Стоит отметить, что этот список далеко не полный, и с каждым годом на российском рынке появляются новые системы онлайн-рекрутмента.

Для дальнейших исследований были отобраны два источника headhunter.ru и superjob.ru. У каждого из этих ресурсов существует свой api: api.hh.ru и api.superjob.ru, соответственно. Они оба возвращают json-объект с данными о вакансиях.

Пример json-объекта вакансии с сайта headhunter.ru представлен в приложении В. Полный текст вакансии, включающий html-разметку, хранится в полях description и richtext для headhunter.ru и superjob.ru, соответственно.

Анализ текстов вакансий из отрасли информационных технологий, собранных в системах онлайн-рекрутмента HeadHunter и Superjob, выявил несколько проблем, которые существенно затрудняют извлечение отдельных сущностей *знаний* и *навыков/компетенций*:

– лишь 20-25% всех текстов вакансий имеют строгую html-разметку на 4 класса: общий текст (общая информация), требования, обязанности и условия. Это обусловлено тем, что при составлении текстов вакансий отсутствует жесткие требования для оформления структуры вакансии, в следствии чего из-за человеческого фактора эти разделы часто имеют произвольный порядок, часть из них может отсутствовать в тексте вакансии или иметь нестандартные формулировки, что существенно затрудняет выделение блока текстов требований

из текстов вакансий для дальнейшего анализа. Примеры вакансий с жесткой и произвольной структурой представлены в приложении Г.

– тексты требований часто представляют собой сложные предложения [28], где отдельный сущностей требований разнесены в тексте и разделены другими словами, что существенно снижает эффективность применения стандартных методов извлечения именованных сущностей. В качестве примера «сложного» текста требования можно привести следующий текст «Описание, моделирование (желательно Bizagi) и/или оптимизация бизнес-процессов». Из этого текста можно выделить следующий список простых сущностей требований: «бизнес – процессов»; «бизнес процессов»; «описание бизнес процессов»; «моделирование бизнес процессов»; «оптимизация бизнес процессов»; «описание процессов»; «оптимизация процессов»; «моделирование процессов»; «желательно Bizagi». Однако, как уже было сказано выше, существующие методы извлечения сущностей не способны извлекать сущности, составные токены которых, разделены другими токенами.

– наличие большого количества синонимичных сущностей, т.е. текстов сущностей, которые, по сути, определяют одно и тоже в реальном мире. В качестве простых примеров можно привести: «бизнес-процессов» и «бизнес процессов», «html5» и «html 5». Кто-то пишет с использование тире или знак пробела в качестве разделителя, а кто-то может написать без них. Проблема поиска и объединения таких сущностей является актуальной.

1.6. Постановка цели и задач исследования

В результате анализа подходов и методов в пунктах 1.1-1.3 сформирована **цель диссертационного исследования** – разработка методов и алгоритмов интеллектуальной поддержки процедур (процесса) формирования требований к вакансиям, которые обеспечат повышение качества анализа современных тенденций рынка труда, повысят эффективность процессов подбора персонала и соответствие разрабатываемых требований в проектах вакансий реальным потребностям рынка труда.

Для достижения указанной цели необходимо решить следующие **задачи**:

1. Разработать модель формализованного описания требований реального рынка труда на уровне отдельных сущностей знаний и навыков/компетенций, которая бы позволила учитывать структурные и семантические отношения между ними.
2. Разработать метод и алгоритм извлечения сущностей знаний и навыков/компетенций из текстов требований вакансий реального рынка труда на основе нейросетевых моделей языка и методов классификации.
3. Разработать интеллектуальный метод поддержки формирования списка требований вакансий на основе семантического сопоставления сущностей знаний и навыков/компетенций предложенной структурно-семантической модели и применения методов кластеризации.
4. Выполнить программную реализацию предложенных методов, моделей и алгоритмов в виде прототипа интеллектуальной рекомендательной системы поддержки формирования требований вакансии.
5. Провести экспериментальную оценку реализованных методов и алгоритмов интеллектуальной поддержки формирования требований на текстовом корпусе проектов вакансий.

Выводы по первой главе

1. Проведен анализ процесса подбора персонала и процесса формирования требований к вакансии. Определены основные группы проблем связанные с человеческими факторами в процессе формирования требований к вакансиям. В результате анализа было выявлено, что человеческий фактор вносит значительные искажения и неэффективность в процессы подбора персонала и формирования требований к вакансиям. Для преодоления этих проблем необходимо внедрение более структурированных и автоматизированных подходов, а также повышение осведомленности и обучение участников процесса.

2. Проведен анализ исследований российских и зарубежных ученых по разработке методов, моделей и алгоритмов интеллектуального анализа рынка труда. Анализ показал, что существующие на сегодняшний день интеллектуальные методы анализа требований рынка труда основываются либо на анализе статистической информации, которые по определению не позволяют оценить изменения в каждой профессии или отрасли на уровне отдельных сущностей требований; либо предлагают малоэффективные подходы анализа текстовой информации на основе векторных представлений прошлого поколения, таких как, TF-IDF, word2vec; либо представляют собой подходы на основе вручную размеченной таксономии знаний и навыков ESCO.

3. Проанализированы современные модели векторного представления текстов, различные меры семантической близости, существующие модели построения векторных индексов. Проведенный анализ показал, что сочетание современных моделей векторного представления текстов, мер семантической близости и векторных индексов является перспективным подходом для эффективного решения задач, связанных с определением семантического сходства текстовых данных.

4. Проанализированы современные методы извлечения именованных сущностей. Определены основные виды проблемы в процессе извлечения

именованных сущностей. В результате анализа было определено, что существующие методы извлечения именованных сущностей являются малоэффективными в задаче извлечения отдельных сущностей требований, и требуют разработки новых методов для решения этой задачи.

5. Проведен анализ существующих систем онлайн-рекрутмента в России. В результате анализа текстов онлайн вакансий были выявлены проблемы, связанные с отсутствием жесткой структуры в текстах вакансий. Также в результате анализа текстов требований из вакансий было выявлено, что большинство из них представляют собой сложно организованные тексты, что делает использование существующих методов извлечения именованных сущностей малоэффективными в задаче извлечения текстов отдельных сущностей требований из текстов вакансий, таких как знания и навыков/компетенции для дальнейшего анализа.

ГЛАВА 2 КОНЦЕПЦИЯ ИНФОРМАЦИОННОЙ ПОДДЕРЖКИ ФОРМИРОВАНИЯ ТРЕБОВАНИЙ ВАКАНСИИ

Во второй главе описывается концепция информационной поддержки формирования требований вакансии на основе семантического сопоставления сущностей структурно-семантической модели требований рынка труда и методов кластеризации. Описывается модель формализованного описания требований рынка труда на уровне отдельных сущностей знаний и навыков/компетенций в виде структурно-семантической модели. Проводится подробный обзор нейросетевых моделей на архитектуре трансформеров, а также процесс их дообучения на текстах предметной области. Предлагается метод извлечения текстов отдельных сущностей требований знаний и навыков/компетенций.

2.1. Концепция информационной поддержки формирования требований вакансии на основе семантического сопоставления сущностей структурно-семантической модели и методов кластеризации

Концепция информационной поддержки формирования требований вакансии на основе семантического сопоставления сущностей требований структурно-семантической модели и методов кластеризации представлена на рисунке 2.

Обычно анализ рынка труда проводится на основе изучения больших статистических данных, включая статистику по занятости, безработице, среднему заработку и другие ключевые показатели. Тем не менее, такой подход не дает полного и подробного понимания потребностей рынка. В основе концепции предлагается переход от общего исследования статистических данных к анализу потребностей рынка труда на более детальном уровне, сфокусированном на отдельных сущностях знаний, навыков и компетенций, а также на частотной информации об этих сущностях и частотной информации об их совместной встречаемости. Такой анализ способствует более эффективному подбору

персонала и позволяет преодолеть проблему обеспечения соответствия потребностей организации с актуальными данными о востребованности отдельных требований реального рынка, за счет более точного определения, какие сущности требования чаще всего используются совместно для успешного выполнения профессиональных задач по конкретной вакансии.

Рисунок 2 – Схема концепции информационной поддержки формирования требований вакансии

В основе концепции лежит структурно-семантическая модель, которая способна учитывать структурные и семантические связи между отдельными сущностями требований. В рамках модели предлагается представлять отдельные сущности требований как точки в многомерном пространстве на основе плотных векторов современных языковых моделей. В силу несовершенства существующих методов извлечения сущностей (см. раздел 1.4) и большого количества синонимичных словоформ, существующих на рынке труда сущностей требований, важным этапом в процессе построения структурно-семантической модели является снижение фрагментированности модели.

Для решения этой проблемы предлагается на основе семантической меры близости определяются группы сущностей требований с высокой степенью близости в некоторой окрестности многомерного пространства, что может свидетельствовать о синонимичности понятий в этой окрестности, а затем на основе их частотности определять гипероним (общее понятие) среди этих синонимичных словоформ сущностей требований. Таким образом объединение синонимов сущностей требований позволит перейти от всего многообразия словоформ сущностей требований, увеличит количество связей между отдельными компонентами модели и снизить эффект фрагментированности модели.

Общая концептуальная схема структурно-семантической модели представлена на рисунке 3.

Методику решения концепции формирования списка требования можно представить в виде последовательности шагов:

– извлечение отдельных сущностей знаний и навыков/компетенций из текстов вакансий;

– построение структурно семантической модели на основе выделенной информации с предыдущего шага;

– построение подграфа исходной структурно-семантической модели с учетом требований пользователя;

– кластеризация и ранжирование сущностей требований смежных с исходным списком сущностей.

Используемые методы

Для реализации этой концепции используется семантическое сопоставление отдельных сущностей требований. Оно основывается на векторном представлении отдельных сущностей требований, полученных с помощью современных нейросетевых языковых моделей. Семантическое сопоставление позволяет определить, насколько близко связанные сущности или термины семантически близки друг к другу имеют отношение к конкретной вакансии. Например, если в требованиях к вакансии указано «опыт работы с Java», семантическое

сопоставление может определить, насколько близко связаны другие языки программирования с Java, и предложить их в качестве дополнительных требований.

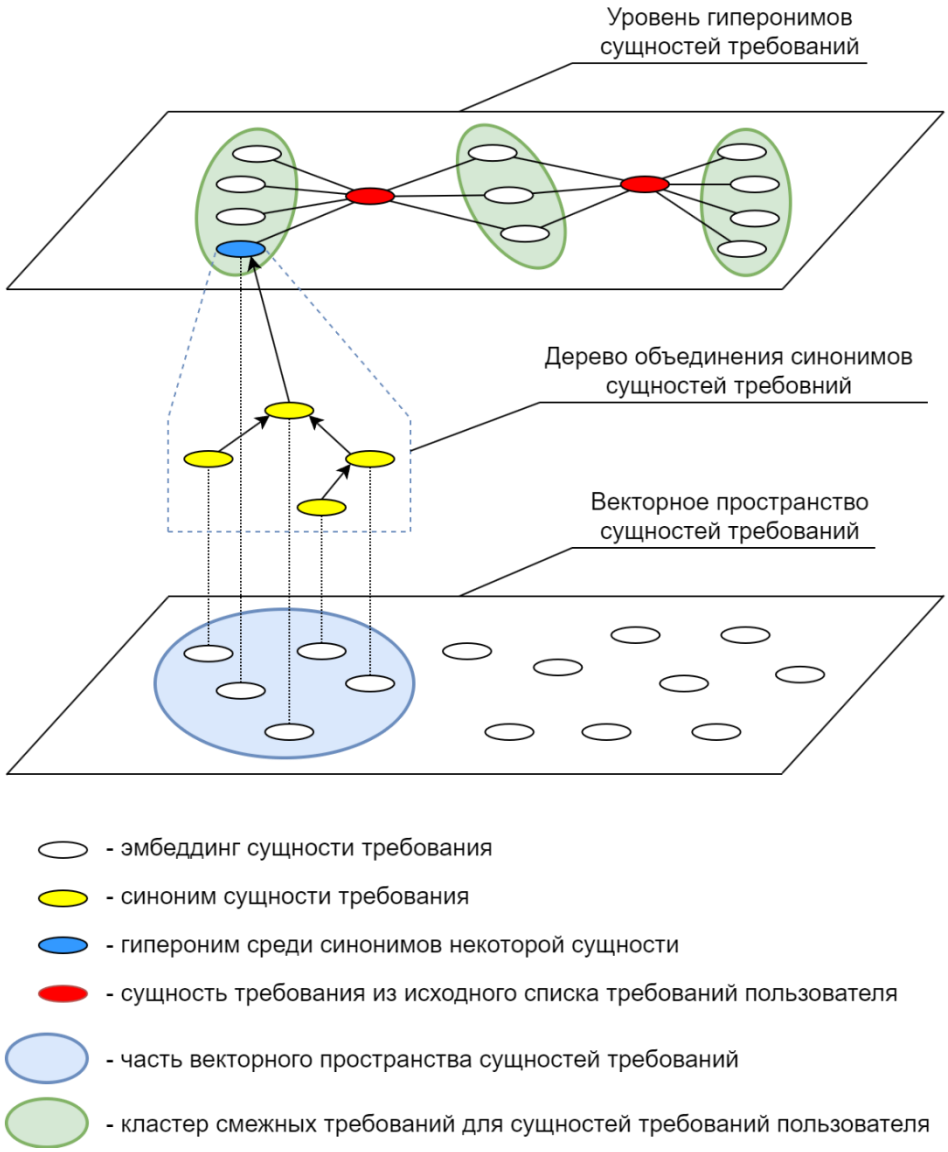

Рисунок 3 – Концептуальная схема структурно-семантической модели требований рынка труда

В дополнение к семантическому сопоставлению требований в концепции используются методы классификации и кластеризации, которые применяются для группировки требований к вакансии на основе их схожести. Это позволяет эффективно организовывать группы требований и упрощает процесс анализа и формирования требований к вакансии. Например, требования, связанные с опытом

работы, могут быть сгруппированы в один кластер, в то время как требования, связанные с навыками и квалификацией, могут быть сгруппированы в другой кластер.

Использование системы

Выделение отдельных сущностей знаний, навыков и компетенций является важным этапом в процессе формулирования требований к вакансиям. Это позволяет специалистам более четко и точно определить необходимые навыки и компетенции кандидатов, а также оценить их уровень. Наличие четких критериев помогает производить более точный подбор кандидатов и повышает эффективность процесса найма. Также это позволяет сократить количество неподходящих кандидатов на стадии отбора резюме, а значит, сократить время и затраты на процесс подбора персонала.

В качестве лица принимающего решения (ЛПР) при работе с системой могут выступать:

– линейные и функциональные руководители служб, отделов, структурных подразделений;

– специалисты по подбору персонала;

– эксперты из профессионального сообщества.

Процесс работы с рекомендательной системой для ЛПР включает в себя два сценария:

– формирование списка требований и уточнение формулировок требований для вакансии;

– сопоставление требований из текстов вакансии с текстами присланных резюме.

Общая структурная схема управления процесса формирования рекомендаций представлена на рисунке 6.

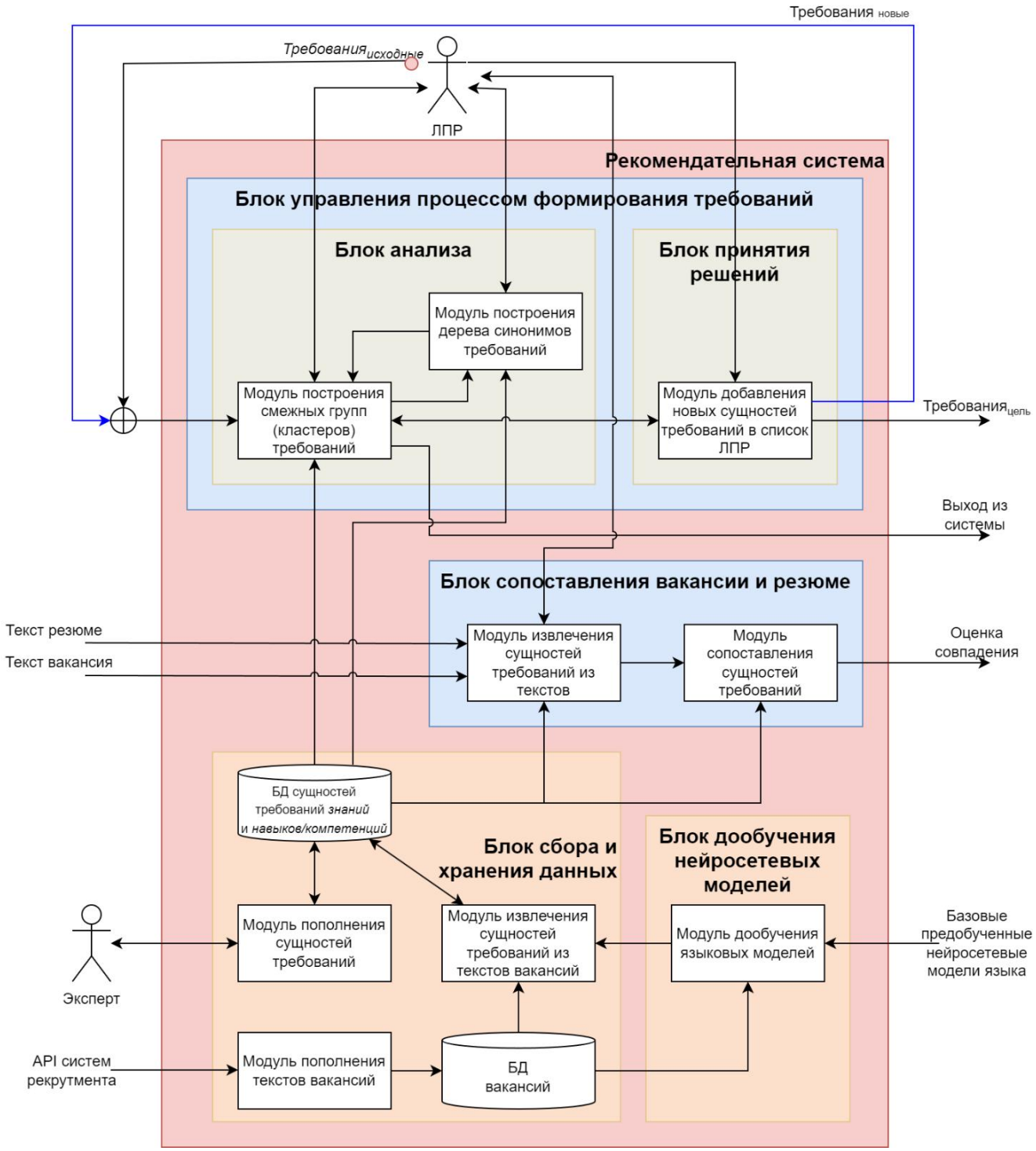

Рисунок 4 – Структурная схема управления процессом формирования рекомендаций

Структура процесса управления

Блок сбора и хранения данных. В этом блоке находятся модуль сбора текстов вакансий из систем онлайн-рекрутмента и модуль пополнения сущностей требований, через который эксперт может добавлять тексты сущностей требований

в база данных. Также в этом блоке находятся базы данных вакансий и сущностей требований знаний и навыков/компетенций.

Блок дообучения нейросетевых моделей состоят из одного модуля дообучения базовых предобученных нейросетевых моделей языка. Предполагается, что по мере обновления базы данных текстов вакансий требуется периодическое дообучение языковых моделей.

Блок управления процессом формирования требований из двух подблоков: блока анализа; блока принятия решений.

Блок анализа состоит из двух модулей:

– Модуль построения смежных группы требований. В этом модуле пользователь получает возможность визуально проанализировать группы требований, которые имеют внутренние устойчивые связи в структуктурно-семантической модели требований рынка труда с сущностями требований из его первоначального списка;

– Модуль построения дерева синонимов требований. В этом модуле пользователь получает возможность проанализировать дерево синонимов формулировок отдельной сущности требования и тем самым выбрать наиболее востребованную формулировку того или иного требования, не меняя при этом смысл требования.

Блок принятия решений состоит из одного модуля добавления новых сущностей требований в исходный список требований, тем самым расширяя контекст поиска требований на следующих итерациях процесса формирования требований (см. обратную связь, выделенную синим цветом).

Блок сопоставления текста вакансии и текста резюме. В этом блоке осуществляется процесс извлечения (модуль извлечения) и сопоставления (модуль сопоставления) отдельных сущностей из текстов вакансий и резюме, в процессе работы которого осуществляется оценка соответствия резюме и вакансии.

Алгоритмы работы с системой

Алгоритм работы системы в процессе формирования списка требований и уточнение формулировок требований для вакансии представлен на рисунке 5 (слева):

1. ЛПР вводит список первоначальных требований.
2. Система выделяет отдельные сущности требований
3. Система сопоставляет отдельные сущности требований с сущностями структурно-семантической модели на основе семантической близости и векторных представлений текстов, используя предварительно обученные модели.
4. На основе анализа, система формирует и ранжирует рекомендации в виде списка групп смежных требований.
5. ЛПР изучает список рекомендаций и решает, какие из них могут быть добавлены к первоначальному списку требований для конкретной вакансии.
6. Шаги с 1-5 могут повторяться несколько раз, пока список требований не станет удовлетворять ЛПР.
7. ЛПР может выбрать отдельное требование и построить для него дерево синонимов.
8. Система анализирует выбранное требование пользователя и строит дерево связанных с ним синонимичных сущностей.
9. ЛПР анализирует полученное дерево и смотрит какие формулировки для данного требования имеют более употребительную форму.
10. ЛПР вносит изменения в список требований, изменяя формулировки требований, которые он считает целесообразными.
11. Шаги с 7-10 могут повторяться несколько раз, пока список требований не станет удовлетворять ЛПР.

Алгоритм работы системы в процессе сопоставления текстов вакансии с текстами резюме представлен на рисунке 5 (справа):

1. ЛПР выбирает вакансию и резюме, которые он хотел бы сравнить.

2. Система извлекает отдельные сущности требований из текста вакансии и из текста резюме, и проводить их сравнительный анализ.

3. Система выдает оценку пересечения отдельных сущностей знаний и навыков в вакансии и в резюме.

4. На основе предложенной оценки ЛПР принимает решение по данному резюме.

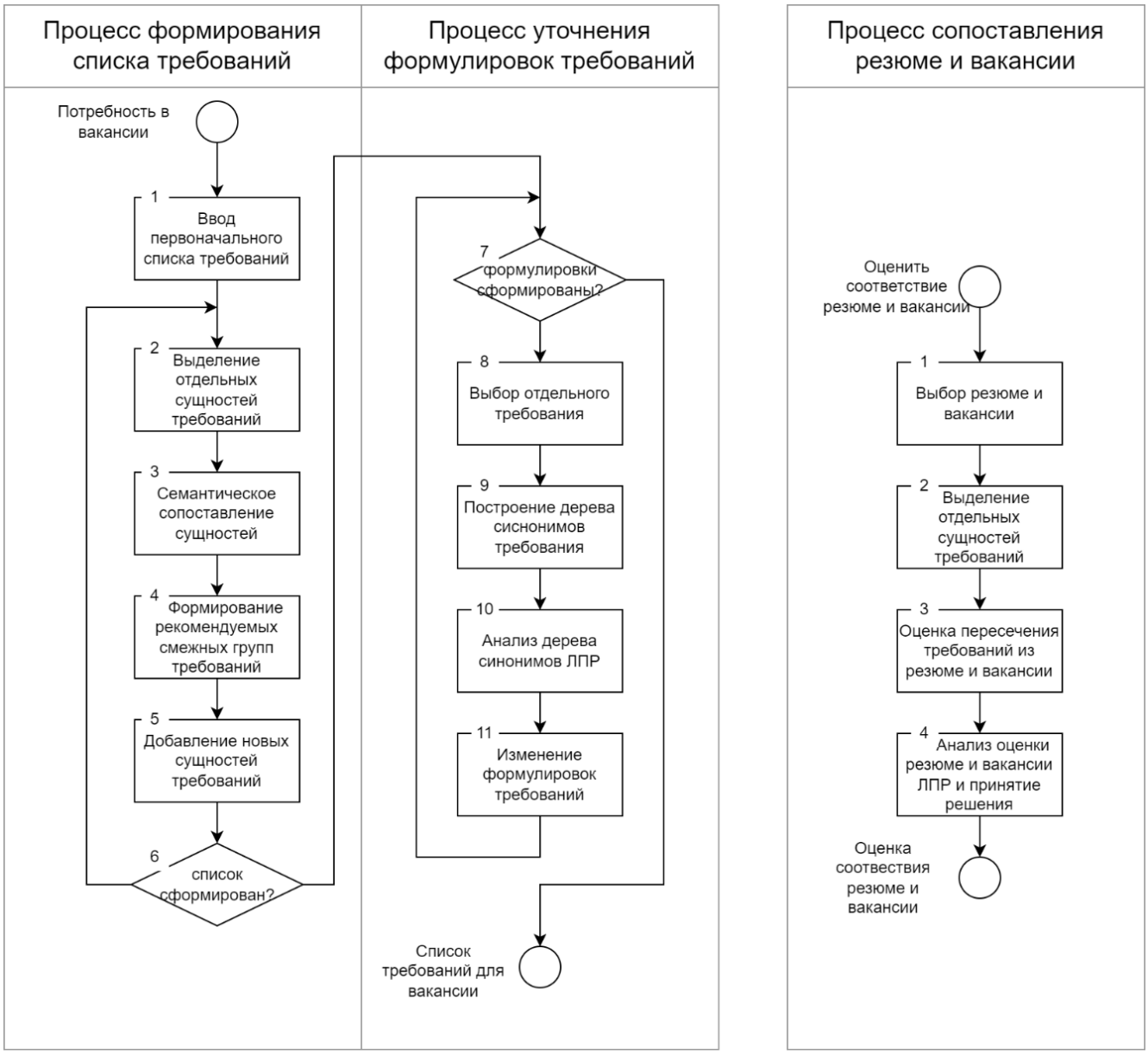

Рисунок 5 – Алгоритмы работы системы: алгоритм формирования списка требований (слева) и алгоритм работы системы сопоставления текстов вакансии с текстами резюме (справа)

Таким образом, ЛПР будет взаимодействовать с системой, вводя первоначальные требования и изучая ее рекомендации. В процессе работы с системой будут анализироваться рекомендации и добавляться новые требования к исходному списку, чтобы получить более полный и точный список требований для конкретной вакансии. Также система будет полезна в процессе сопоставления вакансий и текстов резюме, и позволит быстрее и эффективнее производить первоначальный отбор резюме.

Концепция может быть полезной для ЛПР, связанных с подбором персонала, так как предоставляет им следующие преимущества:

1. Более точные требования. Концепция позволяет более точно исследовать и анализировать требования рынка труда на уровне отдельных сущностей знаний и навыков/компетенций. Это помогает специалистам формировать более точные требования к кандидатам на вакансию, что повышает вероятность успешного подбора персонала.

2. Эффективное сопоставление. Семантическое сопоставление сущностей и методы кластеризации, используемые в данной концепции, позволяют эффективно сопоставлять данные о кандидатах с требованиями вакансии. Это упрощает процесс подбора и помогает выделить наиболее подходящих кандидатов.

3. Улучшенное понимание рынка труда. Концепция предоставляет специалистам более глубокое понимание потребностей рынка труда на уровне отдельных сущностей требований, а также выработки стратегии по подбору квалифицированных специалистов и повышения квалификации имеющихся сотрудников. Это помогает им адаптировать подход к подбору персонала и сориентироваться на актуальные требования рынка.

4. Ускоренный процесс подбора. Благодаря, использованию структурно-семантической модели и методов кластеризации, концепция позволяет автоматизировать и ускорить процесс подбора персонала. Это экономит время и ресурсы специалистов и позволяет им более эффективно выполнять свои обязанности.

Концепция помогает ЛПР, связанным с подбором персонала, повысить эффективность и точность процесса подбора, а также более глубоко понять потребности рынка труда. Это в свою очередь способствует улучшению качества набора персонала и достижению более успешных результатов в сфере подбора персонала.

Анализ рынка труда на уровне отдельных сущностей требований может быть полезен и другим специалистам, не связанным с процессом подбора персонала:

1. Соискатели могут использовать данную концепцию для анализа и изучения требуемых навыков и качеств, указанных в описании вакансий. Это помогает им лучше понять, что именно ищет работодатель, какие навыки и компетенции наиболее ценятся в определенных отраслях или профессиональных областях, и подготовить более целевое и релевантное резюме и мотивационное письмо. Также она может помочь им определить приоритеты в своем профессиональном развитии и сосредоточиться на необходимых навыках для повышения своих шансов на рынке труда.

2. Маркетологи могут использовать эту концепцию для анализа требований вакансий в связи с запуском новых продуктов или услуг. Они могут получить информацию о наиболее востребованных навыках и компетенциях в отрасли, что поможет им разработать более целевые маркетинговые стратегии и сообщения для продукта или услуги.

3. Учебные заведения и академические институты могут применять эту концепцию для анализа потребностей рынка труда и требований к вакансиям в различных сферах. На основе полученных данных преподаватели могут разрабатывать новые и обновлять существующие образовательные программы, и осуществлять более целенаправленную подготовку студентов в соответствии с запросами рынка труда.

