электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 81'322.2

О задаче определения стиля документов на естественном языке

***Ланко А.А.**, студент,*
Россия, 105005, г. Москва, МГТУ им. Н.Э Баумана,
кафедра «Программное обеспечение ЭВМ и компьютерные технологии»

Научный руководитель: Волкова Л.Л., ассистент
Россия, 105005, г. Москва, МГТУ им. Н.Э Баумана,
кафедра «Программное обеспечение ЭВМ и компьютерные технологии»
bauman@bmstu.ru

Данная статья посвящена определению стиля текста на естественном языке. Будет проведено детальное рассмотрение предметной области и результатов ее исследования, а также будет изложена идея, которая положена в основание предлагаемого метода.

Прежде всего, необходимо отметить, что решаемая задача лежит в области компьютерной лингвистики [1], продолжающей набирать популярной как среди обычных пользователей-лингвистов, так и среди профессионалов (программистов-лингвистов) в связи с развитием таких технологий, как распознавание голоса, автоматический разбор текстов, их категоризация, определение стиля и других характеристик текста. Разумеется, задача определения стилистики не является краеугольной, но это не отменяет ее важности, в частности, в вопросах поиска. Когда пользователь ищет информацию определенного рода, поисковая система сортирует результаты, в том числе представляется полезной их категоризация в зависимости от стиля.

Итак, что же такое стиль? Функциональные стили русского литературного языка – это разновидности языка, определяемые сферами деятельности человека и имеющие свои нормы отбора и сочетания языковых средств. Каждый стиль отличается от других следующими признаками: сферой использования, целью общения; формами, в которых он существует; набором языковых средств. Современный русский литературный язык представляет собой систему стилей: научный, официально-деловой, публицистический, художественный, разговорный.

Автор для демонстрации работоспособности разрабатываемого метода выбрал как эталонные лишь три из существующих стилей: научный, публицистический и художественный. Следует сказать несколько слов о каждом из них, чтобы дать им

определения и показать основные различия между ними с точки зрения смыслового содержимого и, как следствие, используемых конструкций русского языка.

На основании материалов [8] можно составить обобщающую таблицу для указанных стилей и провести разграничение, результаты приведены в таблице 1.

Таблица 1

Основные стили и их характерные признаки

Наименование стиля	Область применения	Назначение (задача) стиля	Общие характеристики	Языковые конструкции
Научный стиль	Учебники, пособия, научные труды и доклады	Сообщить научную информацию, объяснить ее, представив систему научной аргументации	Официальность; Логичность; Объективность; Смысловая точность	Редкое использование личных местоимений; Использование причастий и деепричастий, и оборотов с ними; Использование сложных предложений с союзами связи
Публицистический стиль	Газетные статьи; Журналы; Критика (литер.); Митинги	Воздействие на массовое сознание посредством общественно значимой информации	Логичность; Образность; Эмоциональность; Оценочность; Призывность	Широчайшее использование эпитетов; Восклицательные предложения; Неиспользование причастных и деепричастных оборотов
Художественный стиль	Книги; Художествен-ная литература	Нарисовать словами картину, выразить	Высокая образность	Использование эпитетов; Описательные конструкции

		отношение, воздействовать на чувства и воображение		

Корпус – набор размеченных и однозначно разобранных текстов на естественном языке. Это общий результат усилий программистов и лингвистов, рассмотревших, проанализировавших и формализовавших информацию об огромном пласте текстов всех возможных видов и типов. Анализ этих данных позволяет выделять в языках зависимости и связи, которые не были ранее обнаружены при «ручной» обработке текстов.

Автором используется русскоязычный корпус текстов SynTagRus [2, 9] – аббревиатура от английского Syntax Tagged Corpus. В переводе означает – Синтаксически Размеченный Корпус. Для корпусов текстов, например, разработчиками RusCorpora [6], постулируется включение в корпус текстов со всеми рассматриваемыми в статье стилями.

Разрабатываемый метод подразумевает выделение связей между словами в предложении. Эти связи могут быть результатом синтаксического анализа, что более затратно в связи с разрешением неоднозначностей, но более качественно [4], или их можно получить из N-грамм.

N-грамма имеет в основе своего очень широкого использования математическую модель и определяется так: «N-граммой на алфавите V называют произвольную цепочку длиной N, например последовательность из N букв алфавита V одного слова, одной фразы, одного текста или, в более интересном случае, последовательность из грамматически допустимых описаний N подряд стоящих слов» [5].

Для того чтобы положить этот метод в основу дипломной работы, сначала нужно было убедиться, что такой метод еще не реализован. После длительных и не слишком результатных поисков в сети интернет по различным запросам удалось обнаружить весьма интересный ресурс [3], где опубликованы обобщающие таблицы известных программ в области компьютерной лингвистики. Указаны не только авторы и методы, но также указаны ссылки и стоимость продукта. Также неоспоримым показателем качества является то, что любой пользователь может, обнаружив что-то новое, отправить эту информацию создателям ресурса, и после проверки информация появляется в таблице.

