Автоматическое порождение обновления к аннотации новостного кластера

© А.А. Алексеев[1], Н.В. Лукашевич[2]

[1]Факультет вычислительной математики и кибернетики МГУ имени М.В. Ломоносова

[2]Научно-исследовательский вычислительный центр МГУ имени М.В. Ломоносова

a.a.alekseevv@gmail.com, louk_nat@mail.ru

Аннотация

Представлен метод автоматического составления обновления к аннотации новостного кластера. Данный метод основан на выявлении предложений, содержащих новую информацию, и дальнейшем автоматическом аннотировании на основе тематического представления кластера, моделирования лексической связности текста и тезаурусном описании лексических значений. Проведен эксперимент по оценке порожденных аннотаций при помощи выявления информационных единиц, так называемая пирамидная оценка аннотаций (Pyramid Evaluation).

1 Введение

Новостные сервисы в современном мире собирают новостные сообщения от многих тысяч новостных источников. В таких сервисах поступающие новостные сообщения обычно объединяются в кластеры похожих сообщений, для которых создается аннотация – краткий обзор документов кластера. Эта краткая аннотация является одним из основных продуктов работы новостного сервиса для конечного пользователя, который может быстро ознакомиться с основными фактами интересующего его события. Краткую аннотацию совокупности тематически близких документов также называют обзорным рефератом.

При пополнении новостных кластеров новой информацией соответствующая аннотация должна перестраиваться для того, чтобы отразить эту новую информацию. Другим способом отражения новой информации о происходящем событии является создание отдельного обновления к существующей аннотации, сообщающего только новую информацию по сравнению с уже аннотированными документами.

Задача создания обновления к аннотации новостного кластера формулируется следующим образом: пусть имеется новостной кластер, содержащий W новостных сообщений, и через некоторое время данный кластер пополняет ещё N сообщений. Необходимо составить обзорный реферат пришедших N сообщений таким образом, чтобы он содержал только новые факты по отношению к имеющимся W документам кластера. Задача ставится в предположении, что пользователь уже знаком с первыми W документами кластера и хочет ознакомиться только с новыми фактами по данному событию.

В данной работе будет рассмотрен метод порождения обновлений к аннотациям развивающихся новостных кластеров. Он базируется на выявлении предложений, содержащих новую информацию, и дальнейшем автоматическом аннотировании на основе тематического представления кластера, моделирования лексической связности текста и тезаурусном описании лексических значений. Для оценки качества предложенного метода реализован также известный метод аннотирования MMR (Maximum Marginal Relevance [4, 6]).

Сравнение методов производится так называемым методом Пирамид, который позволяет объективно оценить полноту изложения информации новостного кластера [9]. Также построенные аннотации были оценены профессиональным лингвистом с точки зрения читабельности и связности текста.

2 Обзор существующих подходов

Важной подзадачей поставленной задачи создания обновлённой аннотации новостного кластера является задача определения новизны информации. Подобная задача была поставлена в 2002 – 2004 годах на конференции по информационному поиску TREC (Text REtrieval Conference), проводимой Национальным Институтом Стандартов и Технологий США (NIST) [13]. Задача носила название Novelty Track и ставилась следующим образом: даны некоторый топик и набор документов (возможно, содержащий нерелевантные документы). Необходимо указать предложения, релевантные и содержащие новую информацию по отношению к данному топику.

Участники использовали большое количество различных подходов к решению поставленной задачи. Один из лучших результатов показала система Колумбийского университета [12], в которой ис-

Труды 12ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

пользовалось два подхода для определения новизны, а именно:

- представление предложения в виде вектора слов и последующее сравнение предложений посредством скалярного произведения;
- ранжирование предложений в соответствии с весом новых слов, встретившихся в этих предложениях.

Задача создания обновления к аннотации впервые была поставлена на конференции DUC (Document Understanding Conference), и её рассмотрение продолжилось (продолжается и сейчас) в рамках конференции TAC (Text Analysis Conference). Задача носит название Update Summarization и представляет собой аннотирование по запросу [8]. Участникам даются некоторый запрос и два набора документов, релевантных данному запросу и упорядоченных в хронологическом порядке (документы второго набора строго более поздние, чем документы первого). Необходимо составить две аннотации по данному запросу длиной не более 100 слов, такие, что первая является обычной аннотацией первого набора документов, а вторая является аннотацией второго набора документов, но в предположении, что пользователь уже знаком с документами из первого набора (аннотация должна содержать только новые факты).

