Автоматическое составление обзорных рефератов новостных сюжетов♣

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

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

Аннотация

Работа посвящена одной из актуальных проблем автоматического реферирования – составлению обзорных рефератов по набору документов применительно к новостным сюжетам. За рубежом данному направлению исследований придается очень большое значение, однако в России не уделяется достаточного внимания программе исследований по данной тематике. Авторами предложен метод составления обзоров новостных сюжетов, на основе которого разработана система реферирования. Приводятся результаты работы системы, описаны эксперименты по оценке качества реферирования. Эксперименты показали, что в среднем (с покрытием 80 % по всем трем коллекциям предоставленных для исследования документов) обзорные рефераты отражают содержание оригинальных текстов кластера.

1 Введение

Задача автоматизации аннотирования и реферирования текстовой информации не потеряла актуальность, несмотря на огромное количество появившихся в последние годы публикаций. Это вызвано, в первую очередь, необходимостью в условиях постоянного роста информации знакомить специалистов и других заинтересованных людей с необходимыми им документами, представленными в сжатом виде, но с сохранением их смысла. В обзорной статье [9] описывается современное состояние в области автоматического реферирования, а также основные направления и пути развития.

В традиционных методах реферирования чаще всего используются различные модификации подхода Г. Луна, известного с конца 50-х годов прошлого века, который заключается в отборе предложений с наибольшим весом для включения их в реферат, а также подходы, сочетающие традиционный подход с некоторыми новыми элементами, например, в [2] – с запросами пользователей, в [10] – с предварительной кластеризацией предложений. Вес предложения определяется как сумма частот, входящих в него значимых слов. В работе [1] описан метод, в котором в качестве значимых элементов выбираются не слова, а словосочетания.

При формировании и показе сообщений новостных сюжетов приобретает актуальность задача составления обзорных рефератов (по некоторому набору документов), в которых были бы представлены все основные вопросы, затрагиваемые в каждом документе, но в обобщенном виде без повторений информации. Важно также, чтобы обзорный реферат по каждой теме, освященной в нем, содержал ссылки на полные тексты документов, в которых отражена данная тема.

Составление обзорных рефератов относится к новым сферам применения автоматического реферирования, также как и получение одноязычных рефератов, охватывающих источники на разных языках, использование гибридных источников (например, статистической информации и сведений из баз данных), составление мультимедийных рефератов.

За рубежом в рамках конференций по проблемам автоматического аннотирования DUC (Document Understanding Conference) и текстового реферирования TSC (Text Summarization Challenge) данному направлению исследований придается очень большое значение. В Колумбийском университете (США) разработана система Newsblaster для поиска и обработки новостей, в которой есть возможность аннотирования новостной информации. О возможностях этой системы можно узнать из работы [7]. В статье Д. Радева и др. [11] описываются эксперименты по формированию обзорных рефератов с заранее заданным количеством слов (50, 100, 200 и 400). Приводятся экспертные оценки качества реферирования.

В работе [8] описывается метод формирования рефератов, представляющих собой краткое изложение всех сообщений об одном и том же событии из разных источников, и построенных путем выявления и отбрасывания избыточной информа-

Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007.

ции из каждого сообщения. В [5] решается задача выявления общей темы в различных сообщениях и представления результатов в виде диаграмм и графиков.

Однако в России не уделяется достаточного внимания программе исследований по данной тематике. Можно отметить работу [3], в которой предлагается лингвистический алгоритм автоматического построения табличного реферата группы текстов одной тематики (поездки, встречи и т.п.). Система автоматически выделяет главные опорные слова каждого текста и делит их на субъекты, объекты, предикаты, места действия и время действия. Совокупность таких разрядов слов для группы текстов и составляет их реферат.

Важнейшей частью постановки задачи является оценка качества полученного реферата. Основные положения метода оценки реферирования, предложенного в работе [12], выражаются в следующих утверждениях: 1) хорошие рефераты должны быть приемлемой заменой оригинала документа с точки зрения их индексации; 2) если проиндексировать весь текст и составленный по нему реферат или аннотацию, то в тексте и результате реферирования должны быть найдены термины, схожие с точки зрения индексации.

Тесты проверяют, были ли сохранены в реферате наиболее частотные словосочетания и насколько близки распределения относительных частот появления этих словосочетаний в наборе документов и в реферате.

В работе [6] рассматривается проблема связности текста полученного реферата. Предлагается алгоритм для упорядочивания предложений в реферате с помощью построения направленного графа "предшествования", в котором узлы являются темами, а ребра характеризуют отношения предшествования тем. Под темами понимаются отобранные по некоторым критериям ключевые предложения из всего реферируемого набора текстов. Задача упорядочивания предложений в реферате сводится к поиску оптимального пути на графе.

Для целей проводимого исследования важны не только методы, но и словарные средства. Исследования, описанные в работе [4], показали, что в задачах, связанных с кластеризацией новостных потоков, может успешно применяться заранее созданный большой лингвистический ресурс, например, тезаурус.

2 Идея исследования

Новизна настоящего исследования заключается в том, что для русского языка впервые предлагается метод автоматического составления обзорных рефератов документов, относящихся к определенному новостному сюжету. Исследование проводилось с использованием набора данных "Новости", предоставленного компанией «Яндекс».

По данным Информационного бюллетеня Яндекса "СМИ российского интернета" (осень 2006 года) 20% от объема информационных сообщений составляют новости, копируемые одними изданиями у других. Часто сообщения содержат перефразированный текст или являются рефератами более развернутых сообщений. Вследствие этих соображений, нецелесообразно представлять пользователям для просмотра все сообщения сюжета в виде последовательного списка, никак не связанных между собой сообщений.

Целью исследования явилась разработка нового формата представления новостных сюжетов в системе Яндекс.Новости в виде обзора, текст которого разбит на блоки, снабженные ссылками на полные тексты сообщений, в которых имеется близкая по смыслу информация.