4. Государственные организации и регуляторные органы могут использовать эту концепцию для анализа требований вакансий в определенных отраслях и экономических секторах. Они могут использовать эту информацию для

разработки и внедрения политик и программ, направленных на поддержку развития определенных отраслей и создание рабочих мест, соответствующих спросу на рынке.

5. Консультанты, бизнес-аналитики и эксперты в различных сферах бизнеса и менеджмента могут использовать эту концепцию для анализа требований вакансий и определения основных компетенций, необходимых для успешных руководителей и специалистов в определенных областях, а также могут изучить динамику разных отраслей и профессий, и выявить и прогнозировать тенденции, связанные с определенными навыками и компетенциями. Это позволяет им предоставлять более точные и целевые рекомендации клиентам и организациям.

6. Исследователям в области анализа рынка труда. Результаты анализа рынка труда на уровне отдельных сущностей требований могут стать исходным материалом для различных научных исследований. Например, могут быть использованы для проведения исследований, связанных с качеством подготовки работников, проблемами безработицы и поиска работы, оценкой эффективности программ повышения квалификации и многим другим.

Концепция может быть полезна для различных специалистов, помогая им лучше понимать требования и потребности рынка труда в определенных отраслях и сферах, и принимать осознанные решения для достижения лучших результатов.

Ограничения концепции

1. Ограничение данных. Для эффективной работы концепции требуется наличие достаточного объема данных о вакансиях и требованиях к кандидатам. Если данные ограничены или недостаточно репрезентативны, они могут повлиять на точность и надежность результатов семантического сопоставления и кластеризации.

2. Зависимость от качества и точности моделей. Концепция требует построения структурно-семантической модели и использования методов кластеризации. Качество и точность этих моделей и методов могут существенно влиять на качество результатов и, соответственно, на эффективность концепции.

3. Неучтенные контексты. Концепция может столкнуться с проблемой неучета контекста, которая может привести к неправильному пониманию требований вакансий и некорректному сопоставлению с кандидатами. Контекст может включать в себя уникальные особенности компании, культурные и отраслевые специфики, которые может быть сложно учесть в моделях и алгоритмах.

4. Отсутствие субъективности. Модель, основанная на семантическом анализе и кластеризации, может перестать учитывать субъективные оценки и предпочтения работодателей и соискателей. Такие субъективные критерии, как культурная совместимость или коммуникативные навыки, могут быть сложными для учета в модели.

5. Необходимость постоянного обновления модели: Рынок труда, требования и тренды вакансий и навыков постоянно меняются. Это требует регулярного обновления и адаптации модели, чтобы она оставалась актуальной и релевантной.

6. Важно учитывать эти ограничения при разработке и применении концепции, чтобы достичь наилучших результатов и учесть особенности и специфику конкретного контекста.

Концепция информационной поддержки формирования требований вакансии на основе семантического сопоставления сущностей структурно-семантической модели и методов кластеризации позволяет более точно, полно и эффективнее определять требования к вакансиям, учитывая конкретные потребности рынка труда, и как следствие повысить эффективность процесса подбора персонала.

2.2. Структурно-семантическая модель требований рынка труда

Представим модель требований рынка труда в виде неориентированного графа $G_{LM} = (V_{LM}, E_{LM})$, V_{LM} – множество вершин графа, описывающих сущности

рынка труда двух типов: знания и навыки/компетенции, а E_{LM} – множество ребер графа, описывающих связи между сущностями на графе.

Каждая вершина графа $v \in V_{LM}$ определяется следующим кортежем атрибутов $v =< text, pop(v), type(v), emb(v), hyperonim(v) >$:

– $text$ – текст сущности.

– $pop(v)$ – обозначает частотность узла v. Частотность представляет собой числовое значение, отражающее встречаемость данной сущности в текстах вакансий, представленных на рынке труда.

– $type(v)$ – обозначает тип узла v. Тип представляет категорию или класс, к которому относится данная сущность. Может принимать значение «знание» или «навык/компетенция».

– $emb(v)$ – обозначает векторное представление (эмбеддинг) узла v, полученный с помощью нейросетевой модели.

– $hyperonim(v)$ – гипероним (общее понятие) для сущности.

Пример реализации кортежа на примере сущности html 5 представлен на рисунке 6 (справа).

Описание алгоритма поиска гиперонима среди синонимов сущностей (ПГСС):

Пусть $emb(v_i)$ – векторное представление текста i-ой сущности, $pop(v_i)$ – показатель частотности i-ой сущности, E – порог косинусной меры близости для отбора ближайших соседей. Тогда алгоритм ПГСС можно представить в виде последовательности шагов:

Шаг 1. По формуле 2.1. для каждой сущности v_i формируем множество ее ближайших соседей S_i из сущностей v_j в некоторой окрестности E по формуле 1.1 косинусной меры близости

$$S_i = \left\{ j \middle| \cos\left(emb(v_i), emb\left(v_j\right)\right) < E \right\}. \quad (2.1)$$

Шаг 2. Среди множества отобранных вершин S_i определяем вершину-гипероним с наибольшей частотой упоминания, чем у исходной вершины v_i (2.2)

$$hyperonim(v_i) = \begin{cases} S_i = \emptyset \mid null \\ S_i \neq \emptyset \mid v_k, k = \arg\max\limits_{j \in S_i} pop(v_j) \end{cases}. \quad (2.2)$$

Если среди отобранных вершин S_i, есть вершина v_k с большей частотой встречаемости чем у исходной вершины v_i, тогда вершина v_k становится гиперонимом для исходной вершины v_i, если среди S_i нет вершины v_k с большей частотой встречаемости чем у исходной вершины v_i, тогда v_i считается конечной вершиной, и может считаться гиперонимом в данной окрестности E.

Фактически, на данном этапе для каждой вершины на основе векторно-частотного анализа по очень небольшой окрестности определяется наиболее популярный синоним, который в некоторой степени можно считать гиперонимом для рассматриваемой сущности.

Каждое ребро графа $e \in E_{LM}$ определяется следующим кортежем атрибутов $e =< v_s, v_t, w(v_s, v_t) >$:

– v_s, v_t – обозначают концевые сущности, соединенные ребром e. Концевые сущности являются узлами графа.

– $w(v_s, v_t)$ – обозначает вес ребра e. Вес представляет собой числовое значение, которое отражает частоту совместной встречаемости двух связанных сущностей в текстах требований из вакансий.

Таким образом, структурно-семантической модель отдельных сущностей требований реального рынка труда представляет собой сложную сеть, где каждая сущность (например, знания, навыки, компетенции) описывается своим текстом, частотой упоминания в требованиях, типом и векторным представлением, полученным через нейросетевую модель. Эти сущности соединены ребрами, отражающими их совместную встречаемость в текстах вакансий.

Фрагмент структурно-семантической модели отдельных сущностей требований рынка представлен на рисунке 6 (слева).

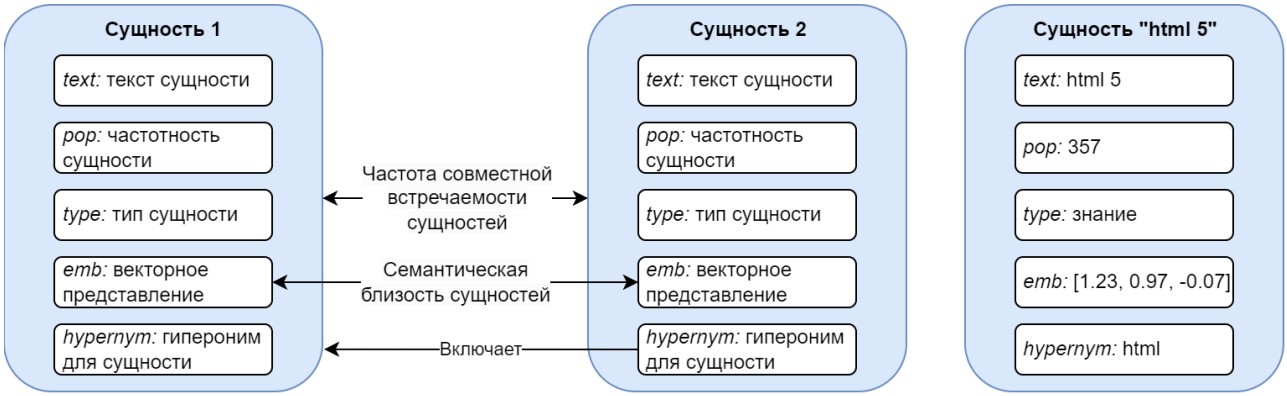

Рисунок 6 – Фрагмент структурно-семантической модели отдельных сущностей требований рынка (слева), пример сущности «html 5» (справа)

2.3. Метод извлечения сущностей знаний и навыков/компетенций из текстов вакансий

Для преодоления проблем, выявленных ранее в главе 1, был разработан метод извлечения отдельных сущностей требований из текстов вакансий реального рынка труда, основанный на нейросетевых моделях языка и методах классификации. Структура метода представлена на рисунке 7.

В предлагаемом методе можно выделить 4 основных этапа:

1. Дообучение языковой модели
2. Выделение текстов требований из текстов вакансий
3. Выделение отдельных сущностей знаний и навыков/компетенций
4. Построения структурно-семантической модели

Рассмотрим каждый из этапов подробнее.

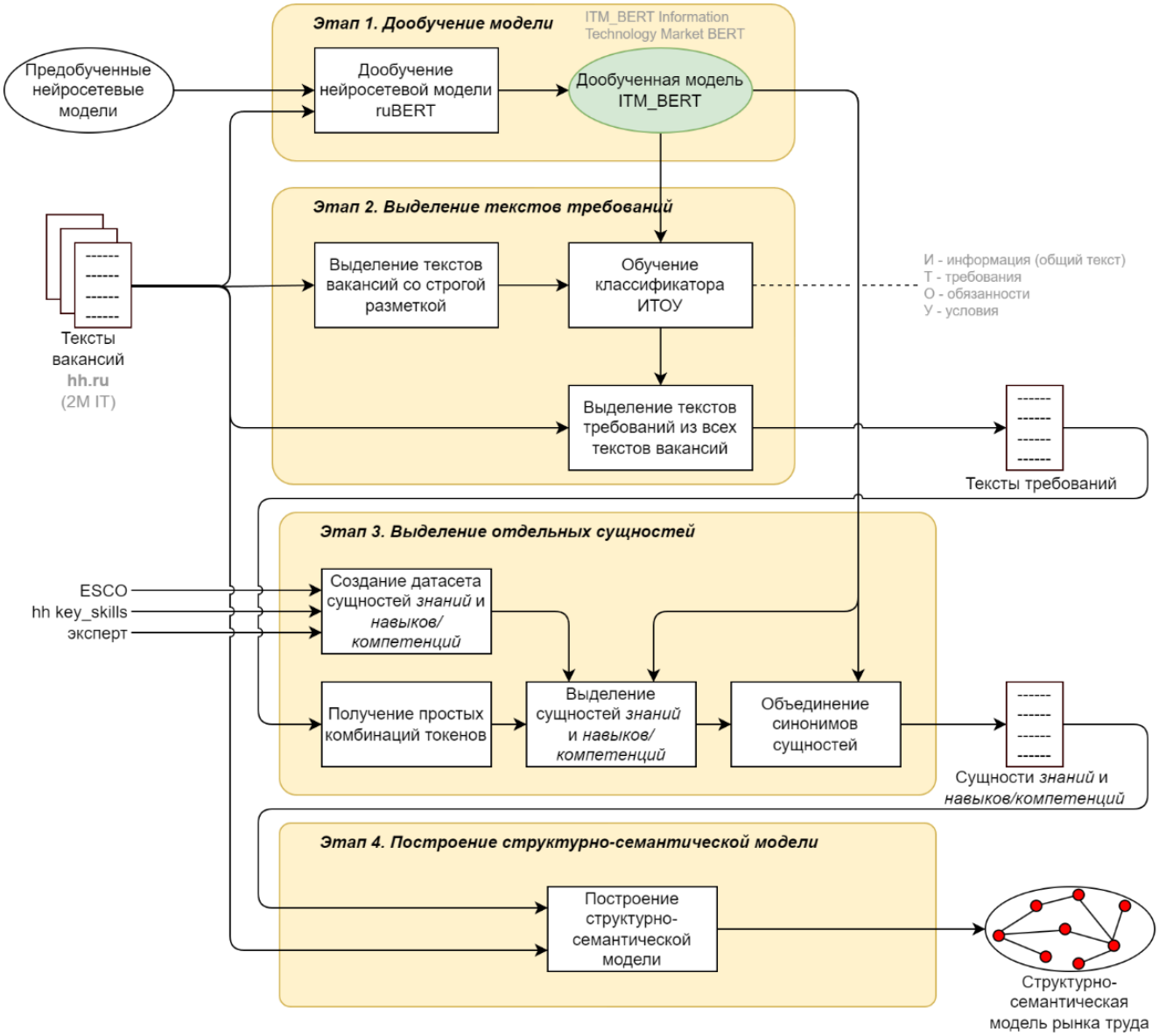

Рисунок 7 – Метод извлечения текстов сущностей отдельных требований из текстов вакансий

Этап 1. Дообучение языковой модели

Модель BERT (англ. Bidirectional Encoder Representations from Transformers) [50]. Разработанная исследователями из Google AI Language в 2018 году, она предназначена для решения разнообразных задач обработки естественного языка (NLP), включая классификацию текста, анализ тональности, извлечение именованных сущностей и множество других приложений.

Отдельные задачи обработки естественного языка традиционно решались с помощью моделей, созданных для каждой конкретной задачи. Модель BERT

произвела настоящую революцию в области обработки естественного языка, и на данный момент является state-of-the-art практически на всех популярных бенчмарках [117, 60, 59].

BERT был обучен на большом текстовом корпусе, состоящим из текстов Википедии (~2,5 миллиарда слов) и Google BooksCorpus (~800 миллионов слов). Эти большие информационные массивы данных способствовали глубокому обучению модели не только для английского языка, но и для 104 других языков. Обучение на таком большом наборе данных заняло достаточно много времени. Обучение модели стало возможным благодаря новой архитектуре Transformer (см. рисунок 8), и благодаря использованию TPU (тензорных процессоров – пользовательской схемы Google, созданной специально для больших моделей ML). В итоге 64 TPU обучали модель в течение 4 дней.

Архитектура Transformer обеспечивает высокую степень параллелизма в процессе обучения машинного обучения, что позволяет BERT эффективно обрабатывать большие объемы данных в короткие сроки.

Механизм внимания, лежащий в основе архитектуры трансформера, значительно улучшает способность модели распознавать контекстуальные зависимости между словами или их частями в тексте. Идея этого механизма, впервые представленная в статье 2017 года «Внимание – это все, что вам нужно» (англ. Attention is all you need) [111], привела к широкому использованию трансформеров в языковых моделях по всему миру. В общем случае архитектура трансформера состоит из двух основных компонентов: кодировочного (encoder), который анализирует входной текст, и декодировочного (decoder), который генерирует прогноз для задачи (см. рисунок 8). Каждый из этих компонентов может быть представлен в виде нескольких слоев (см. рисунок 9).

Поскольку целью BERT является создание языковой модели, способной кодировать текстовую информацию во внутреннее векторное пространство, то ей необходим только кодирующий компонент. Общая структура модели BERT представлена на рисунке 9.

BERT относится к классу т.н. автокодировщиков (англ. autoencoder) – специальная архитектура искусственных нейронных сетей, позволяющая применять обучение без учителя с использованием метода обратного распространения ошибки. В отличие от направленных моделей, кодирующий компонент трансформера BERT считывает всю последовательность слов целиком, а не по порядку. Это позволяет модели анализировать контекст слова, включая все его окружение, а не только левую или правую часть.

Все слои кодирующего компонента BERT имеют одинаковую структуру (см. рисунок 10), но различаются весами. Входная последовательность проходит через слой внутреннего внимания (англ. self-attention), который позволяет кодирующему компоненту анализировать другие слова в предложении при кодировании конкретного слова. Результат слоя внутреннего внимания передается в нейронную сеть прямого распространения (англ. feed-forward neural network), которая применяется к каждому слову в предложении независимо.

Каждый слой кодирующего компонента использует набор векторов размерностью 512, где для первого слоя это векторные представления токенов, а для последующих слоев – выходные вектора нижележащих слоев кодирования.

При вводе текста в сеть сначала происходит его токенизация. В модели BERT часто используется *токенизация на подслова*, где текст преобразуется в последовательность токенов (элементарные части текста – слова или их части), чтобы модель могла его обработать. Токенами служат слова из словаря или их части, если слово не найдено в словаре, оно разбивается на субслова. Для этого могут использоваться алгоритмы WordPiece [114] и Byte Pair Encoding (BPE) [58]. Словарь является частью модели, например, в модели BERT-Base словарь состоит из ~30,000 токенов.

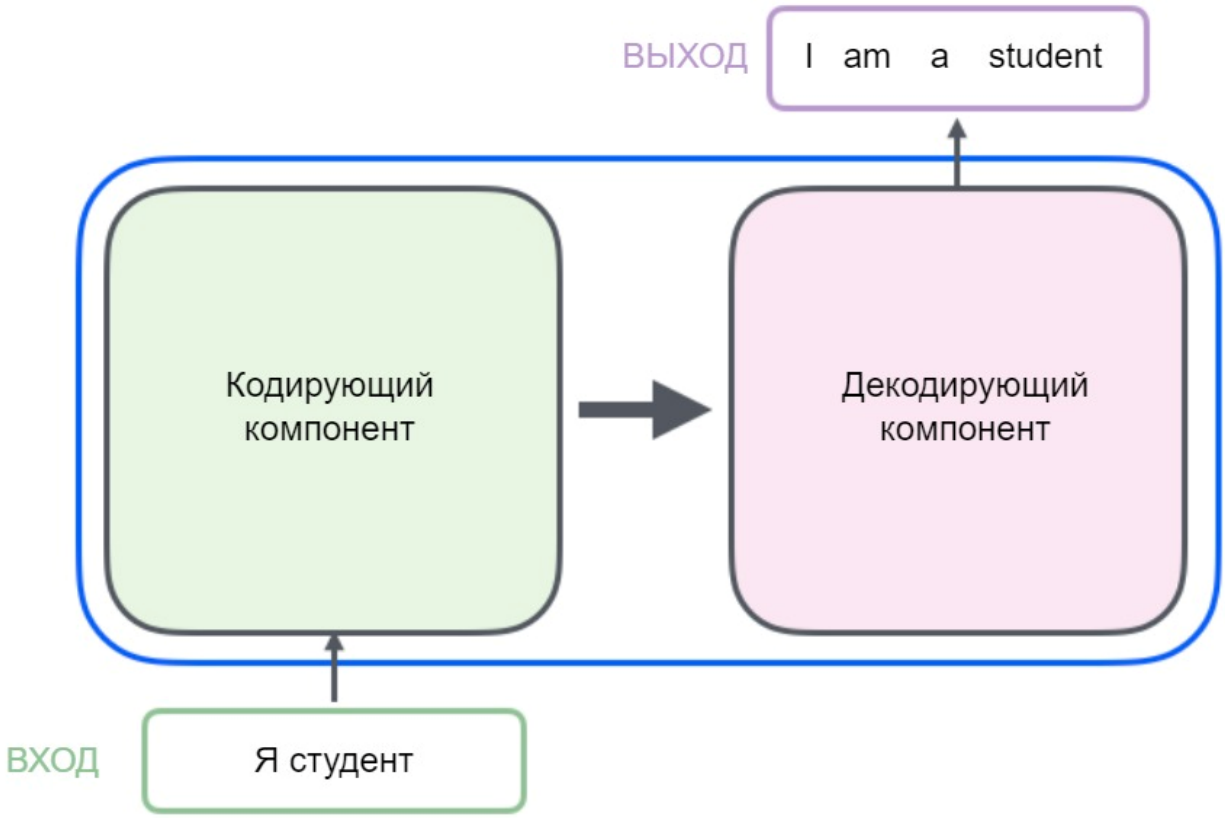

Рисунок 8 – Общая структура трансформеров

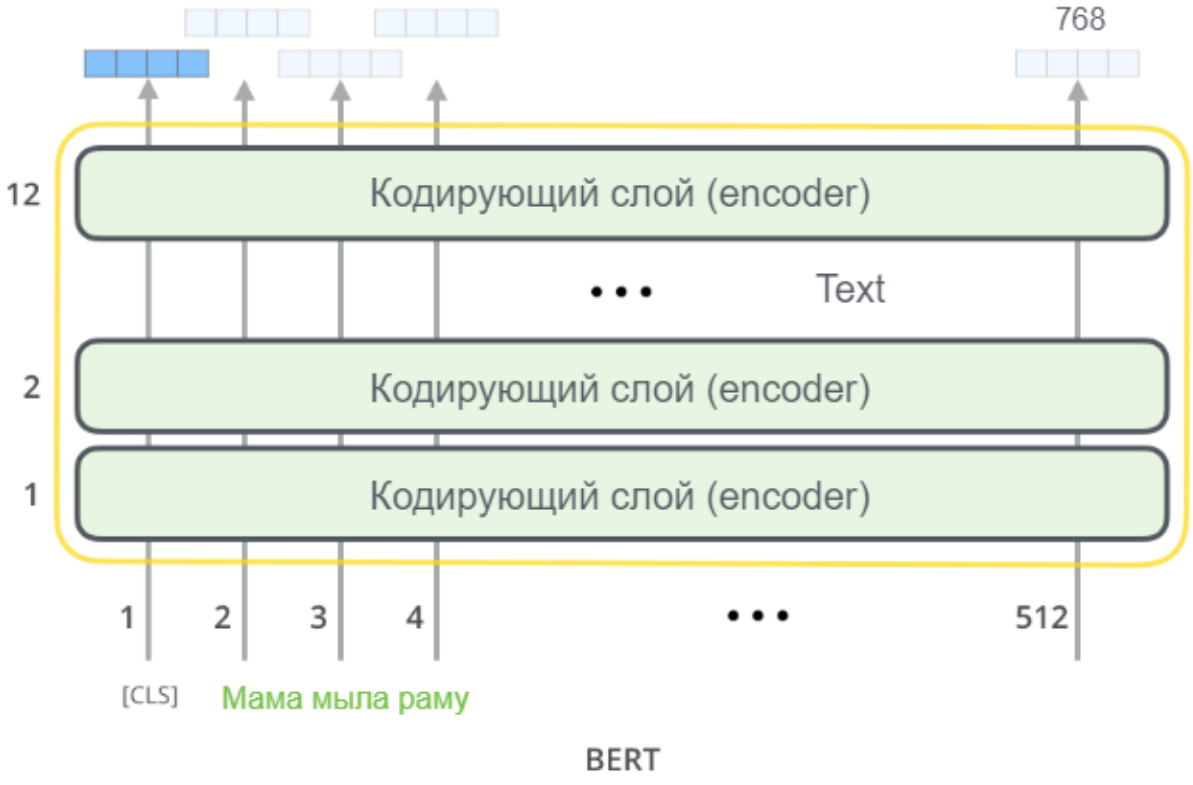

Рисунок 9 – Общая структура BERT

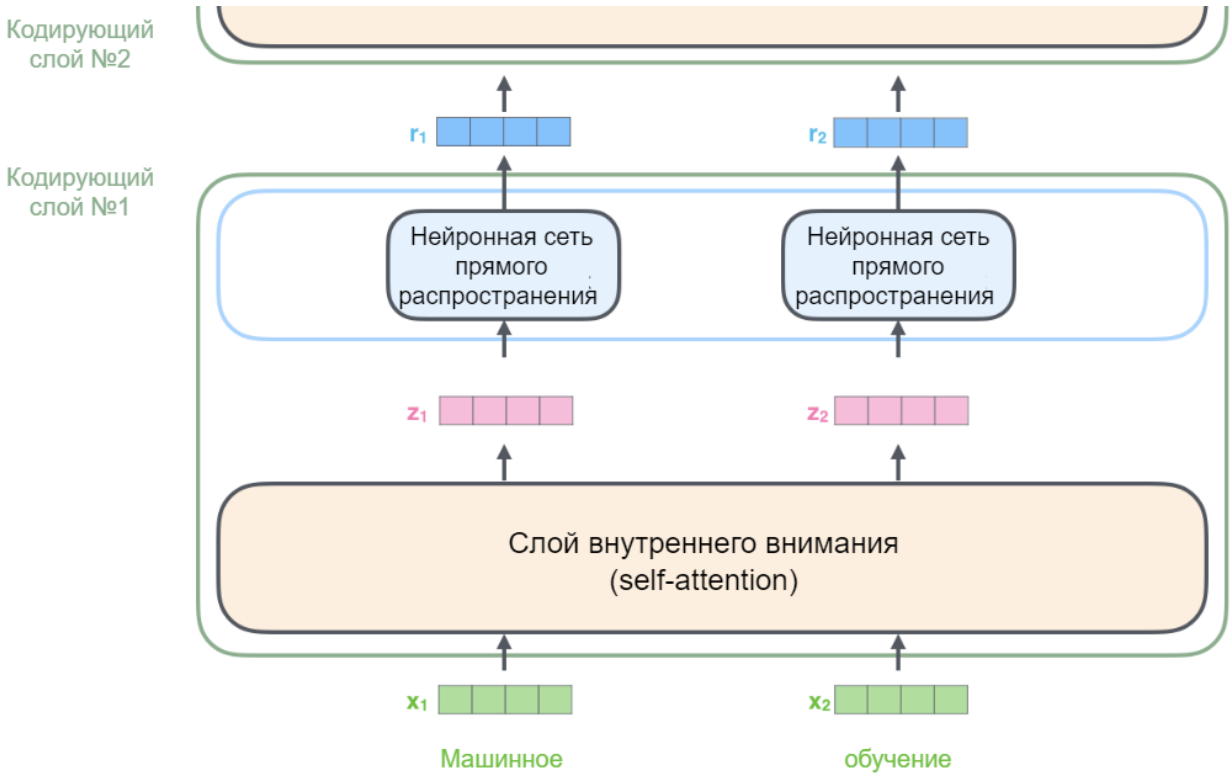

Рисунок 10 – Структура слоя кодирующего компонента (encoder)

В нейронной сети токены кодируются их векторными представлениями, объединяя представления самого токена, его номер в предложении и позицию в предложении. Входные данные обрабатываются параллельно, сохраняя информацию о взаимном расположении слов в предложении и в позиционной части векторного представления каждого токена.

Токенизация позволяет преобразовать произвольный текст в структурированный набор токенов для последующей обработки моделью.

Механизм внутреннего внимания позволяет модели фокусироваться на наиболее важных частях входных данных при генерации выхода.

В трансформерах внутреннее внимание реализуется с помощью механизма самовнимания (self-attention). Каждый токен входной последовательности соотносится с каждым токеном этой же последовательности для вычисления весов внимания. Эти веса внимания затем используются при вычислении представления

каждого токена на выходе: токены с бóльшим весом внимания оказывают большее влияние на результат.

Формально, если есть входная последовательность токенов $x_1, \dots, x_n$, то для каждой пары токенов (x_i, x_j) вычисляется вес внимания α_{ij}, который показывает, насколько важен токен x_j при вычислении представления токена x_i. Этот вес вычисляется на основе соответствия между векторными представлениями токенов.

Затем полученные веса внимания используются для вычисления контекстных векторов каждого токена перед последующими слоями сети. Таким образом модель фокусируется на наиболее релевантных элементах входной последовательности.

На примере предложения «The animal didn't cross the street because it was tootired» (см. рисунок 11) ясно видно двусмысленность, к чему относится слово «оно» (it) – к улице или к животному. Этот простой вопрос для человека является сложной задачей для алгоритма машинного перевода. Механизм внутреннего внимания в архитектуре трансформера помогает модели определить, что «оно» (it) относится к слову «животное» (animal). Анализируя каждое слово входного предложения, модель использует внутреннее внимание для учета контекста и интерпретации текущего слова.

В отличие от рекуррентных нейронных сетей, где при обработке очередного слова учитывается скрытое состояние от предыдущих слов, механизм внутреннего внимания в трансформере анализирует сразу весь контекст предложения. Это позволяет модели точнее определять смысл и связи между словами при переводе.

