О методе автоматического реферирования, основанном на результатах рубрицирования документов

RCDL 2008

7- 11 октября 2008 г.

Абрамова Н.Н., к.т.н., ФГУП «НИЦИ при МИД России»
Абрамов В.Е., ЗАО «ТЭЛКА»
NAbramova@mid.ru **NAbramova@mid.ru,**
AbramVal@yandex.ru

Предлагается рассмотреть

1. Основные цели и задачи исследования.
2. Существующие подходы к решению задачи реферирования.
3. Общую характеристику системы автоматического рубрицирования.
4. Метод автоматического реферирования, основанный на результатах рубрицирования.
5. Примеры реферирования документов.
6. Результаты экспертной оценки.
7. Выводы и направления дальнейших исследований.

Основные цели и задачи

Основная задача:

Разработать метод автоматического реферирования для работающей системы автоматического рубрицирования текстов.

Цели:

1. Максимально использовать результаты обработки, полученные на этапе рубрицирования, и составлять реферат после определения основных тем документа.
2. Оценить качество полученных рефератов.

Основные подходы

Для современных методов, относящихся к направлению квазиреферирования, характерно сочетание традиционного подхода, предложенного Г. Луном, с некоторыми модификациями. Например, в качестве значимых элементов выбираются не слова, а словосочетания, вводятся дополнительные критерии выбора значимых слов: вес слова увеличивается в зависимости от его нахождения в заголовке, в первом и последнем предложениях или выделения шрифтами в тексте или в запросе пользователя.

В России известны методы Белоногова Г.Г., Браславского П.И., Яцко В.А., Мальковского М.Г., Гусева В.Д., Мирошниченко Л.А., Саломатиной Н.В., Ступина В.С. и др.

За рубежом в области автоматического реферирования работают Salton G., Radev D.R., Blair-Goldensohn, Nomoto T., Matsumoto Y. , Nenkova A. , Mani I., Hahn U., Tait J. , Barzilay R. , Ando R.K. , Alonso L. и др.

Система автоматического рубрицирования текстов на разных языках (САРТ)

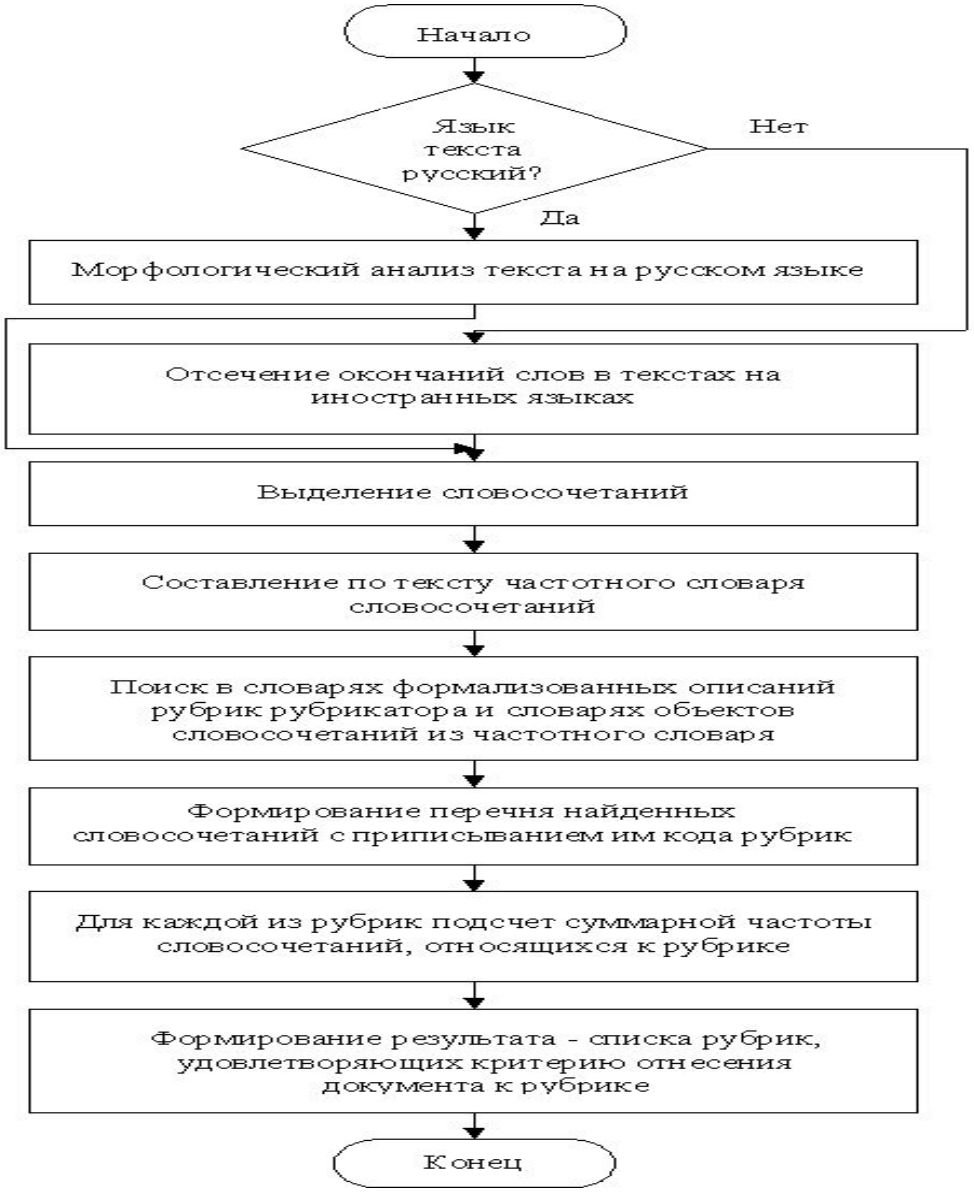
Система "САРТ" обеспечивает выполнение следующих функций в автоматическом режиме:

- определение тематических рубрик документа;
- определение объектов (стран, организаций);
- формирование поискового образа документа;
- формирование частотного словаря ключевых слов и словосочетаний на языке обрабатываемого текста;

и в автоматизированном режиме:

- составление словарей формализованных описаний рубрик по представительным выборкам документов;
- ведение словарей, необходимых для работы программы;
- корректировку результатов автоматического рубрицирования.

Обобщенная схема алгоритма рубрицирования текстов



Принципы выделения именных словосочетаний из русских текстов

1. Слова в словосочетании располагаются контактно.
2. Словосочетание не начинается с предлогов или союзов и ими не оканчивается .
3. Границами словосочетаний являются:

- знаки препинания (исключая запятую между однородными членами и точку после инициалов, цифр и сокращений и в интернет-адресах);
- скобки разного рода (круглые, квадратные, косые, фигурные);
- глаголы и подчинительные союзы;
- существительные или прилагательные в именительном или винительном падеже без предшествующего предлога.

4. Имена и фамилии, географические названия, названия организаций, партий и т.п. должны распознаваться в текстах с помощью специальных словарей.

Алгоритм выделения именных словосочетаний из русских текстов

1. Вычленение предложений в исходных текстах.
2. Определение предварительных границ словосочетаний в пределах предложения.
3. Генерирование всех возможных непрерывных последовательностей слов (двухсловных, трехсловных, четырехсловных и т.д. до десятисловных) из определенных фрагментов текста.
4. Редактирование последовательностей слов.
5. Формирование поисковых образов словосочетаний (ПОС).
   ПОС – это последовательность словоизменительных основ слов, входящих в словосочетание, с сохранением порядка следования. ПОС необходим для отождествления словосочетаний, отличающихся только формами слов, при формировании частотного словаря.
6. Подсчет количества слов в словосочетании.
7. Сортировка списка словосочетаний в алфавитном порядке ПОС-ов и длине словосочетания, т.е. в словаре по алфавиту сначала будут идти самые длинные словосочетания, потом менее длинные и в самом конце однословные словарные единицы.
8. Исключение из списка словосочетаний с совпадающими ПОС-ами, кроме одного из них, которому приписывается частота встречаемости данного словосочетания.

Пример работы программы автоматического рубрицирования на русском языке

Российским миротворцам преградили путь сотрудники МВД Грузии
19.10 2007 // 18:00
18 октября 2007 года в 16 часов 15 минут в районе н.п. Арцеви сотрудниками МВД Грузии были остановлены машины миротворческих сил от РФ с группой военных наблюдателей от трех сторон.
Грузинские полицейские блокировали дорогу на служебном автомобиле «Toyota» гос. номер FZZ 883. После разбирательства машины МС от РФ продолжили движение. Командование ССПМ обращает внимание, что данный факт является прямым вмешательством в деятельность миротворческих сил, грубейшим образом нарушает принятые сторонами договоренности и носит явно провокационный характер.
Источник : ГКИП РЮО.

1502025 Миротворцы России в зоне конфликтов на территории стран СНГ
1505000 Провокационные, недружественные действия по отношению к России
01268 Грузия
01643 Россия

Пример работы программы автоматического рубрицирования на английском языке

Sergey Bagapsh: Abkhazia and South Ossetia will take joint action to release Russian peacekeepers

Abkhazia will do its best seeking release of frontier guards detained by Georgian authorities and condemned to custody, Abkhaz President Sergey Bagapsh announced at a news conference in Sokhumi today, adding it is a priority issue for the Abkhaz leadership.

