Московский государственный университет имени М.В. Ломоносова

На правах рукописи

Алексеев Алексей Александрович

Метод автоматического аннотирования новостных кластеров на основе тематического анализа

05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата физико-математических наук

Научный руководитель
доктор физ.-мат. наук
профессор М.Г. Мальковский

Москва – 2014

Оглавление

Введение

Развитие информационных технологий и появление сети Интернет явились причиной экспоненциального роста объемов электронной информации, начавшегося приблизительно два десятилетия назад и стремительно продолжающегося в настоящее время. Объемы информации уже сейчас достигли таких размеров, что человек не способен самостоятельно ознакомиться с материалами всех информационных источников, зачастую даже в контексте специализированных информационных потребностей. Данный факт обусловил активное развитие исследований в области задачи автоматического аннотирования – представления релевантной и наиболее значимой информации, необходимой пользователю, в сжатом, лаконичном виде.

Методы автоматического аннотирования исследовались в трудах российских и зарубежных ученых, таких как Барзилай Р., Добров Б.В., Лукашевич Н.В., Лун Х., МакКьюин К., Мальковский М.Г., Мани И., Машечкин И.В., Ненкова А., Петровский М.И., Севбо И.П., Тарасов С.Д., Шиффман Б., Эдмундсон Х. и многих других авторов. Спектр областей применения систем автоматического аннотирования является обширным и разнородным, от бытовых информационных потребностей обычных пользователей, до узкоспециализированных аналитических задач. Например, в рамках программы SUMMAC (TIPSTER Text Summarization Evaluation) [43] рассматривалась задача оценки релевантности текстового документа некоторой тематике. Данное исследование предполагало два варианта принятия решения экспертом:

- на основании прочтения всего исходного документа;
- на основании прочтения аннотации исходного документа.

Было установлено, что системы автоматического аннотирования позволяют лучше решать данную задачу - аннотации с максимальной длиной в 17% от исходного документа в два раза уменьшают время принятия аналитиком

решения, без статистически значимого ухудшения точности данного решения.

Подготовка обзорных рефератов для коллекции документов уже давно является одним из ключевых элементов в организации и представлении результатов поиска, основной задачей которого является снижение его общего времени. В работе [46] представлено исследование, в рамках которого пользователям была поставлена задача написания отчетов на фиксированные темы, с использованием наборов новостных документов, которые содержали как релевантные данным темам документы, так и нерелевантные. Установлено, что предоставление пользователям результатов автоматической кластеризации документов по необходимым тематикам и автоматических аннотаций сформированных кластеров позволяет улучшить общее качество отчетов пользователей, а также сократить время подготовки данных отчетов.

Автоматические аннотации также применяются для решения более сложных и комплексных задач, чем задача определения релевантности документов некоторой тематике. Например, при анализе научных статей помимо задачи отбора полезных данных для прочтения (описанная задача определения релевантности документов), перед пользователем также стоит задача определения взаимосвязи с предшествующими работами в исследуемой области, на которые ссылается анализируемая научная статья. Системы автоматического аннотирования могут помочь определить основные идеи и направления, которые подвергаются критике и, напротив, поддерживаются и развиваются в рамках текущей работы.

Системы автоматического аннотирования находят применение и в узкоспециализированных областях. Аннотирование голосовых сообщений может быть полезно для быстрого установления приоритета звонка, номера или имени собеседника; аннотации форумных веток обсуждений позволяют устанавливать значимость и интенсивность обсуждения интересующей темы;

подготовка аннотаций совещаний и встреч может быть полезна для быстрого ознакомления новых участников с результатами прошлых сессий и так далее.

Потенциальный спектр областей применения систем автоматического аннотирования уже сейчас является чрезвычайно широким и продолжает расти, вместе с развитием систем искусственного интеллекта, компьютерной лингвистики и систем автоматической обработки информации в целом. При этом различные задачи и области применения обладают своими особенностями и спецификой, что влечет за собой необходимость разработки индивидуальных решений и алгоритмов для конкретных задач и областей.