В системе Яндекс.Новости каждому сюжету, формируемому автоматическим путем, соответствует кластер, содержащий тематически и хронологически связанные между собой документы.

Определим понятия темы и подтемы новостного сюжета. Тема сюжета – это проблема, которой посвящены все документы кластера (конечно, при условии хорошей кластеризации). Тема уточняется и конкретизируется за счет подтем, т.е. связанных с ней проблем, освещаемых в каждом документе сюжета. Например, тема кластера #15 ("Обычная неделя") – это "Расширение НАТО и планы военного строительства в России". Подтемы перечислены ниже:

1. Реакция России на неприсоединение к ДОВСЕ стран-новичков НАТО.
2. Встреча Совета Россия-НАТО на уровне министров иностранных дел.
3. Соглашение между Россией и НАТО "О статусе сил".
4. Соглашение между Россией и НАТО о сотрудничестве по борьбе с международным терроризмом.
5. Защита воздушного пространства России в связи с приближением к российским границам натовской инфраструктуры.
6. Выход в море флагмана Северного флота "Петр Великий".

Каждая из подтем находит свое выражение в предложениях, извлеченных из текстов документов. Далее подтемы рассматриваются только для отдельных кластеров, поэтому вместо слова «подтема» для краткости изложения будем пользоваться термином «тема».

Исходя из того, что все документы новостного сюжета объединяются общей тематикой, а каждый документ сюжета может характеризоваться некоторым набором тем, для решения поставленной задачи нужно уметь автоматически выделять темы в каждом новостном сообщении, отождествлять близкие по смыслу темы из разных сообщений сюжета и представлять информацию в виде обзора со ссылками на списки сообщений, в которых рассматривается та или иная тема.

Выявление основных тем в каждом документе основано на подходе, описанном в работе [1], в соответствии с которым темы определяются выделенными из текста документа наименованиями понятий, представляемыми частотными словосочетаниями. Предложения, несущие наибольшую смысловую нагрузку, определяются по максимальной сумме частот входящих в них значимых словосочетаний. При этом экспериментально нами было установлено, что, если вес какого-то предложения в 5 раз превышает средний вес всех предложений в документе, то такое предложение не должно выбираться как определяющее тематику. В новостных текстах такого рода предложения включают часто перечни имен и фамилий с указанием места проживания, списки футбольных команд со ссылками на место проведения матча, фамилии кандидатов в выборные органы со ссылками на партийную принадлежность, биржевые котировки и т.д.

Кроме того, при составлении обзорных рефератов невозможно обойти проблему связности текста. Эта сложная задача может быть предметом отдельного исследования [6]. С целью улучшения вида обзора на основе эмпирических наблюдений применительно к специфике новостных текстов нами были разработали алгоритмы для разрешения анафор, сжатия информации (удаления вводных слов и предложений и всякого рода пояснений), определения порядка расположения предложений, характеризующих основные темы, в обзорном реферате.

Еще одной важной задачей исследования была оценка качества реферирования. Тестирование проводилось по всем трем наборам документов, предоставленным компанией "Яндекс", на основе подхода, изложенного в работе [12]. При этом необходимо было выявить зависимости длины реферата от объема кластера и количества документов в кластере, что имеет немаловажное значение при составлении обзоров кластеров, содержащих большое количество сообщений (от 100 и более).

3 Описание методов, алгоритмов, экспериментов

3.1 Исходные данные

На обработку поступило три кластеризованных коллекции документов с условными названиями: "Неделя Шеварднадзе", "Неделя выборов" и "Обычная неделя", содержащих 32107 сообщений за период времени в 24 дня из 16 источников.

Эта выборка содержит примерно 20% от общего числа 163 тыс. документов, поэтому в большинстве кластеров имеется мало документов. Так из 4134 кластеров набора "Обычная неделя" 2598 состоят из одного документа, 913 кластеров – из двух, 215 кластеров – из трех, и только 95 кластеров содержат восемь и более документов. Аналогичная картина наблюдается и по двум другим коллекциям.

Из всех трех предоставленных коллекций для исследования было отобрано 160 кластеров с числом документов более семи:

– "Обычная неделя" – 95 кластеров;

– "Неделя Шеварднадзе"– 69 кластеров;

– "Неделя выборов" – 96 кластеров.

По размеру документов кластеры также не однородны: в некоторых кластерах преобладают документы размером до 1 Кб, в других – до 3-5 Кб. Но все же большинство кластеров содержат документы размером от 1 до 3 Кб. Однако встречались документы размером от 20 до 61 Кб, которые представляли собой, как правило, политематические обзоры СМИ.

Кластеры отличаются также по количеству документов (от 1 до 333). Большинство кластеров (95%) содержат менее 50 документов, и только 3 кластера из всех трех коллекций имеют более 100 документов. В реальности же в системе Яндекс.Новости многие кластеры содержат более сотни документов, поэтому проверить работу системы на кластерах с большим количеством документов мы практически не имели возможности.

В ряде документов имеются ошибки разбиения на предложения, что также создает дополнительные трудности при реферировании.

Помимо ошибок кластеризации, отмеченных в [5], в некоторых кластерах содержится информация, которая в принципе не подлежит реферированию. Например, в кластере #1735 почти все документы представляют собой списки мероприятий и памятных событий, происшедших в какой-либо день.

При составлении обзорного реферата необходимо учитывать отмеченные особенности кластеризации, чтобы минимизировать их влияние на качество реферата.

3.2 Метод обзорного реферирования

Предлагаемый метод опирается на разработанные нами ранее средства морфологического и концептуального анализа неформализованной текстовой информации.

Рассмотрим все этапы работы алгоритма для обработки одного кластера.

Этап 1. Морфологический анализ.

В результате всем словоформам исходных текстов приписывается грамматическая информация, необходимая для дальнейшей обработки.

Этап 2. Автоматическое выделение словосочетаний и формирование частотных словарей в каждом документе кластера.