According to him, Georgia would still pursue the policy towards ousting Russian frontier guards from Abkhazia and South Ossetia. "We shall act together with Tskhinvali," Bagapsh said.

He also mentioned that the Abkhaz side knows everything about the attackers on the camp of the frontier guard service recruits, who were taking part in trainings in Khodjal, Tkvarcheli District. "We know them by names; they are the same people, who were gangsters in Gali District," the president said.

"Under our information, the detained servicemen are now in Zugdidi; six of them are hurt, one is in hospital," Sergey Bagapsh noted.

Exacerbation of tension has been seen in the Georgian-Ossetian and Georgian-Abkhaz conflict zones. According to Tskhinvali, Russian citizens, residents of Kabardino-Balkariya, Biosman Gizhgiyev and Beslan Khaptsev, were detained in South Ossetia. On September 20, two frontier guards were killed and seven were detained by Georgian law enforcers in the Georgian-Abkhaz conflict zone. The detainees were sentenced to two months in custody.

1502025 Миротворцы России в зоне конфликтов на территории стран СНГ
1505000 Провокационные, недружественные действия по отношению к России
01268 Грузия
01269 Абхазия
01270 Южная Осетия
01643 Россия

Исходные данные для задачи реферирования

n- количество документов

D- массив из n документов, $D=\{d_1, d_2, d_3, \ldots\ldots, d_n\}$

Для $\forall\ d_i \in D$ формируется набор тем $T=\{t_1, t_2, t_3, \ldots. t_m\}$ и набор весов каждой темы $P=\{p_1, p_2, p_3, \ldots\ldots\ldots p_m\}$.

$\forall\ t_j \in T$ описывается множествами слов и словосочетаний и частотами их появления в тексте

$W=\{w^{(1)}_j, w^{(2)}_j, w^{(3)}_j, \ldots w^{(l_j)}_j\}$ и $F=\{f^{(1)}_j, f^{(2)}_j, f^{(3)}_j, \ldots f^{(l_j)}_j\}$

$w^{(l_j)}_j$ – слово или словосочетание из документа d_i, определяющее тему t_j;

$f^{(l_j)}_j$ – частота появления в документе d_i слова или словосочетания $w^{(l_j)}_j$;

l_j – количество слов или словосочетаний, описывающих тему t_j.

Алгоритм реферирования

1. Формирование списка предложений, в которые входят слова и словосочетания, характеризующие темы.
2. Удаление из текста каждого предложения неинформативной лексики.
3. Вычисление веса каждого предложения.
4. Удаление примечаний и некоторых оборотов.
5. Проверка предложений на тождественность.
6. Вычисление коэффициента сжатия реферата.
7. Удаление предложения с самым маленьким весом, если был получен реферат с коэффициентом сжатия более заданной величины.
8. Повторение п. 6-7 до тех пор, пока не будет получен реферат, удовлетворяющий критерию сжатия.

Исходный текст для реферирования

В Южной Осетии может пролиться кровь из-за спорных фруктовых садов

Командование Смешанных сил по поддержанию мира (ССПМ) в зоне грузино-осетинского конфликта выступает инициатором встречи представителей администрации Знаурского района Южной Осетии с населением приграничных сел Грузии, в связи с нерешенностью территориальных споров вокруг фруктовых садов, сообщил корреспонденту ИА REGNUM помощник командующего ССПМ по работе со СМИ подполковник Юрий Верещак.

14 октября группой военных наблюдателей от трех сторон совместно с представителем Миссии ОБСЕ был проведен мониторинг в районе населенного пункта Нули (территория Грузии) и населенного пункта Гвертев (Южная Осетия) по факту обострения ситуации в данном районе.

Для предотвращения возможных инцидентов в районе садов выставлен временный наблюдательный пост миротворческих сил от России с наблюдателями от трех сторон.

15 октября для разрешения проблемы была проведена встреча представителей сторон, однако они к взаимоприемлемому решению не пришли. "До настоящего времени вопрос остается открытым" - сказал Верещак.

Как сообщил глава администрации Знаурского района Южной Осетии Заур Цховребов, суть конфликтной ситуации заключается в необоснованных претензиях жителей Нули на яблоневые сады обрабатываемых осетинским населением Гвертев. Сады находятся на территории Южной Осетии и эти претензии мы не понимаем", - сказал Цховребов, добавив, что "на предложенную 17 октября командованием ССПМ и Миссией ОБСЕ повторную встречу, грузинская сторона не явилась". "Чтобы ситуация окончательно не вышла из под контроля, мы предложили провести встречу завтра. Надеемся, что представители от грузинского села все-таки на нее явятся", - сказал глава администрации.