Самым популярным подходом для решения задачи создания обновлённой аннотации стал подход, аналогичный обычному аннотированию документов. В данном случае предложение-кандидат для аннотации не должно быть похоже не только на предложения, уже отобранные в аннотацию, но и на документы из первого набора. Для улучшения результатов своих систем участники использовали различные техники, среди них:

- вычисление расстояний между термами с учётом их места в иерархии WordNet;
- учёт частей речи и обработка именованных сущностей;
- изучение позиции предложения в документе с точки зрения его важности;
- исключение слишком коротких и слишком длинных предложений из рассмотрения, а также предложений, содержащих кавычки, анафору и т. д. [7].

Для отбора предложений в аннотацию использовалось два основных подхода: кластеризация и ранжирование. Первый подход заключается в разделении всех предложений на кластеры похожих предложений и отборе в итоговую аннотацию центральных предложений получившихся кластеров. Второй подход основан на вычислении веса каждого предложения с использованием некоторой метрики и отбора в итоговую аннотацию предложения, имеющего наибольший вес.

Лучшие результаты практически по всем метрикам показал метод Maximal Marginal Relevance (MMR). Данный метод давно известен и успешно используется для запросного аннотирования [6], в работе [5] представлена модификация данного метода для задачи создания обновления к аннотации. Метод MMR является итеративным, на каждом шаге производится ранжирование предложений так, чтобы предложение, во-первых, было как можно ближе к запросу для аннотирования по некоторой метрике, а во-вторых, как можно дальше от предложений, уже отобранных в аннотацию, и от предложений из первого набора документов.

3 Предложенный подход

Сформулируем более подробно поставленную задачу. Дан новостной кластер – набор новостных сообщений по поводу некоторого события, упорядоченный в хронологическом порядке. Новостной кластер представляет собой динамическую структуру, которая постоянно пополняется новыми сообщениями. Такой новостной кластер делится на две части по хронологическому принципу – до и после некоторого момента времени.

Необходимо составить две аннотации: первая представляет собой классическую аннотацию набора документов первой части кластера; вторая аннотация должна отражать факты из второй части кластера, являющиеся новыми по сравнению с первой частью кластера.

Приведём пример. 2 февраля 2007 года в 16.00 начался полуфинал мужского теннисного турнира в Загребе, в котором сыграли россиянин Михаил Южный и хорват Иван Любичич. Все новостные сообщения новостного кластера, описывающего данное событие, которые пришли до 16.00 часов, содержат только информацию о том, что данный полуфинал должен будет состояться, и не содержат информацию о результатах данной встречи.

Таким образом, разделив новостной кластер на две части по временному принципу (до и после 16.00), мы получим «хорошие» входные данные для поставленной задачи. Обновлённая аннотация, как минимум, должна будет содержать информацию о результатах игры (победитель, счёт и т. д.), так как эта информация, во-первых, является новой по отношению к первой части кластера и, во-вторых, самой важной информацией второй части кластера.

Для решения задачи составления обновлённой аннотации разделённого новостного кластера предлагается подход, основанный на выявлении предложений второй части кластера, содержащих новую информацию по отношению к первой части кластера, и дальнейшем аннотировании с использованием только новых предложений.

3.1 Определение новизны предложений

Для выявления предложений, содержащих новую информацию, используется комбинация методов, предложенных командой Колумбийского университета на конференции TREC 2004 [12]. Для того чтобы предложение было сочтено новым, необходимо, чтобы оно было сочтено новым двумя следующими подходами.

При первом подходе все предложения новостного кластера представляются в виде вектора лемм. Исследуемое предложение второй части кластера сравнивается со всеми предложениями из первой части по косинусной мере угла между векторами. Если все получившиеся значения (лежащие в диапазоне от 0 до 1) меньше некоторого порога, вычисляемого эмпирически, то исследуемое предложение объявляется новым. В качестве порога использовалось значение 0.5.

