ОТЗЫВ НАУЧНОГО РУКОВОДИТЕЛЯ

о диссертационной работе Алексеева Алексея Александровича

«МЕТОД АВТОМАТИЧЕСКОГО АННОТИРОВАНИЯ НОВОСТНЫХ КЛАСТЕРОВ НА ОСНОВЕ ТЕМАТИЧЕСКОГО АНАЛИЗА»,

представленной на соискание ученой степени

кандидата физико-математических наук

по специальности 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

В настоящее время одним из активно развивающихся направлений компьютерной обработки текстов является автоматическое аннотирование и, в частности, аннотирование новостных кластеров – выделение наиболее значимой информации из набора новостных документов, посвященных одной теме. Результаты такой обработки востребованы в самых разных контекстах: от задач рядовых пользователей по поиску информации о конкретном событии до специализированных аналитических задач по мониторингу СМИ.

Большинство существующих методов автоматического аннотирования основаны на пословной модели представления текстов, то есть представлении текстов как массивов отдельных слов. В то же время известно, что в текстах на естественном языке широко используется лексико-семантическая вариативность – одни и те же сущности могут выражаться различными языковыми выражениями, а не только отдельными словами. Так, например, одна и та же страна Великобритания может упоминаться в рамках некоторого новостного кластера как *Англия, Британия, Туманный Альбион, страна, Соединенное королевство* и так далее. Подобная вариативность особенно критична для задачи автоматического аннотирования, так как информация о различных упоминаниях сущностей существенна для установления их объективной значимости.

В своей работе А.А. Алексеев предлагает новый метод автоматического построения ***тематических цепочек*** – структур, соответствующих отдельным

сущностям рассматриваемого новостного кластера и содержащих в себе все варианты упоминаний данных сущностей внутри него. Метод основывается на комбинировании набора характеристик схожести языковых выражений различной природы. Математически задача ставится как задача агломеративной кластеризации с ограничениями. Используются как контекстно-независимые характеристики схожести, такие как формальное сходство языковых выражений и информация из предопределенных ресурсов с описанием семантически-связанных выражений, так и контекстно-зависимые характеристики, анализирующие сходство контекстов употребления языковых выражений внутри рассматриваемого новостного кластера. Все разнородные характеристики схожести единообразно используются для выявления связанных слов и выражений в едином алгоритме.

Автоматически создаваемые тематические цепочки затем используются для порождения аннотаций новостных кластеров. Новым в работе А.А. Алексеева является то, что предлагается подход, позволяющий посредством тематических цепочек интегрировать разные типы знаний о словах и выражениях в процедуру автоматического аннотирования. В основе алгоритма интеграции лежит переход от пространства отдельных слов к пространству тематических цепочек – размерность каждого языкового выражения преобразуется в размерности соответствующих ему тематических цепочек, с учетом веса языкового выражения внутри данных цепочек. Подобная интеграция тематических цепочек в методы автоматического аннотирования добавляет информацию о лексико-семантической вариативности, позволяя алгоритмам аннотирования автоматически сопоставлять различные языковые выражения соответствующим им реальным сущностям.

Кроме того в работе предлагаются два новых метода автоматического аннотирования на основе построенных тематических цепочек. Суть этих методов заключается в отборе предложений для аннотаций, включающих в себя наиболее значимые и неупомянутые тематические цепочки в первом алгоритме

и, соответственно, наиболее частотные связи тематических цепочек во втором алгоритме.

В рамках исследования разработан программный комплекс, модули которого реализуют:

- предложенный метод построения тематических цепочек;
- методы аннотирования с интеграцией и без интеграции построенных тематических цепочек (на примере классических методов автоматического аннотирования MMR и SumBasic);
- автоматическую оценку аннотаций официальным пакетом ROUGE (стандартный метод сравнения автоматических аннотаций, суть которого заключается в автоматическом сопоставлении с экспертными аннотациями, составленными лингвистами).

Получены результаты, подтверждающие улучшение качества классических методов аннотирования с интегрированными тематическими цепочками по различным мерам качества ROUGE. Для подтверждения результатов автоматической оценки пакетом ROUGE была проведена дополнительная ручная оценка методом «Пирамиды», который предлагает формализованный подход к оценке автоматических аннотаций путем выделения и сопоставления информационных единиц из экспертных аннотаций с автоматическими. Оценка методом «Пирамиды» подтвердила полученные результаты.

При выполнении диссертационной работы А.А. Алексеев проявил себя вполне сложившимся ученым, способным самостоятельно решать сложные задачи, демонстрирующим высокое профессиональное мастерство, завидную работоспособность. Он активный, целеустремленный и отзывчивый человек, готовый всегда поделиться своими знаниями с окружающими.

Диссертация А.А. Алексеева вносит значительный вклад в решение актуальной задачи автоматического аннотирования новостных кластеров и автоматической обработки новостных кластеров в целом. Ее отличает удачное

и органичное сочетание теоретических результатов и практической направленности.

Работа выполнена на высоком научном уровне. Публикации адекватно отражают содержание диссертации. Материал изложен связно и последовательно. Диссертация соответствует всем требованиям ВАК РФ, предъявляемым к кандидатским диссертациям, а её автор Алексей Александрович Алексеев заслуживает присуждения ему ученой степени кандидата физико-математических наук по специальности 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Заведующий кафедрой алгоритмических языков факультета ВМК МГУ, профессор
доктор физико-математических наук

М.Г. Мальковский

23.05.2014.

Российская Федерация Московский государственный университет вычислительной математики и кибернетики им. М.В. Ломоносова Факультет г. Москва ОТДЕЛ КАДРОВ

ПОДПИСЬ УДОСТОВЕРЯЮ НАЧАЛЬНИК ОТДЕЛА КАДРОВ С.Ф. РЕЖЕТОВ