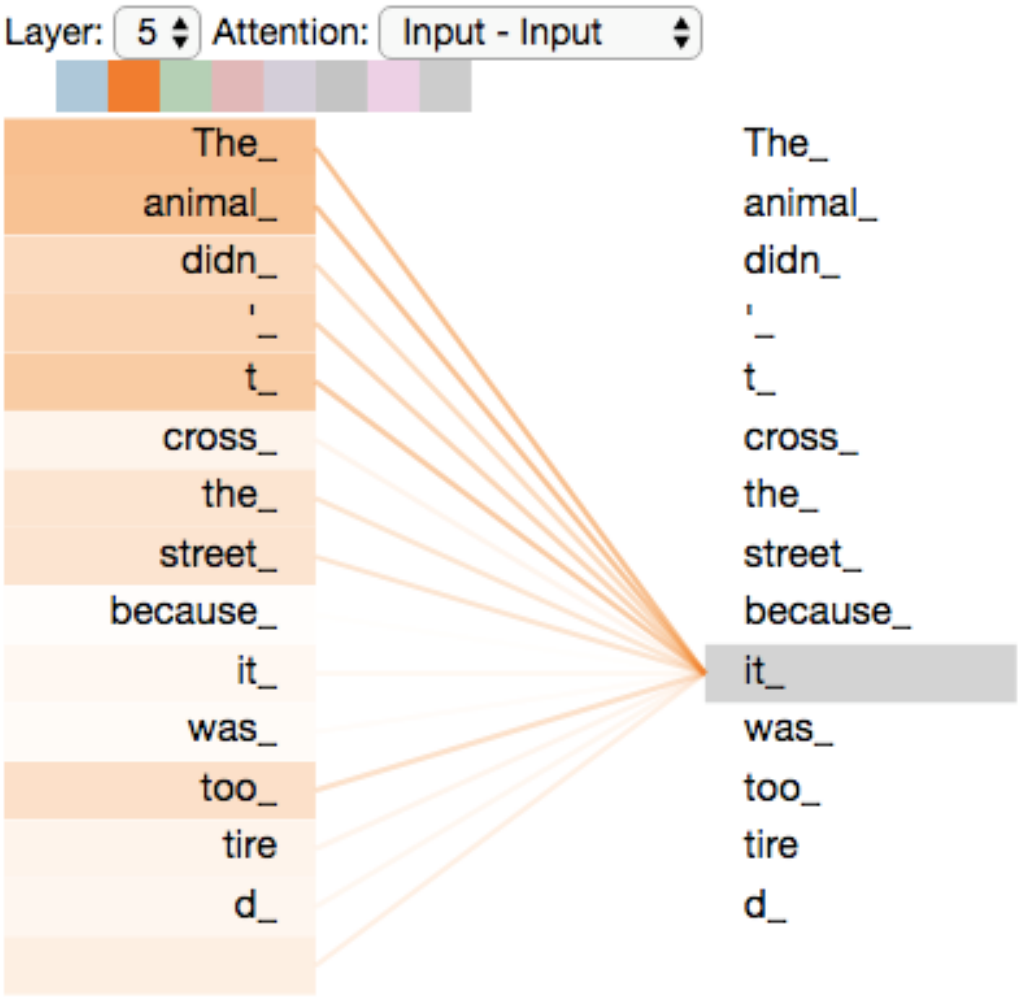

Рисунок 11 – Иллюстрация механизма внутреннего внимания

Предобучение и дообучение нейронной сети BERT.

В нейронных сетях на архитектуре трансформеров обучение можно разбить на две части – «предобучение» (англ. pre-train) и «дообучение» («тонкая настройка», англ. fine-tuning).

Во время этапа «предобучения» модель анализирует общие закономерности и структуру языка, используя обширные объемы текстовых данных без разметки, таких как тексты из Википедии, книг или веб-страницы. В процессе предобучения модель решает задачи предсказания замаскированных слов (MLM) или следующего предложения (NSP), что позволяет ей строить информативные векторные представления слов и предложений с учетом контекста. Предобученная модель способна решать базовые языковые задачи, но не заточена под конкретную

проблему. Предобучение требует большого объема данных и вычислительных ресурсов, занимает продолжительное время (дни и недели), но результатом является универсальная языковая модель.

На этапе дообучения предобученная модель адаптируется для решения конкретной задачи на меньшем объеме размеченных данных. К предобученной модели добавляются слои, специфичные для целевой задачи (например, классификация, извлечение сущностей), и модель дообучается на размеченном датасете. При дообучении настраиваются как новые, так и предобученные слои, что позволяет переносить знания из общей модели на конкретную задачу и получать прирост качества по сравнению с обучением с нуля. Дообучение требует меньше данных и ресурсов, чем предобучение, занимает меньше времени (часы или дни) и дает в результате модель, заточенную под специфику конкретной задачи и домена.

В работах BioBERT [75], LegalBERT [46] и SciBERT [98] показано, что дообучение на специфических текстах предметной области предобученных нейросетевых языковых модели на архитектуре трансформеров существенно повышает качество использования этой модели в задачах NLP при работе с текстами этой предметной области..

Предобучение и дообучение – это два последовательных этапа трансферного обучения трансформеров, которые позволяют эффективно решать широкий спектр задач обработки естественного языка.

Существует два основных метода обучения/дообучения для языковой модели BERT:

1. MLM (англ. Masked Language Modeling) – маскированное моделирование языка. При этом случайно выбираются некоторые токены во входной последовательности и маскируются (заменяются на [MASK]). Модель должна предсказать эти замаскированные токены на основе контекста. Это позволяет BERT научиться контекстному представлению слов.

2. NSP (англ. Next Sentence Prediction) – предсказание следующего предложения. Для обучения берутся пары предложений – настоящие

последовательные предложения текста и случайные. Модель должна определить, являются ли предложения последовательными друг за другом. Это задача «да/нет». Она нужна, чтобы BERT учился моделировать интер-сентенциальные связи в текстах.

Задачи MLM и NSP в совокупности учат BERT понимать контекстные связи и семантику как внутри предложений, так и между последовательными предложениями в текстах.

В статье о BERT [50] описан новый метод предобучения языковых моделей – «маскированное моделирование языка» (англ. MLM – Masked Language Modeling).

Суть его заключается в следующем. Из исходного текста случайным образом выбирается 15% токенов (слов). 80% этих токенов заменяются на специальный символ [MASK], 10% – на случайные слова из словаря, а 10% остаются без изменений. Затем модель BERT должна предсказать те токены, которые были заменены на [MASK], опираясь на окружающий контекст из немаскированных токенов.

Также в статье [50] авторами предлагается идея использования комбинации из 80% масок [MASK], 10% случайных слов и 10% оригинальных слов как наилучший способ при обучении BERT:

– Применение 100% масок [MASK] привело бы к тому, что модель оптимизировалась бы в большей степени только для предсказания замаскированных токенов. При этом контекст из немаскированных токенов использовался бы недостаточно для формирования хороших векторных представлений всех слов.

– Использование в 90% случаев [MASK] и в 10% случайных слов научило бы модель тому, что реальное слово никогда не является правильным ответом. Это неверная установка.

– Применение в 90% масок [MASK] и в 10% оригинальных слов позволило бы модели тривиально копировать векторное представление слова без учёта контекста.

Выбранная комбинация коэффициентов 80%/10%/10%, чтобы одновременно обучить модель восстанавливать замаскированные слова по контексту и формировать качественные векторные представления для всех слов, чтобы соблюсти баланс между хорошим моделированием контекста и качественным векторным представлением всех слов. Возможно, небольшая вариация этих коэффициентов могла бы дать ещё лучшие результаты.

Процесс предсказания масок в BERT использует классификатор, преобразующий выход кодировщика в вероятности слов из словаря с помощью softmax.

Такой подход к предобучению позволяет получить мощные контекстные векторные представления слов и формировать глубокое понимание языка.

Этот метод предварительного обучения позволяет создать сильные контекстные векторные представления слов и развить глубокое понимание языка.

Технически, для предсказания выходных слов требуется (см. рисунок 12):

– добавление классификационного слоя над выходными данными кодировщика;

– умножение выходных векторов на матрицу векторных представлений словаря (англ. embedding matrix) для приведения их к размерности словаря;

– оценка вероятности каждого слова в словаре при помощи softmax.

В расчете функции потерь учитываются прогнозы только замаскированных значений, в то время как прогнозы не замаскированных слов игнорируются.

Модели BERT и их модификации стали очень популярны в последние годы благодаря своей эффективности в решении широкого спектра задач NLP. В России есть несколько проектных групп и компаний, которые занимаются разработкой моделей по типу BERT и их применением для решения различных задач NLP.

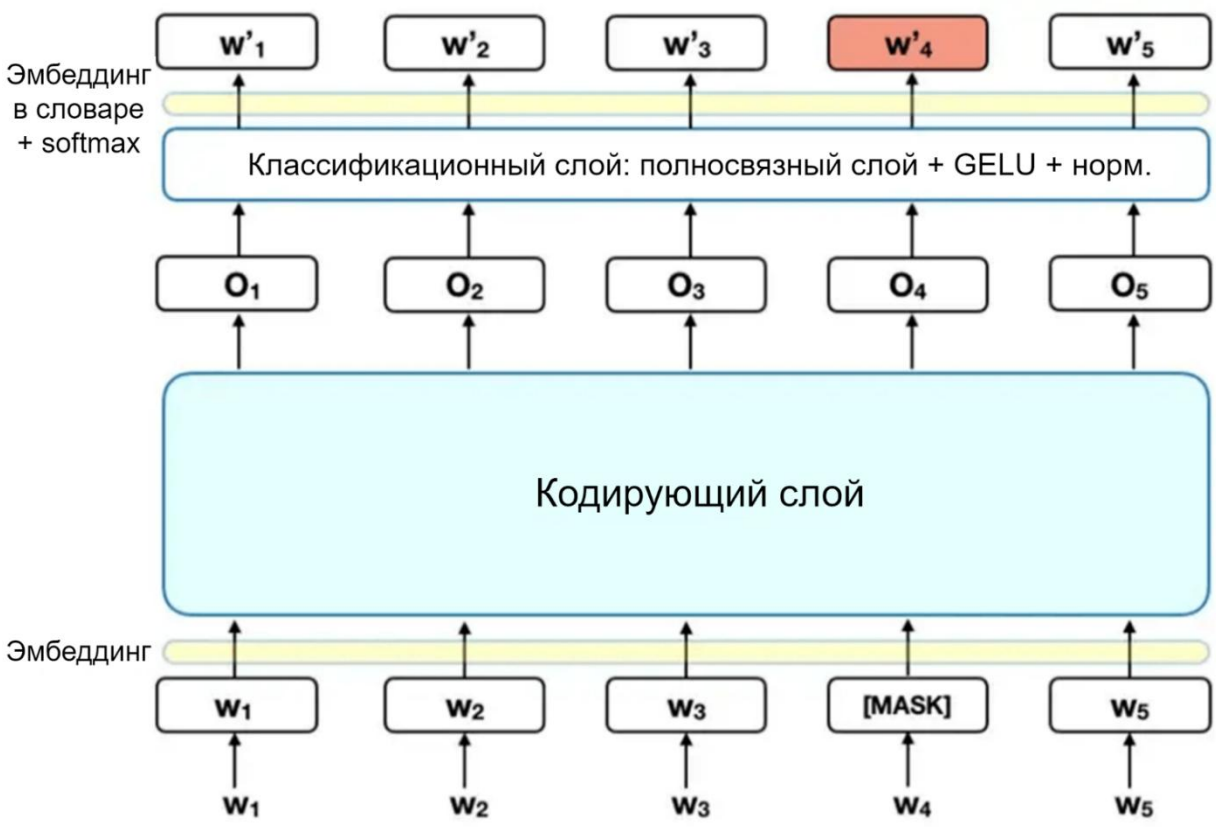

Рисунок 12 – Обучение BERT по методу MLM

Команда Сбербанка (Sberbank-ai) обучила и выложила в открытый доступ модель ruBERT в нескольких модификациях. Эти модели используется в различных проектах компании Сбербанк, связанных с NLP.

В рамках проекта DeepPavlov были разработаны русскоязычные модели BERT, такие как RuBERT и Russian SuperGLUE BERT, которые доступны для свободного использования. Также существует одноименная открытая библиотека для разработки диалоговых систем и решения задач NLP, созданная командой исследователей из МФТИ, МГУ и Сколтеха.

Кроме того, есть и другие компании, исследовательские институты и университеты в России, которые занимаются разработкой и использованием моделей типа BERT, например, Mail.ru Group, Yandex, МФТИ, МГУ, НИУ ВШЭ и др.

Так как процесс предобучения языковых моделей «с нуля» является крайне трудозатратой задачей, в рамках данного исследования использовались

предобученные модели от проектной команды Сбер, как наиболее востребованные и точные по оценке портала huggingface.com [17].

Оценка качества обучения/дообучения модели.

В статье [39] авторы предложили использовать принцип максимальной энтропии для построения вероятностных моделей языка. Они ввели метрику перплексии как способ оценки качества этих моделей.

Перплексия определяется по формуле 2.3 как экспонента средней логарифмической функции правдоподобия модели на тестовых данных:

$$Perplexity(W) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p(w_i|w_1,\dots,w_{i-1})\right) \qquad (2.3)$$

где $W = w_1,\dots,w_{i-1}$ – последовательность слов в тестовом корпусе, $p(w_i|w_1,\dots,w_{i-1})$ – вероятность, присвоенная моделью слову w_i при условии предыдущих слов.

Интуитивно, чем лучше модель предсказывает каждое следующее слово в последовательности (чем выше вероятность), тем ниже перплексия. Модели с меньшей перплексией на тестовых данных считаются обученными лучше для задач моделирования языка.

В настоящее время перплексия стала стандартной метрикой для оценки и сравнения языковых моделей, особенно в задачах распознавания речи, машинного перевода и генерации текста.

Этап 2. Выделение текстов требований из текстов вакансий

На данном этапе предлагается отобрать вакансии с жесткой структурой (см. пример вакансии с жесткой структурой приложении Г). Затем на основе отобранных вакансий сформировать датасет отдельных предложений, где каждое предложение за счет жесткой структуры из вакансий будет относиться к одному из 4 классов: «общий текст», «требования», «обязанности», «условия работы».

Следующим шагом необходимо будет обучить классификатор на основе дообученной на предыдущем этапе нейросетевой модели BERT.

Оценка качества многоклассовой классификации включает рассмотрение различных параметров и метрик, чтобы понять, насколько хорошо модель справляется с классификацией в разных классах. Результаты измерения качества обычно представляются в виде матрицы ошибок (англ. confusion matrix).

Матрица ошибок для многоклассовой классификации представляет собой расширение стандартной матрицы ошибок, используемой для бинарной классификации. Она позволяет анализировать, как классификатор работает по отношению ко всем классам в задаче многоклассовой классификации. Каждая строка матрицы соответствует истинным классам, а каждый столбец – предсказанным классам. Элементы матрицы указывают на количество примеров, которые были классифицированы определенным образом.

Для N-классовой классификации матрица будет иметь размер N×N, где N – количество классов. В такой матрице:

– диагональные элементы (от верхнего левого к нижнему правому) показывают количество правильных предсказаний для каждого класса. Это означает, что для i-го класса элемент матрицы в i-й строке и i-м столбце показывает, сколько раз классификатор правильно определил i-й класс.

– недиагональные элементы указывают на ошибки классификации: элемент в i-й строке и j-м столбце показывает, сколько раз модель неправильно классифицировала примеры истинного i-го класса как принадлежащие j-му классу.

Пример матрицы ошибок для трехклассовой задачи (где классы обозначены как A, B, и C) представлен в таблице 4.

В этом примере:

– класс A был правильно предсказан 5 раз, но ошибочно классифицирован как класс B 2 раза и как класс C 1 раз.

– класс B был правильно предсказан 7 раз, но ошибочно классифицирован как класс C 3 раза (и не был ошибочно классифицирован как класс A).

– класс C был правильно предсказан 5 раз, но ошибочно классифицирован как класс A 1 раз и как класс B 4 раза.

Таблица 4 – Пример таблицы ошибок

	Предсказанный класс A	Предсказанный класс B	Предсказанный класс C
Истинный класс A	5	2	1
Истинный класс B	0	7	3
Истинный класс C	1	4	5

Матрица ошибок позволяет оценить не только общую точность модели, но и то, какие конкретные ошибки и как часто допускает модель, что очень важно для понимания её поведения и для дальнейшего улучшения производительности на конкретных классах.

Используя значения из таблицы ошибок, можно вычислить различные метрики оценки качества, включая точность (precision), полноту (recall) и F1-меру для каждого класса. Ниже приведены формулы для этих метрик (2.4, 2.5, 2.6):

Для некоторого класса A:

$$Precision_A = \frac{TP_A}{TP_A + FP_A}, \tag{2.4}$$

$$Recall_A = \frac{TP_A}{TP_A + FN_A}, \tag{2.5}$$

$$F1_A = \frac{2 \cdot Precision_A \cdot Recall_A}{Precision_A + Recall_A}. \tag{2.7}$$

Аналогичные формулы могут быть применены и для других классов B и C.

Одним из способов оценки качества многоклассовой классификации является усреднение метрик по всем классам. Однако, в зависимости от выбранного подхода, могут быть использованы разные методы усреднения – микро-усреднение и макро-усреднение.

Микро-усреднение (micro-averaging) (2.7, 2.8, 2.9):

$$Precision_{micro} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FP_i}, \tag{2.7}$$

$$Recall_{micro} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FN_i}, \quad (2.8)$$

$$F1_{macro} = \frac{2 \cdot Precision_{micro} \cdot Recall_{micro}}{Precision_{micro} + Recall_{micro}}. \quad (2.9)$$

Макро-усреднение (macro-averaging) (2.10, 2.11, 2.12):

$$Precision_{macro} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i}, \quad (2.10)$$

$$Recall_{macro} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i}, \quad (2.11)$$

$$F1_{macro} = \frac{2 \cdot Precision_{macro} \cdot Recall_{macro}}{Precision_{macro} + Recall_{macro}}, \quad (2.12)$$

где N – количество классов, TP_i – количество истинно положительных примеров для класса i, FP_i – количество ложно положительных примеров для класса i, FN_i – количество ложно отрицательных примеров для класса i.

Основное отличие между микро-усреднением и макро-усреднением заключается в том, как усредняются метрики по классам и как каждый класс вносит свой вклад в итоговый результат.

Микро-усреднение (micro-averaging)

В методе микро-усреднения все примеры и неверные классификации из различных классов объединяются в единый комплекс. При этом, для расчета таких показателей как точность, полнота и F1-мера, используется число корректно определенных примеров в сравнении с общим числом примеров по всем классам. В рамках этого подхода, каждый отдельный пример оказывает равное влияние на конечную оценку, что делает его особенно подходящим для анализа общей эффективности модели без учета размеров классов или их баланса.

Макро-усреднение точности (macro-averaging)

В макро-усреднении каждый класс рассматривается отдельно, а затем метрики усредняются по всем классам. Метрики, такие как точность, полнота и F1-мера, для каждого класса вычисляются отдельно, а затем усредняются по всем

классам. В макро-усреднении каждый класс вносит одинаковый вклад в итоговый результат. Этот подход полезен, если важно оценивать производительность именно каждого класса независимо от общих размеров классов.

Взвешенная F1-мера (Weighted F1-score)

Взвешенная F1-мера усредняет F1-меру для каждого класса с учетом их относительных размеров в наборе данных. Формула 2.13 для расчета взвешенной F1-меры:

$$WeightedF1_{score} = \frac{\sum_{i=1}^{N}(F1\_score_i \cdot count_i)}{\sum_{i=1}^{N} count_i}, \quad (2.13)$$

где N – количество классов, $F1\_score_i$ – F1-мера для класса i, $count_i$ – количество примеров класса i.

Этап 3. Выделение отдельных сущностей знаний и навыков/компетенций

Этот этап включает в себя упрощение сложных предложений до простых словосочетаний, добавляя новые связи между словами.

Последовательность шагов в рамках этапа 3.

Шаг 1. Необходимо подготовить набор размеченных текстовых данных отдельных сущностей требований знаний и навыков/компетенций.

Шаг 2. Применение правил для yargy-парсера для выделения сведений о знаниях и навыках, как показано на рисунке 13.

Шаг 3. Переход от сложного текста требования к набору простых комбинаций токенов. На данном шаге тестировалось два варианта получения простых комбинаций токенов:

– линейная комбинация слов (токенов);

– дополнение дерева синтаксического разбора новыми ребрами + случайное блуждание по дереву.

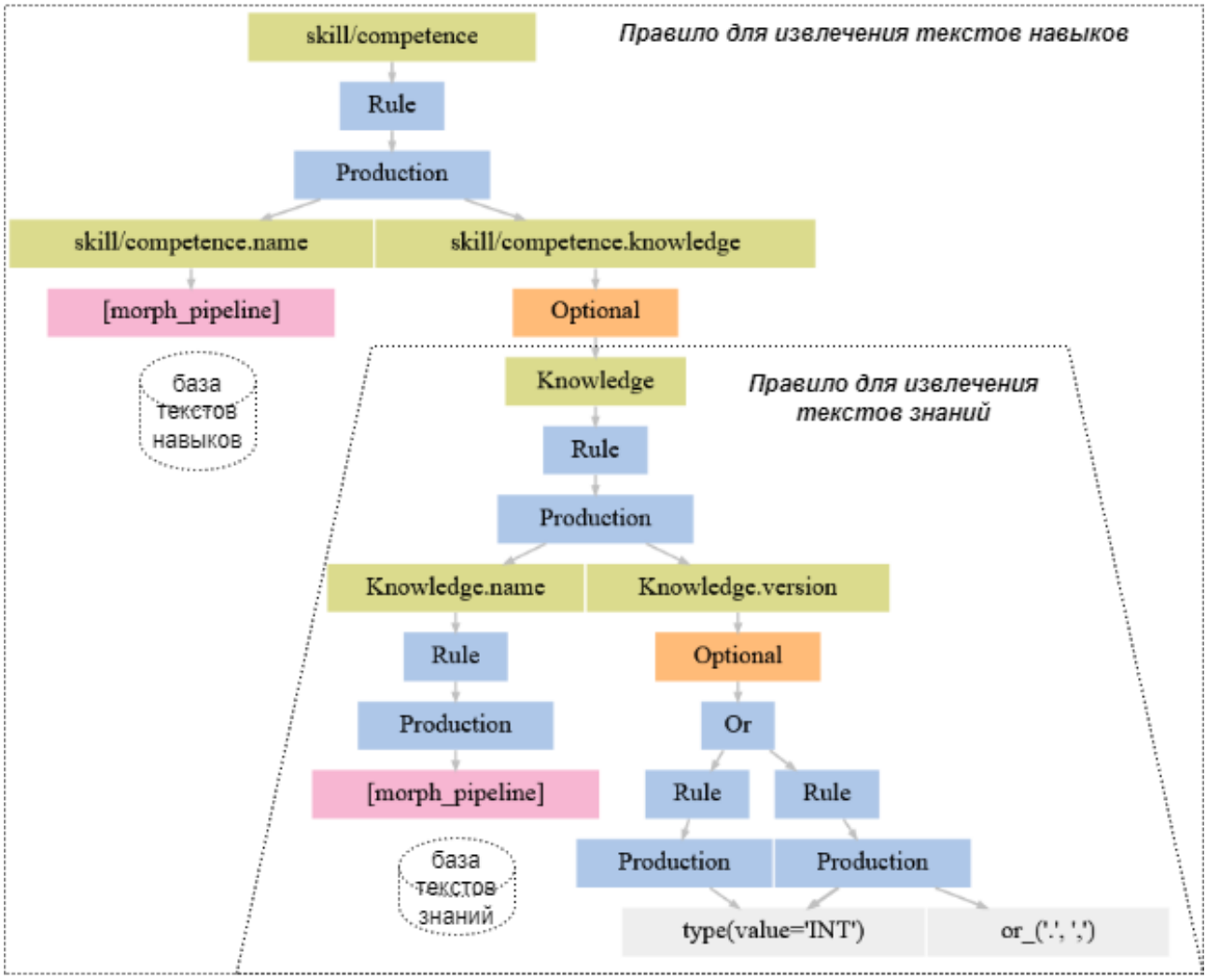

Рисунок 13 – Структура правила yargy-парсера для извлечения текстов знаний и навыков/компетенций

Метод формирования линейных комбинаций токенов основан на создании последовательностей, где токены соединяются в соответствии с их последовательным расположением, при этом каждый следующий токен в последовательности имеет больший порядковый номер, чем предшествующий. Эта техника исходит из идеи, что передача информации в тексте в основном происходит слева направо. В контексте этой методики, токены в тексте пронумерованы, и на основе их номеров создаются линейные последовательности токенов. Пример работы этого метода демонстрируется на иллюстрации под номером 14.

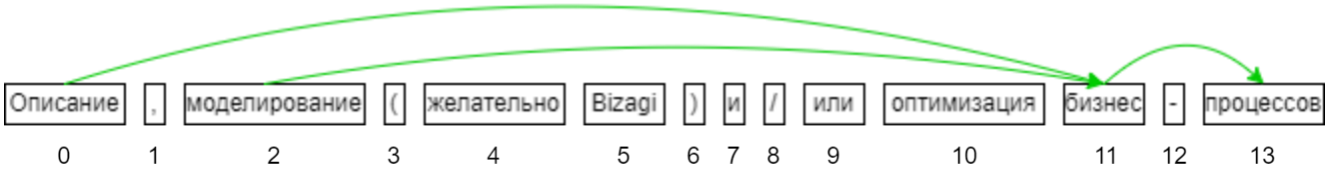

Рисунок 14 – Схема получения линейных комбинаций токенов

Второй вариант получения простых комбинаций токенов – дополнение дерева синтаксического разбора новыми ребрами. Этот подход основан на предположении, что дообученная на текстах некоторой предметной области языковая модель способна находить новые связи между отдельными узлами дерева синтаксического разбора, которые по оценке языковой модели, имеют высокую вероятность совместной встречаемости в текстах этой предметной области.

В ходе экспериментов для создания деревьев синтаксического анализа были применены две различных модели: spacy_syntax_parser [23] и deeppavlov_syntax_parser [24], примеры полученных деревьев представлены на рисунке 15. Когда в эти деревья добавляются новые связи, они преобразуются в графы.

Далее, используя метод случайного блуждания, на этих графах выбираются пути длиной от одного до пяти токенов, на основе которых затем формируются различные комбинации токенов. Случайное блуждание по графу – это процесс, начинающийся с выбора некоторой начальной вершины, из которой на каждом шаге случайным образом выбирается одно из исходящих ребер, и блуждание продолжается до достижения заданной длины пути, после чего сгенерированный путь сохраняется. Каждый сгенерированный путь представляет собой последовательность вершин, которые и являются искомыми комбинациями. Этот процесс повторяется многократно для генерации большого количества путей и комбинаций токенов. При этом длина пути может выбираться случайно из заданного диапазона для каждого блуждания, а стратегии выбора следующей вершины могут учитывать веса ребер, отдавая предпочтение более частым переходам. Сгенерированные комбинации токенов могут дополнительно

фильтроваться или взвешиваться по различным критериям, таким как частота или осмысленность. Случайное блуждание по графу позволяет генерировать множество потенциально значимых комбинаций токенов, учитывая их совместную встречаемость в исходных текстах, что может быть полезно для различных задач обработки естественного языка, таких как извлечение ключевых фраз, тематическое моделирование или расширение запросов.

Шаг 5. Комбинации токенов передаются в yargy-парсер, который, опираясь на правила, определенные на шаге 2, выделяет тексты, относящиеся к отдельным единицам сущностей требований.

Важно отметить, что в будущем для извлечения именованных сущностей, требований, может применяться дообученная для этой задачи нейросетевая модель BERT, что станет альтернативой использованию парсера на основе контекстно свободных грамматик.

Таблица 5 – Пример сложного требования и простых текстов сущностей требований

Текст сложного требования	Сущности знаний и навыков/компетенций	Разметка
Описание, моделирование (желательно Bizagi) и/или оптимизация бизнес-процессов;	описание бизнес-процессов	навык/компетенция
	моделирование бизнес-процессов	навык/компетенция
	оптимизация бизнес-процессов	навык/компетенция
	описание процессов	навык/компетенция
	оптимизация процессов	навык/компетенция
	моделирование процессов	навык/компетенция
	Bizagi	знание

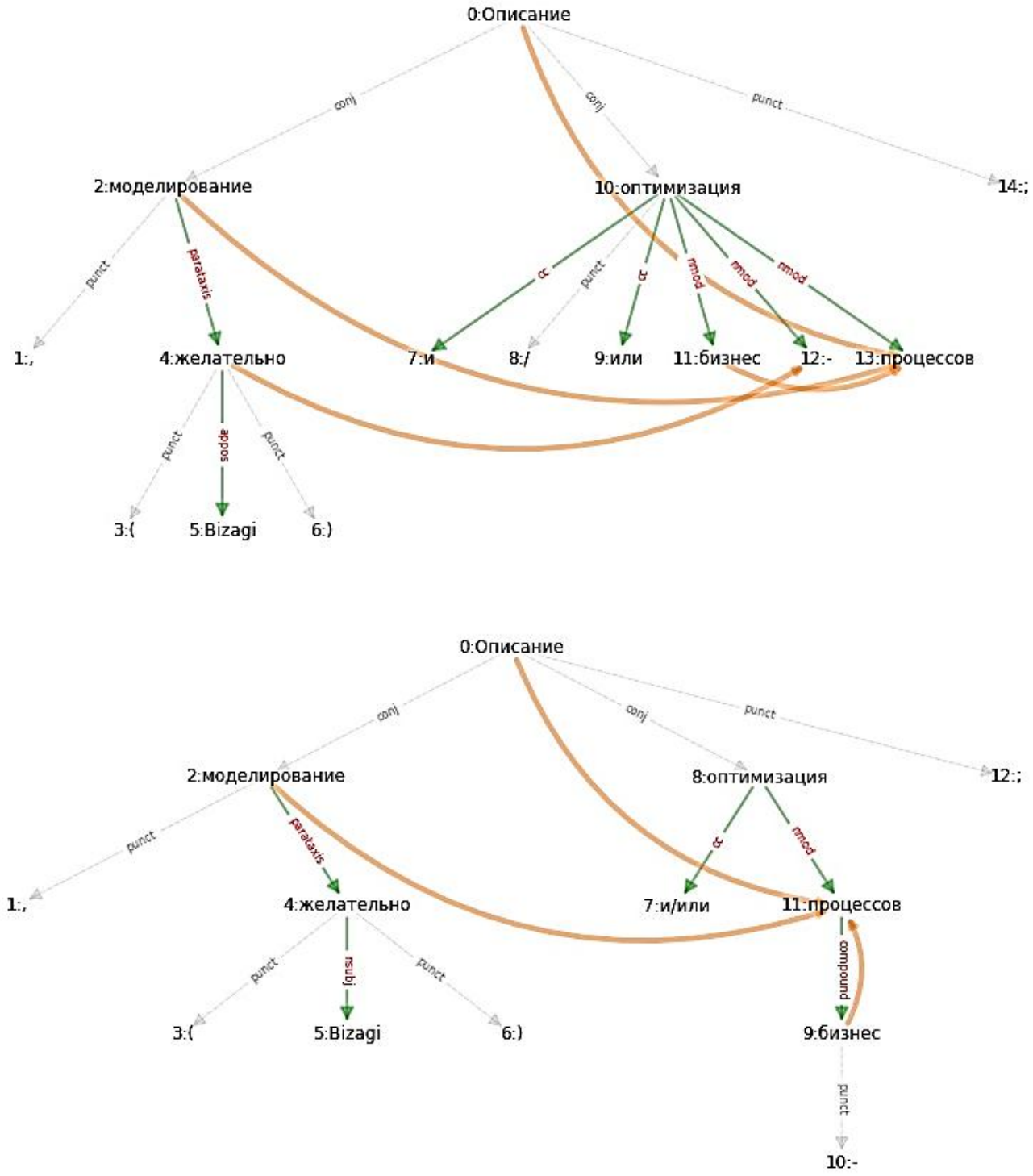

Рисунок 15 – Примеры дополненных деревьев синтаксического разбора (сверху *spacy_syntax_parser*, снизу *deeppavlov_syntax_parser*)

Выводы по второй главе

1. Предложена концепция информационной поддержки формирования требований вакансии на основе семантического сопоставления сущностей структурно-семантической модели и методов кластеризации. Описаны методика решения, использование концепции и ее ограничения.

