ОТЗЫВ официального оппонента
о диссертации на соискание ученой степени кандидата 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

В работе исследовалась задача автоматизированного пополнения графов знаний новыми понятиями и именованными сущностями. Поддержка графов знаний в актуальном состоянии, а также их адаптация на новые предметные области является нетривиальной задачей. Во многих случаях подобные процедуры производят вручную или полуавтоматически. Развитие методов, которые могли бы упростить и ускорить процесс пополнения графов знаний, является важной и актуальной задачей. Сами же графы знаний имеют широкое применение в различных задачах обработки естественного языка. В последнее время наблюдается рост популярности методов, использующих графы знаний, благодаря возможностям редактирования содержимого и связей, а также интерпретируемости подходов, основанных на графах знаний.

В диссертационной работе рассматривались две проблемы. Первая проблема заключается в том, как пополнять таксономии графов знаний, которые формируют каркас подобных ресурсов. Исследованию данного вопроса посвящена глава 2. Вторая проблема заключается в том, как повысить качество извлечения именованных сущностей из текстов, как в случае сложных типов сущностей, так и в случае новых предметных областей.

Содержание работы

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

Во введении автор обосновывает актуальность своего диссертационного исследования, описывает поставленные цели, задачи, новизну, теоретическую и практическую значимость. Представлена также информация об апробации основных результатов, публикациях и достоверности проделанной работы.

В первой главе представлен обзор, который можно разделить на три основные части. В первой части обзора обширно описываются основные методы формирования векторных представлений слов, как классические, так и современные. Во второй части дается представление о задаче предсказания гиперонимии, которая является ключевой для пополнения таксономий графов знаний. Описываются существующие подходы и их ограничения. Третья часть обзора посвящена задаче извлечения именованных сущностей. Охвачены методы, начиная с 1999 года и заканчивая лучшими на момент написания диссертации.

Во второй главе описываются проведенные автором исследования задачи предсказания гиперонимии. В работе предложено два новых метода: комбинированный метод на основе шаблонов и векторных представлений слов и комбинированный метод на основе мета-векторных представлений. Методы были протестированы на общедоступных наборах данных и было произведено сравнение с некоторыми другими подходами. Важной составляющей исследования является то, что была проверена эффективность метода на основе мета-векторных представлений для двух языков и для двух предметных областей (общая предметная область и предметная область информационной безопасности). Набор данных для тестирования качества в области информационной безопасности был сформирован автором.

В третьей главе описываются исследования и эксперименты по задаче извлечения именованных сущностей в узкой предметной области.

Рассматривается случай, когда присутствуют не только три стандартных типа именованных сущностей (локации, организации и персоны), но и более сложные. Одна из трудностей в том, что они недостаточно представлены в наборе данных. Для решения проблемы недостатка данных предложен метод автоматической генерации псевдоразметки и способ учета ее при обучении итоговой модели. Также в работе исследовано влияние дообучения трансформера BERT на текстах предметной области на целевые метрики качества извлечения именованных сущностей.

В четвертой главе описывается система автоматизированного пополнения таксономии. Данная система использует разработанный в главе 2 метод на основе мета-векторных представлений.

В заключении приводятся основные результаты и выводы диссертационной работы.

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

Сформулированные автором научные положения обоснованы проведенными экспериментами и анализом результатов. Предложенный метод на основе мета-векторных представлений является следствием выявленной в предыдущих экспериментах особенности, что итоговое качество решения существенно зависит от использованных векторных представлений слов. Выводы и рекомендации об использовании автоматизированной постановки при использовании разработанного подхода основаны на анализе ошибок, который был проведен в конце второй главы.

Рекомендации автора по использованию смешанной точности и подбора количества эпох в зависимости от размера батча при обучении трансформера BERT для задачи извлечения именованных сущностей обоснованы результатами проведенных экспериментов.

Достоверность и новизна

Достоверность полученных результатов подтверждается использованием общедоступных наборов данных, сравнением с существующими подходами, открытым исходным кодом реализованных методов и алгоритмов. Использованные метрики оценки качества также являются подходящими.

Новизна работы состоит в том, что было предложено два новых метода для задачи предсказания гиперонимии, была проведена серия вычислительных экспериментов на двух языках и двух предметных областях. Был предложен метод автоматической генерации псевдоразметки и способ ее учета при обучении модели для извлечения именованных сущностей, которые позволили улучшить качество целевых метрик.

Замечания

1. Для выявления отношений гиперонимии было бы естественно использовать контекстуализированные векторные представления слов, однако в работе используются в основном статические векторные представления.
2. В главе 3 рассматриваются методы извлечения именованных сущностей, однако в главе 4, при рассмотрении автоматизированного пополнения графов знаний новыми понятиями, именованные сущности и их векторные представления не используются.

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 физико-математических наук, профессор РАН,
профессор кафедры Математических методов прогнозирования факультета Вычислительной математики и кибернетики МГУ имени М. В. Ломоносова
Воронцов Константин Вячеславович.

3 июня 2022 г.

Контактные данные:
тел.: +7(916)333-71-69; e-mail: voron@forecsys.ru.
Специальность, по которой официальным оппонентом защищена диссертация: 05.13.17 – теоретические основы информатики

Адрес места работы: 119991, Российская Федерация, Москва, ГСП-1, Ленинские горы, д. 1, стр. 52, факультет ВМК.
Тел.: +7 (495) 939-30-10; e-mail: cmc@cs.msu.su.

Подпись сотрудника Воронцова Константина Вячеславовича удостоверяю.

Специалист по кадрам