

Автоматическое аннотирование новостных кластеров на основе тематического представления

Summarization of news clusters based on thematic representation

Лукашевич Н. В. (louk@mail.cir.ru), **Добров Б. В.** (dobroff@mai.cir.ru)

Научно-исследовательский вычислительный центр МГУ
им. М. В. Ломоносова; АНО Центр информационных исследований

Представлен метод автоматического построения аннотации для новостного кластера на основе тематического представления кластера, моделировании лексической связности текста и тезаурусном описании лексических значений, что позволяет улучшить связность и полноту аннотации, а также снизить повторы.

1. Введение

Современные технологии обработки новостных потоков обычно включают в себя краткое представление содержания новостного кластера в виде аннотации (обзорного реферата). Так как в широких предметных областях результаты по автоматической генерации текста не являются устойчивыми, то стандартным способом аннотирования кластера является составление обзорного реферата, включающего заголовок и несколько предложений, обычно извлеченных из разных текстов кластера [11].

В настоящее время предложено достаточно большое количество методов автоматического обзорного реферирования [7]. Существенными проблемами при составлении аннотации новостного кластера являются [1, 7, 11]:

- обеспечение полноты представления информации, в том числе наиболее свежей информации,
- снижение повторов при представлении информации,
- обеспечение связности и понятности представляемой информации.

Для определения избыточности в порождаемых аннотациях используются различные меры сходства между предложениями. Одним из распространенных подходов является предварительная кластеризация — выделение близких по содержанию кластеров предложений [10]. Другим подходом для уменьшения избыточности является сравнение предложений-кандидатов с предложениями, уже попавшими в аннотацию, и оценка новой (непохожей) информации, например, подход Maximal Marginal Relevance (MMR) [6].

Для повышения связности аннотации и снижения повторов в работах [5, 8, 12] предлагается использовать построение аннотаций на основе так называемых лексических цепочек, то есть групп слов текста, связанных между собой по смыслу. Связность аннотации повышается на основе моделирования лексической связности текста. Суть подхода заключается в том, что повтор устраняется за счет оперирования не отдельными словами, а набором близких по смыслу слов. Основным источником информации о взаимосвязи слов при анализе англоязычных текстов служит обычно тезаурус WordNet.

В нашей работе [3] 1997 года был предложен подход к построению аннотаций отдельных документов на основе лексических цепочек специального вида — тематических узлов [4]. В качестве лексической базы для создания тематических узлов использовался двуязычный Информационно-поисковый тезаурус для автоматического концептуального индексирования по общественно-политической тематике (далее Тезаурус) [2].

Данная статья посвящена описанию расширения исходного подхода по автоматическому аннотированию одного документа на случай создания аннотации для новостного кластера.

2. Построение тематического представления текста

Основными критериями для построения лексических цепочек в большинстве подходов являются следующие [5, 8, 12]:

- наличие и сила связей между лексемами, описанных в некотором ресурсе,
- расстояние между вхождениями лексем в тексте, измеряемое обычно в предложениях. Если расстояние от текущего слова до предшествующих вхождений лексической цепочки больше некоторого порога, то лексическая цепочка прерывается и начинается новая,
- лексические цепочки строятся по порядку появления слов в тексте,
- для оценки важности лексической цепочки используется ряд параметров, таких как длина цепочки, плотность и др.

В нашем подходе мы обосновываем (подробнее см. [2, 3, 4]), что для того, чтобы лексическая цепочка соответствовала тематической структуре текста, нужно выполнение следующих условий:

- лексическая цепочка должна иметь внутреннюю структуру узла — к одному выделенному элементу относятся все другие элементы лексической цепочки,
- лексические цепочки в виде тематических узлов строятся на основе обработки целого текста, анализа частотностей концептов тезауруса, системы отношений между ними,
- значимость тематического узла для отражения содержания текста определяется не столько длиной, покрытием и другими характеристиками цепочки, а тем, насколько часто элементы этого тематического узла встречались с элементами других тематических узлов в одних и тех же предложениях текста, то есть насколько много пропозиций конкретных предложений текста было посвящено обсуждению отношений между элементами выявленных тематических узлов.

В совокупности построенные тематические узлы составляют тематическое представление текста.

