УДК 004.912

Кожевников Д.В.
Студент магистратуры
Факультет вычислительной математики и кибернетики
МГУ им. Ломоносова
Россия, г. Москва

ТРЕБОВАНИЯ К СОВРЕМЕННЫМ СИСТЕМАМ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ СТИЛИСТИЧЕСКИХ НЕСООТВЕТСТВИЙ В НАУЧНОМ ТЕКСТЕ

Аннотация: *В данной статье приводятся основные применяемые подходы к решению задачи контроля качества текста. Описываются основные направления, требующие решения в современной системе. Демонстрируется пример решающей их экспериментальной системы. Приводятся способы проверки ее корректности.*

Ключевые слова: *компьютерная лингвистика, автоматическая обработка текста, обнаружение ошибок, стилистика текста, научный стиль.*

Annotation: *This article lists most commonly used methods of solving the problem of text quality control. Describes main topics that should be addressed in a modern system. Provides an example of an experimental system, that solves these problems. Proposes methods of its testing.*

Key words: *computational linguistics, automated text processing, error detection, text stylistics, scientific style.*

Вопрос автоматической обработки стилистических несоответствий в текстах на естественном языке всегда был крайне актуальным для научного сообщества. В частности, вопросы автоматического обнаружения и коррекции

несоответствий текста научному функциональному стилю являются ключевыми при разработке программных систем, направленных на облегчение и частичную автоматизацию работы научных редакторов, а также повышение эффективности промежуточного контроля качества студенческих выпускных работ и диссертаций.

Действительно, несмотря на широкую доступность материалов, специфицирующих требования, накладываемые научным сообществом на текст, для человека вполне естественно не владеть этой информацией в полной мере. Для решения этой проблемы существуют программные системы, которые помогают человеку работать в рамках этих требований более эффективно. Тем не менее, каждая из этих систем по отдельности обычно затрагивает лишь небольшую часть требований, предъявляемых к научному тексту.

Рассмотрим, какие существуют подходы к решению данной задачи. Традиционно использовались подходы, основывающиеся на методах синтаксического и семантического анализа текстов ([1], [2], [3]). В настоящее время, все большую популярность набирают подходы, связанные с использованием машинного обучения (например, [4], [5]). Также немаловажной частью работы с научным текстом является анализ его структуры и оформления [6]. И, конечно, требуется формальная проверка корректности структуры текста [7], а также вторичных документов текста: аннотации [8], библиографических ссылок [9]. Разработчику современного программного комплекса, автоматизирующего проверку стилистических несоответствий, должны быть известны эти моменты для корректного проектирования программного продукта.

Существуют общие критерии качества, применимые к научным работам, написанным в любом жанре, на которые следует ориентироваться при разработке [10]. Но также следует учитывать, что каждый тип произведения,

написанного в научном стиле, будь то монография, статья, диссертация или реферат, обладает своими особенностями в отдельных деталях [11].

Проблемы реализации

Многие системы, предназначенные для повышения качества текста, специализируются на некоторых конкретных вопросах: оценке легкочитаемости, проверке оформления, глубокой интеграции с WYSIWYG редакторами. Поскольку написание качественного научного текста, особенно в рамках конкретного жанра — это достаточно узкая задача, не существует текстовых редакторов общего назначения, которые бы могли проверить большую часть требований, накладываемых на подобный текст.

Поэтому для работы с научным текстом создаются узкоспециализированные системы [1]. Но такие системы зачастую обладают наборами слабопересекающихся возможностей, в том время как для облегчения обработки научного текста требуются они все [6]. Разработчик современного программного обеспечения для работы в этой области должен учитывать опыт создания предыдущих специализированных систем подобного класса.

Значительный объем работы требует интеграция специализированного программного средства с текстовыми редакторами общего назначения, чаще всего Microsoft Word. Стремление к подобной интеграции вызвано, в первую очередь, требованиями различных научных издательств к входному формату печатаемых материалов. Тем не менее, отказ от поддержки подобных редакторов и использование вместо них систем компьютерной верстки, основанных на макрорасширениях (таких как TeX, LaTeX [12]), позволяет исключить из рассмотрения целый класс ошибок. Примерами функциональных возможностей, которые автоматически достигаются использованием таких систем, являются:

- соответствие разделов текста разделам оглавления;

- оформление нумерации таблиц, рисунков и приложений;
- соответствие списка литературы используемым в тексте ссылкам;
- корректность ссылок на рисунки, таблицы, формулы;
- а также другие возможности, относящиеся к автоматическому оформлению текста. [13, табл. 1, строки 6–11 и др.]