При выделении словосочетаний мы опирались на концепцию Г.Г. Белоногова [1], заключающуюся в том, что основной смысл текста выражается не отдельными словами, а словосочетаниями. Наименования понятий выражаются именными словосочетаниями, которые представляют собой цепочки связанных по смыслу и контактно распо-

ложенных слов, относящихся к грамматическим классам "существительные", "прилагательные", "предлоги", "наречия" и "сочинительные союзы". Также дополнительно выделялись глаголы, кроме занесенных в список стоп-слов (например, "является", "был", "имеется" и т.д.). Для именных словосочетаний в список стоп-слов заносились все личные местоимения, имена, отчества, а также словосочетания типа "конце концов", "в то же время", "в настоящее время", "в связи" и т.д.

На длину словосочетания накладывались ограничения — до 10 слов. Выделение словосочетаний проводилось по следующему алгоритму:

1) *вычленение предложений в исходных текстах.* Границы предложений определялись по знакам препинания: точке, вопросительному и восклицательному знакам. Если встречалась точка, то проводилась проверка на сокращения, инициалы, интернет-адреса и цифры, использованные для нумерации разделов, после которых граница не ставилась. После первого предложения, являющегося заголовком текста, вставлялась точка;
2) *определение предварительных границ словосочетаний в пределах предложения.* Границы между словосочетаниями проводились по знакам препинания (",", ";","-",":"), скобкам разного рода (круглым, квадратным, косым, фигурным), по глаголам и подчинительным союзам, а также по существительным или прилагательным, стоящим в именительном или винительном падеже без предлога;
3) *удаление стоп-слов из каждого определенного на предыдущем шаге фрагмента и генерирование всех возможных последовательностей слов* (от одного до девяти), выбрасывание цепочек слов, начинающихся или заканчивающихся предлогами и союзами, а также цепочек, в конце которых стоят прилагательные, не согласованные с предшествующим существительным, или наречие. Оставшиеся цепочки слов являются искомыми словосочетаниями. Среди них также могут быть отдельные слова, представленные существительными, поэтому речь идет о словаре слов и словосочетаний;
4) *формирование поисковых образов словосочетаний (ПОС), выделенных в п.3.*
   ПОС — это последовательность словоизменительных основ слов, входящих в словосочетание, с сохранением порядка следования. Например, для словосочетания ″защита воздушного пространства России″ ПОС будет выглядеть так: "защит воздушн пространств росси″. ПОС необходим для отождествления словосочетаний, отличающихся только формами слов, при формировании частотного словаря;
5) *подсчет количества слов в словосочетании*;
6) *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;
7) *исключение из списка словосочетаний с совпадающими ПОС-ами*, кроме одного из них, которому приписывается частота встречаемости данного словосочетания.

Выделенные из текстов глаголы добавлялись к списку словосочетаний, сформированному в п.3, затем проводились операции, описанные в п.4-7.

Этап 3. Определение значимых предложений из каждого документа кластера.

1. Из составленных на предыдущем этапе частотных словарей выбирались слова и словосочетания с частотой более единицы.

2. Всем словарным единицам присваивались веса w, вычисляемые по формуле w=p*n, где p — частота встречаемости словосочетания, а n — количество слов в словосочетании. Для слов n=1, поэтому вес слова всегда совпадает с его частотой.

3. Слова и словосочетания проверялись на вхождение в каждое предложение текста, при этом словарные единицы, входящие на одном и том же отрезке текста в более длинные словосочетания, игнорировались.

4. Вес предложения определялся как сумма весов слов и словосочетаний из словаря, найденных в тексте.

5. В зависимости от количества документов в кластере устанавливался критерий отбора наиболее значимых предложений: выбирались предложения, вес которых составляет от 60% до 100% от максимального веса (см. рис.1). Этот критерий может быть изменен для документов, имеющих большой размер. Например, при количестве документов в кластере, равном 25, для всех документов при отборе значимых предложений установлен порог в 80% от максимального веса предложения, однако, если размер какого-то документа превышает 12 Кб, то происходит корректировка порога в соответствии с установленным коэффициентом изменения критерия только для данного документа. Параметры могут настраиваться в программе, а изменение критерия проводится автоматически в процессе работы.

Установлено ограничение для предложений, имеющих аномально высокий вес: вес значимого предложения не должен превышать более чем в 5 раз, средний вес всех предложений документа.

Также установлено ограничение по длине предложения: чтобы предложения, состоящие менее чем из 4 слов, не попадали в реферат, им присваивается нулевой вес.

6. Среди отобранных предложений могут встретиться предложения с так называемыми "висячими" словами, смысл которых не ясен из контекста. Для этих целей необходим алгоритм распознавания синтаксических анафорических связей.

Рис.1. Критерии отбора наиболее значимых предложений в каждом документе кластера

Чаще всего в анафорической функции выступают местоимения третьего лица, а также указательные, возвратные и относительные местоимения, которые могут быть поняты только тогда, когда соответствующий им референт имеется в предыдущем предложении. Нами рассмотрен только наиболее распространенный случай, когда между анафорой и антецедентом имеется явная кореферентность.

В простых предложениях анафоры находятся просто путем сравнения всех слов предложения с заданным списком. В сложных предложениях отыскиваются анафорические местоимения, находящиеся в первой части предложения, остающейся после удаления стоящих в начале предложения вводных слов и словосочетаний, вводных и вставных предложений, грамматически не связанных с членами предложения.

Автоматическое разрешение анафоры проводилось на основе следующего алгоритма:

1) последовательно выбирались наиболее значимые предложения с весами, удовлетворяющими критерию отбора;
2) в каждом из них удалялись вводные слова и обороты (например, такие как "в самом деле", "благодаря тому", "по словам" и т.п.) с помощью специальных словарей оборотов и вводных слов. Отличие между этими словарями в том, что вводные 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;
3) выделялась начальная часть выбранного предложения до первой запятой или бралось все предложение целиком, если в нем отсутствуют запятые;
4) все слова из данной части предложения проверялись на наличие в них анафорических местоимений, заданных специальным списком;
5) если были найдены анафорические местоимения, то выбиралось новое предложение, которое в этом тексте предшествует рассматриваемому предложению, и заносилось в массив предложений-кандидатов для включения в реферат;
6) в случае отсутствия анафор выбиралось следующее предложение (см. п.1).