Данные, полученные на этапе рубрицирования

Тема	Вес темы	Определяющие ключевые слова с частотой встречаемости
Политические представители сторон и посредников урегулирования конфликта	15	наблюдатели 2, наблюдательный пост 1, миссия ОБСЕ 2, представители сторон 1, грузинская сторона 1, Россия 1, Южная Осетия 5, Грузия 2
Угроза применения силы, санкций и блокады, выдвижение ультиматумов	12	инцидент 1, конфликт 1, конфликтная ситуация 1, кровь 1, обострение ситуации 1, Южная Осетия 5, Грузия 2
Территориальные споры в населенных пунктах зоны конфликта	11	спорные фруктовые сады 1, спор вокруг фруктовых садов 1, претензии 1, территориальный спор 1, Грузия 2, Южная Осетия 5
Миротворцы России в зоне конфликтов на территории стран СНГ	5	Командование Смешанных сил по поддержанию мира 1, командование ССПМ 1, конфликт 1, миротворческие силы 1, Россия 1

Веса предложений

Номер пред-я	Вес пред-я	Определяющие ключевые слова и словосочетания с частотой встречаемости
1	7	Южная Осетия 5, кровь 1, спорные фруктовые сады 1
2	11	Командование Смешанных сил по поддержанию мира 1, конфликт 1, Южная Осетия 5, Грузия 2, территориальный спор 1, спор вокруг фруктовых садов 1
3	12	наблюдатели 2, миссия ОБСЕ 2, Южная Осетия 5, Грузия 2, обострение ситуации 1
4	6	инцидент 1, наблюдатели 2, наблюдательный пост 1, миротворческие силы 1, Россия 1
5	1	представители сторон 1
6	0	–
7	7	Южная Осетия 5, конфликтная ситуация 1, претензии 1
8	5	Южная Осетия 5
9	4	командование ССПМ 1, миссия ОБСЕ 2, грузинская сторона 1
10	0	–
11	0	–

Реферат текста

В Южной Осетии может пролиться кровь из-за спорных фруктовых садов.

Командование Смешанных сил по поддержанию мира в зоне грузино-осетинского конфликта выступает инициатором встречи представителей администрации Знаурского района Южной Осетии с населением приграничных сел Грузии, в связи с нерешенностью территориальных споров вокруг фруктовых садов.

14 октября группой военных наблюдателей от трех сторон совместно с представителем Миссии ОБСЕ был проведен мониторинг в районе населенного пункта Нули и населенного пункта Гвертев по факту обострения ситуации в данном районе.

Методика экспертной оценки

Трем экспертам были предложены 10 текстов документов и их рефераты. Эксперты отвечали на следующие вопросы, выбирая ответ из шкалы оценки:

1. Насколько полно реферат отражает содержание документа?
(0 – не отражает, 1 – недостаточно полно, 2 – удовлетворительно).
2. Присутствует ли избыточность в реферате?
(0 – да, много, 1 – да, не слишком много, 2 – нет).
3. Удовлетворяет ли реферат представлению о связности текста?
(0 – нет, 1 –не совсем, 2 – да).
4. Оцените длину реферата
(0 – слишком длинный, 1 – очень короткий, 2 – оптимальный).

Результаты экспертной оценки

N текста	*Оценки экспертов* 1	2	3	
1	7	8	6	21
2	6	5	6	17
3	6	7	5	18
4	5	8	6	19
5	8	6	4	18
6	5	7	6	18
7	5	6	6	17
8	7	7	8	22
9	6	5	5	16
10	6	7	7	20
Итого	61	65	59	186

Математическое ожидание оценки текстов всеми экспертами μ =6,16

Среднеквадратическое отклонение σ = 1,85

Выводы

1. Предложенный метод составления рефератов, рассматриваемый в данной работе, может быть с успехом применим в современных информационных системах при обработке больших информационных потоков, когда проводится автоматическое рубрицирование документов.

2. По сравнению с системами реферирования, в которых проводится полный цикл обработки документов, данный метод позволяет значительно сократить временные затраты на составление реферата.

3. Проведенная независимыми экспертами оценка качества реферирования показала, что метод, в целом, дает удовлетворительные результаты.

Планы дальнейших исследований

1. С целью сокращения объема реферата кроме вводных слов и предложений и оборотов из текста предложений можно удалять распространенные дополнения, причастные и деепричастные обороты в том случае, если они не включают частотную лексику. Но для этого нужны более сложные алгоритмы синтаксического анализа.
2. Проблема связности текста реферата требует разработки более совершенного алгоритма распознавания анафор, обеспечивающего замену анафорических слов и групп на их антецеденты, что позволило бы не вносить дополнительные вышестоящие предложения в реферат.