В рамках второго подхода критерием наличия новой информации в предложении является наличие в нём новых слов. Разные слова вносят разный вклад в новизну предложения, данный вес определяется на основе данных о частотных характеристиках слов в новостных коллекциях и вычисляется по формуле

$$Вес\_слова = \frac{1}{\log(df_{set})},$$

где df_{set} – частота употребления слова в новостной коллекции. Частота слов, которые не встречались в коллекции, считается равной 10. Для каждого предложения второй части кластера вычисляется некоторый вес – сумма весов слов, входящих в данное предложение и не входящих во множество слов первой части кластера, то есть слов, являющихся новыми. Предложения, набравшие вес больше некоторого эмпирического порога, объявляются новыми. Наилучшие результаты достигаются при величине порога 0.3.

3.2 Аннотирование на основе тематических узлов

В качестве базового метода аннотирования используется метод, основанный на тезаурусных знаниях и тематическом представлении новостного кластера [1, 2].

Построение тематического представления документа состоит в разбиении всех понятий, упомянутых в документе, на группы близких по смыслу понятий – тематических узлов. Для этой процедуры используются описания понятий в Тезаурусе русского языка РуТез.

Для автоматического построения тематического представления текст обрабатывается морфологическим анализатором и сопоставляется с Тезаурусом. В результате сопоставления синонимы сводятся к одним и тем же понятиям Тезауруса. Для многозначных слов производится автоматическая процедура разрешения многозначности, в результате которой разные значения слов соотносятся с разными понятиями Тезауруса.

Для идентифицированных в тексте понятий из Тезауруса извлекаются приписанные этим понятиям взаимные отношения, и в итоге выявляется сеть понятий, которая необходима для интерпретации текста. Эта сеть понятий автоматически разбивается на совокупности близких по смыслу понятий – тематические узлы.

Тематические узлы в тематическом представлении разделяются на две категории: основные тематические узлы, которые соответствуют сущностям из основной темы документа, а также локальные тематические узлы, соответствующие побочным темам документа. Для построения тематического представления новостного кластера такой кластер объединяется в единый документ [2].

Аннотирование новостных кластеров на основе тематического представления и тезаурусных знаний позволяет снижать значимость известных проблем построения обзорных рефератов, таких, как:

- обеспечение полноты представления информации;
- снижение повторов при представлении информации;
- обеспечение связности и понятности представляемой информации.

Полнота изложения содержания кластера обеспечивается тем, что для аннотации отбираются предложения, содержащие пары основных тематических узлов, – именно тогда эти предложения будут описывать взаимоотношения между основными тематическими элементами кластера.

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

Для решения проблем обеспечения полноты, снижения повторов, обеспечения связности используются не только повторы слов текстов, но и разнообразная информация о синонимах, родовидовых и других типах отношений слов.

Приведём примеры тематических узлов, созданных в процессе обработки описанного выше новостного кластера про полуфинал теннисного турнира в Загребе (главное понятие тематического узла выделено сдвигом влево; указана частота употребления понятия в тексте):

ТЕННИСНЫЙ КОРТ	14
ТЕННИС	12
АВСТРИЙЦЫ	12
АВСТРИЯ	6
КИПРИОТЫ	16
КИПР	11
ХОРВАТЫ	10
СЕТ (ПАРТИЯ В ТЕННИСЕ)	6
ИГРОВАЯ ПАРТИЯ	5
ЧЕТВЕРТЬФИНАЛ	10
МАТЧ	12
ПОЛУФИНАЛ	29
ПОЛУФИНАЛИСТ	2
СПОРТИВНЫЙ ФИНАЛ	36
СПОРТИВНОЕ СОРЕВНОВАНИЕ	54
СПОРТ	8
СПОРТСМЕН	2
ФИНАЛИСТ	1
ЮЖНЫЙ, МИХАИЛ	23
РОССИЯНЕ	12
РОССИЙСКАЯ ФЕДЕРАЦИЯ	10
ТЕННИСИСТ	6

ЗАГРЕБ	70
ХОРВАТИЯ	36
ТЕННИСНЫЙ КОРТ	14
ТЕННИС	12
АВСТРИЙЦЫ	12
АВСТРИЯ	6
КИПРИОТЫ	16
КИПР	11
ХОРВАТЫ	10
СЕТ (ПАРТИЯ В ТЕННИСЕ)	6
ИГРОВАЯ ПАРТИЯ	5
ЧЕТВЕРТЬФИНАЛ	10
МАТЧ	12
ПОЛУФИНАЛ	29
ПОЛУФИНАЛИСТ	2
СПОРТИВНЫЙ ФИНАЛ	36
СПОРТИВНОЕ СОРЕВНОВАНИЕ	54
СПОРТ	8
СПОРТСМЕН	2
ФИНАЛИСТ	1
ЮЖНЫЙ, МИХАИЛ	23
РОССИЯНЕ	12
РОССИЙСКАЯ ФЕДЕРАЦИЯ	10
ТЕННИСИСТ	6
ЗАГРЕБ	70
ХОРВАТИЯ	36

После построения тематического представления документа аннотация новостного кластера строится следующим образом. Аннотация должна состоять из заголовка и нескольких предложений из разных документов новостного кластера.