Ниже приводятся примеры работы алгоритма.

Правильно распознанные анафоры:

*12 сентября 2003г. недалеко от жилой резиденции М. Зязикова было обнаружено мощное взрывное устройство. **Оно** было спрятано в перекрытой решеткой водосточной канаве у дороги, по которой ежедневно передвигается глава республики.*

Неправильно распознанные анафоры:

*Вместе с Аннаном во встрече участвует **его** заместитель, гендиректор Европейского отделения Оон Сергей Орджоникидзе.*

Работа алгоритма оценивалась на случайной выборке из 160 рефератов, составленных по исходным наборам документов. Из каждого набора выбиралось по 20 документов и в каждом из них подсчитывалось число правильно и неправильно определенных анафорических связей. Для набора данных "Обычная неделя" был получен наименьший процент правильно разрешенных анафор - 60%, для двух других наборов он равнялся примерно 65%.

Этап 4. Формирование общего списка тем кластера.

Предложения, выбранные из каждого текста кластера, объединялись в один список с сохранением информации о номере документа. Затем выявлялись близкие по смыслу предложения, выделенные из разных сообщений кластера. С этой целью использовался тезаурус.

Тезаурус по общественно-политической тематике представляет собой словарь условных синонимов (10 тысяч понятий), в качестве которых могут выступать существительные, субстантивированные прилагательные, именные и глагольные словосочетания. Иерархия между терминами не рассматривается. Словарные статьи тезауруса преобразуются к виду с одним текстовым входом, затем формируются поисковые образы заглавных дескрипторов и наименований понятий, представленных текстовыми выражениями. Определение поискового образа словосочетания было дано ранее (см. этап 2, п.4). Оно подходит и для данного случая, поскольку понятия в тезаурусе выражаются с помощью словосочетаний. Слово можно рассматривать как однословное словосочетание.

Словарные статьи в преобразованном тезаурусе отсортированы в алфавитном порядке ПОС-ов, выражающих наименования понятий, и по количеству основ слов в ПОС-е. Фрагмент тезауруса приведен ниже:

балансировани на гран войн#холодн войн
билет на одн поездк#разов билет

борьб за отмен рабств#аболиционизм
возрастн состав рабоч сил#возрастн дифференциаци населени
встреч на высш уровн#встреч в верх
.....
неорганическ мир#нежив природ
неправомерн действи#неправомерн поведени
непреодолим сил#форс-мажор
.....
глубинк#провинци
гнев#возмущени
гносеологи#теори познани
гомилетик#проповедническ искусств
гомогенност#однородност
города-побратим#породненн город
горсуд#суд
госаппарат#государственн аппарат
госбанк#государственн банк
.....

Во всех предложениях, относящихся к списку тем кластера, проводился поиск слов и словосочетаний из тезауруса. Найденные словарные единицы заменялись на заглавные дескрипторы тезауруса, обеспечивая снятие синонимии в списке тем кластера.

Отождествление предложений проводилось после описанных уже действий по сокращению текста предложений и тезаурусной обработки. Так как сравнение предложений сложно проводить на уровне словосочетаний из-за перестановок слов, употребления различных определений и дополнений и т.д., мы сравнивали в предложениях только существительные и субстантивированные прилагательные, предварительно составив эти списки, упорядочив их по алфавиту и выбросив дубли.

Предложения считались тождественными при полном вхождении одного из списка в другой. Из двух тождественных предложений выбрасывалось то, которое полностью входило в другое предложение, или любое из них – если списки, по которым проводилось сравнение, у обоих предложений совпадали. При этом ссылки на номера документов, из которых были извлечены выбрасываемые предложения, добавлялись к ссылкам оставленного предложения.

Этап 5. Построение итогового обзорного реферата.

Обычно при реферировании отдельных документов задают коэффициент сжатия реферата относительно объема исходного текста. Однако при составлении обзорного реферата нельзя задать одинаковый коэффициент для всей коллекции, так как нужно учитывать не только суммарный объем исходных текстов, но и количество документов в кластере. Мы ввели параметр длины реферата в килобайтах (в требованиях DUC принято выражать его в словах), установив по умолчанию 3 Кб (~400 - 420 слов). Реферат большего объема практически бесполезен, так как пользователи просто не будут читать его до конца.

После формирования окончательного списка предложений для включения в реферат проверялся его объем, и если он превышал установленное значение, то происходил возврат к этапу 3 – выбору наиболее значимых предложений с новым критерием отбора. Например, если был установлен критерий включения предложений с весами более 60 % от максимального веса предложения в документе, то автоматически выбирался следующий более сильный критерий – 70%. Если снова не удавалось достигнуть установленного объема реферата, итерация повторялась с еще более сильным критерием и т.д. Теоретически и при критерии в 100% можно получить реферат, превышающий установленный размер. Однако, как уже было отмечено, мы не располагали исходными данными, чтобы проверить это.

Логично предположить, что в подобных случаях можно провести реферирование уже полученного реферата, заново пересчитав веса предложений. Однако критерий выбора значимых предложений должен отличаться от описанного выше.

В процессе формирования окончательного текста реферата возникает задача расстановки предложений, выражающих темы, в логически связанной последовательности. Нами был разработан алгоритм для упорядочивания предложений.

Обозначим документы – D_i, темы – T_j. Тогда каждому документу D_i будет соответствовать упорядоченное множество (далее будем называть его списком) содержащихся в нем тем C_i: $C_i = (T_{j1}, T_{j2}, ..., T_{jk})$, в котором темы находятся в порядке их расположения в документе; количество тем в списке C_i обозначим через n_i. Каждой теме T_j соответствует список документов, содержащих эту тему: $K_j = (D_{j1}, D_{j2}, ..., D_{jm})$.

