ОТЗЫВ

о выпускной квалификационной работе
Щербининой Арины Александровны
«Вычисление сходства русских текстов на основе синтаксических структур»

Актуальность темы выпускной квалификационной работы Арины Александровны Щербининой не вызывает вопросов. Автоматическое обнаружение парафразов востребовано и в информационном поиске, и в задачах извлечения фактов и мнений, и в задачах обнаружения плагиата; практически во всех областях, где требуется автоматическое понимание текстов, требуется и отождествление различных фраз, имеющих одно и то же значение; количество исследований в данной области растёт.

Результаты, полученные в ходе исследования Арины Александровны Щербининой, обладают обоснованной **достоверностью**: в качестве материала был взят фрагмент НКРЯ, объём которого составил 5 миллионов словоупотреблений, а обучающие и тестовые выборки составили в сумме 14 тысяч словоупотреблений.

Не вызывает сомнений и **теоретическая значимость** данной работы: выявлены семантико-синтаксические критерии для определения сходства между текстами, установлены весовые коэффициенты для значимых семантико-синтаксических фрагментов текста, влияющие на степень сходства между текстами. Выработан метод и алгоритм сравнения текстов; более того, представлена программная реализация данного, вне всяческих сомнений, оригинального алгоритма, что говорит о **практической значимости** и **научной новизне** данных результатов.

Текст работы состоит из Введения, трех глав, заключения, списка использованной литературы, перечня электронных источников и пяти приложений.

По представленным на защиту результатам возникают некоторые вопросы.

1. В работе активно используется понятие «чанка», на нём строится и предлагаемый алгоритм, однако определения для данного термина не приводится. В информатике данный термин обозначает произвольный фрагмент строки, файла или иного потока данных; в данной работе под ним явно подразумевается некий аналог синтагмы или иного синтаксически связного фрагмента высказывания, однако, что точно подразумевается под данным термином, из работы неясно; указано лишь то, что был некий «золотой список» чанков из корпуса изложений, на основании которого эти чанки выделялись по грамматическим признакам. Возникает вопрос: откуда взялся этот «золотой» список? Каким конкретно был алгоритм выделения этих «чанков»? Программная реализация чанкера, доступная по приведённой в приложении ссылке, а также модуль, содержащий линейные схемы «золотых» чанков, увы, не обладают документацией, которая была бы достаточной для ответа на данные вопросы.
2. Можно ли считать вообще корректным выбор школьных изложений как материала для составления параллельного корпуса парафразов? Ведь даже в приведённом фрагменте корпуса присутствуют фразы вроде *если честь шла о речи чести*, характеризующиеся явными нарушениями как в синтаксической, так и в семантической структуре!
3. Из текста работы следует, что изначальная оценка сходства фрагментов в параллельном корпусе выставлялась автором данной работы. Насколько правомерно считать достоверными результаты этой оценки и основанной на ней оценки качества работы предлагаемого метода на тестовом подкорпусе корпуса изложений как «хорошего» (параграф 3.5)? На чём вообще основано данное утверждение — ведь в

работе не приводится данных ни о точности, ни о полноте, ни о каких-либо иных показателях качества результатов работы алгоритма на данном подкорпусе?

4. При оценке качества работы алгоритма на новостном корпусе в качестве Baseline были взяты показатели, выдаваемые системой ROUGE на основе n-грамм, приведены значения F-меры, которые данная система присваивает одним новостям, как если бы они были автоматическими рефератами или переводами других, однако совершенно не ясно, на каком основании утверждается, что «Очевидно, что мера ROUGE-2 показала плохие результаты и в данном случае для оценки непригодна», и что «ROUGE-1 и ROUGE-L в целом соотносятся как с человеческими оценками, так и с показателями сходства, полученными при помощи» предлагаемого алгоритма. На основе каких конкретно статистических критериев сделаны данные утверждения, как и основной вывод о том, что «алгоритм, представленный в данной работе, выдает справедливые показатели сходства, наиболее близкие к человеческому восприятию текстов»? Требовали ли эти критерии нормальности распределения, производилась ли соответствующая проверка? По какой причине оценки мер ROUGE производились только на новостном корпусе и не производились на корпусе изложений?

5. Можно ли называть алгоритм объединения «чанков» в «топологические поля» алгоритмом синтаксического анализа? Доступная по ссылке программная реализация даёт лишь косвенные представления о том синтаксическом формализме, который стоит за данным алгоритмом; в тексте работы данный алгоритм не описан; программный код, безусловно, отвечающий большинству стандартов качества, снабжён довольно редкими комментариями и не имеет даже юнит-тестов, которые могли бы служить примерами, проясняющими его работу. Не вполне ясно, какие именно структуры выстраиваются на выходе: структуры непосредственных составляющих, структуры зависимостей, комбинированные структуры? Чем обусловлен в данном случае выбор именно такого — не вполне традиционного — формализма?

Сформулированные вопросы носят сугубо дискуссионный характер; в данном случае, как и во многих подобных, сущность работы, её результаты — это, безусловно, та программная реализация и то лингвистическое обеспечение, которое было создано в рамках исследования, а не текст, который не в полной мере отражает эти результаты и потому вызывает столько вопросов; эти вопросы ни в коей мере не снижают несомненно высокий уровень данной работы как в теоретическом, так и в практическом отношении; заявленная цель исследования достигнута, поставленные задачи — решены; исследование, бесспорно, является самостоятельным, оригинальным и соответствующим требованиям, предъявляемым к выпускным квалификационным работам, а его автор достоин самой высокой оценки.

К.ф.н., ст. преп. А.В. Добров