Отзыв

официального оппонента кандидата физико-математических наук Турдакова Дениса Юрьевича на диссертацию Усталова Дмитрия Алексеевича на тему «Модели, методы и алгоритмы построения семантической сети слов для задач обработки естественного языка», представленную на соискание учёной степени кандидата физико-математических наук по специальности 05.13.17 — «Теоретические основы информатики»

Актуальность темы диссертации. Сегодня наблюдается взрывной рост количества информации, создаваемой людьми и машинами на естественном языке. Постоянное увеличение интенсивности потока входящей текстовой информации делает все более важной задачу математического моделирования естественного языка, в частности — русского языка. Разрешение лексической многозначности является важнейшей проблемой, для решения которой необходимы формализованные знания о значениях слов и их связях с контекстом и предметной областью, в которых употребляется каждое многозначное слово. Такие сведения представляются в семантических сетях — специальных базах знаний, представляющих машиночитаемые сведения об окружающем мире в виде понятий и связей между ними. Связи между понятиями задают семантическую иерархию, которая позволяет решать различные задачи машинного понимания естественного языка и является критически важным элементом семантических сетей. Создание высококачественных баз знаний вручную является длительной и ресурсоемкой задачей, поэтому исследователи уделяют большое внимание вопросу автоматического построения семантических ресурсов, таких как семантические сети. Проблема доступности и качества машиночитаемых семантических ресурсов осложняется наличием ошибок или пропущенными данными в существующих словарях. Методы машинного обучения, особенно — методы обучения без учителя, позволяют обнаруживать скрытые закономерности в неструктурированных данных. Применение таких методов позволяет повысить полноту доступных семантических ресурсов.

Диссертационная работа Д. А. Усталова посвящена задаче развития методов автоматического построения семантических сетей за счет структурирования и расширения существующих слабоструктурированных словарей, не содержащих сведений о значениях слов. Д. А. Усталов успешно использовал математический аппарат для формализации процесса построения семантической сети на основе слабоструктурированных словарей. Автором предложены оригинальные модели, методы и алгоритмы, позволяющие автоматизировать процесс построения семантической сети путём разрешения многозначности и связывания лексических значений слов в специальный вид языкового ресурса — семантическую сеть слов, понятиями которой являются лексические значения слов, связанные иерархическим отношением. На основе предложенных моделей, методов и алгоритмов разработан комплекс программ построения семантической сети слов, позволяющий осуществлять построение такого языкового ресурса на основе словарей синонимов для построения понятий и словарей с иерархическими связями между словами для построения связей между понятиями. Приводятся результаты вычислительных экспериментов по исследованию разработанных автором моделей, методов и алгоритмов в сравнении с мировыми аналогами. Вычислительные эксперименты свидетельствуют об эффективности предложенных автором подходов.

Обоснованность и достоверность полученных научных результатов подтверждается экспериментами, проведенными в соответствии с общепринятыми стандартами. Стоит отметить предельно аккуратный подход Д. А. Усталова к дизайну экспериментов.

Научная новизна работы определяется разработанными автором оригинальными моделями, методами и алгоритмами обнаружения лексических значений слов и их связывания в виде семантической сети слов. По сравнению с ранее известными методами построения семантических сетей, предложенный подход обнаруживает лексические значения слов без использования учителя, осуществляет построение однозначных иерархических связей между значениями слов, а также расширение связей между словами при помощи

плотных векторных представлений слов с использованием обучения с учителем.

Теоретическая значимость работы состоит в том, что в ней дано формальное описание методов, алгоритмов и архитектурных решений, позволяющих производить автоматическое построение семантической сети слов на основе слабоструктурированных языковых ресурсов.

Практическая значимость работы заключается в том, что на базе разработанных моделей, методов и алгоритмов разработан комплекс программ автоматического построения семантической сети слов, позволяющий повысить полноту сведений о семантических связях. Разработанные методы, алгоритмы и программное обеспечение могут применяться для построения интеллектуальных поисковых систем, систем машинного понимания текста, систем общения, и других информационных систем, основанных на знаниях. Стоит отметить, что практический результат работы оформлен в виде свободного программного обеспечения.

В качестве замечаний, не снижающих общего высокого уровня работы, можно отметить следующее:

1. В формальных определениях окрестности вершины (формулы 2.2 и 2.3) и графа значений слов (формула 2.7) не указано, что соответствующие графы является взвешенным, хотя веса далее используются.
2. Автор указывает, что "если лексическое значение слова в иерархическом контексте определить не удалось, то данное слово исключается из иерархического контекста" (стр. 77). Однако в работе не приводится исследования, показывающего, как часто возникает такая ситуация и как это влияет на общий результат.
3. Предложенный метод использует в качестве основы словари синонимов. Это накладывает ограничение на применимость метода, например, к неформальным текстам, написанным пользователями социальных сетей, которые часто используют сленг и несловарные слова. Кроме того,

не приводится экспериментальное исследование зависимости результатов работы методов от исходных словарей.

4. Значения некоторых параметров приводится без объяснения. Так, в работе используются 500-мерные векторные представления слов (стр. 87), из тестового набора данных YARN исключаются синсеты, которые редактировались менее восьми раз (стр. 88), шаблоны должны появиться не менее тридцати раз (стр. 98).

Заключение. Считаю, что диссертация Усталова Дмитрия Алексеевича представляет собой самостоятельную и законченную научно-квалификационную работу, в которой решена актуальная научно-практическая задача построения семантических сетей на основе неструктурированных данных, имеющая существенное значение в области обработки естественного языка. Диссертационная работа Д. А. Усталова в полной мере отвечает требованиям Положения о порядке присуждения учёных степеней, а её автор заслуживает присуждения учёной степени кандидата физико-математических наук по специальности 05.13.17 — «Теоретические основы информатики».

Официальный оппонент
заведующий отделом информационных систем,
ФГБУН Институт системного программирования
им. В.П. Иванникова Российской академии наук,
к.ф.-м.н Турдаков Денис Юрьевич

31 января 2018 г.

E-mail: turdakov@ispras.ru
Тел.: (495) 912-56-59 (доб. 461)

Адрес организации:
109004, Российская Федерация, г. Москва, ул. А. Солженицына, д. 25,
Федеральное государственное бюджетное учреждение науки Институт системного программирования им. В.П. Иванникова Российской академии наук (ФГБУН ИСП РАН)

Подпись Д. Ю. Турдакова заверяю:
Директор ИСП РАН, д.ф.-м.н., член-корр. РАН А.И. Аветисян