ОТЗЫВ официального оппонента
на (о) диссертацию(и) на соискание ученой степени
физико-математических наук Тихомирова Михаила Михайловича
на тему: «Методы автоматизированного пополнения графов знаний на основе векторных представлений»
по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Актуальность избранной темы

Работа направлена на исследование методов автоматизированного пополнения графов знаний новой терминологией и именованными сущностями. Графы знаний являются полезным инструментом в задачах обработки естественного языка, в частности, из-за их интерпретируемости, но построение таких ресурсов, а также их поддержка и адаптация на предметные области представляет проблему и требует существенного количества человеческого труда. Поэтому, исследование подобных методов несомненно необходимо.

Автор исследовал две подзадачи: пополнение существующей таксономии графа знаний новыми понятиями (задача предсказания гиперонимии) и извлечение именованных сущностей из текстов предметной области для последующего пополнения графа сущностями. Обе подзадачи представляют интерес, так как даже относительно исследованная задача извлечения именованных сущностей недостаточно проработана в рамках конкретных предметных областей и сложных типов именованных сущностей. Задача предсказания гиперонимии также представляет проблему, так как многие слова имеют несколько смыслов, и помимо всего прочего, зависят от исследуемой предметной области. По этим причинам, избранная автором тема является актуальной.

Содержание работы

Диссертационная работа состоит из введения, четырех глав, заключения и списка литературы.

Во введении автор обосновывает актуальность работы, ставит цель, задачи. Также определяется научная новизна, описываются публикации и апробация работы.

В первой главе диссертации представлен обзор векторных представлений слов, методов для предсказания гиперонимии, извлечения именованных сущностей.

Во второй главе диссертант описывает проведенные исследования для задачи предсказания гиперонимии с целью пополнения таксономии графа знаний новыми абстрактными понятиями. Предлагается два новых метода: метод основанный на шаблонах и векторных представлений слов и метод основанный на мета-векторных представлениях. Подходы были протестированы на общедоступных наборах данных и было показано, что предложенный подход на основе мета-векторных представлений, позволяет достичь на них наилучшего на данный момент качества. Помимо этого, подход на основе мета-векторных представлений был протестирован на предложенным автором наборе данных, содержащим документы из узкой предметной области информационной безопасности.

В третьей главе изложены исследования задачи извлечения именованных сущностей в узкой предметной области. Автор поднимает проблему недостатка данных для обучения, особенно касаемо сложных типов именованных сущностей, а также проблему адаптивности существующих контекстуализированных векторных моделей. В работе предложен автоматический метод порождения псевдоразметки, а также двухэтапный подход к обучению модели BERT с использованием автоматически порожденных данных.

В четвертой главе автор описывает реализованную систему для автоматизированного пополнения таксономии графа знаний.

В заключении приводятся основные результаты и выводы диссертационной работы.

Степень обоснованности научных положений, выводов и рекомендаций, сформулированных в диссертации

Автор, решая задачу предсказания гиперонимии (для пополнения таксономии новыми понятиями), проверяет предложенные гипотезы и их влияние на итоговый результат. Также, идея основного метода для предсказания гиперонимии на основе мета-векторных представлений слов следует из предыдущего разработанного автором метода, как итог анализа выявленных проблем. Подход к автоматическому пополнению тренировочных данных для задачи извлечения именованных сущностей в предметной области напрямую следует из озвученных автором проблем недостатка данных по некоторым типам именованных сущностей. Таким образом, предложенные автором научные положения обоснованы, как и итоговые выводы.

Достоверность и новизна

Автор оценивает качество разработанных методов на общедоступных наборах данных, используя подходящие для задач меры качества: F1 для задачи извлечения именованных сущностей и MAP вместе с MRR для задачи предсказания гиперонимии. В работе присутствуют сравнения с другими подходами, что повышает достоверность результатов. Сами эксперименты подробно описаны, а программный код системы выложен в открытый доступ.

Новизна подхода к пополнению таксономии заключается в предложенном гибридном алгоритме вместе с идеей использования мета-векторных представлений слов, вместо обычных векторных представлений. Используя мета-векторные представления слов, автор комбинировал векторные представления слов с графовыми векторными представлениями, а также векторные представления общей предметной области и конкретной предметной области.

Предложенный двухэтапный подход к обучению трансформера для задачи извлечения именованных сущностей и сам метод формирования дополнительных данных являются новыми и показали улучшение качества целевых метрик. Также был показан ожидаемый, но требующий проверки

результат о том, что использование дообученного на предметную область трансформера положительно сказывается на метрики качества. Сама же обученная модель также представляет пользу, так как обучение таких моделей требует существенных вычислительных ресурсов.

Замечания

1. Обобщение или унификация результатов для других предметных областей, помимо информационной безопасности, повысили бы практическую значимость результатов.
2. Слишком краткое описание разработанного программного комплекса и результатов его работы. Дополнительные примеры работы позволили лучше понять достоинства и возможности программного комплекса.
3. Использование не описанных и не определенных ранее терминов, таких как «внешние» и «внутренние» модели, «сильные» и «слабые» модели, «гиперонимы золотого стандарта». Их значение интуитивно понятно, но было бы лучше давать четкое определение применяемым определениям.
4. Некоторая обособленность глав, а также отсутствие выводов к первой и четвертой главам немного затрудняет восприятия работы в целом. Четвертая глава исследования заканчивается внезапно, есть ощущение недосказанности.

Вместе с тем, указанные замечания не умаляют значимости диссертационного исследования. Диссертация отвечает требованиям, установленным Московским государственным университетом имени М.В.Ломоносова к работам подобного рода. Содержание диссертации соответствует паспорту специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» (по физико-математическим наукам), а также критериям, определенным пп. 2.1-2.5 Положения о присуждении ученых степеней в Московском государственном университете имени М.В.Ломоносова, а также оформлена, согласно приложениям № 5, 6

Положения о диссертационном совете Московского государственного университета имени М.В.Ломоносова.

Таким образом, соискатель Тихомиров Михаил Михайлович заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Официальный оппонент:

доктор технических наук,
профессор кафедры «Вычислительные системы и технологии»
Института радиоэлектроники и информационных технологий
ФГБОУ ВО «Нижегородский государс ...ный технический университет им. Р.Е. Алексеева»

Суркова Анна Сергеевна

31.05.2022 г.

Контактные данные:

тел.: 7 (904) 7871575, e-mail: ansurkova@yandex.ru

Специальность, по которой официальным оппонентом защищена диссертация: 05.13.01 – «Системный анализ, управление и обработка информации (в науке и промышленности)»

Адрес места работы:

603095, РФ, г. Нижний Новгород, ул. Минина, д. 24,

ФГБОУ ВО «Нижегородский государственный технический университет им. Р.Е. Алексеева», кафедра «Вычислительные системы и технологии»
Тел.: 7 (831) 4368228; e-mail: vt@nntu.ru

Подпись сотрудника Сурковой А.С. удостоверяю:
Заведующий кафедрой
«Вычислительные системы и технологии» Д.В.Жевнерчук