В качестве примера рассмотрим автоматически создаваемое тематическое представление для следующего текста (представлен начальный фрагмент):

Китай и Тайвань установили авиасообщение после 60-летнего перерыва

После почти 60-летнего перерыва открылось регулярное авиасообщение между Тайванем и материковым Китаем. Первый чартерный рейс с 250 пассажирами уже прибыл в столицу Тайваня из китайского города Гуанчжоу, передает «Би-би-си». Ожидается, что аэропорты острова будут принимать рейсы из пяти китайских городов: Пекина, Шанхая, Гуанчжоу, Сямэня и Нанкина. Договоренность о прямых регулярных авиарейсах была достигнута в середине июня 2008 года на переговорах между руководством Тайваня и Китая. Восстановление авиасообщения

произошло не в последнюю очередь благодаря победе на выборах главы администрации Тайваня в марте 2008 года сторонников тесного сотрудничества с материковым Китаем...

Приведем примеры тематических узлов, созданных в процессе обработки этого текста (главное понятие тематического узла выделено сдвигом влево; указана частота употребления понятия в тексте):

КИТАЙ	8
ГАНЧЖОУ	2
ШАНХАЙ	2
НАНКИН	1
ПЕКИН	1
ТАЙВАНЬ	7
ТАЙБЕЙ	1
АВИАЦИОННЫЕ ПЕРЕВОЗКИ	2
АВИАРЕЙС	1
АЭРОПОРТ	1
ПОЛИТИЧЕСКАЯ ПАРТИЯ	1
КОММУНИСТ	1
ПРАВИТЕЛЬСТВО	1
ПУБЛИЧНАЯ ВЛАСТЬ	1

Для многих приложений важно определение основной темы документа, которая моделируется совокупностью основных тематических узлов. Если рассматривать основную тему документа как пропозицию, то каждый основной тематический узел соответствует отдельному тематическому элементу, входящему в состав этой пропозиции.

В нашей модели мы предполагаем, что понятия основных тематических узлов постоянно встречаются рядом друг с другом (связаны по тексту). Для оценки совместной встречаемости понятий (текстовых связей понятий) используется линейное окно понятий длиной 3.

После того как созданы тематические узлы, текстовые связи понятий каждого тематического узла суммируются и определяются текстовые связи между тематическими узлами.

Основными тематическими узлами, которые соответствуют основной теме документа, в первую очередь, являются такие тематические узлы, которые:

- все связаны между собой текстовыми связями;
- сумма частот текстовых связей между ними максимальна.

В рассматриваемом примере тематического представления основными тематическими узлами стали узлы с главными понятиями: *КИТАЙ, ТАЙВАНЬ, АВИАЦИОННЫЕ ПЕРЕВОЗКИ, ГОРОД, РЕЙС*

3. Использование тематического представления для составления аннотации одного текста

Для составления аннотации отдельного документа используются два принципа.

Во-первых, важными (информативными) и, следовательно, возможно включенными в аннотацию считаются те предложения текста, которые содержат по крайней мере два понятия, входящих в состав разных основных тематических узлов текста.

Во-вторых, для каждой пары выявленных основных тематических элементов текста в аннотацию выбираются предложения, содержащие первое вхождение этой пары, следуя по порядку текста.

Не все основные темы начинают обсуждаться в тексте сразу, с первого предложения -- часть из них возникает в последующих предложениях. Чтобы сохранить связность и последовательность изложения текста автор именно в этом первом предложении новой темы должен наиболее точно указать связь новой темы со всем предшествующим текстом. Таким образом, следуя за автором вводе нового тематического элемента, можно повысить общую связность аннотации.

Нужно отметить, что при хорошем покрытии предметной области тезаурусом появление в очередном предложении новой темы выявляется весьма точно, а это означает, что связность получаемой аннотации в среднем весьма высока.

Построение аннотации реализуется следующим образом:

- 1) Для построения аннотаций сначала формируется множество «аннотационных» фрагментов, которые являются целыми предложениями исходного текста, содержат в своем составе глагол в изъявительном наклонении или краткое прилагательное, и не являются вопросительными или восклицательными предложениями.
- 2) Перед построением аннотации создается таблица всех возможных пар основных тематических узлов.
- 3) Начиная с начала текста, отбираются такие предложения, которые содержат еще не упомянутую пару разных тематических узлов.