Что касается непосредственно возможностей, которые требуются человеку при работе со стилистическими несоответствиями в научном тексте, то, на основании существующих исследований, можно выделить следующие требования:

1. Обнаружение орфографических и пунктуационных ошибок, проверка грамотности речи. [10, 11]
2. Обнаружение ошибок в оформлении. [1, 2, 6, 13]
3. Обнаружение ошибок в структуре работы. [1, 2, 6, 13]
4. Проверка корректного употребление терминологии. [10, 14]
5. Обнаружение стилистических ошибок в употреблении слов и словосочетаний. [10]
6. Обнаружение ошибок в построении предложений. [10]
7. Обнаружение ошибок в логике изложения работы. [10, 11]
8. Обнаружение фактических несоответствий в работе. [3, 11]

Автоматическая обработка пунктов 7–8 на данный момент является практически неразрешимой задачей и в большинстве случаев выполняется человеком — научным редактором или специалистом подобного профиля. Однако, все остальные пункты являются важными для любой современной программной системы, предоставляющей пользователю функционал для повышения качества научного текста.

Для демонстрации приведенных пунктов была разработана система КОНТРОЛЬ, демонстрирующая основные обозначенные положения. Она

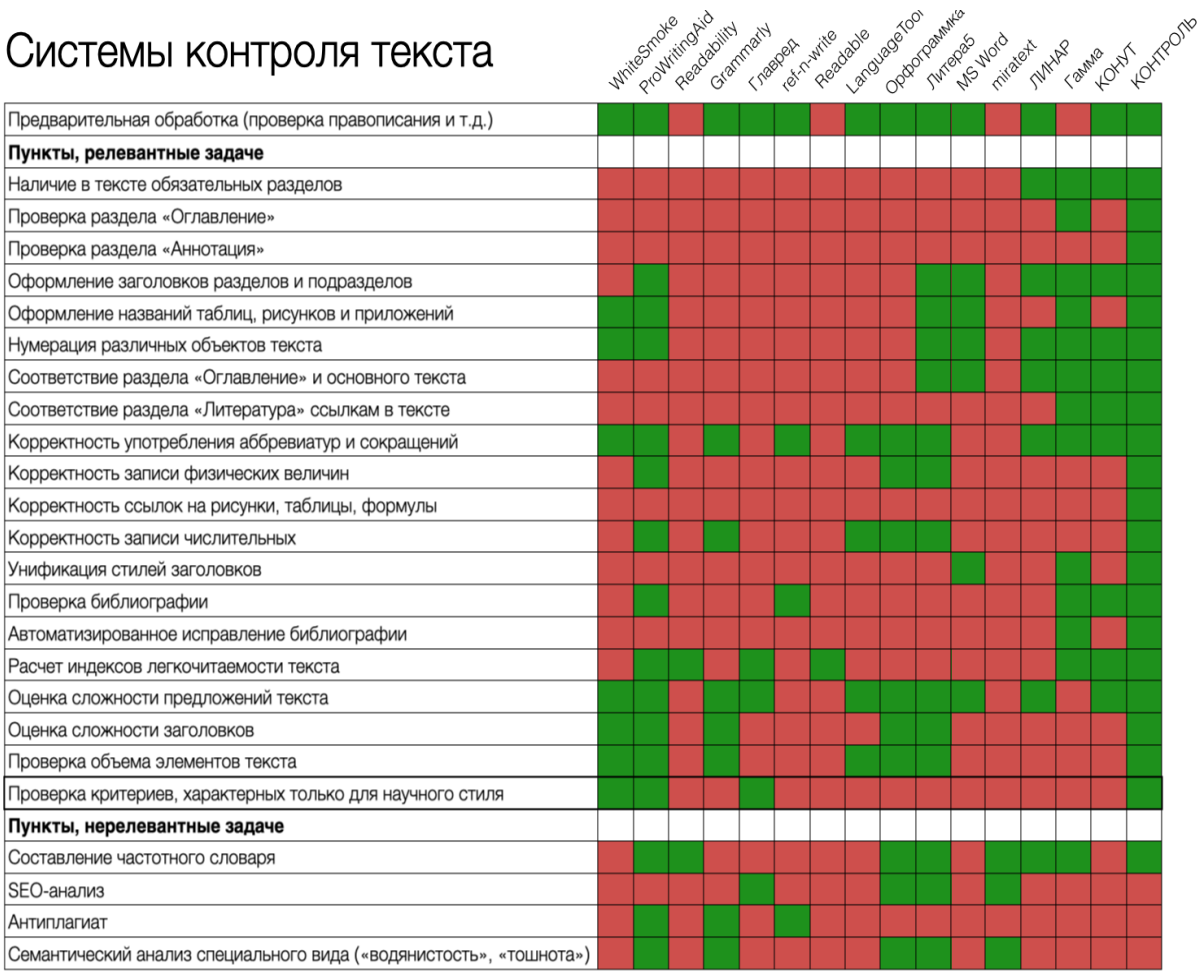

Рисунок 1. Сравнение систем контроля текста

построена на основании перечисленных выше принципов, а частичное сравнение ее функционала можно посмотреть на рисунке №1.

