ФУНКЦИОНАЛЬНЫЕ СТИЛИ РУССКОГО ЯЗЫКА И ИХ ВЛИЯНИЕ НА ЗАДАЧУ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ ТЕКСТА

FUNCTIONAL STYLES OF RUSSIAN LANGUAGE AS APPLIED TO AUTOMATIC TEXT SUMMARIZATION

Емашова О.А. *(olga.emashova@gmail.com)*
Мальковский М.Г. *(malk@cs.msu.su)*
Московский Государственный Университет им. М.В. Ломоносова,
факультет Вычислительной Математики и Кибернетики

В работе предложен новый настраиваемый алгоритм реферирования текстов на русском языке. Отличительной чертой и основной особенностью предложенного алгоритма является гибкая настройка метода реферирования на функциональный стиль текста.

Введение

В современном мире самым критическим ресурсом является время. Каждый день человек вынужден сталкиваться с огромным количеством информации, которая требует своевременной обработки. Значительная часть этой информации представлена текстами на естественном языке. Зачастую документов оказывается так много, что физически невозможно внимательно прочитать их в отведенное для этого время. В этом случае на помощь человеку приходят системы автоматического реферирования текстовых документов.

Система автоматического реферирования позволяет составить реферат текста на одном из естественных языков. Уже довольно длительное время проводятся исследования, посвященные проблеме автоматического составления рефератов текстов [1]. Результатом исследований явились несколько методов, активно используемых в современных системах автоматического реферирования. Методы общего назначения не могут обеспечить высоких результатов при реферировании широкого класса текстов.

Выходом служит разбиение всего класса обрабатываемых текстов на несколько подклассов, внутри каждого из которых тексты обладают схожими свойствами [2]. Как правило, в качестве параметра разбиения выбирается предметная область [3], для которой строиться узкоспециальный алгоритм реферирования. Однако такие алгоритмы трудно поддаются модификации при смене предметной области.

Функциональный стиль (ФС) текста является одним из эффективных параметров классификации текстов русского языка в задаче автоматического реферирования. Известно нескольких эффективных алгоритмов реферирования для текстов информационно-аналитических и научных статей [4, 5]. Логическим продолжением этих работ является построение системы, которая может реферировать тексты разных функциональных стилей. В данной работе предлагается новый метод автоматического квазиреферирования, учитывающий стилистические особенности текстов.

Функциональные стили русского языка

За каждой сферой общественной жизни закреплен свой подтип русского литературного языка, имеющий ряд отличительных черт на всех языковых уровнях. Эти черты определяют подтипы русского языка, которые называются функциональными стилями. В русской письменной речи принято выделять четыре функциональных стиля [6]:

- научный,
- официально-деловой,
- публицистический,
- художественный.

Тексты, относящиеся к разным функциональным стилям, обладают сходными характеристиками, учет которых позволяет точнее оценивать информативность слов и отрывков текста. Рассмотрим те свойства текстов соответствующих функциональных стилей, которые существенным образом влияют на выбор правильной стратегии реферирования.

Научный стиль (НС) характеризуется логичностью и последовательностью изложения, обилием терминов.

Именной характер научной речи приводит к десемантизации глаголов [7] и повышению информативной нагрузки имен (существительных, прилагательных) и причастий. Характерная для научного стиля специфическая организация текста выражается в том, что каждый логически законченный отрывок текста обрамляется вступительными и заключительными предложениями, содержащими основную информацию в кратком виде. Эту особенность, а также информационную насыщенность заголовков параграфов и всего текста целесообразно учитывать при реферировании документов научного стиля.

Для официально-делового стиля (ОДС) отличительными чертами являются четкость формулировок и однозначность толкования. Особое внимание уделяется выбору слов еще на этапе написания исходного текста, в результате чего в нем практически отсутствует побочная, необязательная информация, что позволяет говорить о равномерном распределении важных предложений по тексту и информативности по частям речи. Официально-деловой стиль диктует регулятивный, предписывающий характер речи [6]. В текстах часто встречаются параллельные синтаксические конструкции, оформленные в виде нумерованного списка. Списочные структуры часто могут быть сокращены без опасения за нарушение синтаксической корректности предложения.