Таким образом для текста примера получаем следующую аннотацию, в которой упомянуты все основные тематические узлы данного документа:

Китай и Тайвань установили авиасообщение после 60-летнего перерыва.

Первый чартерный рейс с 250 пассажирами уже прибыл в столицу Тайваня из китайского города Гуанчжоу, передает Би-би-си. Ожидается, что аэропорты острова будут принимать рейсы из пяти ки-

тайских городов Пекина, Шанхая, Гуанчжоу, Сямэня и Нанкина.

Отметим, что в аннотации пропущено первое предложение, которое не содержит новой пары тематических узлов по сравнению с заголовком текста.

4. Построение тематического представления для новостного кластера

Новостной кластер представляет собой совокупность тематически близких документов.

Поэтому тематическую структуру новостного кластера так же, как и отдельного элемента можно выявить за счет построения тематического представления этого кластера, и это представление можно будет использовать для управления набором предложений в аннотацию кластера, а именно для обеспечения полноты, снижения повторов, а также обеспечения связности аннотации кластера.

Построение тематического представления новостного кластера осуществляется простым способом: все тексты кластера склеиваются в единый текст, для которого производится стандартный тематический анализ одного документа и строится тематическое представление.

Результат этой процедуры, а затем и результат построения аннотации в некоторой степени зависит от порядка просмотра документов в кластере. Мы используем следующий метод объединения документов кластера в единый текст, используемый для построения аннотации.

Сначала в новостном кластере определяется «центр кластера» — документ, наиболее близкий к центру тяжести множества документов кластера в метрическом пространстве нормализованных лемматическом и концептуальном (по тезаурусу) индексов. Определяется «ядро» кластера — документы достаточно близкие к центру (по некоторому порогу). Затем «центр кластера» сдвигается в документ из ядра кластера, который был опубликован последним по времени. Пересчитываются веса связей документов кластера к новому центру. С учетом задаваемого интервала времени по убыванию веса сначала заполняются документы за последнее время, затем все остальные. Так как отбирается всего несколько предложений, то имеется общее ограничение на количество отбираемых в «единый документ» документов.

Приведем примеры тематических узлов для кластера, в который входит текст примера (справа указана частотность концепта в кластере):

КИТАЙ	103
ПЕКИН	21
ГУАНЧЖОУ	13
ГОСУДАРСТВО	9
ЮАНЬ	7
ШАНХАЙ	6
КИТАЙЦЫ	5
НАНКИН	5
ГУАНДУН	1
АВИАЦИОННЫЕ ПЕРЕВОЗКИ	33
АВИАЦИОННАЯ КОМПАНИЯ	9
САМОЛЕТ	9
АВИАРЕЙС	7
ТРАНСПОРТНАЯ СФЕРА	4
АЭРОПОРТ	3
ТРАНСПОРТНЫЕ ПЕРЕВОЗКИ	2
АЭРОБУС	1
АВИАЛИНИЯ	1
ТУРИСТ	12
ЧЕЛОВЕК	62
ТУРИЗМ	2
ПОЕЗДКА	1
ПРАВИТЕЛЬСТВО	6
РУКОВОДИТЕЛЬ	6
ОРГАН ПУБЛИЧНОЙ ВЛАСТИ	3
РУКОВОДСТВО	2
ОРГАН ИСПОЛНИТЕЛЬНОЙ ВЛАСТИ	2
ПУБЛИЧНАЯ ВЛАСТЬ	1

Таким образом, по основным тематическим узлам тематического представления могут быть определены основные элементы, обсуждаемой в кластере темы.

Как видно, тематические узлы включают понятия достаточно разной частотности.

Низкочастотные концепты тематического узла могут быть ошибочно включены в тематический узел, кроме того, представительность ими основной темы документа невелика. Поэтому можно задать выделение ядра тематических узлов, которое определяется как коэффициент от 0 до 1. Этот коэффициент определяет, какая доля от общей частотности тематических узлов будет включена в ядро. Для построения аннотации кластера используются только ядра основных тематических узлов (используется коэффициент 0,7).