2. Предложена структурно-семантическая модель описания требований реального рынка труда на уровне отдельных сущностей знаний и навыков/компетенций, включающая формализованное описание сущностей в виде графа с учетом структурных и семантических отношения между сущностями.

3. Разработан метод извлечения текстов отдельных сущностей знаний и навыков/компетенций из текстов требований вакансий реального рынка труда, на основе нейросетевых моделей языка и методов классификации, который в отличие от существующих методов не требует, чтобы искомые сущности были представлены в виде последовательности подряд идущих синтаксических или лексических конструкций.

ГЛАВА 3 ИНТЕЛЛЕКТУАЛЬНЫЙ МЕТОД ПОДДЕРЖКИ ФОРМИРОВАНИЯ ТРЕБОВАНИЙ ВАКАНСИИ

В третьей главе описывается интеллектуальный метод поддержки формирования требований вакансии на основе семантического сопоставления сущностей знаний и навыков/компетенций предложенной структурно-семантической модели и методов кластеризации. Метод включает в себя три этапа. Описываются основные датасеты и поэтапное проведение экспериментов. каждого этапа. Описывается методика оценки качества результатов рекомендательной системы.

3.1. Метод поддержки формирования требований вакансии

Далее описывается интеллектуальный метод поддержки формирования требований вакансии на основе семантического и частотного сопоставления сущностей структурно-семантической модели и кластерного анализа. Схема метода представлена на рисунке 16.

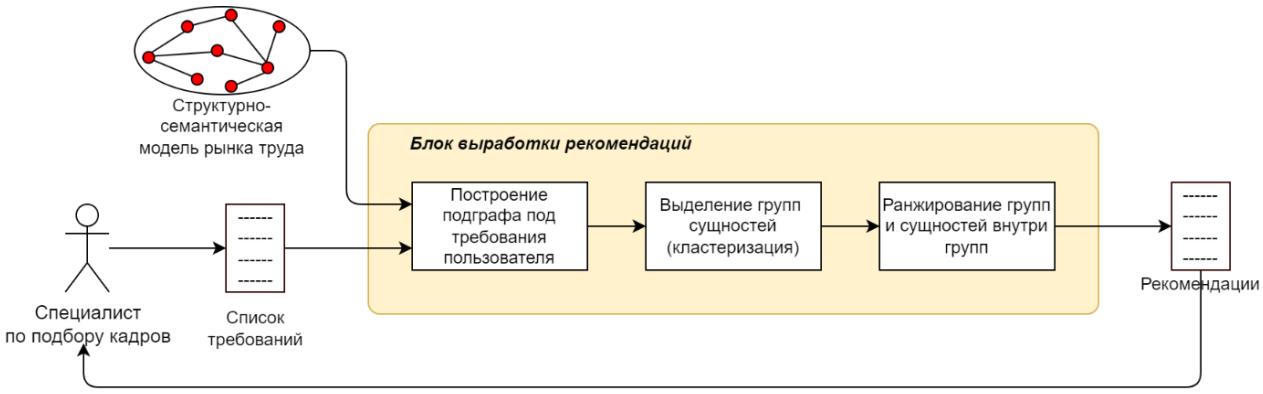

Рисунок 16 – Схема метода выработки рекомендаций

В предлагаемом методе можно выделить три основных этапа:

1. Построение подграфа структурно-семантической модели сущностей требований с учетом требований пользователя.

2. Выделение смежных групп сущностей требований на основе кластерного анализа

3. Ранжирование групп смежных сущностей требований и ранжирование сущностей требований внутри групп

Этап 1. Построение подграфа структурно-семантической модели сущностей требований с учетом требований пользователя

Этот этап направлен на выбор категорий требуемых знаний и навыков/компетенций, которые имеют связь с указанными пользователем в первоначальном списке, для их дальнейшего расширения новыми пунктами. Основная задача - создать подграф из структурно-семантической модели, включающий только те элементы, которые связаны в исходном графе, то есть те, что по крайней мере однажды были упомянуты вместе в текстах вакансий.

Пусть $G_{LM} = (V_{LM}, E_{LM})$ – структурно-семантическая модель требований рынка трыда в виде графовой структуры, описанной в разделе 2.2.

Пусть U – список сущностей *знаний* и *навыков/компетенций*, которые вводит пользователь.

Каждый элемент $u \in U$ и $v \in V_{LM}$ имеет свой собственное векторное представление $emb(u)$ и $emb(v)$, соответственно.

Пусть $sim(u, v)$ обозначает семантическую меру близости между векторными представлениями узлов u и v. В данном случае, в силу большого количества сущностей и попарных сравнений между ними, мы используем классическую косинусную меру близости, определенную с помощью HNSW из библиотеки faiss.

Инициализация

Создаем пустое множество S для синонимов сущностей для каждого $u \in U$.

Создаем пустое множество V' для списка соседних с $s \in S$ сущностей.

Создаем пустое множество E' для ребер между узлами S и V'.

Отбор синонимов сущностей требований, заданных пользователем:

Для каждого $u \in U$ находим узел $v \in V$ с наибольшей семантической мерой близости к u, такой что (3.1):

$$v = argmax_{v \in V} sim\big(emb(u), emb(v)\big). \quad (3.1)$$

Добавляем узел v в S.

Отбор соседних сущностей

Для каждого узла $s \in S$ добавляем все узлы $v \in V$, такие что (s, v) является ребром графа G_{LM}, в V'. Добавляем ребро (s, v) в множество ребер E'. В результате у нас получается подграф исходного графа $G' = (V', E')$ под исходные требования пользователя.

Этап 2. Выделение смежных групп сущностей требований на основе кластерного анализа

Суть этого этапа заключается в том, чтобы отобранные на этапе 1 сущности разбить на семантические группы (кластеры). Для вычисления наилучшего количества кластеров используется комплексная оценка на основе двух метрик оценки качества кластеризации оценка силуэта и индекс Девиса-Болдина.

Оценка силуэт.

Оценка качества кластеризации силуэт (англ. silhouette) [100] основана на расчете меры схожести объекта с объектами своего кластера по сравнению с объектами других кластеров. Оценка силуэта позволяет оценить внутреннюю когерентность кластеров, исходя из значений расстояний между объектами внутри кластера и между объектами разных кластеров, и тем самым, понять, насколько хорошо элементы распределены по кластерам.

Алгоритм вычисления оценки силуэта:

1. Для каждого объекта данных i вычисляется среднее расстояние $a(i)$ до всех объектов внутри его кластера. Это показатель того, насколько объект похож на другие объекты своего кластера.

2. Затем для этого же объекта i находится ближайший к нему кластер из остальных и вычисляется среднее расстояние $b(i)$ до объектов этого ближайшего кластера. Это показатель различия объекта с другими кластерами.

3. Далее рассчитывается силуэт объекта i как отношение разницы этих двух расстояний $a(i)$ и $b(i)$ к их максимуму (3.2):

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)}. \quad (3.2)$$

Коэффициент $s(i)$ показывает, как объекты внутри кластера сравниваются с объектами в ближайшем кластере по средним расстояниям. Чем выше s(i), тем лучше объекты отделены друг от друга по кластерам.

4. Для общей оценки качества кластеризации используется среднее значение силуэта по всем объектам данных, которые находятся в диапазоне от -1 до +1. Положительные (высокие) значения силуэта указывают на то, что объекты находятся ближе к другим объектам в своем собственном кластере, чем к объектам в других кластерах, что говорит о хорошей плотности кластеров и их четком разделении. Отрицательные (низкие) значения, наоборот, означает, что объект может быть классифицирован неправильно, и указывает на неопределенность разделения объектов на кластеры или на наличие перекрывающихся групп.

Оценка силуэт является одним из инструментов для выбора наилучшего числа кластеров или оценки эффективности алгоритма кластеризации и позволяет количественно оценить и сравнить насколько хорошо объекты распределены по найденным кластерам. Она позволяет сравнивать различные методы кластеризации и выбирать наиболее подходящий для конкретной задачи.

Индекс Дэвиса-Болдина.

Оценка качества кластеризации с использованием индекса Девиса-Болдина (Davies-Bouldin index) [49, 63] основана на соотношении между дисперсией кластеров (компактность кластеров) и расстоянием между кластерами (разделимость кластеров). Индекс Девиса-Болдина помогает измерить, насколько хорошо объекты организованы внутри кластеров и насколько различные кластеры хорошо разделены друг от друга, основываясь на межкластерных и внутрикластерных расстояниях.

Для каждого кластера C_i вычисляется его диаметр S_i – среднее расстояние между всеми парами точек внутри кластера (3.3):

$$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} d(x, c_i), \tag{3.3}$$

где c_i – центроид кластера C_i, $d(x, c_i)$ – расстояние между точкой x и центроидом c_i.

Затем для каждой пары кластеров C_i и C_j вычисляется их разделимость R_{ij} (3.4):

$$R_{ij} = \frac{S_i + S_j}{d(c_i, c_j)}, \tag{3.4}$$

где $d(c_i, c_j)$ – расстояние между центроидами кластеров C_i и C_j.

Индекс Дэвиса-Болдина определяется как среднее значение максимальных R_{ij} для каждого кластера (3.5):

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}, \tag{3.5}$$

где k – количество кластеров.

Значение индекса Девиса-Болдина может быть отрицательным, нулевым или положительным. Высокое отрицательное значение DBI, тем лучше разделение на кластеры и тем меньше «пересечений» кластеров. Высокое положительное значение указывает на то, что межкластерные расстояния меньше внутрикластерных, что свидетельствует о плохом разделении кластеров. Значение близкое к нулю или положительное говорит о наличии перекрытий между кластерами.

Индекс Дэвиса-Болдина позволяет численно оценить качество полученной кластеризации по критерию плотности и разделимости кластеров.

Совместное использование оценки силуэт (silhouette) и индекса Дэвиса-Болдина (Davies-Bouldin) для определения наилучшего количества кластеров встречается в ряде исследований и практических задач кластеризации.

Основная идея такого подхода заключается в том, чтобы найти количество кластеров, при котором достигается максимум оценки силуэт и минимум индекса

Дэвиса-Болдина. Это соответствует ситуации, когда кластеры являются плотными и хорошо разделенными.

В работе [44] авторы анализируют обе эти метрики для различного числа кластеров и выбирают значение, обеспечивающее наилучший баланс для определения оптимального числа кластеров при разделении изображений.

В работе [117] представлен новый алгоритм автоматического определения числа кластеров для временных рядов. Авторы демонстрируют, что такой подход превосходит использование каждой метрики по отдельности.

В статье [16] рассматривается задача кластеризации профилей пользователей в рекомендательных системах. Для выбора наилучшего числа кластеров авторы используют взвешенную сумму оценки Силуэт и индекса Дэвиса-Болдина, подбирая веса эмпирически.

Совокупное использование оценки Силуэт и индекса Дэвиса-Болдина является признанным подходом к определению наилучшего количества кластеров, особенно в случаях, когда отдельные метрики дают неоднозначные результаты.

Выделение групп сущностей

Пусть EMB – это множество векторных представлений сущностей из полученного подграфа, которые мы хотим разбить на семантические кластеры.

Пусть k – параметр, представляющий количество кластеров, которое мы рассматриваем от 2 до k_{max},. k_{max} – гиперпараметр, который задается пользователем. По результатам проводимых экспериментов было принято решение ограничить $k_{max} = 100$, как наиболее рациональное, т.к. в процессе анализа поведения оценок силуэта и индекса Девиса-Болдина первый максимум разницы двух этих оценок гарантированно располагался в интервале [0, 100].

Пусть $C = \{C_1, C_2, \dots, C_k\}$ – множество кластеров, где каждый кластер C_i содержит подмножество, где $v \in V'$.

Пусть $SIL = \{sil_2, sil_3, \dots, sil_{k_{max}}\}$ – множество значений оценок силуэта для каждого значения параметра k.

Пусть $DBI = \{dbi_2, dbi_3, \dots, dbi_{k_{max}}\}$ – множество значений индекса Девиса-Болдина для каждого значения параметра k.

Алгоритм выделения групп можно описать следующим образом:

Для каждого значения от 2 до k_{max} :

– Кластеризуем множество EMB методом K-Means [65] на k кластеров.

– вычисляем оценку силуэта sil_k для полученных кластеров C и сохраняем в SIL;

– вычисляем индекс Девиса-Болдина dbi_i для полученных кластеров C и сохраняем в DBI;

Вычисляем покомпонентную разность между списками SIL и DBI (3.6):

$$diff = |SIL - DBI|. \quad (3.6)$$

Выбираем значение k, которое соответствует наибольшей разности $diff$ (см. формулу 3.7).

$$k = argmax(diff). \quad (3.7)$$

Алгоритм позволяет найти рациональное количество кластеров, как наибольшую разность между оценками силуэта и индекса Девиса-Болдина при последовательном переборе параметра k в алгоритме кластеризации.

Этап 3. Ранжирование групп смежных сущностей требований и ранжирование сущностей требований внутри групп

Суть этого этапа заключается в том, чтобы сформировать ответ системы в виде ранжированного списка групп требований, в котором группы, которые имеют более явную совокупную семантическую связь с исходным списком требований отображались в верхней части, а группы, которые имеют менее выраженную семантическую связь – в нижней части. Списки сущностей требований внутри кластеров (групп) тоже ранжируются, но уже по частоте упоминания в текстах вакансий (по популярности).

Ранжирование кластеров (групп требований)

Для каждого кластера $C_i \in C$ вычисляем сумму совместной встречаемости по всем узлам в кластере с узлами из множества синонимов S (3.8):

$$score_i = \sum_{(u,v)\in E', u\in V', v\in C_i} w(u,v)\,. \quad (3.8)$$

Сортируем кластеры в C по убыванию суммы совместной встречаемости $score_i$ (3.9).

$$C' = \left(C_{i_1}, C_{i_2}, \dots, C_{i_k}\right), score_{i_1} \geq score_{i_2} \geq \dots \geq score_{i_k}\,. \quad (3.9)$$

Ранжирование сущностей требований внутри кластеров

Ранжирование сущностей требований внутри групп происходит по параметру частотности этих сущностей (3.10):

$$C_i{}' = \left(v_{i_1}, v_{i_2}, \dots, v_{i_m}\right), p(v_{i_1}) \geq p(v_{i_2}) \geq \dots p(v_{i_m})\,. \quad (3.10)$$

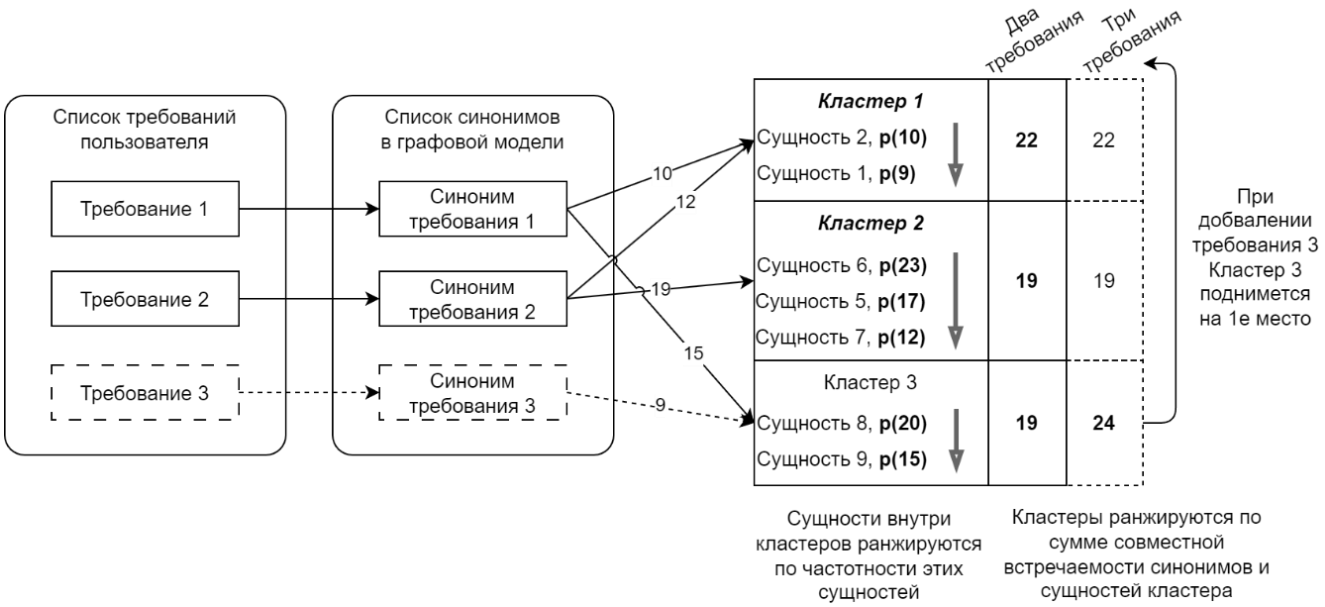

Рисунок 17 – Схема изменения групп рекомендуемых навыков при добавлении нового требования в список требований пользователя

Алгоритм позволяет найти семантически ближайшие группы требований к заданному списку требований пользователя, а также позволяет ранжировать сами группы и отдельные сущности знаний и навыков/компетенций в группах относительно друг друга в зависимости от степени совместной встречаемости в вакансиях реального рынка труда.

3.2. Методика оценка качества результатов выдачи рекомендательной системы

Mean Average Precision (MAP) – это метрика, используемая для оценки качества ранжирования результатов поиска или рекомендательных систем.

Формально MAP рассчитывается по формуле 3.11:

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP@q\,, \tag{3.11}$$

где:

– Q – число запросов

– $AP@q$ – средняя точность для запроса q

Средняя точность для одного запроса (3.12):

$$AP@q = \frac{1}{M}\sum_{k=1}^{M} P(k)\cdot rel(k)\,, \tag{3.12}$$

где:

– M – число документов в результатах поиска

– $P(k)$ – точность для первых k документов

– $rel(k)$ – бинарный индикатор релевантности k-го документа

MAP усредняет значения средней точности для отдельных запросов. Чем выше MAP, тем выше релевантность начальных позиций в упорядоченных результатах.

Рассмотрим пример вычисления метрики MAP для оценки качества ранжирования.

Предположим, у нас есть три поисковых запроса. Для каждого запроса система выдала упорядоченный по релевантности список из 5 документов.

Эксперты оценили релевантность каждого документа бинарно: 1 – релевантный, 0 – не релевантный (см. таблицу 6).

Таблица 6 – Пример результатов оценки экспертов

Запрос	Документ 1	Документ 2	Документ 3	Документ 4	Документ 5
1	1	1	0	1	0
2	0	1	1	0	1
3	1	0	1	0	1

Вычислим среднюю точность (AP) для каждого запроса:

Запрос 1:

P@1 = 1/1 = 1

P@2 = 2/2 = 1

P@3 = 2/3 ≈ 0.67

P@4 = 3/4 = 0.75

P@5 = 3/5 = 0.6

AP = (1 + 1 + 0.67 + 0.75 + 0.6) / 5 = 0.8

Запрос 2:

AP = 0.6

Запрос 3:

AP = 0.75

MAP = (0.8 + 0.6 + 0.75) / 3 = **0.72**

Ключевым параметром здесь является количество рассматриваемых документов. Чем выше число документов в выдаче, тем выше вероятность попадания нерелевантных документов в выдачу, и тем ниже оценка MAP.

Усредненная метрика MAP позволяет оценить общее качество ранжировки результатов по всем запросам.

Методика оценки качества результатов выдачи рекомендательной системы

Поскольку в нашей системе по заданному списку требований система возвращает множество кластеров, в которых требования уже ранжированы по

совместной встречаемости (см. рисунок 17), тогда общую методику оценки качества результатов рекомендательной системы можно представить в виде следующей последовательности шагов.

Шаг 1. Специалист по подбору персонала отбирает множество тестовых список требований (запросов к системе).

Шаг 2. Пропускаем запросы через рекомендательную систему и получаем набор кластеров для каждого запроса.

Шаг 3. Группа экспертов размечает каждый кластер для каждого запроса (по примеру из раздела 3.2).

Шаг 4. Получаем усредненную оценку MAP по всем кластерам каждого эксперта.

Шаг 5. Усреднение оценок всех экспертов.

3.3. Оценка качества метода извлечения отдельных сущностей требований из текстов вакансий

3.3.1. Описание датасетов

Для дообучения нейросетевых моделей языка и построения структурно-семантической модели требований были собраны текстовые корпуса вакансий, описанные в таблице 7. Корпус вакансий был собран через открытые api за период с 2010 по 2020 год с двух источников headhunter.ru и superjob.ru.

Корпус отдельных сущностей требований был собран из трех источников (см. таблицу 8):

– ESCO. В таксономии ESCO собраны тексты требований в виде отдельных сущностей знаний и навыков/компетенций размеченные экспертами. Преимущество данного источника заключается в том, что формулировки требований имеют альтернативные названия. Данные этого датасета не требуют дополнительной разметки.

– headhunter.ru (key_skills). Вакансии этого источника хранятся в формате json (пример вакансии в формате json представлен в приложении В). В некоторых вакансиях имеется параметр key_skills, который представляет собой список формулировок требований, полученных внутренними алгоритмами сервиса headhunter.ru. Недостатком этого источника является тот факт, что формулировки требований получены автоматическими средствами и требуют дополнительной ручной экспертной разметки.

– znantrend.ru. Ресурс в сети интернет, в котором собраны информация по профессиям, и связанными с этими профессиями навыками и знаниями.

Примеры отдельных сущностей знаний и навыков компетенций, с указанием источников, представлен в таблице 9.

Таблица 7 – Содержание текстовых корпусов для экспериментов

Корпус вакансий	Общее число вакансий	Число IT-вакансий	Число вакансий со строгой html-разметкой	Число извлеченных текстов требований
Корпус вакансий с headhunter.ru	35 млн.	1.8 млн.	~423 тыс.	4.9 млн
Корпус вакансий с superjob.ru	7 млн.	0.3 млн.		1.5 млн.
Итого	42 млн.	2.1 млн.	423 тыс.	6.4 млн.

Таблица 8 – Структура корпуса текстов сущностей требований из ИТ-отрасли

Ресурс	Количество уникальных текстов	Количество текстов знаний	Количество текстов навыков/компетенций
ESCO	9936	2411	7525
key_skills hh.ru	50135	20342*	29793*
znantrend.ru	19123	6342*	12781*
Итого	79194	29095	50099

*дополнительно размечались экспертом

Таблица 9 – Примеры текстов знаний и навыков

Тип	Текст	Ресурс
знание	ASP.NET 2.0	ESCO
знание	Adobe Illustrator CC	ESCO
знание	яндекс.Директ	headhunter.ru
знание	numPy	headhunter.ru
знание	1С-Битрикс	headhunter.ru
навык/компетенция	анализировать бизнес-требования	ESCO
навык/компетенция	диагностика потребностей клиента	ESCO
навык/компетенция	документировать разработку	headhunter.ru
навык/компетенция	разработка распределенных систем	headhunter.ru
навык/компетенция	написание кода	headhunter.ru

3.3.2. Дообучение нейросетевых моделей языка

Для эксперимента по дообучению языковых моделей были проанализированы существующие предобученные модели для русского языка (см. таблицу 10). В настоящий момент наиболее популярными считаются модели от двух проектных групп DeepPavlov и sberbank-ai (сменили название на ai-forever).

Были отобраны две языковые модели от команды sberbank-ai: ruBert-base (178M) и ruBert-large (427M). Обучение проводилось методом MLM, где 15% токенов исходного текста были скрыты маской. Основной метрикой качества дообучения служила перплексия. В качестве данных для дообучения были отобраны 500 тысяч текстов вакансий из области информационных технологий за последние 10 лет.

Таблица 10 – Базовые языковые модели для русского языка

Название модели	Слоев / Головы внимания	Скрытых измерений	Размер словаря	Кол-во параметров	Данные
DeepPavlov/ rubert-base-cased	12/12	768	119547	180M	*получена из мультиязычной версии BERT путем transfer learning [25]
sberbank-ai/ruBert-base	12/12	768	120138	178M	16 млрд. токенов из различных датасетов 300 GB
sberbank-ai/ruBert-large	24/16	1024	120138	427M	

Изменение перплексии, показывающий улучшение качества в процессе дообучения моделей, представлен на рисунке 18. График демонстрирует, что уже к концу пятой эпохи обучения (150000 шаг обучения) перплексия модели достигает насыщения и перестает снижаться. Это указывает на то, что дальнейшее продолжение обучения не приводит к существенным улучшениям результатов.

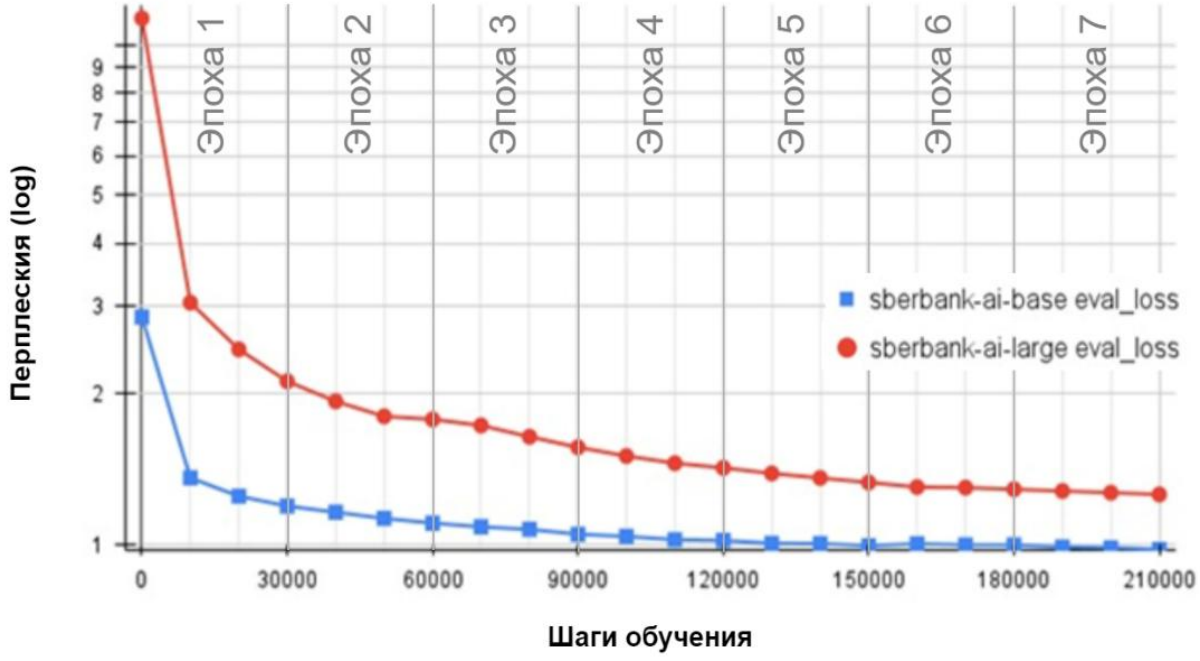

Рисунок 18 – Изменение перплексии в процессе дообучения моделей

Таблица 11 – Оценка лучшей дообученной модели на примерах с помощью инструмента заполнение маски

Пример 1: «сетевой [MASK]»		Пример 2: «выявлять [MASK]»	
Базовая модель без дообучения	Модель после дообучения	Базовая модель без дообучения	Модель после дообучения
сетевой. сетевой! сетевой " сетевой) сетевой » сетевой интернет сетевой? сетевой суд сетевой фронт сетевой анализ	сетевой трафик сетевой мониторинг сетевой Интернет сетевой контроль сетевой безопасности сетевой пакет сетевой интернет сетевой опыт сетевой рост сетевой инженер	выявлять. выявлять! выявлять? выявлять ... выявляться выявлять : выявлять и выявлять » выявлять " выявлять,	выявлять неисправность выявлять ошибки выявлять неисправности выявлять неполадки выявлять закономерности выявлять уязвимости выявлять информацию выявлять их выявлять проблемы выявлять ответственных

Для визуальной оценки качества процесса дообучения базовых моделей на текстах вакансий обученной модели использовался метод заполнения маски. В этом подходе одно из слов в последовательности скрывается под токеном маски, и модель должна предсказать наиболее вероятное слово для данного места. Результаты этого сравнительного теста для базовой модели и модели, дообученной на текстах вакансий, представлены в таблице 10. Можно видеть, что дообученная модель справляется с задаче заполнения маски лучше базовой модели, что может говорить о том, что в процессе дообучения модель приобретает специфические знания о текстах предметной области.

