О методе автоматического реферирования, основанном на результатах рубрицирования документов

© В.Е. Абрамов
ЗАО СКБ «ТЭЛКА»
abramval@yandex.ru

Н.Н. Абрамова
НИЦИ при МИД России
nabramova@mid.ru

Аннотация

Работа посвящена методу автоматического реферирования, в котором используются результаты предварительно проведенного рубрицирования текстовой информации.
Обычно реферирование и рубрицирование рассматривают как самостоятельные задачи, хотя они взаимосвязаны: опираются на одни и те же процедуры автоматической обработки текстов (морфологический и синтаксический анализ). Идея метода заключается в том, чтобы максимально использовать результаты предшествующих этапов обработки и составлять реферат после определения основных тем документа. Полученные результаты показывают, что предложенный метод дает приемлемое качество реферата.

1 Введение

В настоящее время рубрицирование и реферирование информации широко используется в автоматизированных системах при формировании баз данных. Обычно автоматическое рубрицирование и реферирование рассматривают как самостоятельные задачи, хотя некоторые исследователи [10] указывают на взамозависимость этих задач.

Методы автоматического рубрицирования имеют много общего с методами реферирования, хотя присутствует специфика, характерная для каждого из этих двух классов задач.

Основной целью рубрицирования является определение тем документа. Предложения, в которых выражены темы документа, можно рассматривать как краткое изложение смысла документа, то есть его реферат. Составление реферата можно начинать, используя темы, определенные при рубрицировании.

Одна из целей исследования – показать, что использование результатов автоматического рубрицирования дает приемлемое качество реферата независимо от языка реферируемого текста. Преимуществом метода реферирования, использующего результаты автоматического рубрицирования, является сокращение времени обработки документов.

2 Существующие подходы

За прошедшие годы появилось много публикаций в области автоматического реферирования. Расширилось даже само понятие «автоматическое реферирование», которое относится не только к текстовой информации, но и к мультимедийной. Наряду с развивающейся традиционной задачей составления реферата отдельного документа появились новые сферы применения, к которым можно отнести мультиязычное реферирование, составление обзорных рефератов по набору документов, реферирование гибридных источников, содержащих текстовую и фактографическую информацию, создание видеорефератов на основе анализа видеообъектов, относящихся к какому-либо видеофайлу, а также реферирование интернет-сайтов. В обзорной статье [14] дан подробный анализ положения дел в области автоматического реферирования. На сайте постоянно действующей конференции DUC можно найти последние публикации в области автоматического реферирования, сравнительные оценки методов реферирования и результаты тестирования различных систем [13].

Так как целью нашей работы является решение традиционной задачи реферирования, то остановимся на подходах, которые в настоящее время наиболее распространены в этой сфере. Все методы, применяющиеся в реферировании текстовой информации, можно разделить на два направления – квазиреферирование и генерирование рефератов. Методы, относящиеся к направлению квазиреферирования, основаны на выделении из текстов наиболее информативных фрагментов (предложений), передающих основной смысл текста документа. Методы второго направления основаны также на выделении из текстов документов наиболее информативной информации и генерирования с помощью ее новых текстов. Практически все современные системы реферирования относятся к направлению квазиреферирования, хотя в последние годы появи-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

лись интересные работы, развивающие второе направление [2, 6, 11, 17].

Многие известные методы реферирования базируются на подходе, предложенном Г. Луном [18] в 50-х годах прошлого века, который заключается в выделении в тексте частотных слов, вычислении весов предложений с помощью суммирования частот (весов) входящих в их состав слов и включения в реферат предложений с наибольшими весами.

Для современных методов характерно сочетание традиционного подхода с некоторыми модификациями. Например, в качестве значимых элементов выбираются не слова, а словосочетания [3], 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 [5].

В работе [4] предлагается эффективный метод реферирования на основе машинного обучения.