Шаг 1. Сортируем все документы в порядке уменьшения количества тем в документах n_i , а в случае одинакового количества – в порядке уменьшения размера документа. Таким образом, получаем упорядоченный список документов; обозначим его через S.

Шаг 2. Из списка S выбираем первый документ (пусть это будет D_1) и выводим в реферат темы из списка C_1 в порядке их расположения в нем.

Шаг 3. Для каждой темы из выбранного документа D_1 находим все документы, в которых встречается данная тема. Объединяем все найденные документы в один список. Сортируем все документы в этом списке в порядке уменьшения количества их повторений и убираем дубли. Тем самым мы переходим к документам, наиболее близким по тематике к документу D_1. Из этого списка исключаем уже рассмотренные документы. Просматриваем по очереди все документы из полученного списка и, если находим темы, которые еще не встречались, то выводим их в реферат в порядке расположения этих тем в исходном документе.

Шаг 4. Переходим к следующему документу из списка S и для него повторяем шаги 2 и 3.

Темы, выводимые в реферат из документов списка S, можно назвать темами первого уровня, а темы, выводимые из документов, содержащих темы первого уровня, – темами второго уровня.

При таком подходе, наряду с выводом в реферат тем первого уровня в порядке их расположения в документе, делается один шаг в глубину, чтобы добавить темы второго уровня. Однако продвижение далее вглубь нецелесообразно из-за ухудшения связности реферата.

3.3 Программная реализация

В среде Builder 6 C++ разработана экспериментальная система реферирования набора документов. Она позволяет обрабатывать как отдельный кластер, выбранный из списка, так и проводить пакетную обработку всех кластеров по каждой коллекции или сразу по всем коллекциям. В системе можно настраивать параметры отбора наиболее значимых предложений для включения в реферат, задавать ограничение объема реферата, проводить оценку покрытия оригиналов текстов словосочетаниями из реферата, строить графики зависимости длины реферата от объема кластера и количества документов в кластере.

В системе реферирования используются внешние программные модули:

- морфологический анализатор (разработчик Абрамов В.Е.), который базируется на обратном словаре словоформ объемом 70 тыс. лексических единиц (ЛЕ) и словаре служебных и коротких слов объемом 5 тыс. ЛЕ;
- система автоматического составления частотных словарей слов и словосочетаний (разработчики Глобус Е.И. и Абрамова Н.Н.).

3.4 Результаты

Результаты работы метода рассмотрим на примере кластера #101 с заголовком "Шииты захватили в плен солдат коалиции", относящегося к «обычной неделе». Кластер содержит 34 документа. Суммарный объем всех документов кластера составляет 50,5 Кб.

В каждом документе кластера были автоматически выделены словосочетания и сформированы частотные словари. На рис.2 приводится составленный по тексту частотных словарь слов и словосочетаний (с частотой f>=2) одного из документов кластера, а на рис.3 – исходный текст, в котором слева около каждого предложения указаны веса предложений.

8	ас-садр	2	город
4	ирак	2	насилие
4	лидер	2	не намерены мириться
4	сша	2	радикальный лидер
3	багдад	2	смерть
3	заявил	2	столкновения
3	понедельник	2	шиит
2	буш		

Рис. 2. Частотный словарь слов и словосочетаний

В соответствии с описанным в п. 3.2 (см. этап 3) алгоритмом подсчитаем, например, вес первого предложения. В тексте предложения встретилось 3 слова ("США", "шиитов", "ас-садра") и 2 словосочетания ("не намерены мириться", "радикальным лидером") из словаря, содержащего ЛЕ с частотой f>=2.

Вес предложения определяется как сумма частот слов и произведений частот словосочетаний на количество слов в словосочетании, т.е. для нашего примера – 4+2+8+2*3+2*2=24. По критерию отбора предложений (п. 3.2, этап 3) для включения в обзорный реферат проходит 3 предложения с весами 24 и 19.

24 Сша больше не намерены мириться с радикальным лидером иракских шиитов Моктадой ас-садра.
0 Багдад.
24 В связи с ожесточенными уличными столкновениями между оккупационными силами и сторонниками радикального лидера иракских мусульман-шиитов Моктады ас-садра, Сша более не намерены мириться с деятельностью этого человека.
19 Президент Сша Джордж Буш заявил, что ас-садр борется против демократии и делает ставку на насилие.
10 В то же время Буш указал, что Сша по-прежнему намерены передать власть в Ираке в запланированный срок - к 30 июня.
12 Глава временной американской администрации в Ираке Пол Бремер обвинил Моктаду ас-садра в беззаконии.
17 В понедельник стало известно, что уже несколько месяцев тому назад был выдан ордер на арест ас-садра в связи с убийством другого шиитского лидера в апреле прошлого года.
8 В конце минувшей недели Моктада ас-садр призвал своих единомышленников к активному сопротивлению оккупационным войскам.
7 В результате ожесточенных столкновений в Багдаде, Басре и в ряде других городов за последние дни погибли уже около 50 иракцев и 12 солдат коалиции.
6 В понедельник вечером ожесточенные уличные бои в Багдаде продолжались.
2 Корреспонденты агентства Франс пресс сообщают , что центральные районы города сотрясали мощные взрывы.
18 Сам Моктада ас-садр в понедельник заявил , что он намерен и впредь бороться против оккупации Ирака.
13 В заявлении, переданном по каналу арабской телекомпании "Аль-джазира", ас-садр заявил : "Мы не боимся смерти , для нас честь принять мученическую смерть во имя Аллаха".
12 Он отверг призыв высшего духовного лидера шиитов Ирака аятоллы Систани отказаться от насилия.

Рис. 3. Текст новостного сообщения из кластера #101 "Шииты захватили в плен солдат коалиции"

Аналогичным образом из всех остальных документов кластера были выбраны наиболее значимые предложения и сведены в общий список.