5. Построение аннотации новостного кластера

Аннотация новостного кластера состоит из заголовка и нескольких предложений из разных документов новостного кластера.

Зная ядра тематических узлов, полноту изложения содержания кластера мы обеспечиваем тем, что должны отбирать для аннотации предложения, содержащие пары этих тематических узлов — именно тогда эти предложения будут описывать взаимоотношения между основными тематическими элементами кластера.

При отборе заголовка для аннотации ищется заголовок, содержащий пару наиболее частотных тематических узлов. Если таких заголовков нет, то ищутся заголовки, содержащие понятия из одного наиболее частотного тематического узла.

Для выбора очередного предложения в списке основных тематических узлов отмечаются все тематические узлы, которые уже были упомянуты. Очередное предложение должно содержать пару основных тематических узлов.

Для обеспечения связности требуется, чтобы очередное предложение содержало либо уже упомянутый тематический узел, либо уже упоминавшееся слово с большой буквы.

Кроме того, делается ряд дополнительных проверок:

- предложение не должно являться вопросительным или отрицательным предложением;
- предложение не должно содержать в заданном числе первых слов местоимение;
- начало предложения не должно совпадать с началами заголовка и предложений, уже взятых в аннотацию;
- число слов предложения, совпадающего со словами предшествующих предложений, не должно превышать некоторой доли длины предложения.

Понятно, что даже при проверке вышеупомянутых условий может найтись еще достаточно много подходящих предложений-кандидатов. Кроме того, оценка предложений на основе понятий тезауруса не является достаточной без учета упоминаемых именованных сущностей, которые могут быть и не описаны в тезаурусе.

Поэтому вводится еще и общая оценка предложения с помощью вычисления веса предложения, которая складывается из двух компонентов: весов упомянутых понятий Тезауруса, которые были получены в тематическом представлении [2], а также весов содержащихся в предложении слов с большой буквы, не считая первого слова предложения.

Для вычисления весов слов с большой буквы (далее Слов), сначала вычисляется вес самого частотного Слова W_{max_word} в документе кластера:

$$W_{max_word} = \min (1, W_{max_conc} * (Fr_{max_word} / Fr_{max_conc}))$$

где W_{max_conc} — максимальный вес понятия тезауруса в тематическом представлении, Fr_{max_conc} — частотность в тексте понятия тезауруса с максимальным весом, Fr_{max_word} — частотность самого частотного Слова.

Остальные веса Слов (W_{word}) вычисляются пропорционально их частотности:

$$W_{word} = W_{max_word} * (Fr_{word} / Fr_{max_word})$$

Просмотр предложений-кандидатов начинается с начала документа кластера, то есть предложения набираются сначала из главного документа кластера и наиболее близких к нему по содержанию. Каждое следующее предложение берется из другого документа.

Для кластера примера была получена следующая аннотация (в скобках указан источник новости и время публикации):

Предложения	Тематические узлы
Китай и Тайвань установили авиасообщение после 60-летнего перерыва (Новые Известия — лента новостей, 04.07.2008 11:08:45)	<u>КИТАЙ</u> , <u>ТАЙВАНЬ</u> , <u>АВИАЦИОННЫЕ ПЕРЕВОЗКИ</u> (авиасообщение)
Первый чартерный рейс с 250 пассажирами уже прибыл в столицу Тайваня из китайского города Гуанчжоу. (Lenta.ru — главные новости, 04.07.2008 9:47:25)	<u>КИТАЙ</u> , <u>ЧАРТЕРНЫЕ ПЕРЕВОЗКИ</u> (чартерный рейс), <u>ГОРОД</u> , <u>ПАССАЖИР</u>
С 4 июля самолеты с материкового Китая на остров Тайвань и обратно будут летать каждую неделю с пятницы по понедельник. (РезKURSCITY.RU — Курс, 04.07.2008 9:35:34)	<u>КИТАЙ</u> , <u>ТАЙВАНЬ</u> , <u>АВИАЦИОННЫЕ ПЕРЕВОЗКИ</u> (самолет), <u>ОСТРОВ</u>
Перед прибывающими в ближайшие выходные 600 туристами из Китая будет расстилаться красная ковровая дорожка. (BBCRussian.com (Главная), 04.07.2008 1:18:25)	<u>КИТАЙ</u> , <u>ТУРИСТ</u>
По завершении в 1949 году гражданской войны в Китае и изгнания правительства Гом-Инь-Дана на Тайвань, отношения между двумя сторонами Тайваньского пролива были заморожены. (РезЛПГАБизнесИнформ — Украины — Новости за рубежом, 04.07.2008 9:14:00)	<u>КИТАЙ</u> , <u>ТАЙВАНЬ</u> , <u>ПРАВИТЕЛЬСТВО</u>