При отборе заголовка для аннотации выбирается один из заголовков документов кластера, имеющий наибольший вес по тематическим узлам и именованным сущностям (см. далее). 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;
- 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, совпадающих со словами предшествующих предложений, не должно превышать некоторой доли длины предложения.

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 [1], а также весов содержащихся в предложении слов с большой буквы, не считая первого слова предложения. Для вычисления весов слов с большой буквы сначала вычисляется вес самого частотного Слова W_{max_word} в документе кластера:

$$W_{\max\_word} = \min\left(1, W_{\max\_conc} * \left(Fr_{\max\_word} / Fr_{\max\_conc}\right)\right),$$

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 с большой буквы. Остальные веса слов с большой буквы (W_{word}) вычисляются пропорционально их частотности:

$$W_{word} = W_{\max\_word} * \left(Fr_{word} / Fr_{\max\_word}\right).$$

3.3 Реализация метода

Предложенный подход для решения задачи создания обновлённой аннотации состоит из трёх этапов.

На первом этапе происходит предварительная обработка документов кластера. Текст кластера разбивается на слова, выделяются границы предложений, производится лемматизация слов. Полученное морфологическое представление документа сопоставляется с Тезаурусом и строится тематическое представление кластера. В данной задаче тематическое представление строится для кластера в целом, отдельно – для первой части кластера, и в одном из вариантов метода – для второй части кластера.

В рамках второго этапа производится выделение предложений, содержащих новую информацию. Предложение считается новым, если оно сочтено новым обоими методами обнаружения новой информации – при помощи векторно-пространственной модели и обнаружения новой информации по новым словам (см. п. 3.1).

На заключительном, третьем этапе происходит формирование обновлённой аннотации. Аннотирование производится описанным выше методом на основе тематического представления текста, но с дополнительным условием – предложение, отбираемое в аннотацию, должно быть признано содержащим новую информацию в рамках второго этапа.

Применительно к задаче создания обновлённой аннотации исследованы два варианта аннотирования на основе тематического представления текста с использованием тематических узлов:

- по всему новостному кластеру;
- только по второй части новостного кластера.

Каждая аннотация для начальной части кластера и для второй части кластера должна содержать не более 100 слов.

Приведём пример начальной аннотации и обновления к аннотации новостного кластера про теннисный турнир, описанный в начале данного раздела. Аннотации порождены автоматически программой, реализующей предложенный подход.

Начальная аннотация:

За выход в финал турнира в Загребе Михаил Южный поспорит с Иваном Любичичем. ТЕННИС – БОЛЬШОЙ Загреб (Хорватия).
1.Четвертьфинал Любичич (Хорватия, 1) – Юханссон (Швеция) – 7 : 6 (10 : 8), 6 : 7 (8 : 10), 7 : 6 (7 : 4) Багдатис (Кипр, 2) – Клеман (Франция, 8) 6 : 2, 6 : 7 (2 : 7), 7 : 6 (7 : 5).
2. Вслед за россиянином Михаилом Южным и австрийцем Александром Пейя в *1/2 финала пробились хорват Иван Любичич и киприот Маркос Багдатис.*
3. В субботу на турнире ATP в Загребе пройдут полуфинальные встречи.
4. Михаил Южный вышел в полуфинал турнира в Хорватии, обыграв в двух сетах француза Марка Жискеля.
5. Самым серьезным соперником в Загребе станет для Южного хозяин корта Иван Любичич, посеянный под первым номером.
6. Во втором полуфинальном матче встретится киприот Маркос Багдатис с австрийцем Александром Пийя, пробившимся из квалификации.