Яцко В.А. предложил метод симметричного реферирования [12], в котором вес предложения определяется количеством связей между данным предложением и предложениями, находящимися слева и справа от него. Для этого в каждом предложении определяется список ключевых слов, входящих в предварительно составленный тематический словарь, а затем в предложениях, расположенных слева и справа, подсчитывается количество найденных в них ключевых слов (связей) из определенного ранее списка. Сумма лево- и правосторонних связей определяет вес предложения.

Известны подходы к реферированию на основе предварительно проведенной тематической кластеризации документа с дальнейшим выделением ключевых предложений из каждого кластера [21] и с предварительной разбивкой документа на части (с учетом его структуры), построением реферата для каждой части и отбором наиболее важных фрагментов [8].

Важнейшей частью задачи реферирования является оценка качества полученного реферата. Этой проблеме посвящено много работ [15, 19, 20, 22]. Помимо традиционных методик, связанных с экспертными оценками качества рефератов, в последние годы развиваются автоматические методы оценки. Например, в работе [22] предлагается оценивать реферат по наличию в нем частотных словосочетаний из оригинала и близости распределения относительных частот появления этих словосочетаний в документе и в реферате.

3 Идея исследования

Большинство исследователей рассматривают автоматическое рубрицирование и реферирование как самостоятельные задачи без связи между ними. Однако во многих информационных системах при формировании баз данных проводится не только рубрицирование информации, но и ее реферирование, поэтому наряду с полным текстом документов дается их сжатое содержание в виде реферата или аннотации. В разработанной авторами системе автоматического рубрицирования текстов (САРТ), которая работает в промышленном режиме полтора года, имеется функция реферирования [7].

Как известно, методы рубрицирования и реферирования опираются на одни и те же процедуры автоматической обработки текстов: морфологический и концептуальный анализ, т.е. обработку документа можно проводить однократно, увеличивая производительность системы.

Преимущества метода заметны при обработке больших потоков информации в системах, работающих в режиме мониторинга.

Для оценки качества реферирования необходимо разработать методику экспертной оценки и провести анализ ее результатов, что позволит найти пути улучшения работы алгоритма реферирования.

4 Составление реферата отдельного текста документа на основе результатов автоматического рубрицирования

4.1 Описание метода реферирования

После рубрицирования массива из n документов $D=\{d_i,\ i= 1, \ldots\ldots, n\}$ каждому документу d_i приписывается кортеж

$(T,P)=\{t_1,p_1,t_2,p_2,\ldots\ldots t_j,p_j,\ldots..t_m,p_m\}$,

где $j = 1,\ldots., m$ – номер темы,

T – набор тем в документе d_i,

P – набор весов тем в документе d_i (суммарных частот появления в тексте словосочетаний, имеющихся в описании тем из набора T).

В свою очередь каждая из тем $t_j \in T$ определяется множеством ключевых слов и словосочетаний, имеющимися в тексте документа d_i и в описании темы t_j, и частотами появления этих словосочетаний. Каждая тема t_j описывается с помощью кортежа

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

$w^{(l_j)}_j$ – 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.

Выбор наиболее информативных предложений из текста документа d_i проводился по следующему алгоритму:

1. Для каждой темы формировался список предложений, которые ее характеризуют.
2. Далее все предложения подвергались обработке с помощью словаря стоп-слов, включающего

служебные части речи, а также неинформативные слова и словосочетания.

3. Определялся вес каждого предложения путем суммирования частот появления в нем слов и словосочетаний, определяющих темы.
4. В каждом из отобранных предложений удалялись примечания (слова в круглых, косых или угловых скобках) и некоторые обороты с помощью специального словаря. Отличие между словарем оборотов и словарем стоп-слов в том, что неинформативные слова целиком задаются списком, а в словаре оборотов имеется только начальная часть фразы, например, «Как сообщается», «Как стало известно». После распознавания в тексте удаляется из предложения не только эта часть, но и весь оборот до знака препинания вместе с ним.
5. Все предложения с удаленными стоп-словами, примечаниями и оборотами, являвшиеся кандидатами для включения в реферат, последовательно проверялись на тождественность. Предложения считались тождественными при совпадении 80 % слов. Из всех близких предложений в реферат включалось предложение с максимальным весом.
6. Вычислялся коэффициент сжатия реферата как отношение объема реферата к объему документа (в байтах). Коэффициент сжатия обычно задается параметрически. Кроме этого определялось количество предложений в реферате в начале работы и после каждого удаления предложений.
7. Если был получен реферат с коэффициентом сжатия более заданной величины, то удалялось предложение с самым маленьким весом. Затем опять оценивалась длина реферата (п. 6) и в случае неуспеха выбиралось следующее предложение с минимальным весом и т.д. В том случае, если имелось несколько предложений с одинаковым минимальным весом, оставлялось предложение с наименьшей длиной (длина равнялась количеству информативных слов в предложении).
8. Процесс останавливался, если был получен реферат, удовлетворяющий критерию сжатия, или осталось только одно предложение (такая ситуация не встречалась).