Оценка качества

После создания программного комплекса, выполняющего поставленную задачу, требуется исследование качества его работы. Естественно, о *точном*, безусловном соответствии текста на естественном языке некоторому функциональному стилю не приходится говорить. Во-первых, решение подобной задачи потребовало бы неограниченного количества ресурсов: в первую очередь времени работы экспертов, затрачиваемого на разметку и оценку используемых для анализа корпусов текстов [15]. Во-вторых, для ее решения потребовалось бы введение полной формализации естественного языка. [3]

Одним из способов тестирования узкоспециализированной системы является сравнение результатов ее работы с результатами другой системы. Однако, поскольку каждая такая система имеет достаточно узкую специализацию, часто возникает ситуация, когда в данный момент времени нет *ни одной* практически применимой системы, решающей ту же задачу.

Второй вариант, обозначенный выше — инженерный подход с привлечением профессионалов по работе с текстом к разметке некоторого корпуса для проверки. Такой подход отлично подходит для решения поставленной задачи, однако при этом имеет высокую стоимость и некоторую субъективность составления. Поэтому иногда достаточно ограничиться тестированием отдельных возможностей разработанной системы на наиболее характерных случаях несоответствий.

Заключение

Разработчик современных программных систем, направленных на помощь в исправлении стилистических несоответствий в научных текстах, с одной стороны, имеет доступ ко всему массиву практических знаний, накопленных его предшественниками. Но, с другой стороны, должен учитывать тот факт, что требования к узкоспециализированным системам данного класса значительно возросли.

При проектировании программных продуктов в этой области, следует помнить, что желательно максимально учитывать все предъявляемые к научному тексту требования, не допуская ошибок, свойственных предыдущим системам. Но отсюда следует и основная потенциальная проблема данного класса задач — невозможность полностью объективной оценки качества ее решения. Поэтому при решении таких задач всегда следует помнить о том, что каждая такая программа, в первую очередь, должна облегчать работу своего пользователя — студента, ученого, научного редактора.

Литература:

1. Мальковский М.Г., Большакова Е.И. Интеллектуальная система контроля качества текста // Интеллектуальные системы. — Т. 2, вып. 1–4. — Москва, 1997. — с. 149–155.
2. Мальковский М.Г., Большакова Е.И., Волкова И.А. и др. Эксперименты с системой ЛИНАР // Труды машинного фонда русского языка. — 1991. №1. — с. 51–71.
3. Ихсанов Н.Х. Формальное моделирование подмножеств естественного языка: дис. канд. техн. наук: 05.13.18. — Ульяновск, 2000. — 167 с.
4. Raheja V., Alikaniotis D. Adversarial Grammatical Error Correction // The 2020 Conference on Empirical Methods in Natural Language Processing. — 2020. — 13 с.
5. Napoles C., Sakaguchi K., Tetreault J. JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction // 15th Conference of the European Chapter of the Association for Computational Linguistics. — 2015. — 6 с.
6. Большакова Е.И., Баева Н.В. Написание и оформление учебно-научных текстов (курсовых, выпускных, дипломных работ). Составление презентаций: Учебно-методическое пособие. — М.: Издательский отдел факультета ВМиК МГУ имени М.В. Ломоносова (лицензия ИД No 05899 от 24.09.2001); МАКС Пресс, 2012. — 64 с.
7. ГОСТ Р 7.0.11-2011. Система стандартов по информации, библиотечному и издательскому делу. Диссертация и автореферат диссертации. Структура и правила оформления. — Государственный комитет Российской Федерации по стандартизации, метрологии и сертификации, 2011.
8. ГОСТ 7.9-95. Система стандартов по информации, библиотечному и издательскому делу. Реферат и аннотация. Общие требования. — Государственный комитет Российской Федерации по стандартизации, метрологии и сертификации, 1997.

9. ГОСТ Р 7.0.5-2008. Система стандартов по информации, библиотечному и издательскому делу. Библиографическая ссылка. Общие требования и правила составления. — Государственный комитет Российской Федерации по стандартизации, метрологии и сертификации, 2009.
10. Сенкевич М.П. Стилистика научной речи и литературное редактирование научных произведений. — М.: Высш. школа, 1976. — 263 с.
11. Рыжиков Ю.И. Работа над диссертацией по техническим наукам. — СПб.: БХВ-Петербург, 2006. — 496 с.
12. Braams J., Carlisle D., Jeffrey A., Lamport L. The LaTeX 2e Sources. — LaTeX Project team, 2020. — 955 с.
13. Баева Н.В., Большакова Е.И. Проблемы автоматизации контроля учебно-научных текстов. — М.: МГУ им. М.В. Ломоносова, факультет вычислительной математики и кибернетики, 2012. — 6 с.
14. Ефремова Н.Э. Методы и программные средства извлечения терминологической информации из научно-технических текстов: дис. канд. физ.-мат. наук: 05.13.11. — Москва, 2013. — 135 с.
15. Швец А.В., Кузнецова Ю.М., Осипов Г.С., Латышев А.В. Метод и алгоритм обнаружения признаков лингвистических дефектов в научно-технических текстах // Информационные технологии и вычислительные системы. — 2013. №2. — с. 79–87.