Обновление к аннотации:

1. Южный уступил дорогу в финал любимцу публики, ОРЕАНДА / SPORT. RU.
2. Михаил Южный не смог выйти в финал турнира АТП в Загребе, проиграв прошлогоднему победителю турнира и первому сеяному хорвату Ивану Любичичу.
3. Турнирный путь Южного: 1-й круг: Николя Маю (Франция) – 7:6, 6:3; 2-й круг: Томаш Чакль (Чехия) – 6:4, 6:2. Четвертьфинал: Марк Гиквел (Франция) – 7:5, 6:2. Полуфинал: Иван Любичич (Хорватия, 1) – 6:2, 3:6, 5:7.
4. В финале хорватский теннисист сыграет с киприотом Маркосом Багдатисом, выбившим из борьбы австрийца Александра Пейю – 6:4, 6:3.
5. Полуфинал Маркос Багдатис (Кипр, 2) – Александр Пейя (Австрия, Q) – 6:4.
6. Первый сет Михаил взял довольно легко, во втором проиграл, а в третьем вел со счетом 4:2, однако уступил.

Отметим, что аннотация содержит только новую информацию – о результатах игры, а не о том, что игра должна была состояться (первая часть кластера практически полностью посвящена анонсу игры), хотя во второй части кластера содержится достаточное количество новостных сообщений, содержащих устаревшую, по сравнению с первой частью кластера, информацию. Это связано с тем, что сообщения от некоторых новостных агентств поступают со значительным запозданием.

4 Оценка предложенного подхода

Оценка качества порождаемых аннотаций является достаточно сложной процедурой. Несомненно, наиболее правдоподобные оценки можно получить при помощи ручной оценки путём привлечения большого количества экспертов. Но данный метод является очень дорогим и трудоёмким.

В работе [10] предложен алгоритм автоматической оценки автоматических аннотаций – ROUGE. Данный метод основан на автоматическом сравнении автоматически порожденной аннотации с эталонной аннотацией, созданной экспертом. Существуют различные модификации алгоритма, связанные с различными способами сравнения:

- сравнение *n*-грамм (ROUGE-N);
- сравнение максимальных общих подстрок (ROUGE-L и ROUGE-W);
- сравнение пропусков монограмм и биграмм (ROUGE-S и ROUGE-SU).

В работе [3] предложена модификация алгоритма оценки русскоязычных аннотаций – ROUGE-RUS. Метод учитывает русскую морфологию, стоп-слова, а также синонимы (с учётом концептов Тезауруса).

Автоматические ROUGE-метрики позволяют быстро и с малыми трудозатратами производить оценку автоматических аннотаций. Однако оценки, полученные данным алгоритмом, зачастую сильно отличаются от человеческих оценок. Это связано в первую очередь с тем, что человек может использовать разные формы одних и тех же слов, синтаксические структуры, различный порядок слов и парафразы для описания одних и тех же событий. Автоматический учёт всех приведённых сложностей является на данный момент нерешённой задачей. Рассмотренный далее метод оценки автоматических аннотаций при помощи пирамид нивелирует описанные выше проблемы существующих методов оценки.

4.1 Метод пирамид

Метод пирамидной оценки автоматических аннотаций предложен группой Колумбийского университета [9] и успешно применяется при масштабной оценке конкурсных систем аннотирования [11]. Данный метод заключается в выделении из нескольких эталонных аннотаций так называемых информационных единиц (Summary Content Unit – SCU). Информационная единица представляет собой некоторый факт, который отражает эталонная аннотация и соответственно должна отражать порождаемая аннотация. Приведём пример информационной единицы и её вхождений в различные тексты новостного кластера:

SCU: *Мини-субмарина попала в ловушку под водой*.

1. *мини-субмарина... была затоплена... на дне моря...*
2. *маленькая... субмарина... затоплена... на глубине 625 футов.*

3. *мини-субмарина попала в ловушку... ниже уровня моря.*
4. *маленькая... субмарина... затоплена... на дне морском...*

Каждая информационная единица получает вес, равный количеству эталонных аннотаций оцениваемого кластера, где она встречается, то есть для оценки автоматической аннотации необходимо составить несколько эталонных аннотаций (на конференции DUC их было 4) и выделить из них информационные единицы, дифференцировав при этом различные информационные единицы по весу – количеству эталонных аннотаций, где они встречаются. Все найденные информационные единицы образуют пирамиду. На верхних уровнях обычно расположено небольшое количество самых «весомых» информационных единиц. На нижних уровнях – большое количество менее важных информационных единиц. Оценка автоматической аннотации состоит в выявлении в ней найденных информационных единиц и получением итоговой оценки по следующей формуле:

$$Оценка\_аннотации = \frac{[Find\_SCU\_Weight]}{[Sum\_SCU\_Weight]},$$

где Find_SCU_Weight – суммарный вес всех найденных информационных единиц, Sum_SCU_Weight – суммарный вес всех информационных единиц, определённых для данного топика. Данная оценка показывает, какую часть от общей массы информационных единиц отражает автоматическая аннотация, с учётом веса различных информационных единиц.