4.2 Примеры работы системы

Рассмотрим работу алгоритма на примере русского текста. Исходный текст приводится на рис. 1. Он сопровождается результатами автоматического рубрицирования: дается перечень тем и объектов.

После этапа рубрицирования были переданы следующие данные: темы, их веса и ключевые слова и словосочетания, определяющие темы, с частотой их встречаемости в тексте. Эти данные представлены в табл. 1.

1. В Южной Осетии может пролиться кровь из-за спорных фруктовых садов.
2. 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.
3. 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.
4. Для предотвращения возможных инцидентов в районе садов выставлен временный наблюдательный пост миротворческих сил от России с наблюдателями от трех сторон.
5. 15 октября для разрешения проблемы была проведена встреча представителей сторон, однако они к взаимоприемлемому решению не пришли.
6. До настоящего времени вопрос остается открытым" - сказал Верещак.
7. 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.
8. "Сады находятся на территории Южной Осетии и эти претензии мы не понимаем", – сказал Цховребов,
9. добавив, что "на предложенную 17 октября командованием ССПМ и Миссией ОБСЕ повторную встречу, грузинская сторона не явилась".
10. "Чтобы ситуация окончательно не вышла из под контроля, мы предложили провести встречу завтра.
11. Надеемся, что представители от грузинского села все-таки на нее явятся", – сказал глава администрации.

Темы:

Политические представители сторон и посредников урегулирования конфликта

Миротворцы России в зоне конфликтов на территории стран СНГ

Угроза применения силы, санкций и блокады, выдвижение ультиматумов

Территориальные споры в населенных пунктах зоны конфликта

Объекты:

Грузия

Южная Осетия

Организация по безопасности и сотрудничеству в Европе

Рис. 1. Исходный текст документа (рус. яз.)

Таблица 1

Исходные данные для реферирования

Тема	Вес темы	Определяющие ключевые слова с частотой встречаемости
Политические представители сторон.....	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.....	12	инцидент 1, конфликт 1, конфликтная ситуация 1, кровь 1, обострение ситуации 1, Южная Осетия 5, Грузия 2
Территориальные споры.....	11	спорные фруктовые сады 1, спор вокруг фруктовых садов 1, претензии 1, территориальный спор 1, Грузия 2, Южная Осетия 5
Миротворцы России.....	5	Командование Смешанных сил по поддержанию мира 1, командование ССПМ 1, конфликт 1, миротворческие силы 1, Россия 1

Таблица 2

Веса предложений

Номер пред-я	Вес пред-я	Определяющие ключевые слова 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	

Далее подсчитывались веса предложений, равные суммарной частоте встречаемости в них определяющих темы ключевых слов и словосочетаний из табл. 1.

Веса предложений вместе с найденными в этих предложениях ключевыми словами и словосочетаниями, определяющими темы, приведены в табл. 2.

Из всех предложений удалялись стоп-слова, примечания и обороты, которые определялись с помощью соответствующих словарей.

В соответствии с выбранным коэффициентом сжатия реферата 0,3 было отобрано три предложения со следующими номерами: 1, 2, 3.

Окончательный текст реферата приведен на рис.2.