3.3.3. Выделение текстов требований из текстов вакансий

Не во всех вакансиях есть строгая разметка, которая бы позволила извлекать тексты требований. Как следствие, разные вакансии из-за человеческого фактора могут иметь нестрогую структуру, и разделы внутри вакансий могут находиться на разных позициях, что существенно затрудняет извлечение текстов требований из текстов вакансий. Пример вакансий со строгой и нестрогой структурой представлены в приложении Г.

Для решения задачи извлечения текстов требований из текстов вакансий был обучен классификатор текстов. Для обучения классификатора были отобраны 523 тыс. текстов вакансий из датасета вакансий (см. таблицу 7), в которых присутствовала строгая html-разметка на 4 класса: общий текст (common), требования (requirements), обязанности (duties), условия работы (conditions).

Для обучения классификатора были выбраны две модели предобученная модель sberbank-ai/ruBert-base и дообученная на текстах вакансий sberbank-ai/ruBert-base-finetuned. Результаты работы классификатора представлены в таблице 12.

Результаты оценки качества моделей в задаче классификации отдельных предложений из текстов вакансий представлены на рисунке 19. Оценки и матрица ошибок по классам для лучшей модели представлены в на рисунке 19 и на рисунке 20, соответственно.

По результатам анализа матрицы ошибок, представленной на рисунке 20, можно отметить следующее, класс, который изначально был размечен в вакансиях как «общий текст» по результатам классификации содержал в себе большое количество представителей других классов, это можно видеть в графе «Предсказанные значения / общий текст» на рисунке 20. Данное обстоятельство указывает на то, что в процессе составления текста вакансии достаточно много информации о требованиях к знаниям и навыкам/компетенциям помещается составителями вакансий в «общий» раздел.

Таблица 12 – Результаты выделения текстов требований из текстов вакансий по взвешенной F1-мере

Наименование	Класс «Общий текст»	Класс «Требования»	Класс «Обязанности»	Класс «Условия работы»
sberbank-ai/ruBert-base + classification_layer	0.72	0.70	0.71	0.73
sberbank-ai/ruBert-base-finetuned + classification_layer	0.83	0.89	0.89	0.91

	precision	recall	f1-score	support
common	0.97	0.72	0.83	3583
requirements	0.84	0.96	0.89	2399
duties	0.83	0.96	0.89	2086
conditions	0.86	0.98	0.91	1932
accuracy			0.88	10000
macro avg	0.87	0.90	0.88	10000
weighted avg	0.89	0.88	0.87	10000

Рисунок 19 – Результаты обучения лучшей модели классификатора

Рисунок 20 – Матрица ошибок по классам

3.3.4. Эксперимент извлечения текстов отдельных сущностей требований

Для исследования извлечения текстов отдельных сущностей требований был проведен эксперимент с использованием метода, описанного в разделе 2.3. Было отобрано 2000 текстов требований из вакансий длинной от 8 до 15 токенов. Выбор этого диапазона был обоснован тем, что более длинные тексты часто содержат сложные взаимосвязи, что усложняет извлечение коротких сущностей требований.

Далее было отобрано восемь комбинаций моделей, синтаксических анализаторов и методов для выявления дополнительных связей между токенами. В эксперименте использовались следующие модели и синтаксические анализаторы.

Модели:

– sberbank-ai/ruBert-base;

– sberbank-ai/ruBert-base-finetuning.

Синтаксические анализаторы:

– spacy_base – токенизатор и синтаксический анализатор spacy на модели для русского языка ru_core_news_lg;

– deeppavlov_syntax_parser – токенизатор и синтаксический анализатор deeppavlov на модели для русского языка syntax_ru_syntagrus_bert.

Способ получения дополнительных связей:

– линейная комбинация токенов (см. описание алгоритма, блок 2.а);

– добавление новых связей в дерево синтаксического разбора (см. описание алгоритма, блок 2.б).

Для оценки предложенного метода эксперты тщательно проанализировали каждое текстовое описание требований. Они пометили все содержащиеся в нем понятия, относящиеся к знаниям и навыкам/компетенциям. Примеры такой разметки приведены в таблице 13.

На заключительном этапе количество автоматически извлеченных текстов сущностей, относящихся к знаниям и навыкам/компетенциям, было сопоставлено с экспертной разметкой с использованием метрики F1. В таблице 13 представлены

результаты эксперимента по автоматическому извлечению отдельных сущностей требований с помощью разработанного метода.

В дальнейших экспериментах предполагается перейти от парсеров на контекстно-свободных грамматиках к нейросетевым моделям, использующим технологии поиска именованных сущностей. Такие модели позволяют лучше обрабатывать тексты на естественном языке с разной степенью формальности, поскольку они не опираются на жесткие структуры грамматик. Это позволяет обрабатывать тексты с разными стилями и тонами, а также с опечатками и неточностями.

Таблица 13 – Пример экспертной разметки отдельных сущностей требований

Текст требования	Сущности требований	Категория
Описание, моделирование (желательно Bizagi) и/или оптимизация бизнес-процессов;	Описание бизнес-процессов	навык/компетенция
	Описание процессов	навык/компетенция
	Моделирование бизнес-процессов	навык/компетенция
	Моделирование процессов	навык/компетенция
	Оптимизация бизнес-процессов	навык/компетенция
	Оптимизация процессов	навык/компетенция
	Bizagi	знание
Чувство вкуса и стиля	Чувство вкуса	навык/компетенция
	Чувство стиля	навык/компетенция
Разработка WEB-ориентированных, распределенных приложений на языке Java с применением технологии Adobe Flex на стороне клиента.	Разработка WEB-ориентированных приложений	навык/компетенция
	Разработка web-приложений	навык/компетенция
	Разработка распределенных приложений	навык/компетенция
	Adobe Flex	знание
	Разработка приложений на Java	навык/компетенция
	Java	знание

Таблица 14 – Результаты сравнения различных комбинаций инструментов

Модель	Токенизатор и синтаксический парсер	Схема получения комбинаций токенов	F1
yargy-парсер (baseline)			0.31
sberbank-ai-base	spacy_base	линейная комбинация	0.65
sberbank-ai-base	spacy_base	дополнение дерева	0.33
sberbank-ai-base	deeppavlov_syntax_parser	линейная комбинация	0.58
sberbank-ai-base	deeppavlov_syntax_parser	дополнение дерева	0.35
sberbank-ai-base-finetune	spacy_base	линейная комбинация	0.81
sberbank-ai-base-finetune	spacy_base	дополнение дерева	0.69
sberbank-ai-base-finetune	deeppavlov_syntax_parser	линейная комбинация	0.83
sberbank-ai-base-finetune	deeppavlov_syntax_parser	дополнение дерева	0.71

3.3.5. Эксперимент по объединению синонимов сущностей

В результате анализа извлеченных текстов отдельных сущностей требований было выявлено, что многие сущности представляют собой синонимы требований.

Согласно алгоритму, описанному в разделе 2.2, для каждой отдельной сущности требования осуществлялся поиск более популярного соседа в определенной окрестности. Для этого сначала необходимо было найти всех соседей некоторой сущности требования, а затем найти самого популярного из них. Такой поиск строится на попарном определении степени семантической близости между векторами исходной сущности и векторами всех остальных сущностей. Из-за высокой сложности данной задачи вместо стандартного метода оценки схожести по косинусу были применены векторные индексы из библиотеки FAISS, подробно описанные в разделе 1.3.2.

Для проведения эксперимента произвольным образом отбирались 100 отдельных сущностей требований, затем для каждой сущности при помощи различных векторных индексов осуществлялся поиск ее ближайших соседей в некоторой окрестности E, которые в определенном смысле можно считать ее синонимами, т.к. размер окрестности E достаточно мал. Результаты сравнения векторных индексов представлены в таблице 15.

Далее проводился эксперимент относительно определения лучшего значения окрестности E. Произвольным образом было отобрано 100 сущностей. Экспертами

для каждой сущности среди ее ближайших соседей в окрестности E подсчитывалось количество правильных синонимов, т.е. терминов, которые в полной мере были синонимами исходной сущности, и количество ошибочных сущностей, которые не имели или имели отдаленное отношение к исходной сущности. Затем вычислялось среднее значение по общему количество отобранных сущностей. Зависимость среднего числа правильных синонимов и среднего числа ошибок по наибольшему индексу предыдущего этапа *faiss_IndexHNSWFlat* показана на рисунке 21. Наилучший результат показали значения окрестности в районе 12-14 по метрике расстояния для данного индекса. При этих значениях отношение среднего числа правильных синонимов по отношению к среднему числу ошибок – максимально.

Таблица 15 – Результаты сравнения векторных индексов

Тип индекса	Средняя скорость обработки 1 запроса
faiss_IndexFlatL2	803 ms
faiss_IndexFlatIP	574 ms
faiss_IndexLSH	198 ms
faiss_IndexIVFFlat	52.8 ms
faiss_IndexHNSWFlat	25 ms

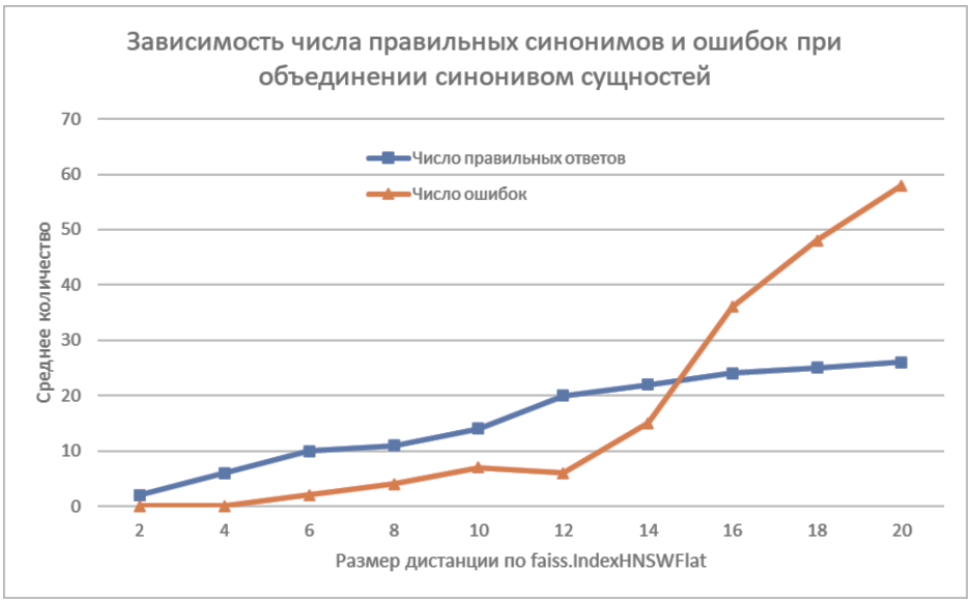

Рисунок 21 – Зависимость числа правильных синонимов и ошибок при объединении синонимов сущностей

Если среди ближайших синонимов некоторой сущности находилась сущность с большей частотностью чем исходная, тогда данная сущность признавалась гиперонимом (более общим понятием) для исходной сущности. Пример работы алгоритма поиска гиперонима среди синонимов сущностей для сущности навык/компетнция – «описание бизнес процессов» представлен на рисунке 22.

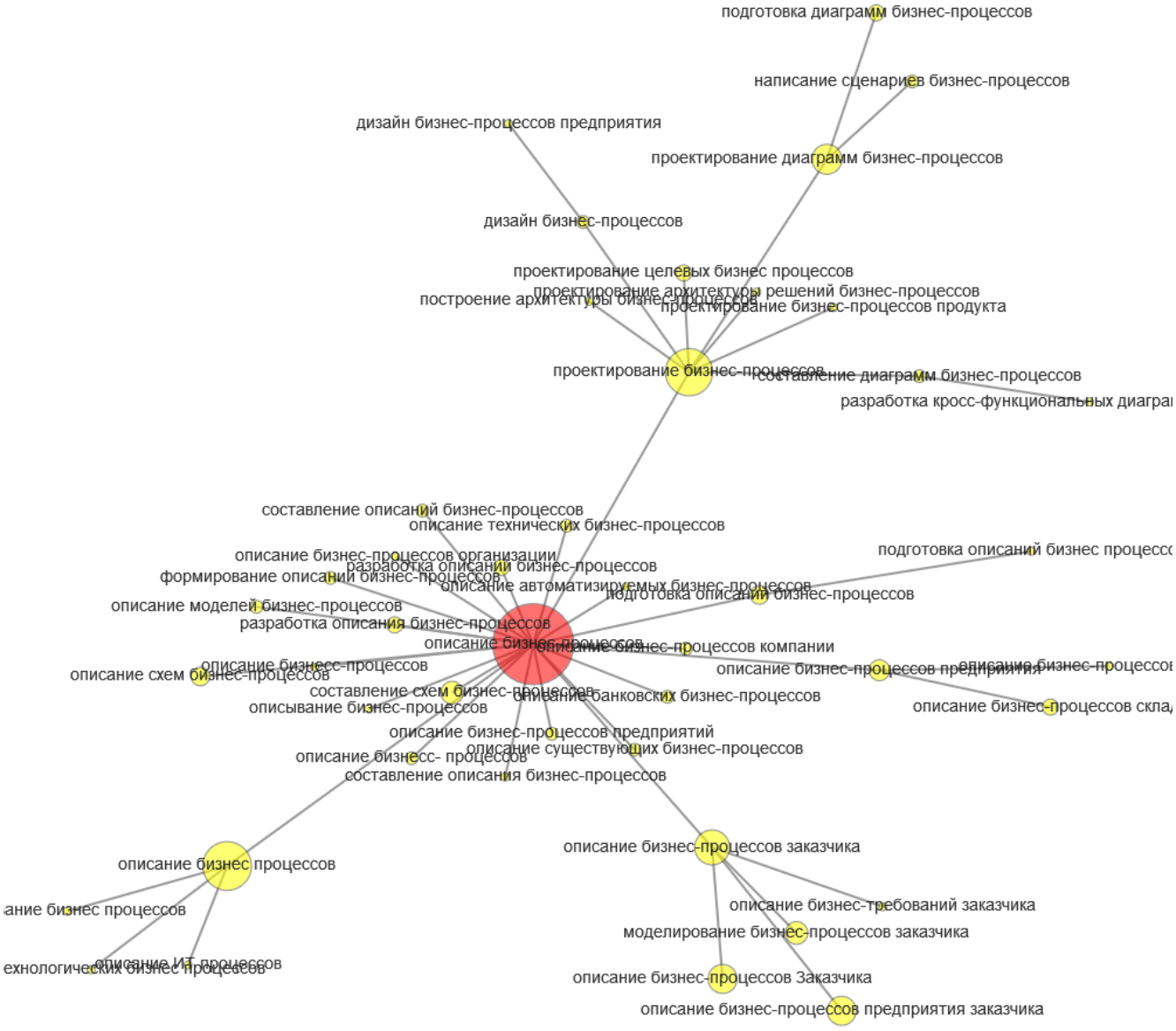

Рисунок 22 – Иллюстрация работы алгоритма поиска гиперонима среди синонимов сущностей (ПГСС)

На рисунке видно, что синонимичные сущности объединяются в дерево синонимов. Размер узлов говорит о частотности их встречаемости в текстах требований вакансий. Связи на рисунке указывают на связь гипоним-гипероним (частное-общее). Рисунок также иллюстрирует, что сущности имеющие более конкретные формулировки, но с меньшей частотой встречаемости группируются вокруг синонима этой сущности с более высокой встречаемостью и имеющего более общую формулировку.

3.4. Оценка точности семантического поиска групп требований под заданный список требований пользователя

Для оценки точности семантического поиска рекомендуемых групп требований среди исходного списка требований, предоставленного специалистом по подбору персонала, 9 экспертов, включая ведущих преподавателей университета и представителей работодателей, дали оценки релевантности на 133 запросах. Эксперты оценивали релевантность выдачи до 10 требований в каждом кластере с использованием методики, описанной в разделе 3.2.2, для каждой векторной модели. Результаты эксперимента представлены в таблице 16.

Таблица 16 – Результаты эксперимента по оценке точности семантического поиска под исходный список требований пользователя.

Векторное представление	MAP@1	MAP@3	MAP@5	MAP@10
Базовая модель ruBERT	0,725	0,673	0,625	0,504
Дообученная модель ruBERT	0,855	0,822	0,793	0,654

По результатам оценки точности представленных в таблице 16 можно отметить, что на глубине в 5 ближайших значений релевантность элементов в выдаче рекомендаций начинает существенно падать. Также можно наблюдать, что дообученная модель демонстрирует более релевантную выдачу рекомендаций, по сравнению с моделью без дообучения.

Выводы по третьей главе

1. Разработан метод поддержки формирования требований вакансии на основе семантического сопоставления отдельных сущностей требований предложенной структурно-семантической модели и методов кластеризации, который обеспечивает соответствие разрабатываемых требований в проектах вакансий реальным потребностям рынка труда.

2. Предложена методика оценки точности семантического поиска групп требований под заданный список требований пользователя.

3. Проведена оценка качества всех этапов предложенного метода извлечения отдельных сущностей требований из текстов вакансий.

4. Проведена оценка точности семантического поиска групп требований под заданный список требований пользователя по предложенному методу поддержки формирования требований вакансии на основе семантического сопоставления отдельных сущностей требований предложенной структурно-семантической модели и методов кластеризации.

ГЛАВА 4 РАЗРАБОТКА ПРОТОТИПА ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ПОДДЕРЖКИ ФОРМИРОВАНИЯ СПИСКА ТРЕБОВАНИЙ ВАКАНСИИ

Четвертая глава посвящена описанию процесса разработки и апробации прототипа интеллектуальной рекомендательной системы поддержки формирования списка требований к вакансиям на основе предложенных методов и алгоритмов. Описаны функциональные требования и определена общая структура прототипа и структура базы данных. Подробно рассмотрены особенности программной реализации различных модулей системы, а также используемые технологии и библиотеки. Приведены результаты апробации прототипа рекомендательной системы, дана комплексная оценка его эффективности в процессе формирования требований при подготовке вакансии.

4.1. Требования к прототипу интеллектуальной системы

На основе анализа процесса подбора персонала в главе 1 была разработана функциональная модель процесса подбора персонала в нотации IDEF0 (см. рисунок 23).

По результатам анализа функциональной модели (см. рисунок 23), были определены основные функциональные требования к прототипу интеллектуальной рекомендательной системы поддержки формирования списка требований вакансии:

– формировать смежные группы требований под исходный список требований пользователя;

– ранжировать смежные группы требований;

– ранжировать требования внутри групп;

– строить двумерную карту смежных групп требований;

– строить дерево объединения синонимов сущностей требований с указанием их частотности;

– периодически, с заданным интервалом, проводить обновление базы данных вакансий, автоматически извлекая свежие объявления с рекрутинговых онлайн-платформ headhunter.ru и superjob.ru.

– периодически, согласно установленному графику, проводить дополнительное обучение нейросетевых языковых моделей на основе обновленного корпуса текстов вакансий.

– поддерживать обновление базы отдельных сущностей требований знаний и навыков/компетенций с определенной периодичностью из нескольких источников ESCO, headhunter_keyskills, znantrend.ru;

– поддерживать ручное редактирование отдельных сущностей требований: добавление, изменение, удаление;

– автоматически с определенной периодичностью пересчитывать частоту упоминания и частоту совместной встречаемости отдельных сущностей требований;

– должна предоставлять функционал поиска вакансий, соответствующих списку требований пользователя, с возможностью дальнейшего анализа найденных вариантов.

– пользователь должен иметь возможность применять гибкие фильтры по отраслям, профессиям, регионам и периоду размещения вакансий для сужения результатов выдачи.

– обеспечивать возможность сопоставления компетенций. указанных в резюме соискателя, со списком требований вакансии и давать оценку степени соответствия.

Также была подготовлена диаграмма прецедентов работы с рекомендательной системой (см. рисунок 24).

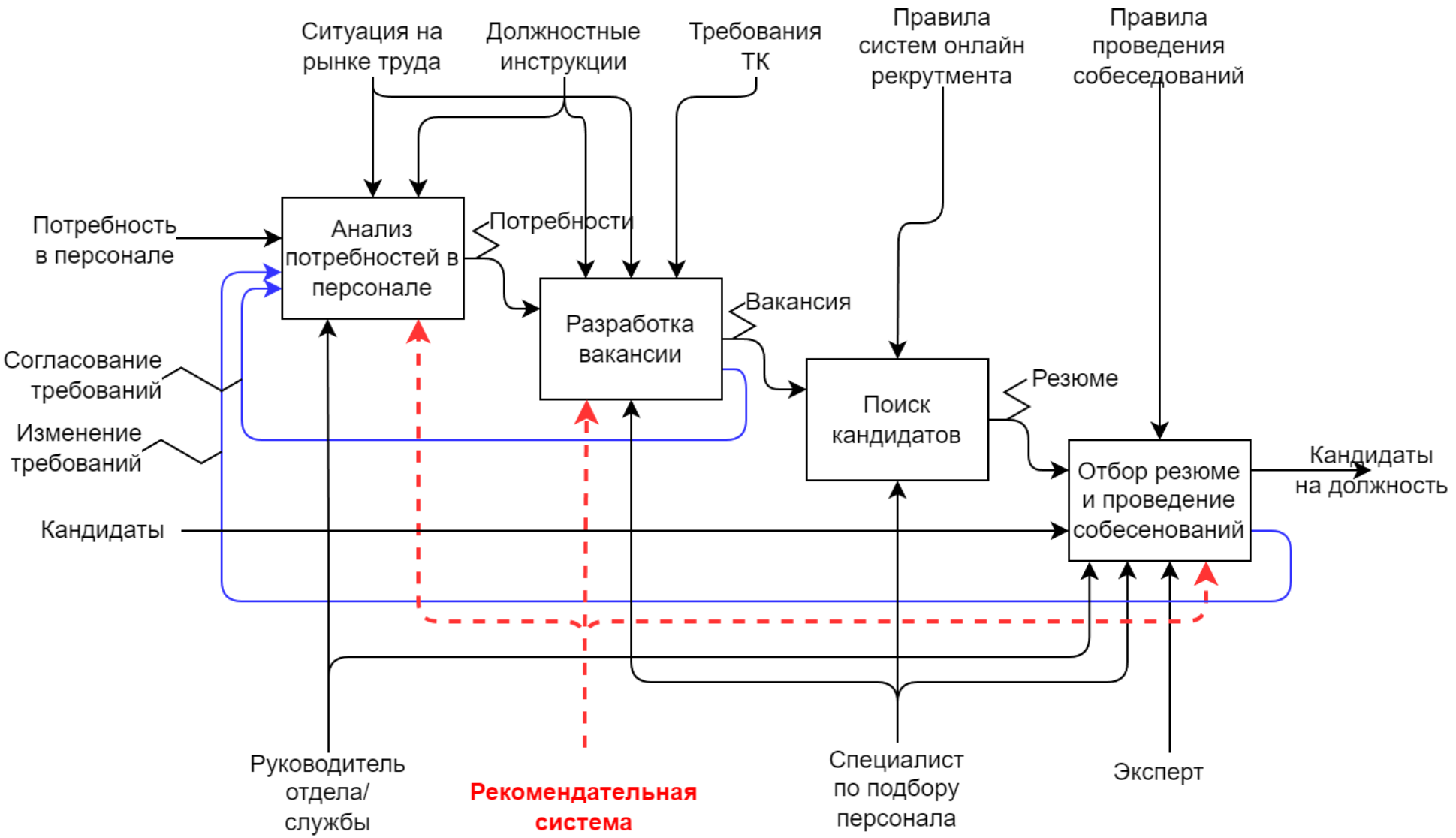

Рисунок 23 – Функциональная модель процесса подбора персонала в формате IDEF0

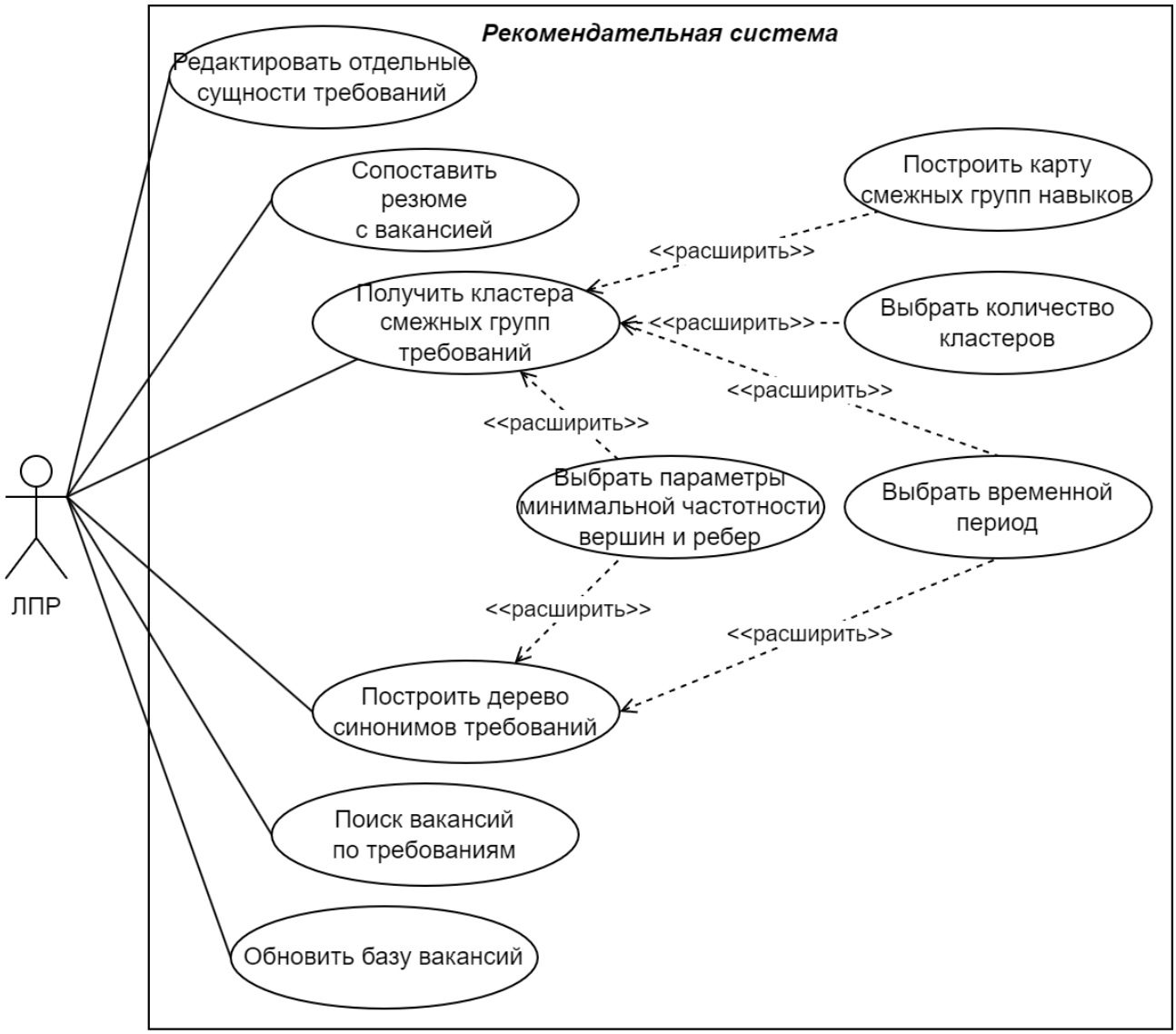

Рисунок 24 – Диаграмма прецедентов

В качестве ЛПР (лица принимающего решения) могут выступать:

– линейные и функциональные руководители служб, отделов, структурных подразделений;

– специалиста по подбору персонала;

– эксперты из профессионального сообщества.