Современные технологии автоматической обработки новостных потоков основаны на тематической кластеризации новостных сообщений, т.е. выделении совокупностей новостей, посвященных одному и тому же событию – ***новостных кластеров*** [78]. Одной из важных и актуальных специализированных задач аннотирования является ***автоматическое аннотирование новостных кластеров***. Новостной кластер и методы автоматического аннотирования новостных кластеров являются основными объектами рассмотрения данной кандидатской диссертации, в рамках которой будет предложен метод выявления скрытой информации, заложенной внутри структуры новостного кластера, а также методы применения данной информации для улучшения методов автоматического аннотирования новостных кластеров.

Кластер документов должен соответствовать ситуации или совокупности связанных ситуаций (обладать основной темой кластера, [5], [78]). В описываемой ситуации есть набор участников, которые в исходном кластере:

- Могут быть выражены не только словами, но и словосочетаниями,
- Могут выражаться не одним, а совокупностью различных выражений. Так, акции некоторой компании могут выражаться в текстах одного новостного кластера, как собственно акции

компании, контрольный пакет акций, контрольный пакет, акционер компании, владелец компании, состав владельцев и др.

Например, международный аэропорт «Внуково», расположенный в Москве, может упоминаться в рамках некоторого новостного кластера как *московский международный аэропорт Внуково*, *московский аэропорт*, *столичный аэропорт*, *аэропорт Внуково*, *международный аэропорт* и так далее.

Можно предположить, что качественное выделение участников ситуации, включая различные варианты их наименования в различных документах кластера, может помочь лучше определять основную тему новостного кластера, и, таким образом, позволит повысить качество различных операций с новостными кластерами, таких как автоматическое аннотирование, определение новизны информации и других автоматических операций.

В данной работе предлагается модель представления содержания новостного кластера, описывающая основных участников ситуации с учетом вариативности их именования – тематических цепочек новостного кластера. Рассматриваются методы улучшения качества извлечения основных участников новостного события, что включает нахождение совокупности слов и выражений, с помощью которых тот или иной значимый участник события именовался в документах новостного кластера. Предлагаемый подход основан на совместном использовании совокупности факторов, в том числе разного рода контекстов употребления слов в документах кластера, информации из предопределенных источников (тезаурус русского языка), а также особенностях построения текстов на естественном языке.

Цель диссертационной работы

Целью данной диссертационной работы является разработка методов и программных средств построения модели основных участников новостного кластера с учетом вариативности их именования на основе комбинации разнородных факторов схожести, и интеграция построенной модели в

методы автоматического аннотирования. Разрабатываемые программные средства и полученная модель должны удовлетворять следующим требованиям: высокая точность выявления различных вариантов именования основных участников; возможность интеграции построенной модели в другие задачи автоматической обработки текста; независимость от предметной области.

Для достижения этой цели были решены следующие задачи:

1. исследование и построение модели основных участников новостного кластера с учетом вариативности их именования и специфики внутреннего устройства текстов на естественном языке;
2. разработка методов интеграции построенной модели в методы автоматического аннотирования, а также разработка двух новых метод на основе построенной модели;
3. разработка и реализация программного модуля для построения тематических цепочек новостного кластера;
4. разработка и реализация программного модуля автоматического аннотирования новостного кластера, реализующего методы аннотирования на базе построенных тематических цепочек.

Основные положения, выносимые на защиту:

1. Предложен и реализован новый метод автоматического построения модели основных участников новостного кластера (тематических цепочек), основанный на комбинировании разнородных признаков схожести;
2. Предложен метод применения построенной модели в существующих методах автоматического аннотирования;
3. На основе построенной модели предложены и реализованы два новых метода автоматического аннотирования;
4. Показано улучшение качества работы алгоритмов аннотирования на основе тематических цепочек;

Научная новизна

Новизна настоящей диссертационной работы заключается в том, что предложен новый метод построения модели совокупности участников новостного кластера, основанный на комбинации признаков различной природы: как статистических контекстных, так и априорных. Применимость данного метода обоснована теоретически, на основе анализа полезности отдельных признаков для определения близости языковых выражений, а также численно, на основе экспериментов по интеграции в методы автоматического аннотирования. Разработанная модель не зависит от предметной области и может применяться в различных задачах автоматической обработки новостных кластеров.

