Отзыв официального оппонента

на диссертационную работу Алексеева Алексея Александровича

«Метод автоматического аннотирования новостных кластеров на основе тематического анализа», представленную на соискание ученой степени кандидата физико-математических наук по специальности

05.13.11 — «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Диссертация А.А. Алексеева посвящена задаче автоматического аннотирования новостных кластеров, а именно выявлению дополнительной информации о внутреннем устройстве новостного кластера – вариативности именования основных участников новостного сюжета. Подобная информация действительно важна для качественного решения задачи автоматического аннотирования, так как лексико-семантическая вариативность является одной из основ естественного языка, вследствие чего построение полных и не избыточных автоматических аннотаций невозможно без её учета. В свою очередь задача автоматического аннотирования новостных кластеров востребована в самых различных приложениях по анализу новостного потока и мониторингу СМИ. Поэтому тема диссертации А.А. Алексеева весьма **актуальна**.

Диссертация состоит из введения, 4 глав, заключения, списка литературы, содержащего 82 наименования, и двух приложений.

Во введении обосновывается актуальность темы исследования, формулируется задача диссертации, описываются основные принципы

подхода, методика исследования, апробация работы, и приводится краткое описание основных результатов диссертации.

В первой главе рассматривается задача автоматического аннотирования, приводится классификация типов аннотаций и обзор соответствующих областей их применения. Кроме того, в данной главе описываются базовые идеи для формирования автоматических аннотаций, наиболее распространенные модели представления данных, а также методы сравнения и оценки качества автоматических аннотаций. Особое внимание уделяется вопросу учета лексико-семантической вариативности в существующих моделях представления данных, используемых методами автоматического аннотирования.

Во второй главе вводится формальная модель для выявляемого типа вариативности – тематическая цепочка. Тематическая цепочка представляет собой совокупность всех вариантов именования некоторого участника новостного сюжета, описываемого в рассматриваемом новостном кластере. Одной из ключевых особенностей данной модели является однообразный учет как отдельных слов, так и многословных выражений. Также во второй главе описывается алгоритм автоматического построения тематических цепочек, который объединяет набор из шести характеристик схожести различной природы в рамках единой процедуры. Каждая из характеристик схожести достаточно подробно рассмотрена в тексте диссертации, включая вопрос значимости для задачи выявления лексико-семантической вариативности.

В третьей главе описывается алгоритм интеграции модели тематических цепочек в методы автоматического аннотирования на основе пословной модели представления, которая лежит в основе подавляющего большинства существующих методов. Также предлагается два новых метода автоматического аннотирования на основе учета упоминаний тематических цепочек и связей между ними. В главе приведены

результаты оценки всех методов аннотирования по методике ROUGE, включающей как оценки аннотаций на основе пословной модели представления, так и модели, обогащенной тематическими цепочками. Кроме того, лучшие методы дополнительно оценены полуавтоматическим методом «Пирамиды». Получены результаты, демонстрирующие, что методы на основе обогащенной модели представления показывают лучшие результаты по сравнению с исходными методами, что подтверждает ценность разработанной модели и предложенного способа интеграции данной модели в существующие методы автоматического аннотирования.

В четвертой главе дается описание разработанного в рамках диссертационного исследования программного комплекса, реализующего предложенный алгоритм построения тематических цепочек, различные методы автоматического аннотирования, а также автоматическую оценку аннотаций по методике ROUGE. Разработанный комплекс может быть полезен как для решения прикладных задач, так и для исследователей в данной области.

Основными преимуществами предложенного подхода являются: объединение разнородных характеристик схожести языковых выражений в едином алгоритме и универсальный способ интеграции модели в существующие методы аннотирования.

По диссертационной работе имеются следующие замечания:

1) Положения, выносимые на защиту, сформулированы, не совсем корректно, в частности, в четвертом положении, отмечается улучшение качества аннотирования, что, скорее, является обоснованием качества работы разработанных методов, а не самостоятельным результатом;

2) В ряде случаев, используемые обозначения требуют дополнительных комментариев, в частности, в разделе 2.4.1 не

определяется содержание элементов a_{ij} и параметр M, в разделе 3.1.1 не определяется параметр N;

3) При рассмотрении сборки многословных выражений в разделе 2.4.3.1 не упоминается возможность формирования словосочетаний путем проверки лингвистической согласованности, а также не даются рекомендации по заданию параметров предложенного алгоритма;

4) При выполнении оценки качества разработанных алгоритмов автоматического аннотирования в разделе 3.3 используется только один массив из 11 новостных кластеров небольшого размера, не оценивается надежность результатов оценки качества, не выполняется сравнение с открытыми сервисами обработки новостей типа Google или Yandex.

Отмеченные замечания в целом не снижают качества проведенного диссертационного исследования. Хочется отметить, что система извлечения лексико-семантической вариативности, подобная предложенной в данной работе, должна стать необходимым компонентом систем содержательного анализа новостных потоков и других текстовых коллекций на естественном языке.

Диссертация Алексеева Алексея Александровича является законченной самостоятельной научно-исследовательской работой, совокупность результатов которой можно квалифицировать как существенное продвижение в решении актуальной научной проблемы разработки методов и программных средств обработки новостных потоков. Результаты работы опубликованы в 14 печатных работах (три работы опубликованы в изданиях из перечня ВАК), которые с достаточной полнотой отражают основное содержание диссертационной работы. Автореферат достаточно полно отражает содержание диссертации.

Принимая во внимание актуальность темы диссертации, научную новизну и практическую значимость ее результатов, считаю, что диссертационная работа «Метод автоматического аннотирования новостных кластеров на основе тематического анализа» удовлетворяет всем требованиям ВАК РФ, предъявляемым к кандидатским диссертациям, а ее автор, Алексеев Алексей Александрович, безусловно, заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 05.13.11 – "Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей".

Научный консультант
Департамента исследований и разработок
ООО «ЛАН-ПРОЕКТ»,
кандидат технических наук, доцент

_____ В.Г. Васильев

Адрес: ООО «ЛАН-ПРОЕКТ»,
Проспект Маршала Жукова, д. 74, корп. 1,
123103, г. Москва
Электронная почта: v.vasiliev@lan-project.ru
Моб.тел.: +7-916-141-52-22

01.09.2014