Метод оценки автоматических аннотаций при помощи пирамид позволяет оценить полноту отражения информации в аннотации вне зависимости от использованных в документах синонимов и перефраз.

4.2 Метод Maximal Marginal Relevance (MMR)

Для сравнения результатов предложенного подхода к созданию автоматических аннотаций реализован альтернативный метод, предложенный командой канадского университета Монреаль [5]. Данный метод представляет собой модификацию классического метода MMR [6] для создания обновлений к аннотациям. Алгоритм представляет собой итеративный метод, на каждом этапе происходит ранжирование предложений по некоторой оценке. Данная оценка вычисляется таким образом, чтобы предложение-кандидат, с одной стороны, как можно лучше подходило к запросу для аннотирования (в случае обзорного реферирования – всему объединенному документу кластера) и как можно меньше пересекалось с предложениями, уже отобранными в аннотацию (и также с предложениями из первой части кластера, применительно к задаче обновления аннотации). Классическая формула метода MMR выглядит следующим образом:

$$MMR = \arg\max_{s \in S} \left[\lambda \cdot Sim_1(s,Q) - (1-\lambda) \cdot \max_{s_j \in S} Sim_2(s,s_j) \right],$$

где Q – запрос к системе, S – множество предложений кандидатов, s – рассматриваемое предложение-кандидат, E – множество выбранных предложений. Формула, предложенная канадским университетом применительно к задаче создания обновлённой аннотации, следующая:

$$S_{MMR}(s) = \underset{s \in S}{Sim_1(s,Q)} \cdot \left(1 - \max_{s_h \in H} Sim_2(s,s_h) \right)^{f(H)},$$

где H – документы первой части кластера (документы, с которыми пользователь уже знаком), $f(H) \to 0$ при увеличении H. В качестве метрик сравнения предложений использовались: $Sim_1(s,Q)$ – стандартная косинусная мера угла между векторами; $Sim_2(s,s_j)$ – максимальная общая подстрока (Longest Common Substring – LCS).

Описанный метод MMR для создания обновлённой аннотации показал один из лучших результатов на конференции TAC'2008 [8] по всем оцениваемым метрикам, поэтому и был выбран в качестве альтернативного метода.

В нашем случае задача состояла в общем аннотировании документа. Как известно, аннотирование по запросу можно свести к задаче общего аннотирования путём использования всего набора документов в качестве запроса.

Также проведён эксперимент по использованию метода MMR совместно с описанными выше методами обнаружения новой информации. Для аннотации отбираются только те предложения, которые признаны новыми по векторной модели и методу обнаружения новой информации по новым словам. Использование данных методов позволило заметно улучшить результаты метода MMR (более подробно см. в следующем разделе).

4.3 Эксперименты

Таким образом, для оценки качества аннотирования мы имеем два основных метода: метод, основанный на тематическом представлении кластера, и метод MMR.

Каждый метод имеет две модификации для построения обновления к аннотации. Модификации метода, основанного на тематическом представлении, различаются использованным тематическим представлением. В первом случае используется тематическое представление всего кластера, во втором случае – тематическое представление только второй части кластера.

Метод MMR тестируется в версии работы [5]. Второй вариант метода MMR состоит в том, что для построения обновления к аннотации используются только предложения, сочтенные новыми по обоим условиям, описанным в п. 3.1.

Все описанные выше методы были оценены методом пирамид. В качестве базы для оценки были использованы ручные аннотации (2 – 4 на кластер),

созданные двумя профессиональными лингвистами. Из этих аннотаций вручную были выделены информационные единицы, для каждой информационной единицы вычислен её вес, равный количеству эталонных аннотаций, где она встречается.