4.2. Проектирование структуры прототипа интеллектуальной рекомендательной системы

На основе описанных требований в разделе 4.1. была спроектирована структура интеллектуальной рекомендательной системы поддержки формирования списка

требований вакансии на основе предложенных моделей, методов и алгоритмов (см. рисунке 25).

Модули интеллектуальной рекомендательной системы:

– *Модуль работы с вакансиями*. Данный модуль отвечает за работу с объектами вакансий: блок поиска текстов вакансий под сформированный список требований, блок обновления базы вакансий.

– *Модуль обучения нейросетевых моделей*. Данный модуль включает блок дообучения языковой модели на текстах вакансий, которая в последующем используется в блоке дообучения модели классификатора отдельных предложений из текстов вакансий на четыре класса: общий текст, требования, обязанности, условия работы.

– *Модуль извлечения текстов требований из текстов вакансий*. Используя классификатор текстов, обученный в предыдущем модуле, из текстов вакансий извлекаются только тексты требований для дальнейшего анализа.

– *Модуль работы с отдельными сущностями требований*. В данном модуле реализован функционал работы с отдельными сущностями требований:

- блок обновления базы отдельных сущностей требований из нескольких источников. Обновление базы данных отдельных сущностей может происходить автоматически из нескольких источников или может быть отредактирована экспертом вручную.
- блок автоматическое извлечение отдельных сущностей требований знаний и навыков/компетенций из текстов требований, извлеченных в предыдущем модуле. Реализует метод, описанный в разделе 2.3.
- блок объединение синонимов отдельных сущностей требований (поиск гиперонимов сущностей). Реализует метод, описанный в разделе 2.2.

– Модуль пользовательского интерфейса (на рисунке 25 представлены стрелками красного цвета).

- блок построения смежных групп требований, включающий построение карты групп смежных групп.

- блок построения дерева синонимов отдельных сущностей требований
- блок сопоставления текста вакансии и текста резюме.

Программной реализации прототипа рекомендательной системы выполнена с применением трехуровневой архитектуры: веб-интерфейс, бизнес-логика, хранилище данных.

В качестве основного языка программирования был использован язык Python. Выбор обусловлен наличием большого количества специализированных библиотек и удобством разработки прототипов на данном языке.

Для реализации интерфейса пользователя и взаимодействия с прототипом был выбран веб-фреймворк Flask. Flask является функциональным и гибким фреймворком для разработки веб-приложений на языке Python и позволяет создавать масштабируемые веб-серверы.

В качестве основных СУБД для хранения и управления информацией о вакансиях, требованиях, отдельных сущностях требований и других необходимых данных использовались MySQL и SQLite. MySQL использовалась для долгосрочного хранения данных, а SQLite для временного хранения промежуточных результатов.

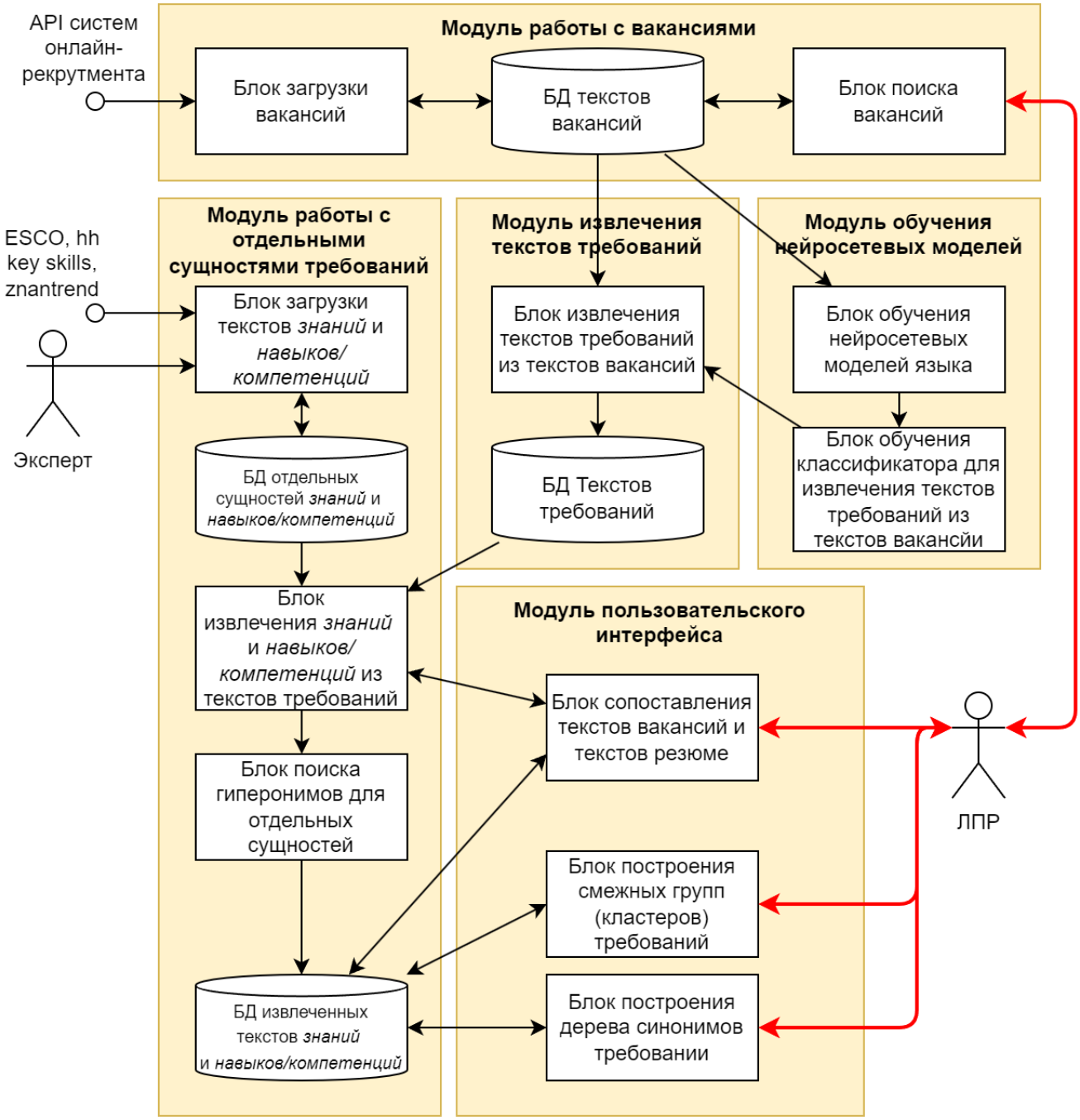

Рисунок 25 – Структура прототипа интеллектуальной рекомендательной системы

Для работы с нейросетевыми моделями использовалась библиотека для работы с нейросетевыми моделями обработки естественного языка (NLP) transformers от компании HuggingFace [117]. Она предоставляет широкий спектр инструментов для обучения и использования самых разных нейросетевых моделей NLP, включая модели BERT, GPT и др. В данном прототипе библиотека HuggingFace используется для обучения нейросетевых моделей BERT и классификаторов.

Для осуществления математических расчетов и реализации алгоритмов классификации и кластеризации, а также вычисления метрик оценки качества на разных этапах использовалась библиотека scikit-learn [3]. Она предоставляет множество инструментов для классификации, регрессии, кластеризации и других задач машинного обучения.

Для анализа текстов вакансий и извлечения сущностей по контекстно-свободным грамматикам была использована библиотека Yargy [88], доступная в открытом доступе и написанная на языке Python. Эта библиотека предназначена для обработки грамматик и извлечения структурированной информации из текстовых данных.

Для построения графов использовалась библиотека NetworkX [90], а для их отрисовки и визуализации библиотека plotly [97]. NetworkX предоставляет широкий набор функций и алгоритмов для анализа, создания и манипуляции графовыми структурами, поддерживает различные типы графов, включая направленные, ненаправленные, взвешенные и ориентированные. Plotly – это библиотека Python для визуализации интерактивных графиков и диаграмм. Сочетание NetworkX и Plotly открывает широкие возможности для изучения и представления сетевых структур данных в удобном и понятном формате, позволяет создавать интуитивно понятные и визуально информативные графики для анализа сетевых структур, связей и влияний. NetworkX предоставляет данные о графах, а Plotly позволяет их визуализировать с помощью интерактивных возможностей, таких как навигация по графу, отображение подробной информации при наведении, переключение отображения элементов и т.д.

Для позиционирования узлов на двумерных отображениях графов для построения карты смежных требований и для визуализации дерева синонимов сущностей требований используются методы понижения размерности.

1. PCA (англ. Principal Component Analysis) [113]. PCA является одним из наиболее распространенных и часто применяемых алгоритмов для уменьшения размерности данных. Алгоритм находит линейные комбинации исходных

признаков, которые наилучшим образом описывают изменчивость данных, а затем строит новые признаки (главные компоненты), которые упорядочены по объясненной ими дисперсии данных. PCA хорошо работает с линейными данными и является хорошим выбором для сокращения размерности больших данных.

2. t-SNE (англ. t-Distributed Stochastic Neighbor Embedding) [109]. t-SNE является алгоритмом нелинейного понижения размерности, который хорошо подходит для визуализации данных в двумерном или трехмерном пространстве. Алгоритм сохраняет близкие точки в исходном пространстве как близкие точки в новом пространстве и наоборот. Хорошо работает для визуализации сложных и высокоразмерных данных, но может быть времязатратным при работе с большими наборами данных.

3. UMAP (англ. Uniform Manifold Approximation and Projection) [87]: UMAP также является алгоритмом нелинейного понижения размерности, похожим на t-SNE, но он имеет более высокую скорость работы и масштабируемость. Алгоритм стремится сохранить локальные структуры данных в новом пространстве, сохраняя близость соседних точек. Хорошо работает как для визуализации, так и для кластеризации данных и может быть эффективным методом для анализа больших наборов данных.

4. Алгоритм Фрухтермана-Рейнгольда [6]. В библиотеки networkx этот алгоритм реализован под названием spring_layout. Это метод визуализации графов, который позволяет представить связи между узлами в двумерном пространстве. В процессе работы алгоритм учитывает силы притяжения и отталкивания между узлами, чтобы распределить их таким образом, чтобы узлы, связанные друг с другом, были близко, а не связанные – далеко. Алгоритм отталкивает узлы друг от друга, если они далеко друг от друга, и притягивает, если они близко. В результате получается графическое представление с наглядными связями между узлами.

Данный стек технологий, алгоритмов и открытых библиотек обеспечивает разработку прототипа рекомендательной системы в эффективной и

масштабируемой форме, обеспечивая удобство использования и достоверность результатов.

Для систематизации данных, связанных с вакансиями, требованиями к кандидатам и частотой упоминания отдельных сущностей требований, используется реляционная база данных. Схема базы данных, описывающая структуру таблиц и связи представлена на рисунке 26 в нотации IDEF1X, которая является стандартом для документирования реляционных баз данных.

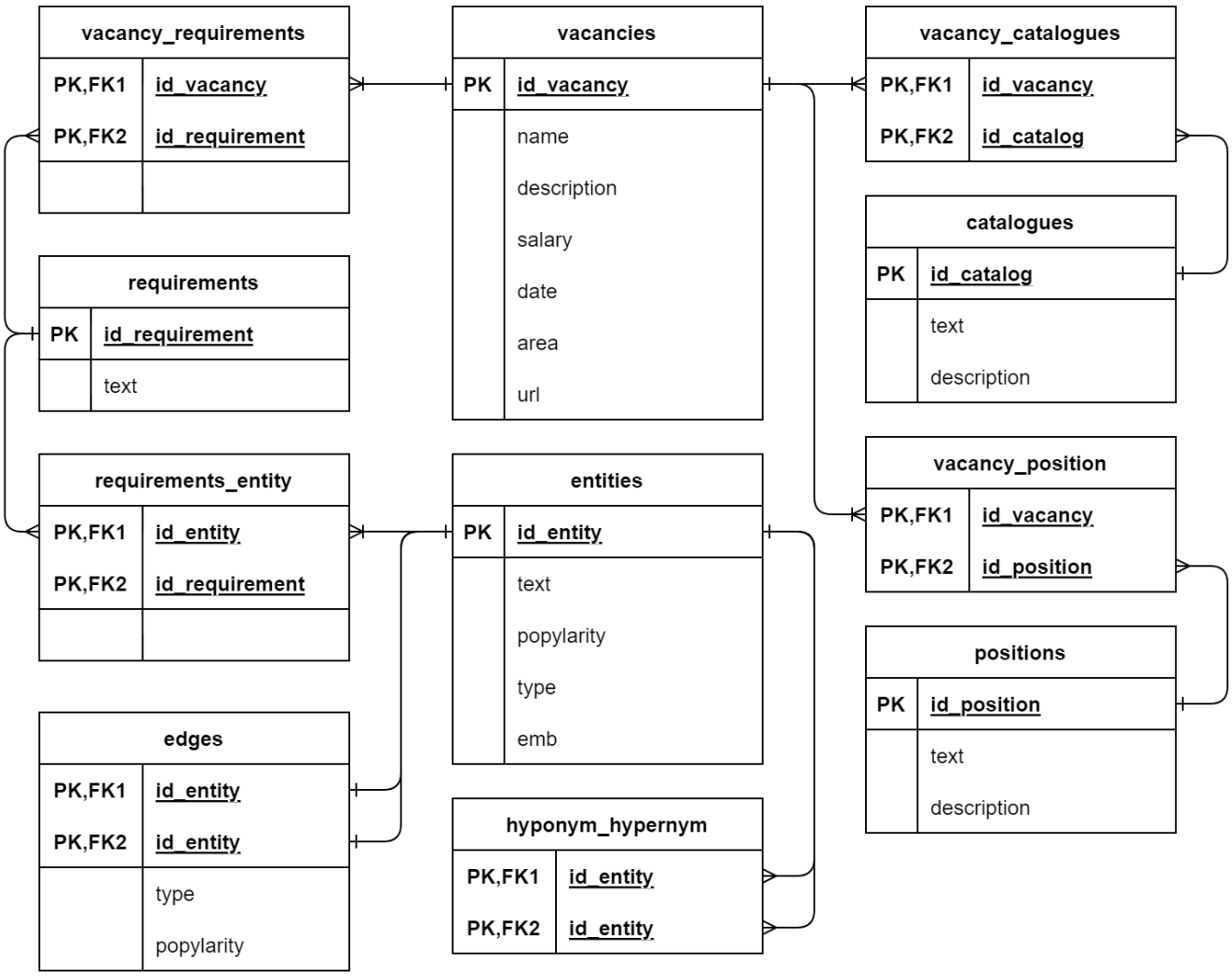

Рисунок 26 – Схема базы данных рекомендательной системы

Для моделирования физического развертывания прототипа интеллектуальной рекомендательной системы использовалась диаграмма развертывания, приведенная на рисунке 27.

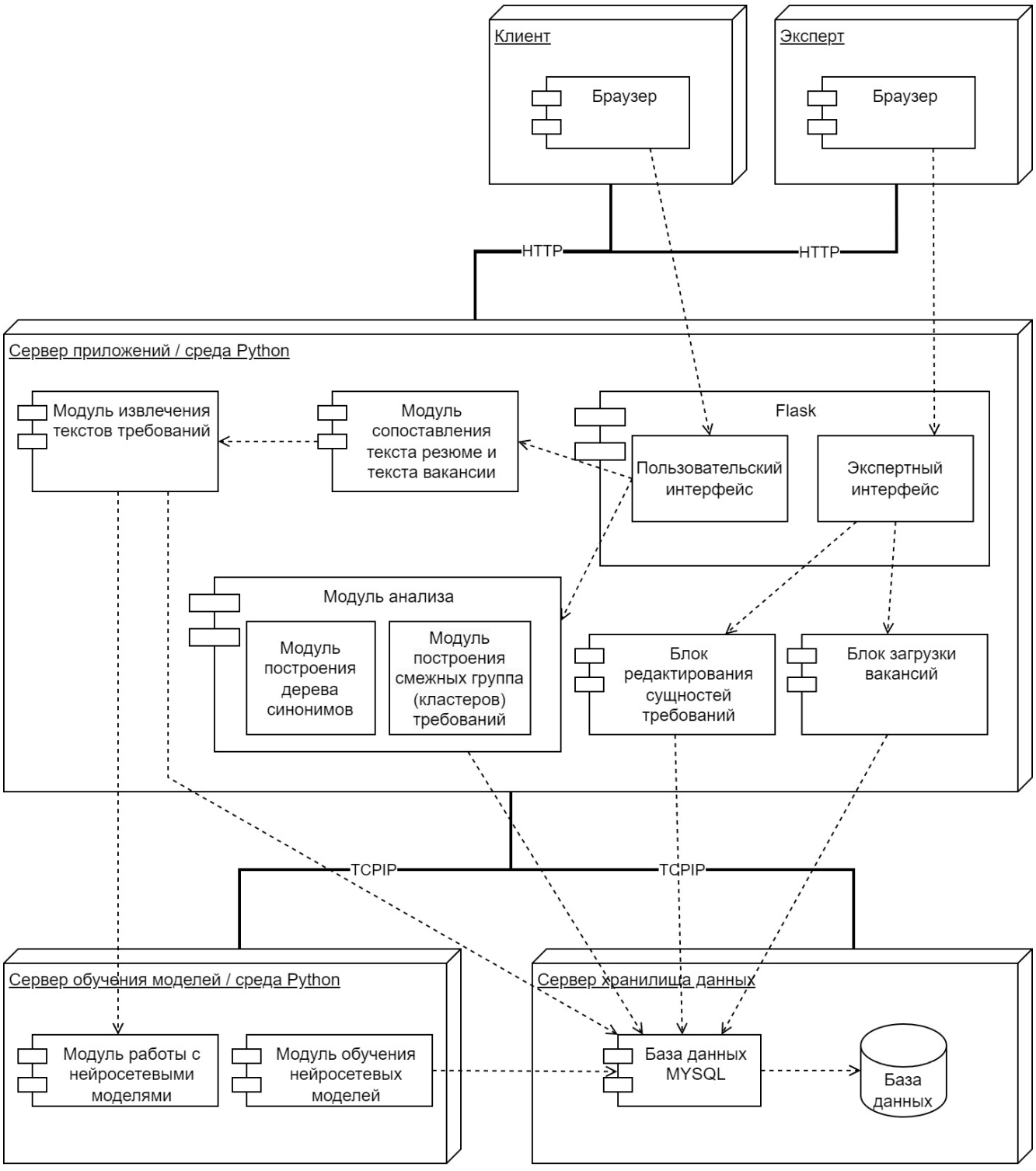

Рисунок 27 – Диаграмма развертывания

4.3. Использование прототипа интеллектуальной системы

Рассмотрим использование прототипа системы на примере вакансии web-разработчика. Для этой профессии зададим следующий список первоначальных требований: 'работа с PHP', '1С-Битрикс', 'опыт работы с различными API', 'разработка технической документации', 'работать с технической документацией'.

Веб-разработчики обычно работают с языками программирования, такими как PHP, для создания динамических веб-сайтов и веб-приложений. 1С-Битрикс – это популярная CMS, которая широко используется для разработки сайтов, в том числе для управления контентом и функциональностью. Опыт работы с различными API может быть необходим для интеграции веб-приложений с внешними сервисами или платформами. Разработка технической документации и работа с технической документацией также важны для веб-разработчиков, чтобы документировать процессы, код и инструкции по использованию и обслуживанию созданных ими веб-приложений.

После нажатия на кнопку «построить список рекомендуемых требований» система анализирует список требований пользователя и на основе структурно-семантической модели формирует группы (кластера) смежных групп требований. Результат работы представлен на рисунке 28.

В результате работы автоматически сформировано 10 кластеров требований, на рисунке 28 представлены 6 из них, которые ранжированы согласно алгоритму из раздела 3.1. В каждом кластере представлено по 5 требований, ранжированных по частоте. Число требований в группе можно задать соответствующим параметром.

Пользователь также может построить карту смежных требований, нажав на кнопку «построить карту требований». При работе с картой требований пользователь может выбирать различные схемы позиционирования узлов (spring_layout, PCA, TSNE, UMAP, см. раздел 4.2) и различные схемы раскраски узлов (см. рисунок 29 и 30). Позиционирование и раскраска узлов позволяет пользователю визуально анализировать группы смежных требований.

На рисунке 30 показан пример карты, где представлены смежные группы требований для web-разработчика, созданные с использованием метода позиционирования узлов UMAP и раскраски по группам. Такие карты позволяют анализировать, какие элементы требований формируют семантически близкие группы, и исследовать, какие из них часто встречаются вместе в вакансиях. Это помогает получить более полное представление о необходимых качествах для данной должности и выявить ключевые требования, которые могут быть недооценены, но важны для успешной работы на данной позиции.

Примеры карты совместной встречаемости смежных сущностей требований под исходный список требований для профессии web-разработчик представлен на рисунке 31. На рисунке 31 красным цветом отмечены сущности из исходного списка требований пользователя, синим цветом отмечены сущности, которые встречаются совместно как минимум с двумя сущностями из исходного списка, а желтым цветом все остальные сущности. Такой вид диаграмм позволяет оценить наиболее приоритетные требования для конкретного списка исходных требований, и сконцентрироваться в первую очередь на требованиях, который связаны с большим числом исходных сущностей требований.

Для обоих типов диаграмм пользователь может выбирать отдельные узлы и ребра графа или применять фильтры и срезы данных, чтобы сфокусироваться на определенных аспектах и посмотреть подробную информацию только по этой части графа. Например, можно отфильтровать требования по определенному порогу частоты, чтобы отобразить только наиболее распространенные навыки, или выбрать определенный кластер требований и детально изучить связи внутри только этого кластера.

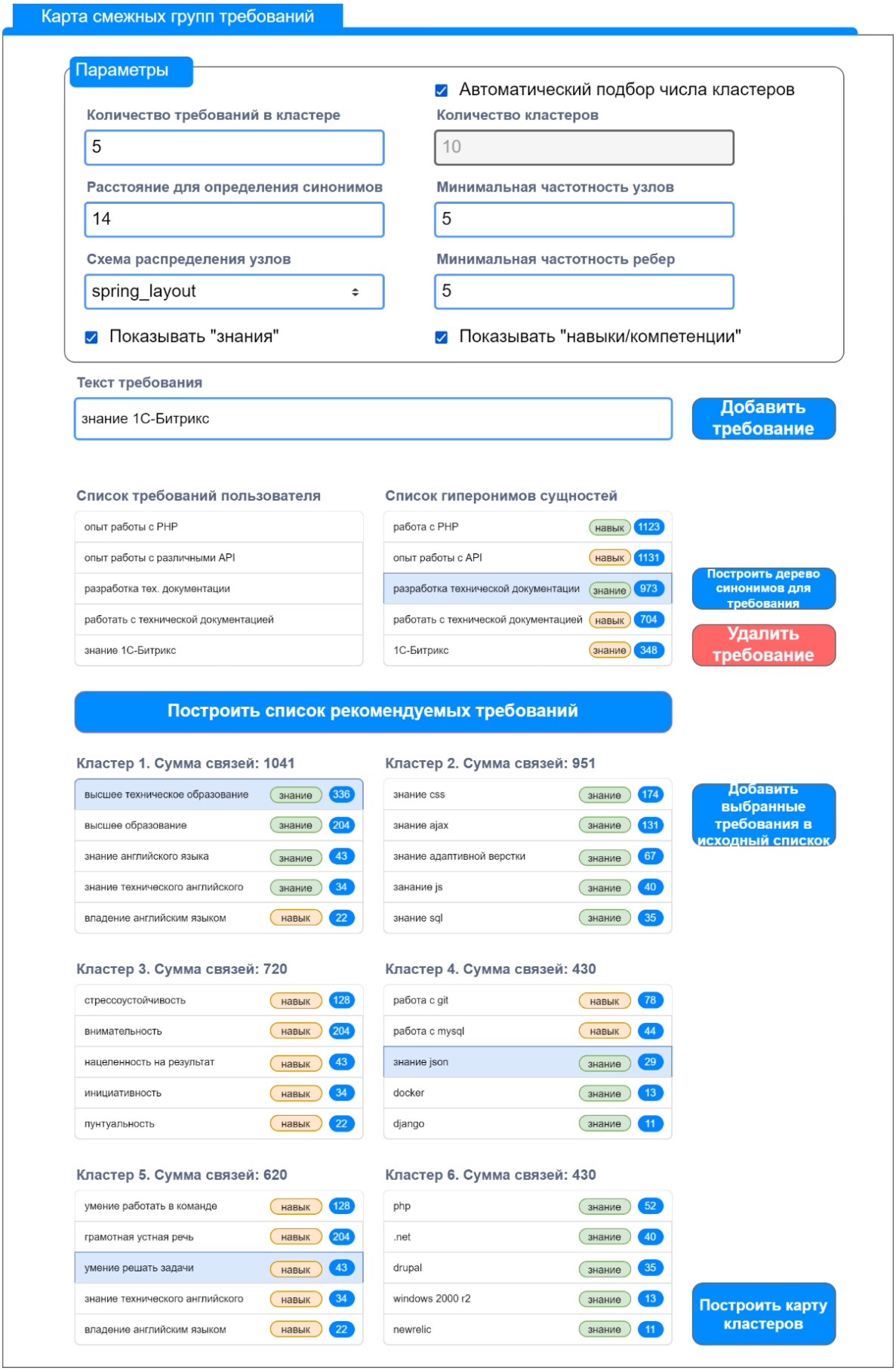

Рисунок 28 – Интерфейс формирования кластеров смежных групп требований

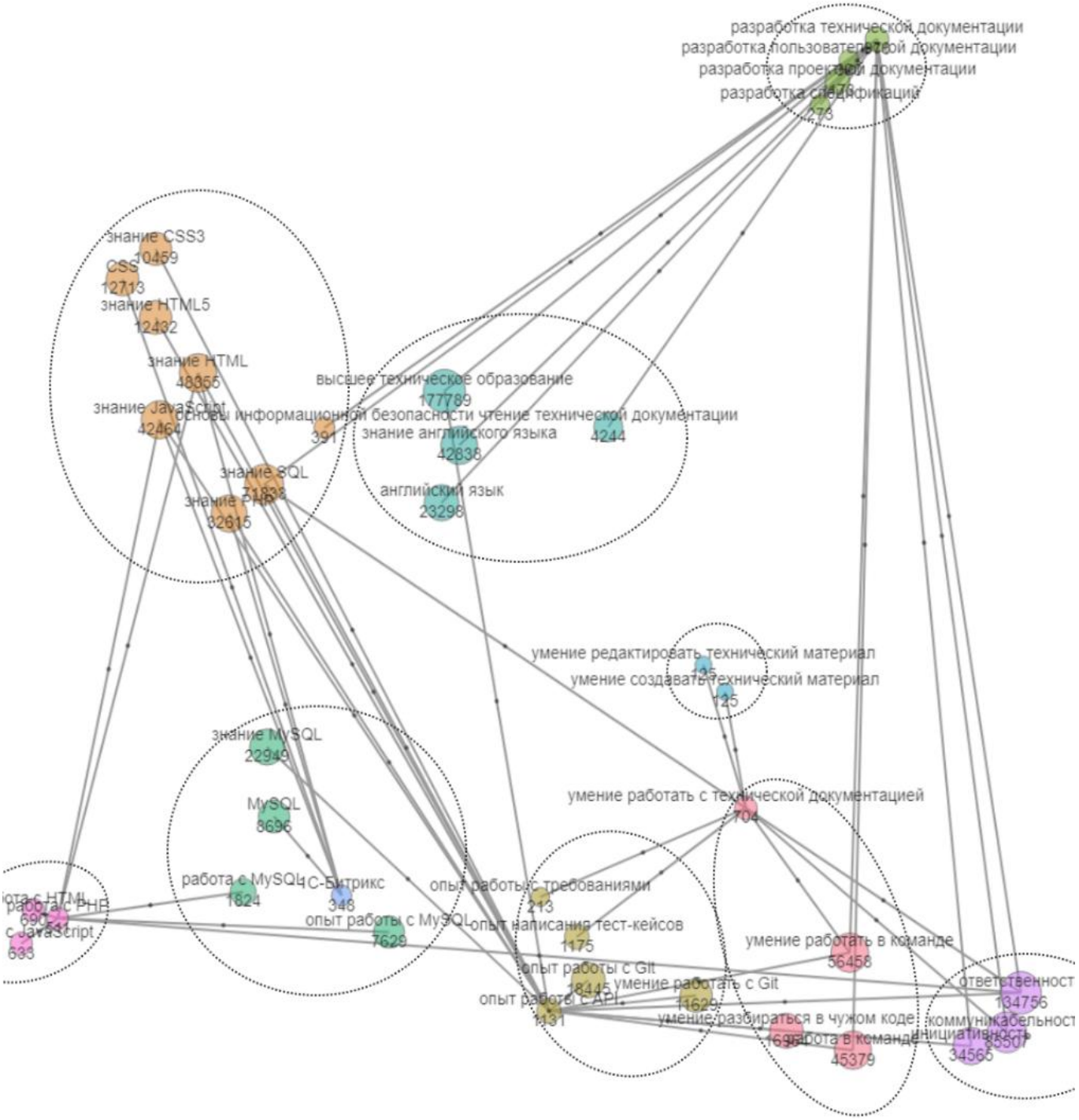

Рисунок 29 – Пример карты смежных групп требований
для профессии web-разработчик

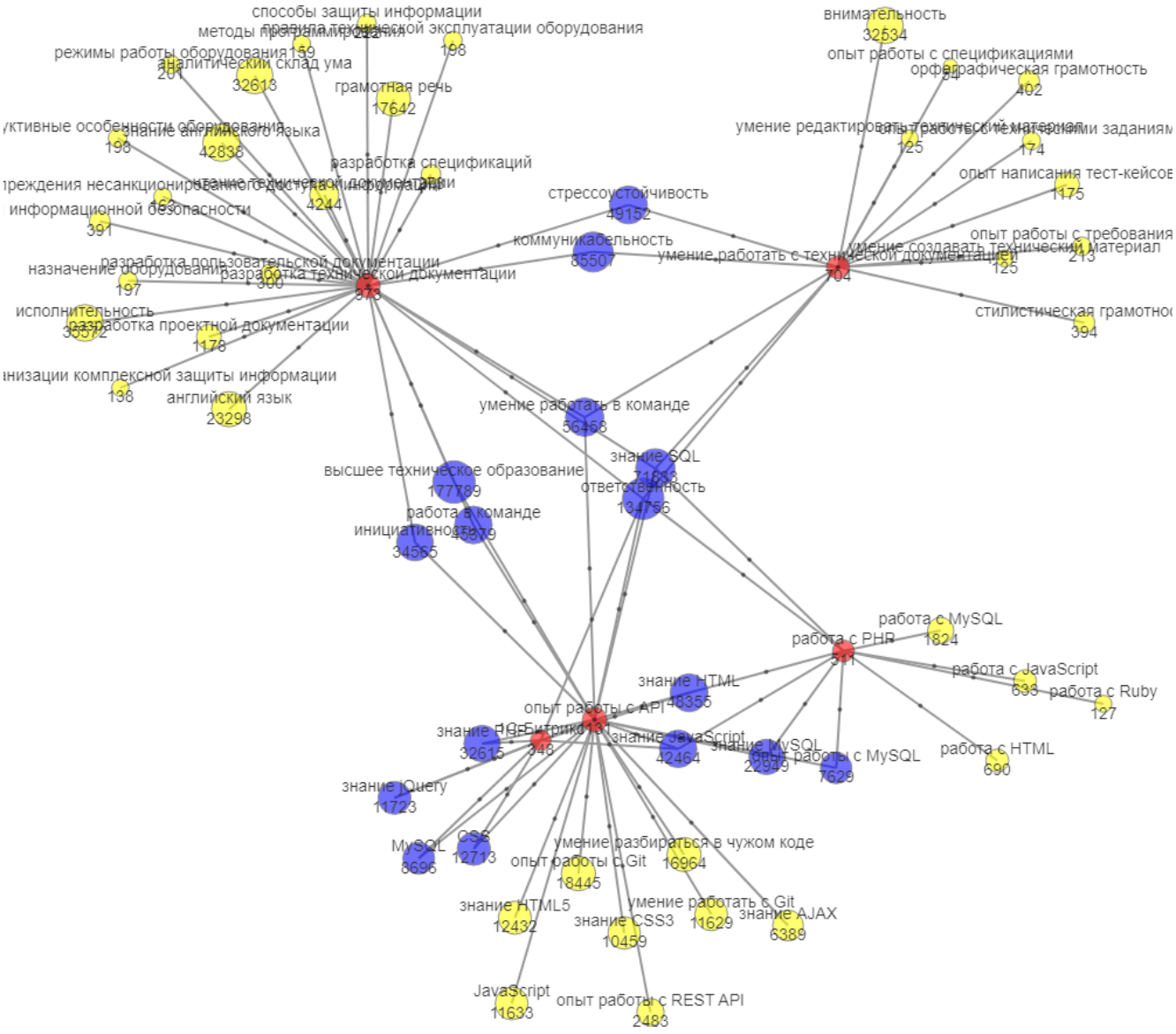

Рисунок 30 – Примеры карты совместной встречаемости смежных сущностей требований для профессии web-разработчик

Прототип системы позволяет пользователю выбрать отдельную сущность требований и построить для нее дерево синонимов. Пример дерева синонимов для сущности навыка/компетенции «описание бизнес-процессов» представлен на рисунке 31.