Приведем также пример работы алгоритма для английского языка. На рис. 3 дается исходный текст, а составленный по нему реферат приводится на рис. 4.

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.

Рис. 2. Текст реферата (рус. яз.)

A ceasefire between Georgia and breakaway South Ossetia signed late on Friday night was ended when shooting broke out in the South Ossetian village of Sarabuk and the Georgian village of Tamarasheni, killing 17 people and injuring 30, MosNews writes.
Of the 17 casualties, 15 were on the South Ossetian side, the Russian Information Agency Novosti reported, citing Georgian regional police.
The shootings, carried out by both sides the very day when Georgian and Ossetian military representatives met in Tskhinvali to establish peacekeeping posts with Russian troops, involved mercenaries from the neighboring North Caucasus region, where Chechnya is fighting its own separatist war with Moscow, Interfax reported, citing regional police.
The Joint Control Commission involving Georgia, Russia and South Ossetia, meeting Friday to regulate the conflict there, signed a preliminary ceasefire at midnight, with pledges from Russia to add peacekeepers to the breakaway region, which has said it wants to join Russia.
Fighting is continuing in the area, the Itar-Tass news agency reports.
Georgian Interior Minister Irakli Okruashvili said there would be no more talks following the recent attacks.
Flying into Eredvi by helicopter to evacuate the wounded, he accused South Ossetia of double standards, saying the authorities in the region's main city Tskhinvali were clearly incapable of sticking to the ceasefire.
As part of the deal, the two sides had agreed to create additional buffer zones between their positions.
These would be patrolled by Russian peacekeepers and monitored by the Organisation for Security and Co-operation in Europe (OSCE).
The commander of Georgia's peacekeeping battalion in South Ossetia, Alexander Kiknadze, said Georgian villages in the region came under heavy artillery fire in the latest attack.
He told Georgia's Rustavi 2 television that his battalion returned fire and had inflicted casualties on the South Ossetian side.
But Irina Gagloyeva, a spokeswoman for the South Ossetian authorities, said Georgian forces were the first to fire.

Темы:
Политические представители сторон и посредников урегулирования конфликта
Миротворцы России в зоне конфликтов на территории стран СНГ

Объекты:
Грузия
Южная Осетия
Россия

Рис. 3. Исходный текст документа (анг. яз.)

A ceasefire between Georgia and breakaway South Ossetia signed late on Friday night was ended when shooting broke out in the South Ossetian village of Sarabuk and the Georgian village of Tamarasheni, killing 17 people and injuring 30.
The Joint Control Commission involving Georgia, Russia and South Ossetia, meeting Friday to regulate the conflict there, signed a preliminary ceasefire at midnight, with pledges from Russia to add peacekeepers to the breakaway region, which has said it wants to join Russia.
Flying into Eredvi by helicopter to evacuate the wounded, he accused South Ossetia of double standards, saying the authorities in the region's main city Tskhinvali were clearly incapable of sticking to the ceasefire.
These would be patrolled by Russian peacekeepers and monitored by the Organisation for Security and Co-operation in Europe.

Рис. 4. Текст реферата (анг. яз.)

4.3 Оценка работы метода

Экспертная оценка проводилась только для текстов на русском языке. Методика оценки заключалась в следующем.

Трем экспертам были предложены 10 текстов документов и их рефераты. Экспертам предстояло ответить на следующие вопросы, выбрав ответ из шкалы оценки:

1. Насколько полно реферат отражает содержание документа? (0 – не отражает, 1 – не достаточно 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 – встречаются не связанные предложения, 2 – да).

4. Оцените длину реферата

(0 – слишком длинный, 1 – очень короткий, 2 – оптимальный).

Результаты экспертных оценок приводятся в табл. 3. По каждому тексту подсчитывались суммарные оценки экспертов, а для каждого эксперта вычислялась его суммарная оценка по всем текстам. Наиболее часто эксперты снижали оценки из-за длины реферата (слишком длинный) и из-за встречаемости предложений, нарушающих картину связности текста. От длины текстов, по которым составлялись рефераты, экспертные оценки практически не зависели.

