Р. М. Ташбулатов, А. Д. Тян, И. Г. Захарова
Тюменский государственный университет, г. Тюмень
УДК 004.89

МЕТОДЫ ФОРМИРОВАНИЯ РЕКОМЕНДАЦИЙ ДЛЯ ОПРЕДЕЛЕНИЯ ТЕМАТИКИ УЧЕБНЫХ ПРОЕКТОВ НА ОСНОВЕ ОНТОЛОГИИ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Аннотация. В статье представлен подход для выработки рекомендаций, где профиль пользователя и рекомендации формализуются с помощью понятий онтологии. Для формирования онтологии решались задачи распознавания терминов и их связывания с сущностями внешней базы знаний на корпусе текстов выпускных квалификационных работ, проведены вычислительные эксперименты и анализ результатов. Описанный подход позволяет формировать рекомендации, учитывая семантическую близость терминов.

Ключевые слова: рекомендательные системы, извлечение информации, распознавание терминов, связывание именованных сущностей, базы знаний.

Введение. В настоящее время для технических специальностей вузов широко распространены учебные проекты, они имитируют будущую профессиональную деятельность и развивают у студентов как универсальные, так и узкопрофессиональные компетенции. Определение темы и научного руководителя является важным этапом любой проектной деятельности, поскольку от этого зависит качество и проработанность самого проекта, а также эффективность роста студенческих компетенций [1].

Как правило, руководители формируют для студентов набор предложенных тем работ. В этом случае студенты сталкиваются с проблемой выбора конкретной темы из-за того, что ее формулировка не отражает суть предстоящей работы в полной мере. Альтернативой является выдвижение инициативной темы, но и в таком случае студенту необходимо найти научного руководителя, имеющего опыт в данной предметной области. Именно такой руководитель способен помочь понять проблему, дать ей научное толкование и направить работу в нужную сторону. В качестве рекомендаций

в работе [1] предлагается принимать во внимание интересы и компетенции студентов при выборе темы и направленности ВКР, а также отмечается важность соответствия тем научным направлениям выпускающей кафедры.

Выпускные квалификационные работы студентов и преподавателей, а также их научно-исследовательские статьи являются одними из наиболее ценных источников данных о их компетенциях и интересах. Учебные заведения ежегодно накапливают данные, которые могут быть полезны для организации проектной деятельности, но не осуществляют их анализ. *Проблема настоящего исследования* заключается в создании методов обработки имеющихся данных для информационной поддержки студентов при выборе темы и руководителя, учитывающей их научные интересы и компетенции.

В качестве исходных данных использованы отчеты выпускных квалификационных работ, которые являются неструктурированными данными, содержащими информацию о способе решения. Предметная область решения может быть описана с помощью набора методов и технологий, а также может характеризовать научные интересы и опыт руководителя.

Цель работы заключается в реализации подхода формирования рекомендаций для выбора научных руководителей и предлагаемых тем согласно запросу студента в виде набора методов и технологий — дескрипторов, множество которых необходимо сформировать на основе исходных данных.

Рекомендательные системы. Рекомендация научных руководителей и предложенных тем, согласно интересам студента, соответствует фильтрации по содержанию. В работе [2] представлена система для рекомендации студентам предложенных тем и выполненных работ на основе ключевых слов. Основным недостатком подхода с использованием ключевых слов является то, что во внимание не принимается их семантическая близость. В статье [3] описана история развития подходов к построению рекомендательных систем. Одной из модификаций, которая способна улучшить качество рекомендаций является использование понятий для описания профиля пользователя, которые в свою очередь устроены иерархически.

Так, в работе [4] представлена система для рекомендации научных статей на основе терминов, используя для семантической близости онтологию предметных областей для формулировки эффективного поискового запроса.

Распознавание терминов. К решению задачи выделения терминологии существует несколько различных подходов. Одним из предложенных подходов является составление правил, основанных на грамматических, морфологических и статистических свойствах. В работе [5] используется лексико-синтаксический язык паттернов, с помощью которых авторы сформировали коллекцию из правил для извлечения терминов из научных статей на русском языке.