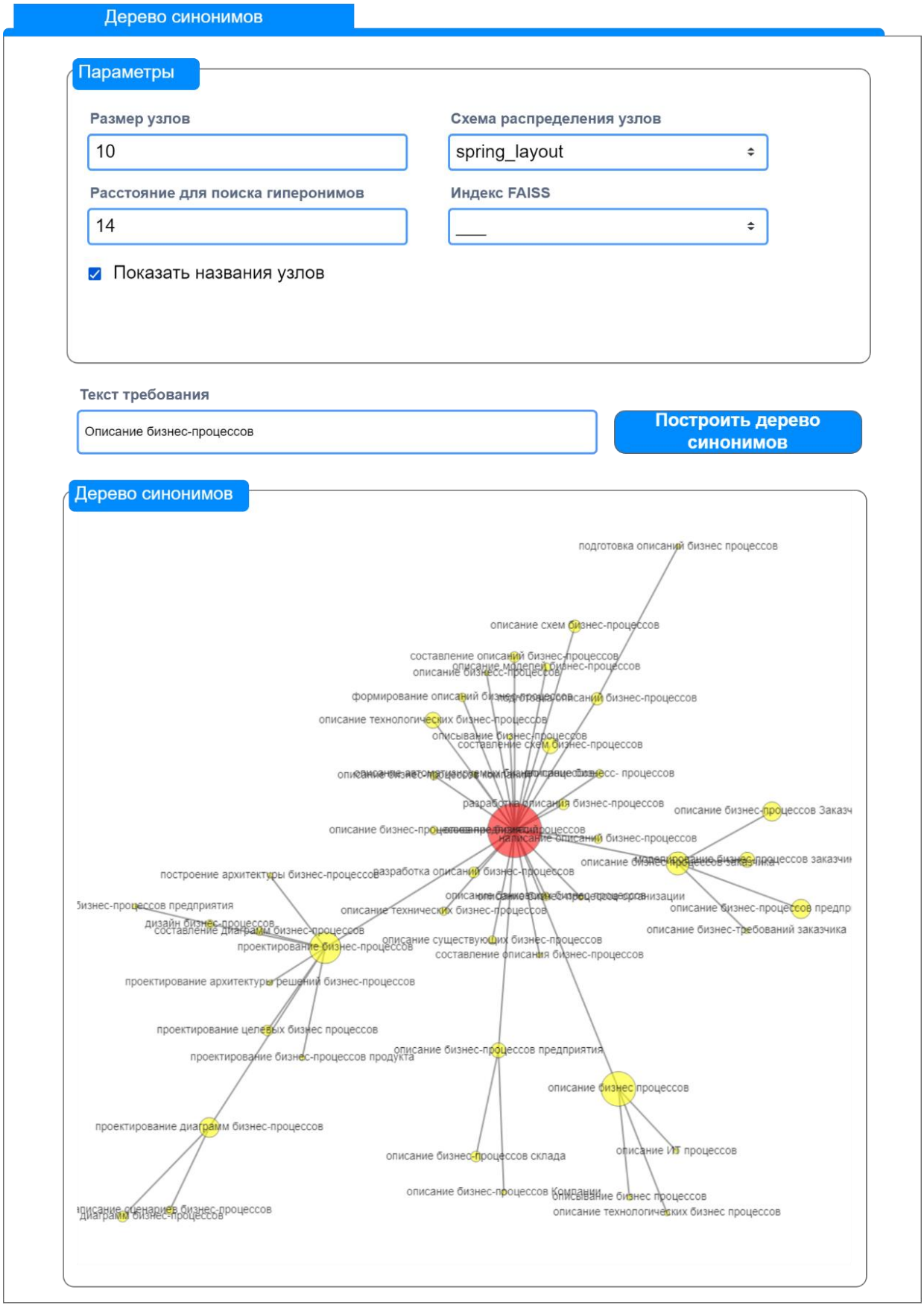

Рисунок 31 – Интерфейс системы формирования дерева синонимов для навыка/компетенции «описание бизнес-процессов»

На примере дерева синонимов можно видеть связи гипоним-гипероним (частное-общее) для синонимов сущностей заданного требования. Красным цветом выделена выбранная сущность, а желтым все остальные сущности. Размер узлов зависит от частоты встречаемости формулировки сущности требования среди всего числа вакансий.

Анализ дерева синонимов позволяет увидеть иерархическую структуру связей семантически близких сущностей требований. Такой вид анализа дает специалисту ряд преимуществ и возможностей:

1. *Унификация терминологии.* Дерево синонимов помогает выявить и связать различные термины, используемые для обозначения одних и тех же или близких по смыслу требований. Это позволяет унифицировать терминологию, используемую в описаниях вакансий, и обеспечить более согласованное и последовательное представление требований к кандидатам.

2. *Расширение семантического поиска.* При поиске и подборе кандидатов на вакансию специалист может использовать дерево синонимов для расширения поискового запроса. Включение в запрос не только основного термина, но и его синонимов позволяет охватить более широкий круг потенциальных кандидатов, которые могут обладать требуемыми знаниями и навыками, но описывать их иными словами в своих резюме.

3. *Выявление скрытых взаимосвязей*. Дерево синонимов может помочь выявить неочевидные взаимосвязи между различными сущносятими требований. Например, он может показать, что определенные сущности часто встречаются вместе или являются смежными, даже если они описаны разными терминами. Это дает специалисту более полное понимание структуры и взаимосвязей требований в конкретной профессиональной области.

Дополнительные параметры и настройки прототипа рекомендательной системы:

Дополнительные параметры и настройки, которые пользователь может изменять на интерактивной карте требований, чтобы сделать ее максимально информативной и адаптированной под свои задачи.

1. Уровень детализации требований:

– Возможность укрупнять или разбивать группы требований;

– Настройка количества отображаемых кластеров и подкластеров;

– Опция "Показать/скрыть" для отдельных групп или узлов.

2. Временной срез:

– Выбор периода, за который анализируются вакансии (за последний месяц, квартал, год);

– Возможность сравнения карт требований за разные периоды.

3. Источники данных:

– Выбор сайтов и платформ, с которых собираются вакансии;

– Настройка приоритета и веса различных источников.

4. Фильтрация по параметрам вакансий:

– Выбор региона или города;

– Фильтрация по профессии.

5. Визуальное оформление:

– Выбор цветовой схемы и стиля отображения карты;

– Настройка размера и формы узлов и связей.

6. Экспорт данных:

– Экспорт карты требований в различных форматах (PDF, PNG, CSV).

Ограничения прототипа рекомендательной системы

Разработанный прототип рекомендательной системы обладает некоторыми функциональными ограничениями и особенностями. Они вызваны тем, что предложенные интеллектуальные методы и алгоритмы могут быть применены

только к определенному набору входных данных, имеющемуся в распоряжении разработчиков.

Рассматриваются только требования и вакансии, связанные с отраслью информационных технологий, и оформленные в электронном виде.

Разработанные методы имеют низкую точность при работе с сущностями знаний и навыков/компетенциями, имеющими низкую частоту встречаемости в текстах вакансий.

Разработанные методы и алгоритмы извлечения сущностей не обладают обобщающей способностью, и не способны извлекать сущности, которые не представлены в исходном датасете отдельных сущностей требований.

Значительные обновления базы данных вакансий требует переобучения используемых нейросетевых моделей для сохранения высокой точности семантического сопоставления сущностей требований, а также актуализации информации о частоте встречаемости отдельных сущностей требований и частоты совместной встречаемости сущностей друг с другом.

Если список первоначальных требований пользователя имеет небольшое количество сущностей, или представленные сущности имеют широкое использование (например, softskills), тогда рекомендации системы могут охватывать очень широкий набор смежных групп требований, что может приводить к снижению эффективности рекомендацию.

4.4. Оценка эффективности прототипа интеллектуальной рекомендательной системы

Для оценки эффективности рекомендательной системы были использованы следующие метрики:

– Оценка изменения среднего времени на формирование описания вакансии;

– Оценка изменения среднего количества соискателей;

– Оценка изменения среднего времени закрытия одной вакансии.

Для вычисления этих трех метрик на базе двух предприятий ЮНИИТ и ООО Фирма «Интерсвязь» в период с 2022 по 2024 год была собрана информация о подготовке 87 вакансий до внедрения системы (контрольная группа), и о 57 вакансиях после внедрения системы (тестовая группа).

Для оценки изменения времени затраченного на формирование описания вакансии проводился сравнительный анализ времени, затраченного на создание одной вакансии до и после внедрения рекомендательной системы, включая время, затраченное специалистом на поиск и анализ информации о требованиях по аналогичным позициям в системах онлайн-рекрутмента, и время, затраченное на создание и редактирование списка требований и текста вакансии.

До внедрения системы у специалистов по подбору персонала уходило в среднем до 4 часов на подготовку одной вакансии, а после внедрения системы это время сократилось до 1,5 часов в среднем на одну вакансию, что свидетельствует о повышении производительности и эффективности процесса подбора персонала после внедрения системы. Процентное снижение среднего времени на подготовку одной вакансии Δt вычислялось по формуле 4.1:

$$\Delta t = \frac{t_{\text{ср.до}} - t_{\text{ср.после}}}{t_{\text{ср.до}}} \cdot 100 \quad (4.1)$$

Где $t_{\text{ср.до}}$ – среднее время на подготовку одной вакансии до внедрения системы, $t_{\text{ср.после}}$ – среднее время на подготовку одной вакансии после внедрения системы.

В результате расчетов снижение среднего времени на подготовку одной вакансии составило 62,5% за наблюдаемый период.

Оценка изменения количества соискателей ΔN проводилась путем сравнения числа соискателей, приглашенных на собеседование до и после внедрения системы по формуле 4.2.

$$\Delta N = \frac{N_{\text{ср.до}} - N_{\text{ср.после}}}{N_{\text{ср.до}}} \cdot 100 \quad (4.2)$$

Где $N_{\text{ср.до}}$ – среднее количество соискателей, приглашенных на собеседование до внедрения системы, $N_{\text{ср.после}}$ – среднее время на подготовку одной вакансии до внедрения системы.

По результатам расчетов, среднее количество соискателей, приглашенных на собеседование до внедрения системы составило 3,7, а после внедрения 4,3, что в процентном соотношении составляет ↑16,2% в среднем на вакансию за наблюдаемый период.

Проведенный анализ показывает, что вакансии, которые были рекомендованы системой в части требований, привлекают больше внимания потенциальных соискателей. Это способствует увеличению числа поданных релевантных резюме и, следовательно, повышает количество проводимых собеседований в компаниях.

Оценка времени закрытия одной вакансии. Для этой оценки в обеих группах контрольной и тестовой фиксировалось общее время, затраченное на составление текста вакансии: подготовка требований, составление текста вакансии, опубликование, анализ резюме и проведение собеседований для каждой вакансии, до момента закрытия вакансии. Затем вычислялось среднее время закрытия одной вакансии в контрольной и тестовой группе. По результатам проведенного анализа время закрытия одной вакансии снизилось с 17,7 до 15,1 рабочих дней в среднем на одну вакансию, что в процентном соотношении составляет ↓14,7% за наблюдаемый период. Снижение среднего времени закрытия вакансии также является косвенным показателем повышения эффективности процесса подбора персонала в компаниях.

Результаты оценки эффективности применения рекомендательной системы в таблице 17.

По результатам, представленным в таблице 17, можно отметить комплексное повышение эффективности процесса подбора персонала после внедрения прототипа рекомендательной системы.

Таблица 17 – Результаты оценки эффективности применения рекомендательной системы

Метрика	Было	Стало	Процент изменения
Среднее время подготовки вакансии (часов)	4	1,5	↓62,5%
Количество соискателей в среднем на одну вакансию (шт.)	3,7	4,3	↑16,2%
Время закрытия одной вакансии (рабочих дней)	17,7	15,1	↓14,7%

Выводы по четвертой главе

1. Сформулированы функциональные требования к прототипу интеллектуальной рекомендательной системы поддержки формирования списка требований для вакансии на основе общей функциональной модели процесса подбора персонала.

2. Спроектирована структура прототипа интеллектуальной рекомендательной системы, включая структуру ее компонентов, пользовательский интерфейс и схему базы данных. Выполнена программная реализация всех модулей системы согласно спроектированной структуре.

3. Проведена апробация прототипа системы при составлении требований к вакансиям IT-специалистов. Апробация проводилась в следующих организациях: Югорском научно-исследовательском институте (Ханты-Мансийский автономный округ) и ООО Фирма «Интерсвязь» (город Челябинск). В ходе тестирования система на основе разработанных интеллектуальных алгоритмов и моделей формировала рекомендации по необходимым требованиям к кандидатам. Это позволило оценить работоспособность прототипа и его способность давать релевантные рекомендации в реальных условиях.

4. Проведена комплексная оценка эффективности использования интеллектуальной рекомендательной системы в процессе подбора персонала по нескольким метрикам: среднее время на подготовку одной вакансии снизилось с 4 до 1,5 часов или ↓62,5%; количество соискателей возросло с 3,7 на 4,3 или ↑16,2%; сокращение времени закрытия одной вакансии с 17,7 до 15,1 рабочих дней или ↓14,7%.

ЗАКЛЮЧЕНИЕ

В рамках диссертационного исследования была достигнута поставленная цель – разработка методов и алгоритмов интеллектуальной поддержки процесса формирования списка требований к вакансиям, которые обеспечиваю повышение качества анализа современных тенденций рынка труда, и повышают эффективность процессов подбора персонала и соответствие разрабатываемых требований в проектах вакансий реальным потребностям рынка труда. Предложенные решения позволяют повысить качество анализа современных тенденций на рынке труда. Кроме того, они способствуют более эффективному подбору персонала, обеспечивая соответствие формируемых требований к вакансиям реальным потребностям, существующим на рынке труда. Внедрение разработанных методов и алгоритмов интеллектуальной поддержки помогает усовершенствовать процесс формирования требований к кандидатам и выбора наиболее подходящих специалистов на вакантные должности.

В ходе исследования были достигнуты следующие результаты:

1. На основе результатов анализа предметной области с учетом выявленных проблем и ограничений существующих методов предложена структурно-семантическая модель формализованного описания требований рынка труда на уровне отдельных сущностей знаний и навыков/компетенций. Данная модель позволяет учитывать структурные и семантические связи между отдельными сущностями требований, и может быть сформирована автоматически.

2. На основе нейросетевых моделей языка и методов классификации разработан метод извлечения отдельных сущностей знаний и навыков/компетенций из текстов требований вакансий реального рынка труда, который в отличии от существующих методов не требует, чтобы искомые сущности были представлены в виде последовательности подряд идущих синтаксических или лексических конструкций. По результатам экспериментов предложенный метод демонстрирует существенный прирост качества извлечения

отдельных сущностей знаний и навыков/компетенций из текстов требований вакансий по метрике F1 по сравнению с существующими методами, основанными на правилах, 0.81 против 0.31, соответственно.

3. Разработан новый метод поддержки формирования требований вакансии на основе семантического сопоставления сущностей знаний и навыков/компетенций предложенной структурно-семантической модели и методов кластеризации, который обеспечивает соответствие разрабатываемых требований в проектах вакансий реальным потребностям рынка труда.

4. По результатам экспериментальной оценки точность семантического поиска под требования пользователей составила ~0.82 по метрике MAP@5, что свидетельствует о достаточно высокой релевантности и адекватности выдаваемых системой рекомендаций по комплексной оценке экспертов, а также о том, что полученные рекомендации могут быть использоваться на практике для принятия решений в процессе формировании списка требований при составлении текстов вакансий.

5. На основе предложенных моделей, методов и алгоритмов разработан прототип интеллектуальной рекомендательной системы поддержки формирования требований в проектах вакансий на основе предложенных моделей, методов и алгоритмов. Проведена апробация прототипа в Югорский научно-исследовательский институт (г. Ханты-Мансийск), в компании ООО Фирма «Интерсвязь» (г. Челябинск) и в Челябинском государственном университете (г. Челябинск). По результатам экспертной оценки, отмечается высокая релевантность предлагаемых системой рекомендаций вариантов требований. Экспертами также отмечается существенное сокращение времени на подготовку и согласование списка требований, среднее время на подготовку одной вакансии снизилось с 4 до 1,5 часов или ↓62,5%; количество соискателей возросло с 3,7 на 4,3 или ↑16,2%; сокращение времени закрытия одной вакансии с 17,7 до 15,1 рабочих дней или ↓14,7%.

Перспективы дальнейших исследований. Дальнейшее развитие интеллектуальной системы предполагает расширение возможностей использования современных нейросетевых моделей, в том числе и больших языковых моделей (англ. LLM – large language model), для получения эмбеддингов текстов и извлечения сущностей знаний и навыков/компетенций из текстов требований. Также рассматривается возможность применения интеллектуальной рекомендательной системы экспертами для актуализации содержания государственных классификаторов занятий и профессиональных стандартов.

Благодарности. Автор выражает признательность коллективу лаборатории машинного обучения и интеллектуального анализа данных Челябинского государственного университета, в частности Д.С. Ботову, А.В. Вохминцеву, Ю.В. Дмитрину, И.А. Рязанову, за неоценимую помощь в проведении экспериментов и подготовке совместных научных публикаций. Автор благодарит руководство федерального интернет-провайдера компании ООО фирма «Интерсвязь» (г. Челябинск) и Югорского НИИ информационных технологий (г. Ханты-Мансийск) за предоставленные вычислительные мощности для проведения экспериментов и апробации прототипа системы. Особую благодарность автор выражает своему научному руководителю, профессору Андрею Витальевичу Мельникову, без помощи и поддержки которого эта работа не была бы завершена.

СПИСОК ЛИТЕРАТУРЫ

1. Антонова, А.Ю. Использование метода условных случайных полей для обработки текстов на русском языке / А.Ю. Антонова, А.Н. Соловьев // Диалог. – 2013. – С. 27-44.

2. Бавыкина, Е.Н. Качественный анализ рынка труда молодых специалистов / Е.Н. Бавыкина, С.С. Гущина, Т.В. Корецкая // Научно-методический электронный журнал «Концепт». – 2017. – Т. 31. – С. 1446-1450.

3. Библиотека scikit-learn. – URL: https://scikit-learn.org/stable/index.html (дата обращения: 01.12.2023).

4. Библиотека transformers от Hugging Face. – URL: https://huggingface.co/docs/transformers/index (дата обращения: 01.12.2023).

5. Ботов, Д.С. Извлечение информации с использованием нейросетевых моделей языка на примере анализа вакансий в системах онлайн-рекрутмента / Д.С. Ботов, Ю.Д. Кленин, И.Е. Николаев // Вестник Югорского государственного университета. – 2018. – №3 (50). – С. 37-48.

6. Волгин, А.Д. Спрос на навыки: анализ на основе онлайн данных о вакансиях / А.Д. Волгин, В.Е. Гимпельсон // Экономический журнал Высшей школы экономики. – 2022. – Т. 26. – №. 3. – С. 343-374.

7. Волошина, И.А. Модель выявления востребованности профессий: ключевые параметры и некоторые особенности / И.А. Волошина, Л.В. Козлова, П.Н. Новиков // Социально-трудовые исследования. – 2019. – 4(37). – 106–119.

8. Гаврилюк, В.В. Профессиональное самоопределение молодежи нового рабочего класса / В.В. Гаврилюк //Известия высших учебных заведений. Социология. Экономика. Политика. – 2019. – №. 1. – С. 43-48.

9. Епанчинцев, А.О. Сетевой анализ в современной социологической теории и его приложение к исследованию рынка труда / А.О. Епанчинцев // Вестник Таганрогского института имени АП Чехова. – 2006. – №. 2. – С. 182-185.

10. Калиновская, И.Н. Анализ представленных на рынке труда компетенций, извлеченных из цифровых источников с помощью искусственного интеллекта / И.Н. Калиновская // Экономика и общество: международный научно-практический журнал. – 2022. – № 04 (22).

11. Корнев, В.А. Основные виды семантических отношений внутри лексико-семантических группировок / В.А. Корнев, О.М. Дедова, О.В. Дедова // Язык и текст. – 2019. – Т. 6. – №. 3. – С. 51-55.

12. Лебедев, П.А. Аналитика данных резюме и вакансий / П.А. Лебедев, Д. А. Шурыгина // Мониторинг общественного мнения: экономические и социальные перемены. – 2015. – №. 3 (127). – С. 150-151.

13. Маматов, А.В. Методы, модели и алгоритмы построения систем поддержки принятия решений в управлении кадровым потенциалом региона на основе ситуационно-поведенческого подхода : дис. ... д-ра. техн. наук : 05.13.10 / А.В. Маматов ; НИУ БелГУ. - Белгород, 2020. - 334 с.

14. Национальная технологическая инициатива «Кадровое обеспечение промышленного роста» // Агентство стратегических инициатив. – URL: https://asi.ru/staffing/ (дата обращения: 01.12.2023).

15. Николаев И.Е. Интеллектуальный метод формирования списка требований профиля должности на основе нейросетевых моделей языка с использованием таксономии ESCO и корпуса онлайн-вакансий / И.Е. Николаев // Бизнес-информатика. – 2023. – Т. 17. – №. 2. – С. 71-84.

16. Николаев, И.Е. Метод извлечения знаний и навыков/компетенций из текстов требований вакансий / И.Е. Николаев // Онтология проектирования. – 2023. – Т. 13. – №. 2 (48). – С. 282-293.

17. Николаев, И.Е. Сравнение нейросетевых моделей на архитектуре трансформеров в контексте задачи оценки компактности векторных представлений семантически близких текстов требований европейской классификации навыков ESCO / И.Е. Николаев, А.В. Мельников // Вестник Южно-Уральского

государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. – 2022. – Т. 22. – №. 3. – С. 19-29.

18. Пахомов, А.В. Эконометрическое моделирование занятости на основе отраслевой специфики / А.В. Пахомов, Е.А. Пахомова, О.В. Рожкова // Национальные интересы: приоритеты и безопасность. – 2017. – Т. 13. – №. 11 (356). – С. 2018-2034.

19. Половинко, В.С. Качественный анализ сегмента «управление персоналом» на рынке труда / В.С. Половинко, Р.А. Долженко, С.Б. Долженко // Вестник Омского университета. Серия «Экономика». – 2023. – Т. 21. – №. 1. – С. 112-123.

20. Программа «Цифровая экономика Российской Федерации». – URL: http://static.government.ru/media/files/9gFM4FHj4PsB79I5v7yLVuPgu4bvR7M0.pdf (дата обращения: 01.12.2023).

21. Сапунов, А.В. Оценка последствий влияния пандемии коронавирусной инфекции на рынок труда в России / А.В. Сапунов // Экономика и бизнес: теория и практика. – 2022. – №. 4-2. – С. 109-112.

22. Свидетельство о государственной регистрации программы для ЭВМ №2023669298 от 13 сентября 2023 г. «Автоматизированная система формирования рекомендаций для составления списка требований вакансии» / Николаев И.Е. М.: Роспатент, 2023.

23. Синтаксический парсер SPACY + UDPipe. – URL: https://github.com/TakeLab/spacy-udpipe (дата обращения: 03.12.2023).

24. Синтаксический парсер от проектной группы DeepPavlov. – URL: https://docs.deeppavlov.ai/en/latest/features/models/syntax_parser.html (дата обращения: 03.12.2023).

25. Ташпулатов, А. Применение социологических исследований в сфере занятости сельского населения / А. Ташпулатов // Экономика и социум. – 2020. – №. 7 (74). – С. 398-404.

26. Толкачева, О.П. Экономико-статистический анализ влияния рынка труда на экономическую безопасность страны / О.П. Толкачева // Известия высших учебных заведений. Серия: Экономика, финансы и управление производством. – 2023. – №. 3 (57). – С. 59-69.

27. Трифонов, А.А. Алгоритмы построения инвертированного индекса для коллекции текстовых данных / А.А. Трифонов // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2013. – №. 3 (27). – С. 52-61.

28. Федеральный проект «Кадры для цифровой экономики» нацпрограммы «Цифровая экономика России 2024». – URL: https://data-economy.ru/education (дата обращения: 01.12.2023).

29. Хакимова, Е.М. Сложные предложения в современном русском языке: ортологический аспект / Е.М. Хакимова // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. – 2013. – Т. 10. – №. 2. – С. 10-14.

30. Хохлова, О.А. Чойжалсанова А.Ц. Разработка алгоритма анализа вакансий на рынке труда по данным из открытых источников / О. А. Хохлова, А. Н. Хохлова // Вопросы статистики. – 2022. – Т. 29. – №. 4. – С. 33-41.

31. Хранилище открытых языковых моделей Hugging Face. – URL: https://huggingface.co/models (дата обращения: 01.12.2023).

32. Цыпин, А.П. Ретроспективный статистический анализ рынка труда на постсоветском пространстве / А.П. Цыпин, М.М. Шайлиева, А.С. Сорокин, И.Б. Хмелев // Вестник евразийской науки. – 2019. – Т. 11. – №. 6. – С. 57.

33. Чеканова, Е.В. Статистический анализ рынка труда в регионах России / Е.В. Чеканова // Социально-экономические исследования, гуманитарные науки и юриспруденция: теория и практика. – 2016. – №. 6. – С. 31-37.