Практическая значимость

На основе предложенного алгоритма спроектирована и реализована многомодульная программная система со следующими функциональными возможностями:

- построения тематических цепочек новостного кластера;
- автоматическое формирование аннотаций новостного кластера различными алгоритмами аннотирования;
- автоматическая оценка конкурсных аннотаций (требуются экспертные аннотации для проведения оценки).

Таким образом, разработанная система может быть использована как для подготовки дополнительной входной информации для других систем автоматической обработки новостных кластеров, так и для формирования автоматических аннотаций новостного кластера несколькими различными алгоритмами.

Апробация работы. Основные результаты работы докладывались на следующих конференциях и семинарах:

- всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Казань, 13-17 октября 2010 г.);

- международной конференции «Математика. Компьютер. Образование» (Дубна, 25-30 января 2010 г.);
- семинаре по поиску концептов в неструктурированной информации (CDUD), проходящему совместно с конференцией RSFDGrC (Москва, 25-30 июня 2011 г.);
- международной конференции «Системный анализ и семиотическое моделирование» (Казань, 24-27 февраля 2011 г.);
- международной конференции «Диалог» (Московская область, 25-29 мая 2011 г.);
- летней школе по информационному поиску RUSSIR (Ярославль, 6-10 августа 2012 г.);
- международной конференции «Spring Researchers Colloquium on Databases and Information Systems» (Москва, 1 июня 2012 г.);
- всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Ярославль, 14-17 октября 2013 г.);

Кроме того результаты обсуждались на семинаре лаборатории анализа информационных ресурсов НИВЦ МГУ, на семинаре в НИУ ВШЭ и на регулярном семинаре ACM SIGMOD в Москве.

Публикации. Основные результаты по теме диссертации изложены в 14 печатных работах, в том числе 3 статьях в журналах из списка ВАК ([68], [72], [74]), 3 статьях, входящих в базу SCOPUS ([1], [3], [4]), 3 – в тезисах докладов ([66], [67], [75]) и 5 в других изданиях ([2], [69], [70], [71], [73]).

Все основные положения, выносимые на защиту, опубликованы в статье [68] журнала, входящего в список ВАК.

Объем и структура данных. Диссертация состоит из введения, четырех глав, заключения и двух приложений. Полный объем диссертации составляет 122 страницы с 15 рисунками и 7 таблицами, объем приложений – 9 страниц. Список литературы содержит 82 наименования.

1. Автоматическое аннотирование

Данная глава посвящена описанию задачи автоматического аннотирования, классификации типов аннотаций и областей применимости систем автоматического аннотирования. Также в данной главе приводится обзор алгоритмов построения автоматических аннотаций, базовых идей, моделей и принципов их построения, а также методов оценки качества и сравнения результатов работы различных систем автоматического аннотирования. Особое внимание уделяется задаче и алгоритмам построения обзорных рефератов (см. Раздел 1.2.1), так как основным объектом исследования данной кандидатской диссертации является новостной кластер. Целью данной главы является анализ достоинств и недостатков существующих методов автоматического аннотирования, а также проблем в данной области, в частности обоснование важности учета лексико-семантической вариативности, широко присутствующей в текстах на естественном языке.

1.1 Задача автоматического аннотирования

Задача автоматического аннотирования – создание краткой версии некоторого текстового документа или коллекции документов, отражающей наиболее значимую информацию исходного документа или документов ([40]). Традиционно в задаче автоматического аннотирования выделяют несколько независимых направлений классификации решаемых задач и типов порождаемых аннотаций ([48], [49], [55]).

Экстрактивные аннотации (*Extractive summaries*) создаются при помощи конкатенации предложений входных текстов документов, без изменения самих предложений. *Аннотации в форме абстракта* (*Abstractive summaries*), напротив, являются авторскими и формируются независимо от текстов исходных документов, хотя могут пере использовать их слова и выражения.