Другим подходом является представление терминов как именованных сущностей, задачу распознавания которых можно свести к решению задачи маркировки последовательностей. В последние годы для задачи распознавания именованных сущностей активно используются глубокие нейронные сети. Распространенными архитектурами для выделения сущностей являются рекуррентные сети с долгой краткосрочной памятью (LSTM) и трансформеры (BERT), где последние имеют чуть лучшие результаты [6] за счет механизма «внимания», позволяющего учитывать контекстуальные отношения между токенами в тексте. BERT представляет из себя предобученную языковую модель, которую можно адаптировать под различные задачи обработки текстов с помощью трансферного обучения.

С помощью использования предобученной модели BERT в работе [7] была дообучена модель для распознавания терминов в научных статьях на русском языке. В данной работе рассмотрены различные подходы к выделению терминов, произведен сравнительный анализ модели BERT и ее модификаций для решения данной задачи.

Связывание сущностей. Традиционно задача связывания сущностей делится на 2 или 3 этапа, в зависимости от необходимости выноса в отдельный этап процесса распознавания сущностей. Основными являются этапы генерации и ранжирования кандидатов.

На этапе генерации кандидатов создается список сущностей базы знаний, потенциально релевантных заданной. Как правило, он генерируется по построчному совпадению с основными и альтернативными названиями сущностей базы знаний. Также используются

различные методы для расширения списка кандидатов, например, с применением n-грамм при поиске и т. д. [8].

Этап ранжирования предполагает определение релевантности каждого кандидата. Оценка того, насколько кандидат является подходящим, определяются с помощью различных способов, таких как бинарная классификация, оценка на основе графов знаний и т. д. В статье [8] описан алгоритм определения релевантности, основанный на информационной насыщенности кандидатов. В работе [9] в свою очередь предлагается использование системы тегов базы знаний, а также контекстов, представленных в виде сущностей, связанных с кандидатами. В статье [10] проведены эксперименты с использованием классических методов машинного обучения и нейронных сетей для определения релевантности кандидатов.

Материалы и методы. В качестве набора данных для формирования онтологической модели и связывания объектов этой модели с научными руководителями использованы 190 выпускных квалификационных работ направления МОиАИС за 2016-2022 гг., представленные в формате PDF. Документ с ВКР содержит изображение отсканированного титульного листа с информацией о теме работы, ее авторах и руководителе. Следующие страницы отчета квалификационной работы являются стандартными для формата PDF с сохранением макета исходного документа.

Информация с титульной страницы использовалась для сопоставления выделенных дескрипторов из текста отчета с конкретным руководителем, а также название темы и год работы для дополнительной информации о его выполненных работах. Для извлечения текста с отсканированного титульного листа использовалась библиотека распознавания символов — Tesseract OCR, с последующим выделением необходимой информации с помощью регулярных выражений, так как структура титульной страницы является стандартизированной. Для извлечения текста из PDF-документов использовалась библиотека pdfminer.six.

Для решения задачи распознавания терминов выбран подход с использованием языковой модели BERT, который требует дообучения для решения конкретной задачи. В качестве обучающего набора

данных был взят датасет RuSERRCv2 [7], который представляет из себя 136 аннотированных научных текстов на русском языке. В каждом тексте размечены термины в BIO-нотации, а также ссылки на сущности в базе знаний Wikidata, что будет использовано при оценке решения задачи связывания сущностей.

В качестве предобученной BERT-модели были использованы две предобученные языковые модели: RuBERT (deeppavlov/rubert-base-cased), BERT Multilingual (bert-base-multilingual-cased) и RuSciBERT (ai-forever/ruSciBERT). С помощью библиотеки transformers от платформы HuggingFace были загружены данные модели, вместе с которыми предоставляются предобученные токенизаторы текста, с возможностью делить текст на подслова.