Сравнительная оценка данного метода с другими широко известными методами реферирования не проводилась на большой выборке тестовых заданий. Однако мы провели эксперимент по реферированию на ряде русскоязычных текстов на основе дан-

ного метода и метода, предложенного Г.Г. Белоноговым [3].

Таблица 3

Экспертные оценки качества рефератов

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**

Этот метод заключается в выделении из текста каждого предложения частотных словосочетаний, вычислении весов словосочетаний, равных произведению количества слов в словосочетании на частоту их встречаемости в тексте, подсчете весов предложений с помощью суммирования весов словосочетаний, входящих в их состав, и включения в реферат предложений с наибольшими весами.

Рефераты, полученные по двум методам, для большинства текстов были похожи.

Например, для текста, рассмотренного в п. 3.2, с помощью метода Белоногова Г.Г. был составлен реферат, приведенный ниже:

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.

Видно, что в реферате, составленном по методу Белоногова Г.Г., отсутствует первое предложение из исходного текста, которое имеется во втором реферате. Однако во втором реферате текст предложений сокращен за счет исключения пояснений, а в первом – предложения включаются в реферат в том виде, в каком они извлекаются из текста (фрагменты, не вошедшие во второй реферат выделены курсивом).

5 Выводы и обсуждение результатов

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

Проведенная независимыми экспертами оценка качества реферирования показала, что метод, в целом, дает удовлетворительные результаты. Отмеченная экспертами в ряде рефератов слишком большая их длина является недостатком, который можно преодолеть за счет укорачивания длинных предложений.

В работе [5] описывается метод выделения по одному фрагменту из каждого предложения реферата вплоть до исчерпывания заданного лимита (в символах). Однако на наш взгляд, при таком подходе невозможно сформировать гладкий связный текст. Здесь требуется другой подход, близкий к предложенному в [9], основанный на синтаксическом анализе текста предложения. Отталкиваясь от результатов синтаксического анализа предложения, можно отсекать распространенные дополнения, причастные и деепричастные обороты. В дальнейших исследованиях мы собираемся опробовать алгоритм синтаксического анализа для реферирования.

В ряде случаев эксперты отметили несвязность текста реферата. Для решения проблемы связности можно использовать методы распознавания анафорических связей. Эта сложная задача может быть предметом отдельного исследования, однако с целью улучшения вида реферата на основе эмпирических наблюдений можно разработать алгоритм для распознавания анафоры. Авторы располагают подобным алгоритмом и программой с точностью распознавания ~60% [1].

Литература

[1] Абрамова Н.Н., Абрамов В.Е. Автоматическое составление обзорных рефератов новостных сюжетов // Интернет-математика 2007. – Екатеринбург, 2007. – С. 1–11.

[2] Алыгулиев Р.М. Автоматическое реферирование документов с извлечением информативных предложений // Вычислительные технологии. – 2007. – Т. 12, № 5. – С. 5–15.