Большинство первых работ по автоматическому аннотированию были посвящены *аннотированию одного документа* (*Single-document summarization*), то есть в качестве входных данных выступает единственный документ, такой как новостное сообщение, научная статья, лекция или т.п. Позже, с развитием исследований в области автоматического аннотирования, а также возникновения большого числа новых источников информации и увеличения информационных потоков в целом, возник новый тип задачи автоматического аннотирования: *подготовка обзорного реферата для коллекции документов* (*Multi-document summarization*). Данный тип аннотирования наиболее востребован при обработке большого количества текстовых документов, связанных некоторой сюжетной линией, темой или каким-либо другим параметром. Особую актуальность данному типу автоматического аннотирования придает развитие сети Интернет, содержащей огромное количество различных текстовых документов. Первые онлайн-системы многодокументного аннотирования применялись в задачах обработки потоков новостей, а именно формирования аннотаций для новостных кластеров [45]. Данная задача сохранила свою актуальность и решается в крупных коммерческих новостных агрегаторах, таких как Rambler.News, Yandex.News, Google.News и других.

Автоматические аннотации также различают по типу содержания. Аннотация, передающая информацию об общем содержании документа, но не сообщающая деталей, называется *индикативной аннотацией* (*indicative summary*). *Информативная аннотация* (*informative summary*), напротив, может быть прочитана вместо исходных документов, то есть должна сохранять информационную ценность входной текстовой коллекции.

Большинство исследований в области автоматического аннотирования связано с подготовкой краткой аннотации, приблизительный размер которой – один абзац текста. Вместе с тем специфичные приложения и/или потребности пользователей приводят к таким задачам, как *аннотирование ключевыми словами* (*keyword summarization*), требующей выделения наиболее

значимых и индикативных слов исходного документа (документов), а также *аннотирование предложениями* (*headline summarization*) – выделение наиболее важного предложения входной текстовой коллекции.

Потребности пользователя в информации формируют ещё одну плоскость для классификации типов автоматических аннотаций. К текущему моменту времени большая работа проделана в области *общего аннотирований* (*generic summarization*), задачей которого является предоставление всеобъемлющей аннотации, охватывающей весь объем информации, содержащейся в исходном документе (документах). Данный тип аннотирования отвечает на вопрос «О чем этот документ (эти документы)?» и должен позволить пользователю быстро войти в тематику входной текстовой коллекции, в идеале полностью избавив пользователя от необходимости ознакамливаться с самими исходными документами.

В отличие от общего аннотирования, задачей *аннотирования по запросу* (*query-focused summarization*) является подготовка аннотации, содержащей наиболее значимую информацию в соответствии с некоторым пользовательским запросом. Данный тип аннотирования отвечает на вопрос «Что в этом документе (этих документах) говорится о <*запрос*>?». Например, в задаче информационного поиска пользовательский запрос превращается поисковой системой в результирующий набор документов, краткая аннотация каждого из которых в результатах выдачи может помочь пользователю быстрее определить релевантность каждого из них. Для подготовки полезной аннотации в данном случае системе автоматического аннотирования необходимо учитывать также запрос пользователя, как дополнение к исходным текстовым документам (самодостаточных в случае общего аннотирования).

Задача подготовки обновленных аннотаций (*update summarization*) покрывает ещё одну возможную информационную потребность пользователя. Это вариация много-документного аннотирования, которая является чувствительной ко времени: обновленная аннотация должна

передавать наиболее важные факты развития интересующего пользователя сюжета, исключая информацию уже известную пользователю (информацию, с которой пользователь уже ознакомлен).

Общая классификация типов автоматических аннотаций может быть представлена следующим образом:

1. **По принципу составления ([48], [49], [55]):**
    - Экстрактивные аннотации (Extractive summaries)
    - Аннотации в форме абстракта (Abstractive summaries)
2. **По типу входной коллекции:**
    - Аннотирование одного документа (Single-document summarization)
    - Формирование обзорного реферата – аннотации набора документов (Multi-document summarization)
3. **По типу содержания:**
    - Индикативные аннотации (Indicative summaries)
    - Информативные аннотации (Informative summaries)
4. **По размеру аннотации:**
    - Аннотации в виде фрагмента текста (Common summarization)
    - Аннотирование ключевыми словами (Keyword summarization)
    - Аннотирование предложениями (Headline summarization)
5. **По потребности пользователя:**
    - Общее аннотирование (Generic summarization)
    - Аннотирование по запросу (Query-focused summarization)
    - Подготовка обновленных аннотаций (Update summarization)

Необходимо отметить, что подавляющее число современных систем автоматического аннотирования работает на основе экстрактивного подхода ([41]), т.е. выбора целых предложений исходной коллекции для автоматической аннотации.

