ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертационную работу Алимовой Ильсеяр Салимовны «Нейросетевой механизм кросс-внимания в задачах извлечения информации из текстов на примере биомедицинских данных», представленную к защите на соискание ученой степени кандидата технических наук по специальности 05.13.11 — «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Актуальность темы

Диссертационная работа Алимовой И.С. посвящена задаче извлечения информации из текста. В настоящее время непрерывно растущие объемы неструктурированной информации, представленной в виде текстов на естественном языке, требуют разработки автоматических методов интеллектуального анализа данных для извлечения структурированной информации. Тексты биомедицинской тематики, рассмотренные в данной работе, представляют особый интерес, поскольку являются важным источником информации при решении задач в области медицины и фармакологии. Вместе с тем разнородность источников и отсутствие единого формата для подобного рода текстовых данных создает множество проблем при разработке автоматических методов обработки.

В работе рассматриваются задачи классификации сущностей и извлечения отношений на примере текстов биомедицинской тематики. Рассмотренные задачи являются базовыми в области извлечения информации из текста и обработки текстов на естественном языке, а также могут быть использованы как компоненты при решении более комплексных задач. В диссертации предлагаются методы на основе нейронных сетей с механизмом кросс-внимания, эффективность которых подтверждается экспериментами на русскоязычных и англоязычных текстах биомедицинской тематики различных источников.

Тема диссертации является актуальной, а модели и алгоритмы, предложенные в работе, являются ключевыми для решения вышеуказанных задач.

Общая характеристика диссертационной работы

Диссертация состоит из введения, четырех глав, заключения, списка литературы, содержащего 182 наименования, и одного приложения. Общий объем работы составляет 146 страниц.

Во **введении** обосновывается актуальность исследований, формулируется цель, ставятся задачи работы, изложены научная новизна и практическая значимость представляемой работы.

В **первой главе** представлено описание существующих методов и подходов для задач классификации текста и извлечения отношений. В дополнение приведен обзор существующих работ, решающих данные задачи в области биомедицины. Проанализированы основные достоинства и недостатки существующих методов. На основании проведенного анализа делается вывод о том, что существующие методы основаны на моделях машинного обучения, где требуется сформировать вектор признаков, или с использованием нейронных сетей. Среди недостатков существующих моделей были выделены следующие: методы разработаны и протестированы для текстов определенной предметной области, модели тестируются в рамках одного корпуса, методы разработаны для английского языка, при этом для русского языка исследования в данной области практически не проводились. Стоит отметить, что Алимова И.С. выполнила большой объем работ при подготовке обзора литературы, собрав обширный список публикаций, посвященных как общим подходам, применяемым в задачах классификации текста и извлечения отношений, так и подходам, применяемым для текстов биомедицинской тематики.

Вторая глава посвящена задаче классификации сущностей. Предложен метод классификации сущностей, основанный на нейронной сети с

механизмом кросс-внимания и набором информативных признаков. Для разработки метода была проведена оценка информативности признаков, выделены наиболее значимые и были интегрированы в нейронную сеть с механизмом кросс-внимания. Хотелось бы отметить обширный объем экспериментов, проведенных для оценки разработанного метода. Эксперименты проводились на пяти существующих англоязычных корпусах, состоящих из текстов биомедицинской тематики, собранных с различных ресурсов, и одного русскоязычного корпуса, состоящего из отзывов пользователей о лекарственных препаратах. В качестве базовых моделей были рассмотрены семь существующих методов, среди которых присутствует современная языковая модель BERT, показывающая высокие результаты в различных задачах обработки текстов на естественном языке. Представленные в данной главе эксперименты показывают улучшение качества классификации сущностей согласно критерию F-меры по сравнению с базовыми рассмотренными моделями в рамках одного корпуса и при обучении и оценке моделей на разных корпусах. Дополнительно была проведена оценка разработанной модели по сравнению с существующими моделями нейронных сетей с механизмом кросс-внимания.