Для текстов публицистического стиля (ПС) характерны социально-оценочный и информационный характер речи, использование широкого спектра выразительных языковых средств, употребление устаревших слов, слов в переносном значении. При написании текста автор стремится к образности и эмоциональной насыщенности. Наиболее информационно значимыми частями речи в текстах публицистического стиля являются глагол и имя существительное, что соответствует классическому для русского языка распределению информационной нагрузки [8]. Для получения лучших результатов, разделим тексты публицистического стиля на две группы: информационно-публицистические жанры (ИПЖ) и аналитико-публицистические жанры (АПЖ). Для текстов первой группы отличительными чертами являются краткость и характерная информационная загруженность первых предложений, в то время как структура текстов аналитико-публицистических жанров роднит их с текстами научного стиля. При реферировании текстов публицистического стиля следует учитывать, к какой из этих двух групп относится текст, и выбирать соответствующее распределение важности предложений в тексте. Автоматическое сокращение предложений требует особой аккуратности, так как может привести к потере или искажению информации.

Художественный стиль (ХС) включает разнообразные по объему, составу, форме, теме и жанру тексты. Реферирование текстов художественного стиля производится с опорой на общие принципы организации текста на русском языке. В предлагаемом алгоритме в качестве базового метода реферирования выбрано цитирование большими отрывками, пользователь может ознакомиться с основными фрагментами произведения и получить представление об особенностях языка автора.

После изучения особенностей функциональных стилей русского языка было решено разбить все тексты на следующие пять групп: тексты научного стиля, тексты официально-делового стиля, тексты художественного стиля, тексты информационно-публицистических жанров, тексты аналитико-публицистических жанров. Такое разделение является интуитивно понятным, выбор группы, к которой следует отнести реферируемый текст, не вызовет затруднения у пользователя даже при условии, что текст не был прочитан. При более детальной классификации пользователю потребуются дополнительные знания, чтобы определить нужную категорию текста.

Настраиваемый алгоритм реферирования для текстов разных функциональных стилей

Для каждой из пяти рассмотренных групп (НС, ОДС, ИПЖ, АПЖ, ХС) разработан собственный сценарий реферирования, позволяющий учитывать такие параметры как распределение важности предложений внутри текста, применимость модуля синтаксического анализа и т.д. Реферирование производится методом автоматической оценки отрывков текста с последующим выбором наиболее представительных отрывков и преобразованием их в конечный реферат. Отличительной чертой и главной особенностью предлагаемого общего алгоритма является гибкая настройка на группу реферируемого текста, что позволяет ему воплощать все пять разработанных сценариев (см. рис. 1).

На первом этапе происходит оценка слов, предложений и абзацев текста. Для назначения весов используются вектор коэффициентов информативности имени существительного, глагола и имени прилагательного/причастия (Inf_Arr[3]) и функция распределения важности предложений внутри абзаца (Inf_Func(m,n)). Обе эти характеристики зависят от функционального стиля исходного текста. Во-первых, для НС и АПЖ функция Inf_Sci(n,m) реализует идею, что начальные и конечные предложения каждого абзаца содержательно важнее, чем внутренние. Для ИПЖ Inf_News(n,m) воплощает следующий подход: начальные предложения текста содержательно важнее, чем остальные, а для ХС и ОДС нет зависимости важности предложения от его положения в тексте. Во-вторых, для ХС и ПС вектор Inf_Arr[3] принимает значение {1.75; 1.5; 1.0}, т.е. самыми информативными частями речи являются имя существительное и глагол. В текстах НС вектор Inf_Arr[3] при-

нимает значение {1.75; 1.0; 1.5}, т.е. самыми информативными частями речи являются имя существительное и имя прилагательное/причастие. Для ОДС все неслужебные части речи одинаково важны. Для оценки важности слов используются статистическая, морфологическая и стилистическая характеристики слов. Вес слова вычисляется по формуле:

$$w_i = \frac{f_i}{N} \cdot Pos(w_i) \cdot Kw(w_i),$$

где f_i – частота слова i, N – количество слов в частотном словаре лексики исходного текста, $Pos(w_i)$ – коэффициент информативности части речи (соответствующая координата вектора R_InfArr[3]), $Kw(w_i)$ – коэффициент, обозначающий принадлежность слова к ключевым и/или тематически важным словам. Далее происходит оценка предложений и абзацев. Вес предложения вычисляется как среднее арифметическое весов слов предложения, увеличенное в K раз, где K – это значение, возвращаемое функцией Inf_Func(m,n). Аналогичным образом вычисляется вес абзацев.