После обработки этого массива с помощью тезауруса, словарей оборотов и стоп-слов, исключения близких по смыслу предложений, разрешения анафор был получен окончательный список предложений для включения в реферат кластера (см.

рис.4). Текст реферата поделен на блоки, и справа указан номер, по которому в таблице ссылок можно найти все документы, имеющие те же темы. Фрагмент таблицы ссылок приводится на рис. 5. Слева от перечня ссылок указан номер блока реферата.

3.5 Оценки работы метода

Мы исследовали только оценку полноты содержания реферата, т.е. какой процент частотной лексики, имеющейся в документе кластера, попадает в реферат. Тестирование проводилось по всем трем исходным коллекциям документов. Для каждого кластера определялась степень покрытия его полученным рефератом, выраженная в процентах.

Методика оценки заключалась в следующем:

1. По тексту реферата составлялся частотный словарь слов и словосочетаний, алгоритм формирования которого описан в п.3.2., этап 2. Учитывались все словарные единицы.

2. Для каждого документа кластера по такому же методу составлялся словарь слов и словосочетаний, и выбирались словарные единицы с частотой f>=2.

3. Пусть для i-го документа кластера c_i - количество совпадений словарных единиц в обоих словарях, k_i - количество словарных единиц с частотой f>=2 в словаре, составленном по тексту документа, p_i - степень покрытия документа рефератом, n - количество документов в кластере. Определим $p_i = \frac{c_i}{k_i}$, где i=1,....,n

4. Для всего кластера степень покрытия p определим по формуле:

$$p = 100 \frac{\sum_{i=1}^{n} \frac{c_i}{k_i}}{n}$$

Из таблицы 1 и рис. 6, характеризующих величину p для трех исходных коллекций данных, видно, что для большинства кластеров степень покрытия находится в интервале 70-90%, однако для четырех кластеров из всех наборов она меньше 50%.

Напротив, в кластере #1735 в каждом документе любое предложение – это сюжет, и обзорный реферат в этом случае должен содержать только одну фразу: «Повестка дня на апрель - Тюменская область, Югра, Ямал». Однако наша система реферирования не рассчитана на обработку такого рода информации. 100%-ое покрытие для кластера #940 можно объяснить тем, что во всех 15 документах кластера содержится практически одна и та же информация.

Таблица 1. Количество кластеров с одинаковой степенью покрытия

Набор данных \ % покрытия	45-50	50-60	60-70	70-80	80-90	90-100
«Обычная неделя»	0	4	6	23	45	17
«Неделя Шеварднадзе»	0	0	4	18	34	13
«Неделя выборов»	1	4	9	11	43	28

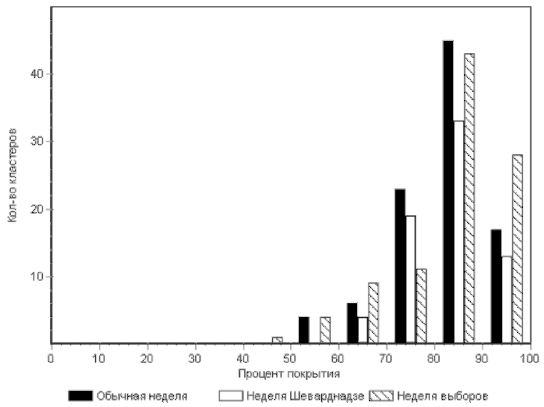

Рис. 6. Распределение кластеров по степени покрытия

Мы пытались выявить зависимость степени покрытия от объема кластера. Данные графика (рис.7) говорят о том, при объеме кластера до 10 Кб наблюдается устойчивое хорошее покрытие, а при больших размерах кластеров зависимость покрытия от объема кластера не прослеживается. Покрытие выше 80% имеют 73% кластеров набора "Неделя выборов", 64% кластеров «Обычной недели» и 68% кластеров "Недели Шеварднадзе". По всем кластерам трех коллекций усредненная величина покрытия равна ~ 80%.

Самый низкий результат покрытия получен для кластеров #8178 ("Неделя выборов") – 47,6% и #1735 («Обычная неделя») – 52,4%. 100%-ое покрытие получено для кластера #940 ("Неделя выборов"). Нами был проведен анализ информации кластеров, имеющих самые низкие и самые высокие показатели.

В документах кластера #8178 отсутствует сюжет – имеются фактографические сведения по раскрытию информации разных ЗАО.