[3] Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии. – М. : Русский мир, 2004. – 246 с.
[4] Браславский П.И., Густелев В. Система автоматического реферирования новостных сообщений на основе машинного обучения // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Девятой Всероссийской научной конференции RCDL '2007 (Переславль–Залесский, Россия, 15–18 октября 2007 г.). – Переславль-Залесский : Изд-во «Университет гор. Переславля», 2007. – С. 142–147.
[5] Браславский П.И., Колычев И.С. Автоматическое реферирование веб-документов с учетом запроса // Интернет-математика 2005. – М. : Яндекс, 2005. – С. 485–501.
[6] Гладун В.П., Святогор Л.А. Конспектирование естественно-языковых текстов с сохранением тематически-связной информации // Горизонты прикладной лингвистики и лингвистических технологий : Труды международной конференции MegaLing'2006 (Украина, Крым, Партенит, 20–27 сентября 2006г.). – http://www.megaling.crimea.edu/archieve/2006/reports.en.htm.
[7] Глобус Е.И., Абрамов В.Е., Абрамова Н.Н. Автоматическое рубрицирование текстовой информации (на русском, английском, немецком и французском языках). Свидетельство об официальной регистрации программы для ЭВМ № 2006613783, 31 октября 2006 г.
[8] Губин М.В., Меркулов А.И. Эффективный алгоритм формирования контекстно-зависимых аннотаций // Компьютерная лингвистика и интеллектуальные технологии : Труды международной конференции «Диалог'2005» (Звенигород, 1–6 июня 2005 г.). – М. : Наука, 2005. – С. 116–120.
[9] Емашова О.А., Мальковский М.Г. Функциональные стили русского языка и их влияние на задачу автоматического реферирования текстов // Компьютерная лингвистика и интеллектуальные технологии : Труды международной конференции «Диалог'2007» (Бекасово, 30 мая – 3 июня 2007 г.). – http://www.dialog-21.ru/dialog2007/ materials/html/25.htm
[10] Журавлев С.В., Добров Б.В. УИС "РОССИЯ". Автоматическое тематическое индексирование полнотекстовых документов // Научно-практическая конференция «Проблемы обработки больших массивов неструктурированных текстовых документов» (21-22 мая 2001 года, г. Москва). – http://www.fep.ru/text/dataarrays03.html.
[11] Ефименко И.В. Лингвистические аспекты кросс-языкового реферирования: синтез текстов под управлением предметных онтологий // Десятая национальная конференция по искусственному интеллекту с международным участием КИИ-2006 (25–28 сентября 2006 г., Обнинск) : Труды конференции : в 3 т. – М. : Физматлит, 2006.
[12] Яцко В.А. Симметричное реферирование: теоретические основы и методика // НТИ. Сер. 2, № 5, 2002. – С. 18–28.
[13] Document Understanding Conferences (DUC) Web site, 2008. http://duc.nist.gov
[14] Hahn U., Mani I. The Challenges of Automatic Summarization // Computer, vol.33, no.11, pp. 29–36, Nov., 2000. http://www.osp.ru/os/2000/12/178370/.
[15] Harman D., Over P. The effects of human variation in DUC summarization evaluation. In Text summarization branches out workshop at ACL'2004.
[16] Lee C.B., Kim M.S., Park H.R. Automatic summarization based on principal component analysis // EPIA 2003 : Portuguese conference on artificial intelligence N°11, Beja, 2003, vol. 2902, pp. 409–413.
[17] Leontyeva N. Semantic dictionary for Text Understanding and Summarization // International Journal of Translation. Vol. 15. No. 1. Ed. M. Blekhman. New Delhi, 2003.
[18] Luhn H. The automatic creation of literature abstracts. In IBM Journal of Research and Development, Vol. 2(2), pp. 159–165, 1958.
[19] Mani I. et al. The Tipster Summac Text Summarization Evaluation // Proc. 9th Conf. European Chapter of the November 2000.
[20] Nenkova A., Passonneau R. Evaluating content selection in summarization: The pyramid method. In Proceedings of HLT/NAACL 2004, Boston, MA, USA, 2004.
[21] Nomoto T., Matsumoto Y. The diversity-based approach to open-domain text summarization. In Information Processing&Management, 2003, 39, pp. 363–389.
[22] Tait J. Making Better Summary Evaluations. Proceedings of the Internatinal Workshop: "Crossing Barriers in Text Summarisation Research" // Horacio Saggion and Jean-Luc Minnel (eds.), Borovets, Bulgaria, Septemeber 2005, pp. 26–31. http://www.dcs.shef.ac.uk/~saggion/ CBTS_Papers/MS08.pdf.

On the Automatic Summarization Method based on the Results of Automatic Text Classification

V.E. Abramov, N.N. Abramova

This paper describes an automatic summarization method using the results of previously conducted classification textual information. Usually, summarization and classification are considered as separate tasks, although they are interrelated: based on the same procedure automatic word processing (morphological and parse). The idea of the method consists in maximum use of the results of previous

phases of processing and compile of summaries after determining basic order paper. Based on the received results it is possible to conclude that the proposed method gives acceptable quality of the summarization.