После вычисл ения всех весов выбираются отрывки текста, обладающие наибольшим весом. Для текстов ИПЖ в качестве отрывков выбираются предложения. Для текстов ХС выбираются параграфы. Для остальных стилей реферирование производится в два этапа: сначала выбираются представительные абзацы, из которых на втором этапе удаляются малосодержательные предложения. Для текстов Н и ОД стилей дополнительно существует возможность сокращения слишком длинных предложений. Сокращение предложений производится после выбора абзацев и перед удалением предложений. Для сокращения предложений требуется дерево предложения, которое строится с помощью системы автоматического синтаксического анализа «Treeton»[9].

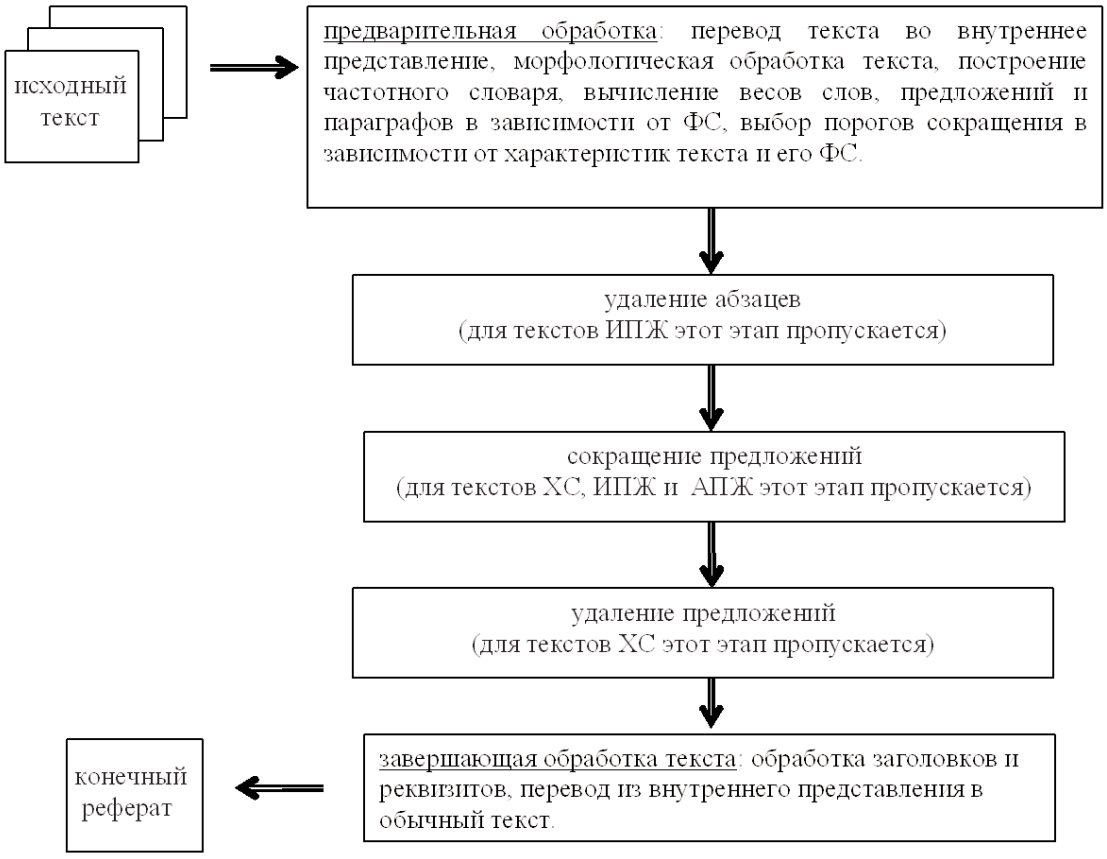

Рис.1. Схема работы настраиваемого алгоритма реферирования

Общий настраиваемый алгоритм предполагает реферирование текста с любым коэффициентом сокращения по выбору пользователя. Группу, к которой относится текст, также указывает пользователь.

Результаты

- Предложен подход, существенным образом учитывающий функциональный стиль реферируемого текста. Выбраны и параметризованы пять групп русского языка. Разработаны методы реферирования, учитывающие относительную информативность частей речи и распределение важности предложений внутри текста, характерные для каждого из выделенных классов текстов.
- Разработан общий алгоритм, настраиваемый на особенности конкретного текста и воплощающий разработанные методы реферирования документов разных функциональных стилей.
- На основе предложенного алгоритма на языке C++ реализован программный продукт, позволяющий реферировать произвольный текст на русском языке. Проведена проверка разработанного алгоритма автоматического реферирования на тестовом наборе текстов.