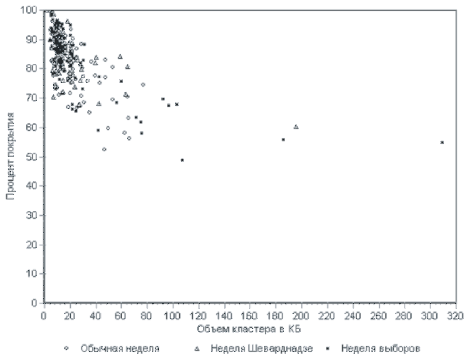

Рис. 7. Зависимость покрытия от объема кластера

Реферат новостного сюжета : сообщений--34

Шииты захватили в плен солдат коалиции

По американцам стреляли пока лишь сторонники 30-летнего Моктады ас-садра, шиитского священнослужителя и сына духовного лидера иракских шиитов великого аятоллы Мохаммеда ас-садра, убитого в 1999 году по приказу Саддама Хусейна.	--> 1
Американские войска в Ираке вступили в самые тяжелые бои за время с окончания войны в Ираке.	--> 2
Если накануне войны и сразу после ее окончания казалось, что наихудшим сценарием развития событий в Ираке станет война между суннитами и шиитами, то американцы наверняка назвали бы этот вариант "наименьшим из зол".	--> 3
Сенатор-демократ Эдвард Кеннеди, представитель самого влиятельного политического клана в Сша, выступая с лекцией в "Брукингс институшин", обратился с резким заявлением в адрес Джорджа Буша. "Ирак - это Вьетнам Буша-младшего", - сказал Кеннеди, обвинив также действующего президента Сша в том, что, "начав войну в Ираке под фальшивым предлогом, он отвлек внимание нации от настоящей войны с терроризмом".	--> 4
На территориях, населенных шиитами, уже третий день продолжается противостояние коалиционных сил и сторонников имама Моктады ас-садра, лидера радикальной группировки "Армия Махди".	--> 5
В ходе боев в Ираке между шиитскими боевиками и оккупационными войсками, за последние двое суток погибли 125 человек и более 550 получили ранения.	--> 6
В связи с ожесточенными уличными столкновениями между оккупационными силами и сторонниками радикального лидера иракских мусульман-шиитов Моктады ас-садра, Сша более не намерены мириться с деятельностью этого человека.	--> 7
В Ираке шиитскими боевиками захвачено в плен несколько солдат коалиции, сообщает агентство reuters со ссылкой на ливанский телеканал al-manar. Об этом было заявлено помощником радикального шиитского лидера Моктада аль-садра на пресс-конференции, прямую трансляцию которой осуществляет телеканал, принадлежащий группировке " Хизбалла ".	--> 8
Восстание сторонников радикального шиитского лидера Моктада аль-садра будет продолжаться, пока оккупационные войска не уйдут из населенных пунктов и не освободят пленных, сказал один из помощников аль-садра Каис аль-хазали на пресс-конференции, сообщает агентство reuters."	--> 9
Сообщения о возможности переброски в Ирак дополнительных войск появились на третий день акций протеста, организованных сторонниками радикального духовного лидера шиитов Муктады ас-садра и направленных против войск коалиции.	--> 10
5 апреля иракские власти выдали ордер на арест радикального лидера шиитов Моктады ас-садра, подозреваемого в причастности к убийству другого шиитского лидера.	--> 11
Духовний лидер иракских шиитов аятолла Али ас-систани подверг критике действия коалиции в последних столкновениях с вооруженными шиитскими отрядами и призвал обе стороны к спокойствию.	--> 12

Рис. 4. Представление обзорного реферата новостного сюжета

6	Бои с шиитами: в Ираке погибло 12 морпехов и 66 иракцев#grani.ru/Politics/World/Iraq/m.66318.html#грани.ру В Багдаде не прекращаются столкновения#www.svoboda.org/hotnews/2004/04/06/22.asp#радио_"свобода" Число погибших в Ираке за посление двое суток достигло 125 человек#www.polit.ru/news/2004/04/07/dead.html#полит.ру В Ираке в городе Рамади по меньшей мере 12 военнослужащих США погибли в боях с местными отрядами, - сообщил представитель Пентагона#www.svoboda.org/hotnews/2004/04/07/6.asp#радио_"свобода" В Ираке убиты 4 американских морских пехотинца#lenta.ru/iraq/2004/04/06/italian/#lenta.ru 12 морских пехотинцев США убиты в Рамади#lenta.ru/iraq/2004/04/07/ramadi/#lenta.ru Шиитский лидер Муктада ас-Садр призывает продолжать восстание#izv.info/world/news79825#известия
7	США больше не намерены мириться с радикальным лидером иракских шиитов Моктадой ас-Садра#www.dw-world.de/russian/0,3367,2207_W_1163540,00.html#dw-world
8	Шииты захватили в плен солдат коалиции#lenta.ru/iraq/2004/04/07/crash/#lenta.ru Иракские шииты захватили в плен несколько солдат коалиции#www.polit.ru/news/2004/04/07/shiit.html#полит.ру Шииты: Мы взяли в плен несколько солдат коалиции#grani.ru/Politics/World/Iraq/m.66428.html#грани.ру
9	Шиитский лидер Муктада ас-Садр призывает продолжать восстание#izv.info/world/news79825#известия Восставшие шииты обещают воевать до полного изгнания американцев из Ирака#lenta.ru/iraq/2004/04/06/continue/#lenta.ru

Рис. 5. Фрагмент ссылок на документы кластера, в которых отражены одни и те же темы

Мы также попытались выяснить, какая связь существует между объемом реферата и объемом кластера (суммарный объем всех документов кластера в Кб), а также зависит ли объем реферата от количества документов в кластере.

Из рис.8 видно, что реферат объемом до 2 Кб (~300 слов) получен только для кластеров, объем которых меньше 20 Кб, однако больше половины кластеров такого объема имеют реферат больше 2 Кб. При объеме кластера от 20 до 80 Кб никакой зависимости не наблюдается, а после 80 Кб объем реферата для всех кластеров приближается к предельному - 3 Кб. Из рис. 9 видно, что объем реферата не зависит от количества документов в кла-

стере, если их меньше 35, при большем количестве документов объем всех рефератов приближается к предельному – 3 Кб.

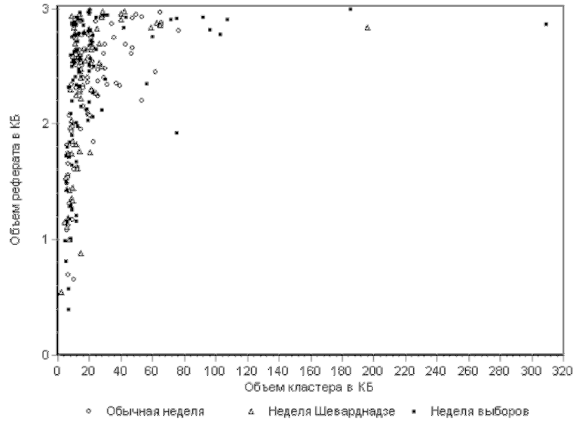

Рис. 8. Зависимость объема реферата от объема кластера

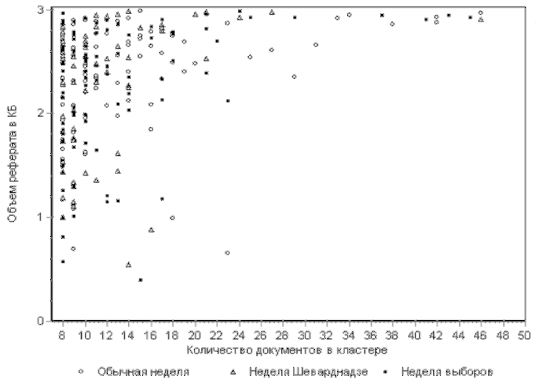

Рис. 9. Зависимость объема реферата от количества документов в кластере

4 Выводы и обсуждение результатов

Метод составления обзорных рефератов, рассматриваемый в данной работе, может быть применим на практике для представления новостных сюжетов в системе Яндекс.Новости. Проведенная автоматическая оценка качества рефератов показала, что они, в основном, отражают содержание кластеров. Однако оценка качества реферирования независимыми экспертами не проводилась в силу объективной причины отсутствия таковых.