Каждое предложение аннотации содержит не менее двух понятий из разных основных тематических узлов, как минимум один из которых новый (выделен подчеркиванием в правом столбце таблицы), а другие были упомянуты ранее.

6. Оценка качества аннотаций новостных кластеров

Оценка качества аннотаций может быть проведена вручную или с помощью автоматических процедур [7].

При ручной оценке относительно аннотации могут быть заданы такие вопросы с оценкой по 5 бальной шкале:

- является ли предложения аннотации грамматически правильными,
- является ли текст аннотации связным,
- содержит ли аннотация все основные обсуждаемые темы исходного документа (документов) и др.

Однако, проведение ручной оценки весьма трудоемко и требует привлечения независимых экспертов.

Для автоматической оценки качества аннотаций может использоваться метод ROUGE (Recall Oriented Understudy for Gisting Evaluation), который подсчитывает число перекрытия (n-граммы слов) автоматической аннотации с «идеальными» аннотациями, составленными людьми [9].

6.1. Оценка качества аннотаций отдельных документов

Качество описываемой технологии автоматического аннотирования англоязычных текстов тестировалось на конференции SUMMAC (Summarization conference), организованной DARPA в 1998 году. Программа использовала английский перевод Общественно-политического тезауруса.

Суть соревнования заключалась в следующем: каждый участник соревнования получил на две недели 1000 документов и представил две аннотации — аннотацию наилучшей длины (то есть система сама определяла длину аннотации) и 10-процентную аннотацию, т.е. аннотацию, составляющую 10 процентов длины исходного текста.

Описываемая технология имела лучший показатель для аннотаций наилучшей длины и показатели на 10-процентную аннотацию были лучше, чем средние [13].

6.2. Оценка качества аннотаций новостных кластеров

В качестве метрики для оценки мы использовали метрики ROUGE-1-cir и ROUGE-2-cir, которые вычисляли следующим образом:

$$ROUGE - N - cir(A_i) = \frac{\sum_{M_{ij}} count(Ngram(A_i) \cap Ngram(M_{ij}))}{\sum_{M_{ij}} count(Ngram(M_{ij}))},$$

где A_i — оцениваемая обзорная аннотация i -того кластера, M_{ij} — ручные аннотации i -того кластера,

$Ngram(D)$ — множество всех n -грамм из лемм соответствующего документа D . При сравнении отдельных документов в расчет берутся только уникальные n -граммы, присутствующие в обоих документах — не поощряется многократный повтор одного и того же предложения. При рассмотрении нескольких аннотаций, наоборот, повторение одинаковых элементов поощряется. Биграмммы в наших оценках учитывались с перестановками.

Для оценки качества построенных аннотаций мы воспользовались данными, любезно предоставленными С.Д. Тарасовым (Военмех, Спб.). В проведенных С.Д. Тарасовым экспериментах группе студентов было предложено построить ручную аннотацию для новостных кластеров, которые брались из системы Google.Новости в период с 01 по 05 декабря 2008 года. Ручная аннотация должна была быть составлена из четырех предложений. Ограничений на выбор предложений из разных текстов не накладывалось.

Мы выбрали достаточно случайным образом из полученных данных 15 новостных кластеров разной тематики, включая новости о погоде, спорте, финансах и политике, для которых имелось от 18 до 40 ручных аннотаций (всего 462).

В качестве «базовой оценки», следуя [7], мы рассматривали следующие варианты искусственных аннотаций:

- первый документ кластера;
- заголовки первых четырех документов;
- первые предложения первых четырех документов;
- последний документ кластера.

В качестве автоматической аннотации рассматривались аннотации из заголовка и трех предложений, взятых из разных текстов.