Пример работы системы

Исходный текст (информационно-публицистические жанры):

“Апофис” - стоит ли верить прогнозам астрономов?

Российские астрономы не слишком верят в вероятность столкновения Земли с астероидом “Апофис-99942” в 2029 году. Как сообщил в четверг на пресс-конференции старший научный сотрудник Главной астрономической обсерватории РАН Сергей Смирнов, малая планета “Апофис” приблизится к Земле в пятницу 13 апреля 2029 года на расстояние приблизительно 0.0002318 астрономических единиц, что составляет примерно 30-40 тыс. км. “Как известно, именно на этой высоте проходит геостационарные орбиты. Находящимся на них спутникам в случае встречи с вышеназванным астероидом грозит поломка, и обломки некоторых из них могут упасть на Землю”, - сказал Смирнов. Между тем, по мнению астронома, астероид пройдет между Землей и Луной, “как маленькая щепка между большим кораблем и катером - не касаясь ни того, ни другого”. Однако у населения России, хорошо знающего цену всевозможным прогнозам, особенно в области природных явлений, есть все основания проявлять если не беспокойство, то осторожность и предусмотрительность. Летящий со страшной скоростью огненный шар может повлечь самые серьезные последствия - достаточно найти на карте мира Мексиканский залив. Астероид, названный в честь древнеегипетского бога тьмы Апофиса, попал в поле зрения астрономов в июне 2004 года. Диаметр его составляет, по разным оценкам, от 400 до 600 метров, а скорость - более 30 километров в секунду.

Реферат с коэффициентом 0.2:

“Апофис” - стоит ли верить прогнозам астрономов ?

Российские астрономы не слишком верят в вероятность столкновения Земли с астероидом “Апофис-99942” в 2029 году . Астероид , названный в честь древнеегипетского бога тьмы Апофиса , попал в поле зрения астрономов в июне 2004 года .

Реферат с коэффициентом 0.6:

“Апофис” - стоит ли верить прогнозам астрономов ?

Российские астрономы не слишком верят в вероятность столкновения Земли с астероидом “Апофис-99942” в 2029 году . Как сообщил в четверг на пресс-конференции старший научный сотрудник Главной астрономической обсерватории РАН Сергей Смирнов , малая планета “Апофис” приблизится к Земле в пятницу 13 апреля 2029 года на расстояние приблизительно 0.0002318 астрономических единиц , что составляет примерно 30-40 тыс. км . Между тем , по мнению астронома , астероид пройдет между Землей и Луной , “ как маленькая щепка между большим кораблем и катером - не касаясь ни того , ни другого “ . Летящий со страшной скоростью огненный шар может повлечь самые серьезные последствия . Астероид , названный в честь древнеегипетского бога тьмы Апофиса , попал в поле зрения астрономов в июне 2004 года .

Список литературы

1. Jones K.S., Endres-Niggemeyer B. Automatic summarizing // Information Processing and Management. 1995, №31(5). С. 625-630.
2. Зубов А.В. Автоматическое построение табличного реферата группы текстов одной тематики // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2005». М.: Наука, 2005.
3. Oakes M.P., Paice C.D. The automatic generation of templates for automatic abstracting // 21st Annual BCS-IRSG Colloquium on IR. Glasgow, 1999.
4. Liang S.F., Devlin S., Tait J. Can automatic abstracting improve on current extracting techniques in aiding users to judge the relevance of pages in search engine results? // 7th Annual CLUK Research Colloquium, University of Birmingham. England, 2004.
5. Seki Y. Automatic summarization focusing on document genre and text structure. Doctoral abstract // ACM SIGIR Forum, №39(1). New York, 2005.
6. Стилистический энциклопедический словарь русского языка. Под ред. Кожиной Н.М. // М.: Флинта: Наука, 2006.
7. Валгина Н.С. Теория текста // Учебное пособие. М.: Изд-во МГУП «Мир книги», 1998.
8. Голуб И.Б. Стилистика русского языка // М.: Рольф, 2001.
9. Старостин С.А., Мальковский М.Г. Модель синтаксиса в системе морфосинтаксического анализа «Treeton» // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». М.: Изд-во РГГУ, 2006. С. 481-492.