34. Шварц, Ю.А. Эконометрическое моделирование рынка труда РФ / Ю.А. Шварц, А.И. Смирнова // Умная цифровая экономика. – 2022. – Т. 2. – №. 2. – С. 81-86.

35. Alam, M. Entity-Based Short Text Classification Using Convolutional Neural Networks / M. Alam, Q. Bie, R. Türker, H. Sack // International Conference on Knowledge Engineering and Knowledge Management. Springer. Cham. – 2020. – С. 136-146.

36. Alekseeva, L. The demand for AI skills in the labor market / L. Alekseeva, J. Azar, M. Gine, S. Samila, B. Taska // Labour economics. – 2021. – Т. 71. – С. 102002.

37. Alsmadi, I.M. Short text classification using feature enrichment from credible texts / I.M. Alsmadi, K.H. Gan // International Journal of Web Engineering and Technology. – 2020. – Т. 15. – №. 1. – С. 59-80.

38. Beltagy, I. SciBERT: A pretrained language model for scientific text / I. Beltagy, K. Lo, A. Cohan // arXiv preprint arXiv:1903.10676. – 2019.

39. Berger, A. A maximum entropy approach to natural language processing / A. Berger, S.A. Della Pietra, V.J. Della Pietra // Computational linguistics. – 1996. – Т. 22. – №. 1. – С. 39-71.

40. Blei, D.M Latent dirichlet allocation / D.M. Blei, A.Y. Ng, M.I. Jordan // Journal of machine Learning research. – 2003. – Т. 3. – №. Jan. – С. 993-1022.

41. Bojanowski, P. Enriching word vectors with subword information / P. Bojanowski, E. Grave, A. Joulin, T. Mikolov // Transactions of the association for computational linguistics. – 2017. – Т. 5. – С. 135-146.

42. Boselli, R. WoLMIS: a labor market intelligence system for classifying web job vacancies / R. Boselli M. Cesarini, S. Marrara, F. Mercorio, M. Mezzanzanica, G. Pasi, M. Viviani //Journal of intelligent information systems. – 2018. – Т. 51. – С. 477-502.

43. Botov, D. Mining labor market requirements using distributional semantic models and deep learning / D. Botov, , J. Klenin, A. Melnikov, Y. Dmitrin, I. Nikolaev, M. Vinel // Business Information Systems: 22nd International Conference, BIS 2019, Seville, Spain, June 26–28, 2019, Proceedings, Part II 22. – Springer International Publishing, 2019. – С. 177-190.

44. Brown, T.B. Language models are few-shot learners / T.B. Brown // arXiv preprint arXiv:2005.14165. – 2020.

45. Celebi, M.E. A comparative study of efficient initialization methods for the k-means clustering algorithm / M.E. Celebi, H.A. Kingravi, P.A. Vela // Expert systems with applications. – 2013. – Т. 40. – №. 1. – С. 200-210.

46. Chalkidis, I. LEGAL-BERT: The muppets straight out of law school / I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos // arXiv preprint arXiv:2010.02559. – 2020.

47. Colace, F. Towards labour market intelligence through topic modelling / F. Colace, M. De Santo, M. Lombardi, F. Mercorio, M. Mezzanzanica, F. Pascale // Proceedings of the 52nd Hawaii International Conference on System Sciences. – 2019.

48. Colombo, E. Applying machine learning tools on web vacancies for labour market and skill analysis / E. Colombo, F. Mercorio, M. Mezzanzanica // Terminator or the Jetsons? The Economics and Policy Implications of Artificial Intelligence. – 2018.

49. Davies, D.L. A cluster separation measure. / D.L. Davies, D.W. Bouldin // IEEE transactions on pattern analysis and machine intelligence. – 1979. – №. 2. – С. 224-227.

50. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding / J. Devlin, M.W. Chang, K. Lee, K. Toutanova // arXiv preprint arXiv:1810.04805. – 2018.

51. Dmitrin, Y.V. Comparison of deep neural network architectures for authorship attribution of Russian social media texts / Y.V. Dmitrin, D.S. Botov, J.D. Klenin, I.E. Nikolaev // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2018». – 2018.

52. Eggertsson T. Concept paper on Labour Market Information System. / T. Eggertsson // Economic Behavior and Institutions. Cambridge Univ. Press. 2011. – URL: http://www.cgsc.in/Concept_Paper_LMIS.pdf (дата обращения: 01.12.2023)

53. European Skills, Competences, Qualifications and Occupations (ESCO) – URL: https://ec.europa.eu/ (дата обращения: 01.12.2023).

54. Ezen-Can A. A Comparison of LSTM and BERT for Small Corpus / A. A. Ezen-Can // arXiv preprint arXiv:2009.05451. – 2020.

55. FAISS (Facebook AI Similarity Search) – это библиотека для эффективного поиска сходных векторов в больших наборах данных. – URL: https://faiss.ai/ (дата обращения: 01.12.2023).

56. Fensel, D. Introduction: what is a knowledge graph? / D. Fensel, U. Şimşek, K. Angele, E. Huaman, E. Kärle, O. Panasiuk, A. Wahler // Knowledge graphs: Methodology, tools and selected use cases. – 2020. – С. 1-10.

57. Flisar, J. Improving Short Text Classification using Information from DBpedia Ontology / J. Flisar, V. Podgorelec // Fundamenta Informaticae. – 2020. – Т. 172. – №. 3. – С. 261-297.

58. Gage P. A new algorithm for data compression / P. Gage // The C Users Journal. – 1994. – Т. 12. – №. 2. – С. 23-38.

59. Garrido-Merchan, E.C. Comparing neural models against traditional machine learning text classification. / E.C. Garrido-Merchan, S. Gonzalez Carvajal // https://arxiv.org/abs/2005.13012. – 2020.

60. Garrido-Merchan, E.C. Comparing BERT against traditional machine learning models in text classification / E.C. Garrido-Merchan, R. Gozalo-Brizuela, S. Gonzalez-Carvajal // Journal of Computational and Cognitive Engineering. – 2023. – Т. 2. – №. 4. – С. 352-356.

61. Giabelli, A. GraphLMI: A data driven system for exploring labor market information through graph databases / A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica // Multimedia Tools and Applications. – 2022. – Т. 81. – №. 3. – С. 3061-3090.

62. Giabelli, A. NEO: A tool for taxonomy enrichment with new emerging occupations / A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, A. Seveso // International Semantic Web Conference. – Cham: Springer International Publishing, – 2020. – С. 568-584.

63. Halkidi, M. On clustering validation techniques / M. Halkidi, Y. Batistakis, M. Vazirgiannis // Journal of intelligent information systems. – 2001. – Т. 17. – С. 107-145.

64. Hao, M. Chinese Short Text Classification with Mutual-Attention Convolutional Neural Networks / M. Hao, B. Xu, J.Y. Liang, B.W. Zhang, X.C. Yin // ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP). – 2020. – Т. 19. – №. 5. – С. 1-13.

65. Hartigan, J.A. Algorithm AS 136: A k-means clustering algorithm / J.A. Hartigan, M.A. Wong // Journal of the royal statistical society. series c (applied statistics). – 1979. – Т. 28. – №. 1. – С. 100-108.

66. Hofmann, T. Probabilistic latent semantic analysis / T. Hofmann // UAI. – 1999. – Т. 99. – С. 289-296.

67. Indyk, P. Approximate nearest neighbors: towards removing the curse of dimensionality / P. Indyk, R. Motwani // Proceedings of the thirtieth annual ACM symposium on Theory of computing. – 1998. – С. 604-613.

68. Kamada, T. An algorithm for drawing general undirected graphs. / T. Kamada, S. Kawai // Information processing letters. – 1989. – Т. 31. – №. 1. – С. 7-15.

69. Kane, L.O. Digitalization in the German Labor Market: Analyzing Demand for Digital Skills in Job Vacancies. / L.O. Kane, R. Narasimhan, J.N. Burning, B. Taska – Bertelsmann-Stiftung, 2020. – С. 1-58.

70. Kertkeidkachorn, N. T2KG: An end-to-end system for creating knowledge graph from unstructured text / N. Kertkeidkachorn, R. Ichise // Workshops at the Thirty-First AAAI Conference on Artificial Intelligence. – 2017.

71. Koncel-Kedziorski, R. Text generation from knowledge graphs with graph transformers / R. Koncel-Kedziorski, D. Bekal, Y. Luan, M. Lapata, H. Hajishirzi // arXiv preprint arXiv:1904.02342. – 2019.

72. Lample, G. Neural architectures for named entity recognition / G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer // arXiv preprint arXiv:1603.01360. – 2016.

73. Le, Q. Distributed representations of sentences and documents / Q. Le, T. Mikolov // International conference on machine learning. – PMLR, 2014. – С. 1188-1196.

74. Le, T.A. Application of a Hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition / T.A. Le, M.Y. Arkhipov, M.S. Burtsev // Artificial Intelligence and Natural Language: 6th Conference, AINL 2017, St. Petersburg, Russia, September 20–23, 2017, Revised Selected Papers 6. – Springer International Publishing, 2018. – С. 91-103.

75. Lee, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining / J. Lee, W. Yoon, S. Kim, D.Kim, S. Kim, C.H. So, J. Kang // Bioinformatics. – 2020. – Т. 36. – №. 4. – С. 1234-1240.

76. Li, S. Constructing Software Knowledge Graph from Software Text. – URL: http://courses.cecs.anu.edu.au/courses/CSPROJECTS/18S1/reports/ u5831882.pdf – 2018. (дата обращения: 01.12.2023).

77. Liu, T. The Influence of Text Length on Text Classification Model / T. Liu, Y. Liang, Z. Yu // International Conference on Green, Pervasive, and Cloud Computing. – Springer Singapore, 2020. – С. 79-90.

78. Liu, Y. RoBERTa: A Robustly Optimized BERT Pretraining Approach / Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov // arXiv preprint arXiv:1907.11692. – 2019.

79. Lu, Z. VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification / Z. Lu, P. Du, J.Y. Nie // European Conference on Information Retrieval. – Springer International Publishing, 2020. – С. 369-382.

80. Luo, A. Deep semantic match model for entity linking using knowledge graph and text / A. Luo, S. Gao, Y. Xu // Procedia Computer Science. – 2018. – Т. 129. – С. 110-114.

81. Malandri, L. Meet: A method for embeddings evaluation for taxonomic data / L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani // 2020 International Conference on Data Mining Workshops (ICDMW). – IEEE, 2020. – С. 31-38.

82. Malkov, Y.A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs / Y.A. Malkov, D.A. Yashunin // IEEE transactions on pattern analysis and machine intelligence. – 2018. – Т. 42. – №. 4. – С. 824-836.

83. Malkov, Y. Approximate nearest neighbor algorithm based on navigable small world graphs / Y. Malkov, A. Ponomarenko, A. Logvinov, V. Krylov // Information Systems. – 2014. – Т. 45. – С. 61-68.

84. Mezzanzanica, M. Big data enables labor market intelligence / M. Mezzanzanica, F. Mercorio // Encyclopedia of Big Data Technologies. Springer International Publishing. – Springer, 2018. – С. 1-11.

85. Mikolov, T. Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean // arXiv preprint arXiv:1301.3781. – 2013.

86. Morwal, S. Named entity recognition using hidden Markov model (HMM) / S. Morwal, N. Jahan, D. Chopra // International Journal on Natural Language Computing (IJNLC) Vol. – 2012. – Т. 1.

87. Narayan, A. Assessing single-cell transcriptomic variability through density-preserving data visualization / A. Narayan, B. Berger, H. Cho // Nature biotechnology. – 2021. – Т. 39. – №. 6. – С. 765-774.

88. Natasha – библиотека для извлечения структурированной информации из текстов на русском языке. – URL: https://natasha.github.io/ (дата обращения: 01.12.2023).

89. Nayak, A. Knowledge Graph from Informal Text: Architecture, Components, Algorithms and Applications / A. Nayak, V. Kesri, R.K. Dubey, S. Mandadi, V.G. Venkoparao, K. Ponnalagu, B.S. Garadi // Applications of Machine Learning. – 2020. – С. 75-90.

90. NetworkX – библиотека для работы с графовыми структурами. – URL: https://networkx.org/ (дата обращения: 01.12.2023).

91. Nikolaev, I. Use of Topic Modelling for Improvement of Quality in the Task of Semantic Search of Educational Courses / I. Nikolaev, D. Botov, Y. Dmitrin, J. Klenin, A. Melnikov // 21st International Workshop on Computer Science and Information Technologies (CSIT 2019). – Atlantis Press, 2019. – С. 104-111.

92. Nikolaev, I. The Comparison of Distributive Semantics Models Applied to the Task of Short Job Requirements Clustering for the Russian Labor Market / I. Nikolaev, I. Ryazanov, D. Botov // 8th Scientific Conference on Information Technologies for Intelligent Decision-Making Support (ITIDS 2020). – Atlantis Press, 2020. – С. 295-301.

93. Ostendorff, M. Enriching BERT with knowledge graph embeddings for document classification / M. Ostendorff, P. Bourgonje, M. Berger, J. Moreno-Schneider, G. Rehm, B. Gipp // arXiv preprint arXiv:1909.08402. – 2019.

94. Panchenko, A. Similarity measures for semantic relation extraction: PhD thesis / A. Panchenko // Université catholique de Louvain, Bauman Moscow State Technical University. Louvain-la-Neuve, Belgium. – 2013.

95. Pennington, J. Glove: Global vectors for word representation / J. Pennington, R. Socher, C.D. Manning // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – С. 1532-1543.

96. Peters, M.E. Deep contextualized word representations / M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer // arXiv preprint arXiv:1802.05365. – 2018.

97. Plotly – библиотека для для визуализации интерактивных графиков и диаграмм. – URL: https://plotly.com/python/ (дата обращения: 01.12.2023).

98. Ponomarenko, A. Approximate nearest neighbor search small world approach / A. Ponomarenko, Y. Malkov, A. Logvinov, V. Krylov // International Conference on Information and Communication Technologies & Applications. – 2011. – Т. 17.

99. Qian, G. Similarity between Euclidean and cosine angle distance for nearest neighbor queries / G. Qian, S. Sural, Y. Gu, S. Pramanik // In Proceedings of the 2004 ACM symposium on Applied computing – 2004. – С. 1232-1237.

100. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis / P. J. Rousseeuw // Journal of computational and applied mathematics. – 1987. – Т. 20. – С. 53-65.

101. Shavrina, T. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark / T. Shavrina, A. Fenogenova, A. Emelyanov, D. Shevelev, E. Artemova, V. Malykh, A. Evlampiev // arXiv preprint arXiv:2010.15925. – 2020.

102. Silva, V.S. Building a knowledge graph from natural language definitions for interpretable text entailment recognition / V.S. Silva, A. Freitas, S. Handschuh // arXiv preprint arXiv:1806.07731. – 2018.

103. Škrlj, B. tax2vec: Constructing interpretable features from taxonomies for short text classification / B. Škrlj, M. Martinc, J. Kralj, N. Lavrač, S. Pollak // Computer Speech & Language. – 2021. – Т. 65. – С. 101-104.

104. Slaney, M. Locality-sensitive hashing for finding nearest neighbors [lecture notes] / M. Slaney, M. Casey // IEEE Signal processing magazine. – 2008. – Т. 25. – №. 2. – С. 128-131.

105. Sparck, J.K. A Statistical Interpretation of Term Specificity and its Application in Retrieval / J.K. Sparck // Journal of Documentation. – 1972. – Т. 28. – №. 1. – С. 11-21.

106. Tayal, K. Short Text Classification using Graph Convolutional Network / K. Tayal, S.R. Nikhil, S. Agarwal, X. Jia, K. Subbian, V. Kumar // NIPS workshop on Graph Representation Learning. – 2019.

107. Ternikov, A.A. Demand for skills on the labor market in the IT sector / A.A. Ternikov, E.A. Aleksandrova // Бизнес-информатика. – 2020. – Т. 14. – №. 2 (eng). – С. 64-83.

108. The importance of LMI // UK Commission for Employment and Skills. 2015. – URL: https://www.gov.uk/government/publications/the-importance-of-labour-market-intelligence (дата обращения: 01.12.2023).

109. Van der Maaten, L. Visualizing data using t-SNE / L. Van der Maaten, G. Hinton // Journal of machine learning research. – 2008. – Т. 9. – №. 11.

110. Vargas, S. Exploiting the diversity of user preferences for recommendation / S. Vargas, P. Castells // Proceedings of the 10th conference on open research areas in information retrieval. – 2013. – С. 129-136.

111. Vaswani, A. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin // Advances in neural information processing systems. – 2017. – Т.30.

112. Vinel, M. Experimental Comparison of Unsupervised Approaches in the Task of Separating Specializations Within Professions in Job Vacancies / M. Vinel, D. Botov, I. Ryazanov, I. Nikolaev // Conference on Artificial Intelligence and Natural Language. – Springer International Publishing, 2019. – С. 99-112.

113. Wold, S. Principal component analysis / S. Wold, K. Esbensen, P. Geladi // Chemometrics and intelligent laboratory systems. – 1987. – Т. 2. – №. 1-3. – С. 37-52.

114. Wu, Y. Google's neural machine translation system: Bridging the gap between human and machine translation / Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, J. Dean // arXiv preprint arXiv:1609.08144. – 2016.

115. YARGY-парсер. – URL: https://github.com/natasha/yargy (дата обращения: 03.12.2023).

116. Zakaria, J. Clustering time series using unsupervised-shapelets / J. Zakaria, A. Mueen, E. Keogh // 2012 IEEE 12th International Conference on Data Mining. – IEEE, 2012. – С. 785-794.

117. Zaland, O. A Comprehensive Empirical Evaluation of Existing Word Embedding Approaches / O. Zaland, M. Abulaish, M. Fazil // arXiv preprint arXiv:2303.07196. – 2023.

118. Zhu Y. Short Text Expansion and Classification Based on Word Embedding / Y. Zhu // International Journal of Social Science and Education Research. – 2020. – T. 8. – C. 92120-92128.

ПРИЛОЖЕНИЕ А

Свидетельство о государственной регистрации программ для ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2023669298

Автоматизированная система формирования рекомендаций для составления списка требований вакансии

Правообладатель: ***Николаев Иван Евгеньевич (RU)***

Автор(ы): ***Николаев Иван Евгеньевич (RU)***

Заявка № **2023668540**

Дата поступления **05 сентября 2023 г.**

Дата государственной регистрации

в Реестре программ для ЭВМ ***13 сентября 2023 г.***

Руководитель Федеральной службы

по интеллектуальной собственности

Ю.С. Зубов

ПРИЛОЖЕНИЕ Б

Акты о внедрении результатов диссертационного исследования

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное
образовательное учреждение
высшего образования
«Челябинский государственный университет»
(ФГБОУ ВО «ЧелГУ»)

ул. Братьев Кашириных, 129, г. Челябинск, 454001
тел. (351) 799-71-01, факс: (351) 742-09-25
E-mail: odou@csu.ru; http://www.csu.ru
ОКПО 05121292, ОГРН 1027402324905,
ИНН/КПП 7447012841/744701001

_____№_____

На № _____от_____

УТВЕРЖДАЮ
Проректор по научной работе
ФГБОУ ВО «Челябинский
государственный университет»
_____ И.В. Бычков
«___» _____ 20__ г.

АКТ

использования в учебном процессе результатов диссертационной работы «Методы и алгоритмы интеллектуальной поддержки формирования требований вакансии на основе нейросетевых моделей языка и актуальных требований рынка труда» старшего преподавателя кафедры информационных технологий и экономической информатики *института информационных технологий* ФГБОУ ВО «Челябинский государственный университет» Николаева Ивана Евгеньевича, представленной на соискание ученой степени кандидата технических наук

Подтверждаем, что результаты, полученные Николаевым И.Е. в диссертационной работе на тему «Методы и алгоритмы интеллектуальной поддержки формирования требований вакансии на основе нейросетевых моделей языка и актуальных требований рынка труда», были апробированы в ФГБОУ ВО «Челябинский государственный университет» (далее - ЧелГУ).

В учебном процессе ЧелГУ при обучении студентов института информационных технологий по направлениям 09.03.04 Программная инженерия и 02.04.02 Фундаментальная информатика и информационные технологии используются следующие результаты диссертационной работы:

1. Модель формализованного описания требований реального рынка труда на уровне отдельных сущностей знаний и навыков/компетенций, которая позволила учитывать структурные и семантические отношения между ними.
2. Интеллектуальный метод извлечения сущностей знаний и навыков/компетенций из текстов требований вакансий реального рынка на основе нейросетевых моделей языка и методов классификации.
3. Метод поддержки формирования списка требований вакансий на основе семантического сопоставления сущностей знаний и навыков/компетенций предложенной структурно-семантической модели и методов кластеризации.
4. Программная реализация предложенных методов, моделей и алгоритмов в виде прототипа интеллектуальной рекомендательной системы поддержки формирования требований вакансии.

Перечисленные результаты используются в лекционных курсах и на практических занятиях по следующим дисциплинам:

- «Машинное обучение и интеллектуальный анализ данных»;
- «Разработка систем искусственного интеллекта на Python»;
- «Глубокие нейронные сети»;
- «Анализ естественного языка методами искусственного интеллекта»;

Директор ИИТ Ю.В. Петриченко

УТВЕРЖДАЮ
Заместитель директора
АУ «Югорского научно-исследовательского
института информационных технологий»
А.Л. Царегородцев

« 13 » 06 2024 г.

АКТ

о внедрении (использовании) результатов

диссертационного исследования

Николаева Ивана Евгеньевича

Ф.И.О.

Настоящим актом подтверждается, что результаты диссертационного исследования Николаева Ивана Евгеньевича «Методы и алгоритмы интеллектуальной поддержки формирования требований вакансии на основе нейросетевых моделей языка и актуальных требований рынка труда» используются в процессе подбора персонала в Югорском научно-исследовательском институте информационных технологий.

Результаты исследования, изложенные в диссертации, обладают актуальностью, имеют научное и практическое значение. Система предназначена для автоматизированной генерации рекомендаций по требованиям к кандидатам на основе анализа вакансий и профессиональных стандартов с использованием моделей искусственного интеллекта.

В ходе опытной эксплуатации системы было обработано более 60 вакансий и сформировано около 150 рекомендаций для различных должностей. Согласно результатам тестирования, точность и полнота рекомендаций превышает 80%.

Результаты пилотного внедрения подтвердили эффективность системы для оптимизации процессов подбора персонала и повышения качества требований к кандидатам на вакансии.

РЕШЕНИЕ:

1. Признать результаты опытной эксплуатации удовлетворительными.

2. Рекомендовать систему к промышленной эксплуатации.

УТВЕРЖДАЮ
Директор центра технологий
искусственного интеллекта
ООО Фирма «ИНТЕРСВЯЗЬ»
г. Челябинск
«12» апреля 2024 г.

АКТ

о внедрении результатов диссертационной работы
Николаева Ивана Евгеньевича
на тему
«Методы и алгоритмы интеллектуальной поддержки формирования требований вакансии на основе нейросетевых моделей языка и актуальных требований рынка труда»

Настоящий акт составлен о том, что в ООО Фирма «ИНТЕРСВЯЗЬ» г.Челябинск были внедрены результаты диссертационной работы на тему «Методы и алгоритмы интеллектуальной поддержки формирования требований вакансии на основе нейросетевых моделей языка и актуальных требований рынка труда».

Разработана методика формирования требований к вакансии на основе нейросетевых моделей языка и данных о текущих потребностях рынка труда. Созданы нейросетевые модели, обученные на корпусе вакансий. Разработаны: структурно-семантическая модель требований рынка труда; методы и алгоритмы формирования рекомендации в процессе составления списка требований к вакансии.

Внедрение результатов диссертационной работы позволило:

– сократить время на создание и редактирование требований к вакансии за счет автоматической генерации требований на основе нейросетевых моделей;
– повысить релевантность требований за счет учета актуальных тенденций рынка труда;

Результаты диссертационной работы внедрены в пилотном/тестовом режиме в ООО Фирма «ИНТЕРСВЯЗЬ» (г.Челябинск) и используются в операционной деятельности в процессе подбора персонала и оценки компетенций кандидатов в сотрудники, что подтверждает высокий научно-технический уровень и практическую значимость проведенного исследования.

Директора центра технологий
искусственного интеллекта
ООО Фирма «ИНТЕРСВЯЗЬ» _____/ Ю.В. Дмитрин

ПРИЛОЖЕНИЕ В

Структура вакансии с сайта headhunter.ru в json-формате


```
{

"id": "7760476",
"premium": true,
"has_test": true,
"response_url": null,
"address": null,
"alternate_url": "https://hh.ru/vacancy/7760476",
"apply_alternate_url": "https://hh.ru/applicant/vacancy_response?vacancyId=7760476",
"department": {"id": "HH-1455-TECH",    "name": "HeadHunter::Технический департамент"},
"salary": {"to": null,    "from": 100000,"currency": "RUR","gross": true},
"name": "Специалист по автоматизации тестирования (Java, Selenium)",
"insider_interview": {
       "id": "12345",
       "url": "https://hh.ru/interview/12345?employerId=777"
},
"area": {
       "url": "https://api.hh.ru/areas/1",
       "id": "1",
       "name": "Москва"
},
"url": "https://api.hh.ru/vacancies/7760476",
"published_at": "2013-10-11T13:27:16+0400",
"relations": [],
"employer": {
       "url": "https://api.hh.ru/employers/1455",
       "alternate_url": "https://hh.ru/employer/1455",
       "logo_urls": {
              "90": "https://hh.ru/employer-logo/289027.png",
              "240": "https://hh.ru/employer-logo/289169.png",
              "original": "https://hh.ru/file/2352807.png"
       },
       "name": "HeadHunter",
       "id": "1455"
},
"response_letter_required": false,
"type": {
       "id": "open",
       "name": "Открытая"
},
"archived": "false",
"working_days": [
       {
              "id": "only_saturday_and_sunday",
              "name": "Работа только по сб и вс"
       }
],
"working_time_intervals": [
       {
              "id": "from_four_to_six_hours_in_a_day",
              "name": "Можно работать сменами по 4-6 часов в день"
       }
],
"working_time_modes": [
       {
              "id": "start_after_sixteen",
              "name": "Можно начинать работать после 16-00"
       }
],
"accept_temporary": false,
"experience": {
       "id": "noExperience",
       "name": "Нет опыта"
},
"employment": {
       "id": "full",
       "name": "Полная занятость"
},
"show_logo_in_search": true

}
```

ПРИЛОЖЕНИЕ Г

Пример вакансии с жесткой стандартной структурой

Тестировщик ПО / QA Engineer

Требуемый опыт работы: 1–3 года
Полная занятость, удаленная работа

Обязанности:

— Анализ требований и документации по проекту;

— Составление и актуализация тест-кейсов;

— Развитие инженерных практик и процессов в QA;

— Выполнение следующих видов тестирования: функциональное, приемочное, регрессионное, исследовательское, тестирование юзабилити;

— Проведение кроссбраузерного и кроссплатформенного тестирования;

— Анализ результатов тестирования;

Требования:

— Умение справляться с большим объемом информации;

— Опыт работы с Git/Gitlab;

— Практические навыки тестирования API;

— Опыт работы с одной или несколькими реляционными базами данных (MySQL, PostgreSQL и др.);

— Знание принципов функционального тестирования и опыт его проведения;

— Можно работать удаленно или из офиса

— Аналитическое мышление, системный подход к решению задач, умение работать в команде, склонность к самообразованию, усидчивость, внимательность;

Условия:

— Возможность принимать решения в реализации технических задач, влиять на проект и процессы разработки в компании;

— Обмен знаниями между смежными командами, помощь в решении сложных задач;

— Отсутствие бюрократизма, что позволяет работать над своими задачами, а не отвлекаться на лишние согласования;

— Имеем гос. аккредитацию как IT-компания

Пример вакансии с произвольной структурой

Тестировщик/QA engineer (Middle, Senior)

до 200 000 ₽ на руки

Требуемый опыт работы: 3–6 лет
Полная занятость, удаленная работа

Нам нужны **профессионалы в области тестирования** веб приложений: функциональное тестирование интерфейсов пользователя, тестирование сервисов бэк энда.

Мы ожидаем от тебя:

— Опыт функционального тестирования (от 2х лет);
— Навыки тестирования веб-приложений;
— Опыт составления сценариев и тест-кейсов;
— Опыт работы с СУБД, базовые знания SQL;
— Опыт тестирования REST API.

Будет дополнительным плюсом:

— Опыт автоматизированного тестирования;
— Опыт нагрузочного тестирования веб-приложений;
— Навыки составления технической документации;
— Опыт работы с Soap UI, POSTMAN, JMeter.

Что надо будет делать:

— Определять область тестирования и описывать ее в виде тест-кейсов;
— Заниматься функциональным тестированием веб приложений и компонентов;
— Совершенствовать навыки в области нагрузочного и автоматизированного тестирования;
— Участвовать в выпусках версий программных продуктов;
— Составлять отчеты по протестированному функционалу.

Мы предлагаем:

— Работу в аккредитованной IT компании;
— Оформление по ТК РФ с первого дня работы (договор вышлем по почте за наш счет, если вы не в Перми);
— Годовой бонус по итогам проекта;