Отзыв официального оппонента

на диссертацию 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

Автор решает проблему пополнения графов знаний новыми словами. Графы знаний предназначены для формализованного описания знаний о мире и предметной области. Они используются в приложениях обработки естественного языка и представляют собой семантические сети с миллионами сущностей. Часть сущностей является абстрактными понятиями, а часть именованными сущностями. Поскольку создание графов знаний и пополнение их новыми сущностями является трудоемкой задачей, требующей большого количества труда экспертов, считаю, что тема работы является актуальной. В своей диссертационной работе автор разделяет задачу пополнения графа знаний на две подзадачи: пополнение *абстрактными понятиями* и пополнение *именованными сущностями*. Для решения первой подзадачи рассматривается постановка, при которой необходимо предсказывать для слов вышестоящее понятие — гипероним. Для решения второй задачи рассматривается случай извлечения сущностей из заданной предметной области для последующего пополнения графа.

Содержание работы

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

Во введении описывается ее актуальность, новизна, практическая значимость, а также ставятся цели и задачи.

В первой главе автор приводит обзор основных методов формирования векторных представлений слов, задачу предсказания гиперонимии, задачу извлечения именованных сущностей.

Во второй главе обосновывается связь между задачей предсказания гиперонимии и задачей пополнения графов знаний. Описываются два новых подхода с использованием мета-векторных представлений слов. Приводятся результаты вычислительных экспериментов с анализом ошибок на английском и русском языках, а также на двух предметных областях: общая область и область информационной безопасности.

В третьей главе рассматривается проблема извлечения именованных сущностей из текстов узких предметных областей и в случае необычных типов именованных сущностей. Рассматривается проблема адаптации трансформера типа BERT на предметную область и производится оценка влияния адаптации предметной области на итоговое качество подхода. Также исследуется важная проблема недостатка данных для обучения и описывается автоматический подход к генерации псевдоразметки для решения этой проблемы. Приводятся результаты вычислительных экспериментов, подтверждающих эффективность предложенных методов для извлечения сущностей.

В четвертой главе описан реализованный автором комплекс программ для автоматизированного пополнения графов знаний новыми понятиями на основе подхода, описанного в главе 2.

В заключении приводятся основные результаты и выводы диссертационной работы.

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

В работе в полной мере обосновываются принимаемые решения, а параметры алгоритмов подбираются таким образом, чтобы максимизировать итоговое качество подходов. Во второй главе присутствует анализ результатов

предсказания гиперонимии для пополнения таксономии графа знаний новыми понятиями, на основании которого делается вывод в необходимости именно автоматизированного формата работы системы, а не автоматического. Помимо этого, автор произвел исследование эффективности подхода на двух языках (английском и русском) и в общей и конкретной предметной области, что позволяет обобщить выводы работы на большее количество случаев.

Достоверность и новизна

Для оценки качества разработанных алгоритмов использовались общепринятые меры качества для рассматриваемых задач. Автор производил сравнение своих подходов с подходами других авторов, что повышает достоверность полученных результатов. Основные публикации работы опубликованы в рецензируемых научных журналах и были представлены на международных и российских конференциях, проходили этапы рецензирования и оценки специалистами.

Разработанные автором методы для пополнения графа знаний новой терминологией включают в себя новые идеи, как например, использование мета-векторных представлений для комбинации векторных представлений разного рода. Также автор самостоятельно разработал модификацию существующего метода для построения мета-векторных представлений, что позволило поднять качество результата. Для проведения экспериментов в предметной области информационной безопасности автор сформировал новый набор данных, что также представляет научную ценность.

Предложенный подход к решению проблем извлечения сложных типов именованных сущностей из текстов предметной области, также является новым и позволяет решать задачу эффективнее. Был предложен оригинальный алгоритм построения дополнительных тренировочных данных за счет псевдоразметки, а также способ эффективного использования этой псевдоразметки.

Замечания

1. Из текста автореферата неясно, в каких приложениях обработки естественного языка применяются онтологии. В тексте диссертации не приведены возможные приложения результатов, полученных автором.
2. В главе 1 многие термины из области машинного обучения используются без определений и должной контекстуализации. Например, перечисляются различные отдельные функции потерь, свертки, слои, но не описывается общая схема подхода с оптимизацией параметров при помощи градиентного спуска или его вариаций.
3. Из раздела 2.2.1 неясно, каким образом выбирались веса для формулы 2.4 и входящих в нее формул. Применялась ли линейная регрессия? В разделе не приводится описание процесса подбора формы зависимости между целевой переменной score и имеющимися факторами.
4. В разделе 3.5 сказано, что в вычислительных экспериментах производится несколько запусков, после чего считается макровзвешенная F-мера, ее среднее и стандартное отклонение. Во-первых, в работе не обоснован выбор именно такого критерия качества вместо более устойчивого к скошенному распределению меток критерию Matthews correlation coefficient (MCC). Во-вторых, в работе не объясняется разница в запусках, не приведен статистический анализ результатов и не упоминается исследование статистической значимости полученных автором результатов.
5. В работе не хватает описания процесса подбора гиперпараметров методов, участвующих в экспериментах, и исследования абляций предложенных автором подходов.
6. В работе не хватает псевдокода для независимого воспроизведения предложенных автором методов.

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:
кандидат физико-математических наук,
руководитель группы исследований краудсорсинга
обособленного подразделения
ООО «Яндекс.Технологии»
в г. Санкт-Петербург
Усталов Дмитрий Алексеевич

2 июня 2022 г.

Контактные данные:
тел.: +7 (961) 574-51-51,
e-mail: dustalov@yandex-team.ru
Специальность, по которой официальным оппонентом защищена диссертация:
05.13.17 — «Теоретические основы информатики»

Адрес места работы:

195027, г. Санкт-Петербург, Пискаревский пр-т, д. 2, к. 2, лит. Щ, БЦ «Бенуа», обособленное подразделение ООО «Яндекс.Технологии» в г. Санкт-Петербург
тел.: (812) 633-36-00, доб. 78327,
e-mail: info@yandex-team.ru

ПРЕДСТАВИТЕЛЬ ПО
ДОВЕРЕННОСТИ
№ Б/Н ОТ 01.11.2021
ШАБАЛКОВА Е.О.

02.06.2022