1.2 Методы автоматического аннотирования

В разделе 1.1 представлено описание различных направлений и подзадач задачи автоматического аннотирования. Наиболее популярным и широко востребованным является направление подготовки общих и запрос-ориентированных экстрактивных аннотаций для коллекции документов (задача подготовки обзорных рефератов). Данная задача особенно актуальна в контексте анализа новостного потока и обработки новостных кластеров. Это направление выбрано в качестве основного в рамках данной кандидатской диссертации, в связи с чем дальнейший обзор методов автоматического аннотирования будет посвящен методам подготовки обзорных рефератов (общих и запрос-ориентированных).

1.2.1 Общая классификация методов

В настоящее время выделяют пять основных классов методов для решения задачи экстрактивного аннотирования ([40], [24]):

I. Использование частотных характеристик слов: аннотирование на основании ключевых слов – topic words (без применения обучения). Более подробная информация о данном классе методов приведена в разделе 1.2.2;

II. Построение тематических моделей текстов. Данная категория методов включает в себя как методы, использующие некоторые предопределенные ресурсы, так и подходы, основанные на вероятностных моделях. Более подробная информация о данном классе методов приведена в разделе 1.2.3;

III. Методы, основанные на графах (без применения обучения). Суть данного направления заключается в адаптации известных алгоритмов на графах для определения центральных и наиболее значимых предложений входной коллекции, и для решения тем самым задачи автоматического аннотирования (см. Раздел 1.2.4);

IV. Подходы, основанные на машинном обучении (machine learning). Данное направление методов автоматического аннотирования

базируется на использовании ручных экспертных аннотаций для предсказания значимости предложений. Более подробное описание методов данного направления находится в Разделе 1.2.5;

V. Стратегии подготовки аннотации. Выделяется две основных стратегии, основанных на ***локальной оптимизации*** (см. раздел 1.2.6.1), объединяющей в себе «жадные» алгоритмы последовательного отбора предложений на основании локальной информации, и алгоритмы ***глобальной оптимизации*** (см. раздел 1.2.6.2), которые осуществляют отбор предложений исходя из качества результирующей аннотации в целом.

1.2.2 Методы, основанные на частотных характеристиках слов

Данный класс методов автоматического аннотирования объединяет в себе широкий спектр подходов, которые имеют значительное количество отличий, но при этом несут единую базу – выделение и использование ***ключевых слов*** (descriptive words) для формирования результирующих аннотаций ([48], [49]).

1.2.2.1 Частоты и вероятности слов

Использование частотности для определения значимости слов было предложено Луном в одной из первых работ по автоматическому аннотированию ([40]). Чем чаще слово употребляется в текстовой коллекции, тем более значимым для данной коллекции оно является. Первым шагом является кластеризация всех слов текстовой коллекции на два класса: описательные (значимые) слова и слова не являющиеся ключевыми. При этом из потенциального списка значимых слов исключаются:

- Стоп-слова - предлоги, союзы и так далее;
- Слова, являющиеся широко употребляемыми в рамках рассматриваемой предметной области (например, слово *клетка* в контексте текстов по биологии);

- Слова с низкой частотностью в рамках рассматриваемой текстовой коллекции.

Следующим шагом эволюции автоматического аннотирования на основе ключевых слов стал уход от жесткого бинарного разбиения слов на «ключевые» и «неключевые» - переход к весам слов. В рамках данной модели каждое слово имеет некоторый вещественный вес, характеризующий значимость данного слова для рассматриваемой коллекции. Наиболее популярными моделями назначения весов являются вероятность слова и $\mathrm{TF}\cdot\mathrm{IDF}$. При этом результаты систем автоматического аннотирования на основании вещественных весов слов могут значительно отличаться в зависимости от выбора конкретных мер схожести ([50]).