Приведем примеры информационных единиц для кластера примера:

1. *PBZ Zagreb Indoors – турнир АТП в столице Хорватии Загребе с призовым фондом 416 миллионов долларов.*
2. *Иван Любичич – хозяин кортов (Хорват).*
3. *Южный сыграет с Любичичем в полуфинале PBZ Zagreb Indoors.*
4. *Южный проиграл Любичичу в полуфинале PBZ Zagreb Indoors.*
5. *В финале Иван Любичич встретится с Маркосом Багдатисом.*
6. *Маркос Багдатис посеян под вторым номером.*
7. *Маркос Багдатис из Кипра.*

и др.

Далее каждая автоматическая аннотация получала оценку в соответствии с числом и весом содержащихся в ней эталонных информационных единиц (см. п. 4.1). Отдельно оценивались начальная аннотация и обновление к аннотации. При оценке обновления к аннотации учитывались только информационные единицы, которые являются новыми по отношению к первой части кластера. Табл. 1 содержит результаты оценки аннотаций методом пирамид.

Таблица 1. Результаты тестирования аннотаций методом Пирамид

Название метода	Начальная аннотация	Обновление к аннотации
MMR	0.643	0.457
MMR+новизна	0.643	0.543
Тематическое представление (по всему кластеру)	0.638	0.630
Тематическое представление (по второй части кластера)	0.638	0.587

Результаты показывают, что предложенный метод создания обновлённой аннотации новостного кластера при помощи тематического представления текста с применением методов определения новой информации приблизительно на 30% лучше (по данной метрике) одного из лучших методов, существующих на данный момент. Также стоит отметить, что дополнение метода MMR методами обнаружения новой информации позволило заметно улучшить результаты данного метода. На начальной аннотации оба основных метода оказались практически эквивалентными.

Для того чтобы определить верхнюю границу качества аннотаций по методу пирамид, мы выполнили оценку ручных аннотаций относительно набора выделенных информационных единиц. В среднем ручные аннотации получили оценку 0.781 по методу Пирамид. Таким образом, по полноте изложения информации автоматические аннотации достигают уровня более 80% от полноты ручных аннотаций, что представляется достаточно высокой величиной.

Для конечного пользователя важными являются не только полнота или новизна предоставляемой информации, но и качество ее изложения.

Для того чтобы понять, каковы «читабельность» и связность представленных аннотаций, мы попросили оценить качество аннотаций профессионального лингвиста, которая не имела информации о том, результаты работы каких методов она тестирует.

Тестирование производилось следующим образом. Лингвист должна была читать каждый вид аннотации, и для всех предложений аннотации, которые казались как бы не на своем месте (не связанными, лишними), начислять штрафной балл. В случае сомнений начислялась половина балла. Таким образом, каждый вид аннотации получил некоторую совокупность штрафных очков (см. табл. 2).

В таблице можно видеть, что на текущий момент «человеческое» качество обновления к аннотации значительно ниже, чем начальной аннотации. Метод MMR+новизна, который получил лучшие оценки по изложению новой информации, оказался хуже по «читабельности» по сравнению с базовым методом.

Табл. 2 Средние штрафные баллы, начисленные каждому виду аннотации за нарушение законов связного изложения

Название метода	Начальная аннотация	Обновление к аннотации
MMR	0.591	1.409
MMR+новизна	0.591	1.900
Тематическое представление (по всему кластеру)	0.318	1.227
Тематическое представление (по второй части кластера)	0.318	1.182

Аннотации, построенные на основе тематического представления, показались лингвисту в среднем заметно более качественными, чем аннотации, построенные по методу MMR.

Хотелось бы еще отметить, что обеспечение связности и читабельности аннотации требует некоторой степени повтора информации в предложениях, что ограничивает возможность повышения полноты изложения в аннотации фиксированной длины. Это частично продемонстрировано в проведенном эксперименте (см., например, соотношение полноты и читабельности в обновлении к аннотации по методу MMR+новизна).

Заключение

В работе мы описали постановку задачи создания обновления к аннотации развивающегося новостного кластера и предложили метод создания такого обновления, основанного на оценке новизны предложений и использовании тематического представления новостного кластера. Для сравнения качества предложенного метода был реализован известный метод аннотирования MMR.

Были проведены два вида оценок качества получаемых аннотаций: на основе метода Пирамид и ручные оценки на качество изложения аннотации.

Показано, что метод создания обновлений в аннотации, основанный на тематическом представлении, существенно лучше отражает новизну фактов. Кроме того, аннотации, порождаемые на основе этого метода, обладают лучшими читабельностью и связностью.