Для связывания терминов-кандидатов с понятиями реального мира и извлечения отношений между терминами информационных технологий была использована база знаний Wikidata, которая покрывает множество предметных областей. Экземпляры данного ресурса содержат информацию о вариантах названия понятия, его описание, атрибуты и отношения с другими понятиями в машиночитаемом формате.

Результаты. Была построена модель для рекомендации научных руководителей и тем согласно запросу студента. Запрос студента представляет из себя множество дескрипторов, имеющих иерархическую структуру. Математическую постановку задачи можно представить следующим образом:

Найти множество $R \subset$ T, при $\forall r \in R$, $\forall t \in T \backslash R$, $sim(S, f(r)) > sim(S, f(t))$, $sim(S, f(r)) \in [0, 1]$, где $R = \{R_c\}$, c = 1, … , n — множество рекомендованных руководителей/тем, $S = \{S_j\}$, $j = 1, \dots,$ q — множество дескрипторов запроса студента, T = $\{T_i\}$, $i = 1, \dots,$ p — множество научных руководителей или предложенных тем, $D = \{D_k\}$, k = 1, … , g — множество дескрипторов, $f: T \rightarrow D$ — принадлежность множества дескрипторов преподавателю или теме, $sim(S, f(T_i))$ — функция определения коэффициента схожести между множествами дескрипторов. Для формирования рекомендаций необходимо сформировать множество дескрипторов D в виде

онтологии, определить принадлежность множества дескрипторов преподавателю f и функцию коэффициента схожести $sim(S,\ f(T_i))$.

Для сопоставления дескрипторов из документа с руководителем было реализовано выделение текстовой информации с титульной страницы и по извлеченному из нее ФИО находится конкретный преподаватель в базе данных.

Был реализован алгоритм формирования онтологической модели, состоящий из решения задач распознавания и связывания сущностей.

Одним из этапов предложенного формирования онтологии информационных технологий на основе документов является выделение терминологии. Тексты документов были предобработаны: удалены формулы, изображения, таблицы и подписи к элементам документа.

Для решения задачи распознавания терминов были дообучены BERT-модели протестированы на тестовой выборке. При трансферном обучении были разморожены 4 последних слоя BERT-модели. Также поверх BERT-слоев был добавлен 1 полносвязный слой для классификации токенов. Обучение производилось на протяжении 10 эпох (полных проходов по обучающим данным), тестовая выборка составила 20% от всех данных. Наилучшими значениями гиперпараметров оказались для RuBERT и BERT Multilingual: скорость обучения — $5 * 10^{-5}$, размер пакета — 32; для ruSciBERT: скорость обучения — $3 * 10^{-5}$, размер пакета — 32.

Для оценивания качества использованы метрики precision, recall и F1-мера. Термины могут состоять из нескольких токенов, только точное совпадение считалось корректным. Результаты представлены в таблице 1.

Таблица 1

Сравнение качества распознавания терминов на RuSERRCv2

	F1-мера	*Precision*	*Recall*
BERT Multilingual	0.49	0.49	0.50
RuBERT	0.50	0.46	0.56
RuSciBERT	**0.57**	**0.56**	**0.59**

В качестве базовых результатов взяты результаты, описанные в статье [7]. В данной работе оценка качества рассматривалась на тестовой выборке первой версии данного датасета. Предобученные нами модели использовали его вторую версию, поэтому для сравнения результатов тестировались на полном датасете.

Таблица 2

Сравнение качества распознавания терминов на RuSERRCv1

	F1-мера	*Precision*	*Recall*
BERT + LSTM + CRF [7]	0.53	-	-
BERT Multilingual	0.42	0.45	0.39
RuBERT	0.47	0.46	0.49
RuSciBERT	**0.57**	**0.55**	**0.59**

Исходя из полученных результатов, можно сделать вывод о том, что среди дообученных моделей лучшие результаты показал RuSciBERT, добившись значение F1-меры 0.53 и 0.57 на RuSERRCv1 и RuSERRCv2 соответственно.