После тщательного изучения всех представленных программ обнаружилось, что задача определения стилистики решается, в основном, основываясь на полном синтаксическом анализе предложений.

И лишь одна программа решала эту задачу весьма интересным образом. Программист Леонид Делицин, автор программы «Худломер»[7], использует очень близкий подход к предложенному решению поставленной задачи, как то, использование размеченного корпуса на русском языке, но проводит анализ стилистики текста исходя из длин слов, наполняющих текст, тогда как здесь предлагается анализировать наиболее характерные для стилистики N-граммы.

Теперь перейдем к рассмотрению самого метода. Подход к анализу текстов на естественном языке приведен на рисунке 1.

Рис. 1. Первый этап решения задачи – анализ текста

Ключевой момент во всем методе – это способ выделения и анализа связей между словами в тексте. В частности, предполагается замена сложного и нетривиального синтаксического разбора методом выделения N-грамм и сравнительный анализ качества работы метода при использовании указанных способов выделения связей.

Синтаксический разбор гарантирует высокое качество анализа, поскольку разбирает предложение «по кирпичикам», устанавливая все зависимости. Но в силу особенностей русского языка – различных неоднозначностей – он зачастую требует корректирующего вмешательства эксперта вдобавок к машинному синтаксическому разбору. Все это отнимает очень много времени. Использование же N-грамм является менее затратной альтернативой, и предполагается изучить влияние подобной замены на качество работы метода.

Формализованные файлы, которые предоставляет корпус SynTagRus, содержат предложения с проведенным синтаксическим разбором, но автор намеревается создать возможность в программе пополнять базу текстов (изначально наполненных

исключительно файлами корпуса). От идеи реализации синтаксического разбора автор отказался в силу указанных выше причин. Именно поэтому во всех файлах корпуса будет проведен N-граммный разбор, который не займет много времени и ресурсов и позволит проводить в дальнейшем не только сравнение и анализ работы метода, но и пополнять базу новыми данными. В силу проводимой замены, по качеству предложенный метод анализа будет уступать синтаксическому, но с учётом того времени, которое метод будет затрачивать на получение конечного результата, быстродействие должно компенсировать потерю качества.

Итак, первый этап – создать основу (в дальнейшем именуемую «Базой») из текстов Корпуса с проведенными в них N-граммными разборами предложений и подсчетом статистического распределения N-грамм, характерных для каждого из стилей. Фактически можно сказать, что корпус необходим исключительно из-за того, что в текстах, которые его наполняют, однозначно и корректно можно определить стилистику, чтобы обеспечить корректность работы всего метода. «База» - представляет собой не набор текстов, а набор статистической информации распределения N-грамм для каждого из трех стилей. Сами тексты для этого не нужны.

Сами N-граммы будут представлять собой связку из частей речи. Пример выделения и заполнения биграмм (выделение цепочек длиной в два слова: N = 2) представлен на рис. 2.

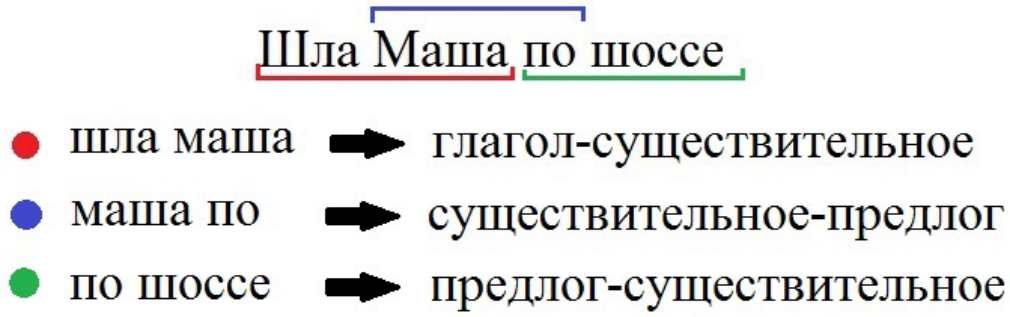

Рис. 2. Пример выделения и заполнения биграмм

Вопрос использования исключительно биграмм еще остается открытым, рассматривается возможность выделения также триграмм, а также и тех и других одновременно с целью параметризации и исследования точности метода на данном корпусе текстов.

Также существенным, но не ключевым моментом является определение частей речи, что не является тривиальной задачей, как и синтаксический разбор, ввиду неоднозначностей в языке (например, «стали» может быть прошедшим временем глагола

или родительным падежом существительного) [1]. Поиск в сети определил достаточно широкий выбор не очень сложных алгоритмов, работающих с погрешностями. Поскольку эта задача заслуживает отдельной статьи, а также в масштабах метода не очень велика, ее подробное и детальное рассмотрение мы опустим.

После просмотра всего текста получается некий набор данных – статистическая характеристика распределения N-грамм, характерная для данного текста. Завершающий этап метода – применение метрики стиля, рассчитанной на основании содержимого «Базы», к полученному результату.