Литература

[1] Добров Б.В., Лукашевич Н.В. Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ// Труды 3ей Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2001. – Петрозаводск, Россия, 2001. С. 78-82. – http://rcdl.ru/doc/2001/dobrov.pdf.

[2] Лукашевич Н.В., Добров Б.В. Автоматическое аннотирование новостного кластера на основе тематического представления// Компьютерная лингвистика и интеллектуальные технологии: труды Межд. конф. Диалог'2009. – Москва, РГГУ, 2009. – С. 299-305. – http://www.dialog-21.ru/dialog2009/materials/html/46.htm.

[3] Тарасов С.Д. Исследование и оптимизация параметров алгоритма Manifold Ranking на основе метрики автоматической оценки качества обзорного реферирования ROUGE-RUS// Труды 11ой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2009. – Петрозаводск, Россия, 2009. – С. 86-93. – http://rcdl.ru/doc/2009/086_093_DIIS-seminar-1-2009-3.pdf.

[4] Boudin F., El-Bize M., Torres-Moreno J.-M. A scalable MMR approach to sentence scoring for multi-document update summarization// Proc. of the 22nd Int. Conf. on Computational Linguistics, Posters and Demonstrations. – Coling, 2008. – P. 23-26. – http://www.aclweb.org/anthology-new/C/C08/C08-2006.pdf.

[5] Boudin F., El-Beze M., Torres-Moreno J.-M. The LIA update summarization systems at TAC-2008// Proc. of the Text Analyze Conference'2008. – Gaithersburg, Maryland USA, 2008. – http://www.nist.gov/tac/publications/2008/participant.papers/LIA.proceedings.pdf.

[6] Carbonell J., Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries// Proc. of the 21st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. – Melbourne, Australia, 1998. – P. 335-336. – http://www.cs.cmu.edu/~jgc/publication/The_Use_MMR_Diversity_Based_LTMIR_1998.pdf.

[7] Dang H.T. Overview of DUC 2006// Proc. of the Document Understanding Conferences'2006. – New York, USA, 2006. – http://duc.nist.gov/pubs/2006papers/duc2006.pdf.

[8] Dang H.T., Owczarzak K. Overview of the TAC 2008 update summarization task// Proc. of the Text Analyze Conference'2008. – Gaithersburg, Maryland USA, 2008. – http://www.nist.gov/tac/ publications/ 2008/additional.papers/update_summ_overview08.proceedings.pdf.

[9] Harnly A., Nenkova A., Passonneau R., Rambow O. Automation of summary evaluation by the pyramid method// Proc. of the Int. Conf. on Recent Advances in Natural Language Processing (RANLP'2005). – Borovets, Bulgaria, 2005. – http://www.cs.columbia.edu/~ani/papers/aabo-ranlp.pdf.

[10] Lin C.-Y. ROUGE: a package for automatic evaluation of summaries// Proc. of the Workshop on Text Summarization Branches Out (ACL'2004). – Barcelona, Spain, 2004. – P. 74-81. – http://acl.ldc.upenn.edu/acl2004/textsummarization/pdf/Lin.pdf.

[11] Passonneau R.J., Nenkova A., McKeown K.R., Sigelman S. Applying the pyramid method in DUC 2005// Proc. of the Document Understanding Conferences'2005. – Vancouver, Canada, 2005. – http://duc.nist.gov/pubs/2005papers/columbiau.passonneau2.pdf.

[12] Schiffman B., McKeown K.R. Columbia University in the novelty track at TREC 2004// Proc. of the Thirteenth Text Retrieval Conf. (TREC'2004). – Gaithersburg, USA, 2004. – http://trec.nist.gov/pubs/trec13/papers/columbiau.novelty.pdf.

[13] Soboroff I. Overview of the TREC 2004 novelty track// Proc. of the Thirteenth Text Retrieval Conf. (TREC'2004). – Gaithersburg, USA, 2004. – http://trec.nist.gov/pubs/trec13/papers/NOVELTY.OVERVIEW.pdf.

Automatic generation of update summaries for news clusters

A.A. Alekseev, N.V. Loukachevitch

In this paper we introduce an approach to automatic generation of update summaries for news clusters. This method is based on the identification of sentences containing new information. Further automatic summarization exploits the thematic representation of a news cluster, lexical cohesion modeling and thesaurus description of lexical senses. The update summaries were evaluated using the Summary Content Units technique – Pyramid Evaluation.