Исходя из результатов исследований [10] в рамках решения задачи связывания сущностей на этапе ранжирования кандидатов была выдвинута гипотеза о том, что чем больше схожесть контекста выделенного термина и контекста, то есть статьи, сущности базы знаний, тем более релевантной является эта сущность. При этом также учитывается схожесть названий этих сущностей. Таким образом, требуется найти кандидата $entity_{\text{i}}$, для которого $f(entity_{\text{i}}) \geq f(entity_{\text{j}})$ для любого $j = 1, \ \ldots, \ n$, где n — количество кандидатов, а функция f для каждого кандидата $entity$ вычисляется по формуле:

$$f(entity_i) = J(d, entity_i) \cdot cos(d, entity_i) \ \cdot weight,$$

где J — коэффициент сходства Жаккара для названий термина d и кандидата $entity_i$, cos — косинусное расстояние между векторизированными контекстами термина и кандидата, а $weight$ — отношение количества совпадающих токенов и общего количества токенов во входной сущности.

В качестве модели для векторизации текста была выбрана TF-IDF. Косинусное расстояние является каноничной оценкой схожести текстов и описывается во многих работах, решающих задачу связывания сущностей как на английском, так и на русском языке.

Для тестирования качества предлагаемой модели использовался вышеописанный датасет RuSERRCv2. Размер выборки составлял 938 сущностей, связанных с сущностями в базе знаний Wikidata. Точность модели определяется как отношение количества верно связанных терминов ко всем терминам, имеющим связь. Описанный подход с точностью 0.43 правильно определяет сущность в базе знаний и среди изученных моделей является вторым по точности (наилучшая точность — 0.54 [10]).

В результате решения задачи связывания сущностей для выпускной квалификационной работы на основе выделенных терминов были получены упоминаемые методы и технологии. Например, для работы с темой «Разработка приложения для кластеризации текстов выпускных квалификационных работ ИТ-направлений» был выделен следующий набор дескрипторов: Data Science, машинное обучение, иерархическая кластеризация, алгоритм кластеризации, метод главных компонент, t-SNE, k-средних, Scikit-learn, Pandas, Numpy, nltk, Matplotlib, QtDesigner, PDF, Python, Cython, C++.

Для каждого извлеченного дескриптора статьи были получена информация о его родительских дескрипторах — связанных с заданным отношениями P31 (instance of) и P279 (subclass of) в базе знаний Wikidata. При формировании онтологии добавлялись только те сущности, родитель которых содержится в определенном наборе сущностей информационных технологий.

На основе полученной онтологии были реализованы методы рекомендации научных руководителей и предлагаемых тем. Для оценки схожести дескрипторов была определена функция:

$$sim(DS_i, DT_j) = \frac{1}{1+d(DS_i,DT_j)},$$

где $d(DS_i, DT_j)$ — расстояние в графе (количество ребер) между дескриптором DS_i и DT_j, $sim(DS_i, DT_j) \in [0, 1]$.

Для формирования рекомендаций была определена функция показателя близости множеств дескрипторов S и T:

$$sim(S,\ T) = \frac{\sum_i^m sim(S_i,T)}{|S|},$$

где $sim(DS_i, DT) = \max(sim(DS_i, DT_1), \dots, sim(DS_i, DT_n))$.

При формировании рекомендации для каждого научного руководителя вычислялось значение показателя близости запросу студента. Чем больше значение схожести, чем более релевантным является научный руководитель для пользовательского запроса.

Обсуждение. На этапе формирования онтологической модели были решены задачи распознавания терминов и их связывания с сущностями базы знаний. Результаты, полученные в рамках решения задачи распознавания терминов с использованием нейросетевых моделей, показали, что они обладают достаточной точностью для использования на практике. Получилось улучшить результаты, полученные в работе [7] за счет выбора модели, обученной на научных статьях, а не на текстах с общей тематикой.

В рамках решения задачи связывания сущностей хочется обратить внимание на положительное влияние контекстной информации на точность решения. Полученная точность (0.43) может быть улучшена путем использования иных методов векторизации текста, таких как FastText, BERT.