Мы получили следующие результаты (в таблице приведены результаты для разных параметров ядра кластера — см.раздел 5):

Следует отметить, что некоторые ручные аннотации совпадали с первым или последним документом кластера. Определенным недостатком исполь-

зуемых данных является то, что некоторые кластеры содержали документы за несколько дней, поэтому ручные аннотации чаще содержали предложения из последних документов кластера.

Существует определенная критика использования метрик ROUGE для оценки качества аннотирования. Метрика чувствительна к длинам сравниваемых документов, не учитывает связность аннотаций. В целом, существует большое разнообразие между ручными аннотациями разных экспертов. В нашем случае нам лишь важно было оценить близость построенных автоматических и ручных аннотаций для оценки перспективности предложенного подхода.

Заключение

В статье представлена технология автоматического построения аннотации для новостного кластера, которая использует свойство лексической связности текста и знания о лексических значениях, которые описаны в Общественно-политическом тезаурусе, для обеспечения связности, полноты и снижения повторов аннотации кластера.

Описанный метод построения обзорных рефератов позволяет в широких пределах варьировать представление кластера при сохранении уровня отображения содержания и связности. Можно задавать как количество документов (исходящих ссылок), отражаемых в аннотации, так и количество предложений из каждого документа. В частности, могут быть смоделированы аннотации, формируемые в ресурсе Яндекс.Новости (до трех-четырёх документов по одному-два предложения), или аннотации, формируемые в ресурсе Google.Новости (три-четыре предложения из одного документа и два заголовка из других документов), или Рамблер.Новости (три предложения из одного документа и два-три предложения из других документов).

Вид аннотации	ROUGE-1-cir	ROUGE-2-cir
первый документ кластера	0,219	0,083
заголовки первых четырех документов	0,162	0,056
первые предложения первых 4 документов	0,269	0,107
последний документ кластера	0,278	0,168
автоматическая аннотация с ядром 0,20	0,331	0,150
автоматическая аннотация с ядром 0,40	0,328	0,140

Литература

1. *Абрамова Н. Н., Абрамов В. Е.* Автоматическое составление обзорных рефератов новостных сюжетов // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2007. Переславль-Залесский: 2007.
2. *Добров Б. В., Лукашевич Н. В.* Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ // Третья Всероссийская конференция по Электронным Библиотекам «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Петрозаводск: 2001. С. 78–82.
3. *Лукашевич Н. В.* Автоматическое построение аннотаций на основе тематического представления текста // Труды международного семинара Диалог'97. М.: 1997 — С. 188–191.
4. *Лукашевич Н. В., Добров Б. В.* Исследования тематической структуры текста на основе большого лингвистического ресурса // Труды международного семинара «Диалог 2000». М.: 2000. Том 2, С.252–258.
5. *Barzilay R., Elhadad M.* Using Lexical Chains for Text Summarization // ACL/ EACL Workshop Intelligent Scalable Text Summarization. Madrid: 1997.
6. *Carbonell J., Goldstein J.* The use of MMR, diversity-based reranking for reordering documents and producing summaries // In Proceedings of the 21st Annual International ACM SIGIR Conference. 1998. pp. 335–336.
7. *Dang H. T.* National Institute of Standards and Technology (NIST) «Overview of DUC 2006» // In Proceedings of DUC 2006.
8. *Doran W., Stokes N., Dunnion J., Carthy J.* Comparing Lexical Chain-based Summarisation Approaches using an Extrinsic Evaluation // In the Proceedings of the Global WordNet Conference(GWC 2004). Brno: 2004.
9. *Lin Chin-Yew.* ROUGE: a Package for Automatic Evaluation of Summaries // In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). Barcelona: 2004.
10. *Radev D., Jing H., Budzikowska M.* Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies // In ANLP/NACCL Workshop on Summarization. Seattle: 2000.
11. *Radev D., McKeown K., Hovy E.* Introduction to the Special Issue on Summarization // Computational linguistics. 2002. 28 (4). P.399–408.
12. *Silber G., McCoy K.* Efficiently computed lexical chains as an intermediate representation for automatic text summarization // Computational Linguistics. 2002. 28 (4). P.487–496.
13. *Tipster* SUMMAC Text Summarization Evaluation. Final report. — MITRE Technical report MTR 98000138. — October, 1998.