Третья глава посвящена задаче извлечения отношений. Предложен метод извлечения отношений, основанный на нейронной сети с механизмом кросс-внимания, принимающий на вход раздельно сущности и контекст между сущностями. Предложенный метод оценивался на четырех англоязычных текстовых корпусах биомедицинской тематики и одном русскоязычном корпусе, состоящем из текстов электронных карточек пациентов. Для оценки модели были также проведены обширные эксперименты, которые показали преимущество разработанной модели по критерию F-меры по сравнению с существующими методами на основе нейронных сетей. В дополнении были также предложены новые признаки, извлеченные автоматически из текста, проведена оценка информативности признаков по сравнению со стандартными признаками, применяемыми в задаче извлечения отношений.

В **четвертой главе** описана архитектура разработанных программных комплексов CAFEC для задачи классификации сущностей и CARE для задачи извлечения отношений. В работе описаны подробные детали реализации методов, приведены оценки времени и памяти, затрачиваемых на обучение моделей, лежащих в основе предложенных методов.

В **заключении** описаны основные результаты диссертационной работы и рассматриваются направления дальнейших исследований. Сформулированные в диссертационной работе научные положения, выводы и рекомендации являются вполне обоснованными.

Научная новизна и практическая значимость.

Научная новизна диссертационной работы И. С. Алимовой заключается в следующем:

1. Предложен и реализован новый метод классификации сущностей, основанный на нейронной сети с механизмом кросс-внимания и набором информативных признаков.
2. Предложен и реализован новый метод извлечения отношений между сущностями, основанный на нейронной сети с механизмом кросс-внимания, разделяющий контекст и сущности на отдельные подсети.

Практическая значимость диссертации И. С. Алимовой заключается в том, что на основе предложенных моделей, разработаны программные комплексы CAFEC для классификации текстов и CARE для извлечения отношений, проведено экспериментальное исследование, обосновывающее улучшение качества предложенных методов по сравнению с существующими алгоритмами в рамках корпусов из одной предметной области и корпусов из разных предметных областей.

Замечания

В целом, диссертационная работа выполнена на высоком уровне, однако имеется ряд недостатков:

1. Во второй и третьей главе при описании результатов экспериментов не представлено исследование статистической значимости в различиях в F-меры с другими методами.
2. Во второй и третьей главе при использовании F-меры для оценки качества следует указывать распределение классов в эталонной выборке.
3. Во второй и третьей главе в экспериментах с использованием английского языка использована специализированная модель BioBERT, а в экспериментах с использованием русского языка использована модель общего назначения RuBERT. Была ли возможна донастройка модели RuBERT на корпусе текстов нужной предметной области?
4. В четвертой главе не описана архитектура программных комплексов CAFEC и CARE.
5. В четвертой главе не исследована теоретическая вычислительная сложность процедуры обучения модели LSTM+CA+feat в программном комплексе CAFEC и модели LSTM+CA в программном комплексе CARE.
6. В работе не приведено обоснование использования вручную составленных тональных словарей вместо модели на основе машинного обучения.
7. В работе вместо термина «предметная область» используется менее распространенный термин «домен».

Указанные замечания, в целом, не снижают качество проведенного диссертационного исследования.

Заключение

Личное участие диссертанта в выполнении теоретических и экспериментальных исследований, разработке программных средств на основе предложенных методов и получении научных результатов подтверждается соответствующими публикациями. Результаты диссертации представлены в 12 статьях и двух тезисах, а также были представлены на российских и международных научных конференциях.

Принимая во внимание актуальность темы диссертации, научную новизну и практическую значимость ее результатов, считаю, что диссертационная работа Алимовой И. С. «Нейросетевой механизм кросс-внимания в задачах извлечения информации из текстов на примере биомедицинских данных» полностью соответствует требованиям ВАК РФ, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук, а Алимова Ильсеяр Салимовна заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.11 — «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Кандидат физико-математических наук,
аналитик-разработчик программного обеспечения,
группа исследований краудсорсинга
Обособленного подразделения
ООО «Яндекс.Технологии»
в г. Санкт-Петербург

Усталов Дмитрий Алексеевич