Конечно, мы не смогли решить все проблемы на должном уровне, но это задача будущих исследований. Остановимся на проблемах, которые считаем важными для дальнейших исследований.

Прежде всего, это совершенствование используемых алгоритмов. Точность алгоритма разрешения анафор должна быть не ниже 90%, так как при неправильном разрешении анафор добавляется избыточная информация в реферат. Еще лучше разработать алгоритм, обеспечивающий замену анафорических слов и групп на их антецеденты, что позволило бы не вносить дополнительные вышестоящие предложения в реферат.

В алгоритм расстановки предложений в обзоре надо включить возможность выбора тематик с учетом хронологии. Было бы хорошо сравнить этот алгоритм с алгоритмом, предложенным в работе [6].

Отождествление предложений, извлеченных из разных документов кластера, очень сильно зависит от полноты тезауруса. Эту зависимость можно значительно ослабить, если предложить другой алгоритм, который, к сожалению, мы не успели проверить.

К нерешенным проблемам следует отнести и некоторую "неестественность" представления информации в реферате, когда в разных предложениях повторяется одно и тоже развернутое название объекта, а краткое название зачастую предшествует полному названию. Например, в первом предложении встретилось Игорь Иванов, а в последующих несколько раз повторяется "Министр иностранных дел России Игорь Иванов". Для решения этой проблемы можно использовать подход, предложенный в статье [6].

Выбор самых значимых предложений в документе не дает возможности выявить другие менее важные темы, на которые желательно также ссылаться в обзорном реферате. В первоначальном варианте нашей системы проводилось автоматическое рубрицирование всех документов кластера. В качестве рубрикатора выступали наиболее значимые предложения, отобранные со всего кластера и прошедшие этап отождествления, а словари слов и словосочетаний, характеризующие рубрики, формировались автоматически по всем близким по смыслу предложениям из разных документов кластера. К этому варианту можно вернуться, однако нужно разработать более точные критерия соответствия документа рубрике.

И, наконец, необходимо проверить использующиеся в данном методе возможности сжатия реферата до приемлемого размера для кластеров, содержащих сотни документов, или доработать метод, если будет получен неудовлетворительный результат.

Литература

[1] Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии. М,. Русский мир, 2004. - 246 с.

[2] Браславский П.И., Колычев И.С. Автоматическое реферирование веб-документов с учетом запроса. Сб. "Интернет-математика 2005", М., ООО "Яндекс", 2005 г.

[3] Зубов А.В. Автоматическое построение табличного реферата группы текстов одной тематики. Материалы конференции «Диалог-2005».
http://www.dialog-21.ru/Archive/2005/Zubov%20A/ZubovA.htm

[4] Лукашевич Н.В., Добров Б.В., Штернов С.В. Обработка потока новостей на основе больших лингвистических ресурсов. Сб. "Интернет-математика 2005", М., ООО "Яндекс", 2005 г.
[5] Ando R.K. et al. "Multidocument Summarization by Visualizing Topical Content," *Proc. ANLP/NAACL 2000 Workshop on Automatic Summarization,* 2000, pp. 79-88; http://www.isi.edu/~cylwas-anlp2000.
[6] Barzilay R. Sentence Ordering in Multidocument Summarization. Computer Science at Columbia University, Web seit, 2007, http://www.cs.columbia.edu/nlp/papers/2001/barzilay_al_01.pdf
[7] Blair-Goldensohn, S.Columbia University at DUC 2005. Публикации конференции DUC2005. http://www-nlpir.nist.gov/projects/duc/pubs.html
[8] Carbonell J.G., Goldstein J. "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," Proc. 21st Int'l ACM and Development in Information Retrieval SIGIR Conf. Research, ACM Press, New York, 1998, pp. 335-336.
[9] Hahn U., Mani I. "The Challenges of Automatic Summarization," Computer, vol.33, no.11, pp. 29-36, Nov., 2000. http://doi.ieeecomputersociety.org/10.1109/2.881692
[10] Nomoto T., Matsumoto Y. The diversity-based approach to open-domain text summarization. In Information Processing&Management, 2003, 39, pp.363-389.
[11] Radev D., Blair-Goldensohn S, Zhang Z. Experiment in Single and Multi-Document Summarization Using MEAD. Radev D.,web seite, 2007, http://tangra.si.umich.edu/~radev/papers
[12] Tait J. Making Better Summary Evaluations. http://www.dcs.shef.ac.uk/~saggion/CBTS_Papers/MS08.pdf.

Automatic compilation of news stories reviews

Abramova N.N., Abramov V.E.

This work deal with one of the topical problems of automatic summarization – multi-document summarization in respect to news stories. Abroad this line of researches is widely developed, however in Russia no is paid to this subject area. Authors propose the method of compilation of news stories reviews, on the basis of which is developed the summarization system. We present the sample summaries and describe experiments of summarization evaluation. The experiments proved that on average (with coating 80% as to three collections of documents provided for the research) survey summaries reflect the content of original texts.

♣ Авторы выражают благодарность компании "Яндекс" за финансовую поддержку проекта. Мы благодарим также наших сотрудников по работе Глобуса Е.И. и Шелимову И.Н., которые оказали содействие при тестировании системы и обсуждении результатов.