При построении рекомендаций для описания профиля преподавателей были использованы понятия из онтологии предметной области. Однако, при этом не учитывались статистические данные, например количество работ у преподавателя с данным понятием, что может улучшить качество рекомендаций. Также существуют альтернативные метрики для оценки близости понятий, что необходимо протестировать в будущем.

Заключение. В данной работе представлен метод для формирования рекомендации для выбора научных руководителей и предлагаемых тем согласно запросу студента. Запрос состоит из набора понятий онтологической модели информационных технологий, благодаря чему рекомендации учитывают семантические связи.

В статье описан алгоритм формирования онтологической модели на основе исходных данных. Для построения были решены задачи распознавания терминов с использованием нейронных сетей и связывания сущностей, учитывая контекстную близость при ранжировании кандидатов.

Представленное решение позволяет извлечь информацию из отчетов выпускных работ и использовать ее для помощи в рекомендациях. В будущих работах планируется улучшить качество решаемых задач при формировании онтологии и расширить методы рекомендаций для учета дополнительных признаков.

СПИСОК ЛИТЕРАТУРЫ

1. Михелькевич В. Н. Выбор тематики, структуры и содержания выпускной квалификационной работы в условиях неопределенности / В. Н. Михелькевич, П. Г. Кравцов. — Текст : непосредственный // Человек в условиях неопределенности: сборник научных трудов в 2-х т. — Самара, 2018. — С. 147-150.
2. Fiarni C. Recommender System of Final Project Topic Using Rule-based and Machine Learning Techniques / C. Fiarni, H. Maharani, B. Lukito. — Direct text // 2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI). — P. 216-221.
3. Городецкий В. И. Онтологии и персонификация профиля пользователя в рекомендующих системах третьего поколения / В. И. Городецкий, О. Н. Тушканова — Текст : непосредственный // Онтология проектирования. — 2014. — № 3 (13) — С. 7-31.
4. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов / А. М. Елизаров, А.Б. Жижченко, Н. Г. Жильцов [и др.]. — Текст : непосредственный // Доклады РАН. — 2016. — Т. 467. № 4. — С. 392-395.
5. Bolshakova E. Terminological information extraction from Russian scientific texts: Methods and applications / E. Bolshakova, N. Efremova, K. Ivanov. — Direct text // Proceedings of 3rd Workshop on Computational linguistics and language science (CLLS 2018), EPiC Series in Language and Linguistics — vol. 4, EasyChair, 2019/ — Ph. 95–106.
6. Hierarchical Transformer Model for Scientific Named Entity Recognition / U. Zaratiana, P. Holat, N. Tomeh, T. Charnois. — URL: https://arxiv.org/pdf/2203.14710.pdf (date of the application 15.05.2023). Text : electronic.

7. Bruches E. Entity recognition and relation extraction from scientific and technical texts in Russian / E. Bruches, A. Pauls, T. Batura, V. Isachenko. — Direct text // 2020 Science and Artificial Intelligence conference (S.A.I.ence). — 2020. — Pp. 41-45.
8. Мезенцева, А. А. Автоматическое связывание терминов из научных текстов с сущностями базы знаний / А. А. Мезенцева, Е. П. Бручес, Т. В. Батура. — Текст : непосредственный // Вестник Новосибирского государственного университета. Серия: Информационные технологии. — Новосибирск, 2021. — Т. 19, № 2. — С. 65-75.
9. Cucerzan S. Large-Scale Named Entity Disambiguation Based on Wikipedia Data / S. Cucerzan. — Direct text // Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). — 2007. — Pp. 708-716.
10. Мезенцева А. А. Методы и подходы к автоматическому связыванию сущностей на русском языке / А. А. Мезенцева, Е. П. Бручес, Т. В. Батура. — Текст : непосредственный // Труды Института системного программирования РАН. — 2022. — Т. 34, № 4. — С. 187-200.