После определения стилистики текста, его можно будет сохранить в «Базу» и отнести к определенному стилю результаты анализа. Таким образом, метод подразумевает дообучение по мере пополнения «Базы».

Итоговая обобщенная (в части анализа и обучения) концептуальная схема работы предложенного метода представлена на рис. 3.

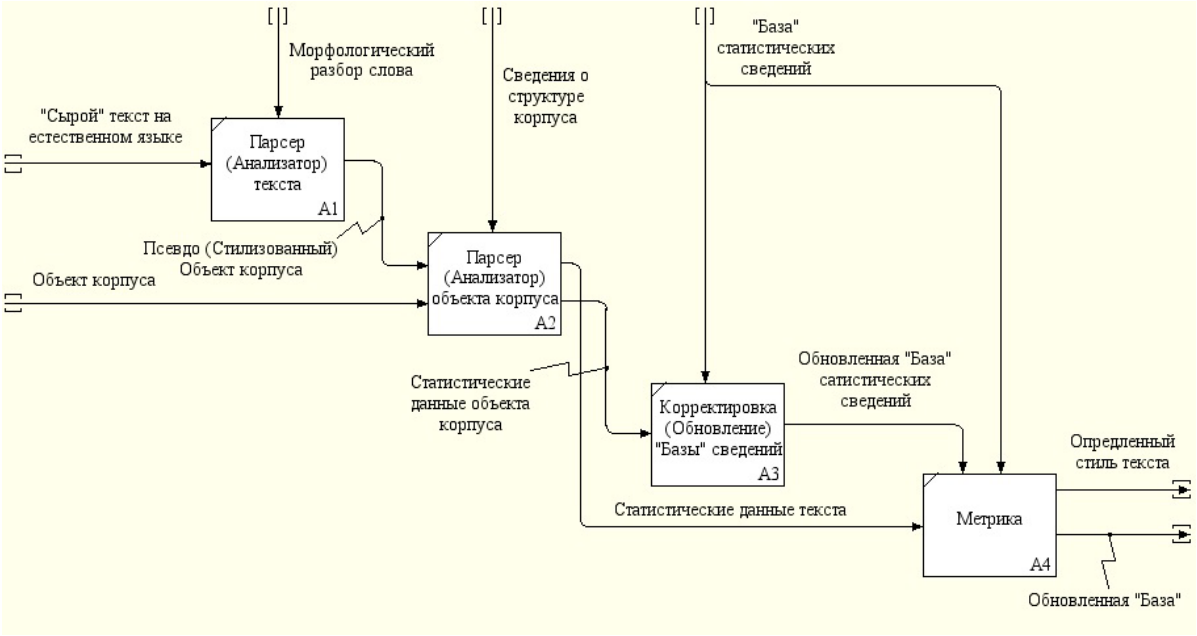

Рис. 3. Концептуальная схема IDEF0 работы метода

Вопрос относительно метрики и обновления «Базы» все еще остается открытым, но можно точно сказать, что будет использован простой частотный анализ, определяющий стиль (качественная оценка) и также некий параметр погрешности (количественная оценка). Возможно дополнительно использовать обучаемую нейронную сеть для обеспечения более точного результата и более эффективного внесения корректировок от вновь полученного анализа.

Список литературы

1. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 272 с.
2. Богуславский И.М., Иомдин Л.Л., Тимошенко С.П., Фролова Т.И. Development of the Russian Tagged Corpus with Lexical and Functional Annotation // Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop. Proceedings. Bratislava, Slovakia. April 15-16, 2009. Bratislava, 2009. P. 83-90.
3. Виртуальная Библиотека. Режим доступа http://www.rvb.ru/soft/catalogue/catalogue.html (дата обращения 25.03.2015).
4. Волкова Л.Л. Машинная лингвистика: от перевода со словарём к нелинейным динамическим системам // Новые информационные технологии в автоматизированных системах: материалы шестнадцатого научно-практического семинара. М.: Моск. Ин-т электроники и математики национального исследовательского университета «Высшая школа экономики», 2013. С. 317-328.
5. Гудков В. Ю., Гудкова Е. Ф. N-граммы в лингвистике // Вестник Челябинского государственного университета. 2011. № 24 (239). Филология. Искусствоведение. Вып. 57. С. 69–71.
6. Официальный сайт корпуса русского языка RusCorpora. Режим доступа http://ruscorpora.ru/en/ (дата обращения 18.03.2015).
7. Последняя версия программы «Хулдомер». Режим доступа http://teneta.rinet.ru/hudlomer/ (дата обращения 25.03.2015).
8. Скаженик Е.Н. Деловое общение: учебное пособие. Таганрог: Изд-во ТРТУ, 2006. 188 с.
9. Nivre J., Boguslavsky I.M., Iomdin L.L. Parsing the SynTagRus treebank of russian // Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), 18-22 August 2008, Manchester, UK. 2008. P. 641-648.