Вероятность слова является простейшим вариантом использования частоты для определения значимости слова ([63]). Она вычисляется как отношения количества вхождений слова к общему количеству слов в документе или коллекции документов. Данная система весов является основой метода автоматического аннотирования SumBasic ([51], [62], [63]), который отбирает предложения для аннотации на основании средней вероятности слов, которые в него входят. Сам алгоритм состоит из пяти шагов. На первом шаге происходит расчет вероятностей слов исходного кластера $p(w_i)$ по следующей формуле:

$$\mathrm{p}(\mathrm{w_i}) = \frac{\mathrm{n}}{\mathrm{N}}$$

где n – число появлений слова w_i в исходной коллекции, N – общее число слов в данной коллекции. Каждому предложению s_j на втором шаге назначается вес, равный средней вероятности слов в данном предложении:

$$weight\ (s_j) = \sum w_i \in s_j \frac{p(w_i)}{|\ \{ w \mid w \in s_j \}\ |}$$

На третьем шаге предложение с наибольшим весом отбирается в итоговую аннотацию. После этого на шаге 4 происходит пересчет вероятностей всех слов, входящих в отобранное предложение, по следующей формуле:

$$p_{new}(w_i) = p_{old}(w_i) \cdot p_{old}(w_i)$$

На пятом шаге проверяется общая длина получившейся аннотации, и если она не превосходит заданного порога, то происходит переход к шагу 2.

Узким местом использования модели вероятностей слов является работа с общеупотребимыми словами. Данная проблема обычно решается использованием списков стоп-слов, но, очевидно, подобное решение не является универсальным. ***Система весов*** TF·IDF (**T**erm **F**requency*Inversed **D**ocument **F**requency) предлагает более гибкий вариант модели весов слов, основанной на использовании дополнительного корпуса для выявления общезначимых слов ([58]). Обычно в качестве подобного корпуса выступает большая коллекция документов той же тематики, что и рассматриваемая входная коллекция. Расчет TF·IDF происходит по следующей формуле:

$$TF \cdot IDF_w = c(w) \cdot \log\left(D \big/ d(w) \right)$$

где *c(w)* – частота слова w в рассматриваемой коллекции, *d(w)* – число документов фоновой коллекции, где встретилось слово *w* и *D* – размер фоновой коллекции. Соответственно, ключевыми словами (словами, которые получают высокие веса) в данной модели являются те слова, которые часто встречаются в рассматриваемой коллекции и редко в фоновой. Данная модель относительно проста для расчета и в том или ином виде используется в большинстве существующих систем автоматического аннотирования ([25], [14], [12]).

1.2.3 Тематические модели для автоматического аннотирования

1.2.3.1 Лексические цепочки (Lexical Chains)

Модели выделения значимой информации на основе пословного представления, такие как вероятность слова и TF·IDF , содержат принципиальную проблему: объекты могут описываться в текстовых коллекциях не только одним словом, но и наборами связанных слов и выражений. Формирование ***лексических цепочек*** ([35], [38], [6], [7], [26])

является одним из вариантов решения данной проблемы. Лексическая цепочка представляет собой совокупность языковых выражений текста, близких по смыслу. Для построения лексических цепочек используются лингвистические ресурсы, называемые тезаурусами. Тезаурус в современной лингвистике - это особая разновидность словарей общей или специальной лексики, в которых указаны семантические отношения (синонимы, антонимы, паронимы, гипонимы, гиперонимы и т. п.) между лексическими единицами. Наиболее популярными ресурсами для построения лексических цепочек являются тезаурусы Wordnet [6] для английского языка и РуТез ([39], [80]) для русского языка.

В работе [6] утверждается, что применение лексических цепочек в задаче автоматического аннотирования может играть важную роль для определения наиболее значимых и обсуждаемых тематик, по сравнению с частотами отдельных слов, а также предлагается алгоритм построения аннотации на основании лексических цепочек. Данный алгоритм является итеративным, в рамках каждой итерации отбирается по одному предложению, содержащему упоминание узлового элемента наиболее значимой лексической цепочки.

Другое направление автоматического аннотирования на основе лексических цепочек связано с отбором предложений не по наиболее значимым лексическим цепочкам, а по их отношениям ([80]). Критерием включения предложения в результирующую аннотацию в данном случае является наличие двух (или более) наиболее значимых лексических цепочек - обсуждение отношений между участниками ситуации, моделируемыми данными цепочками. Данный алгоритм автоматического аннотирования на основе тематического представления по тезаурусу РуТез показал лучшие результаты на первом крупномасштабном независимом тестировании методов автоматического аннотирования The TIPSTER Text Summarization Evaluation (SUMMAC, [42]):

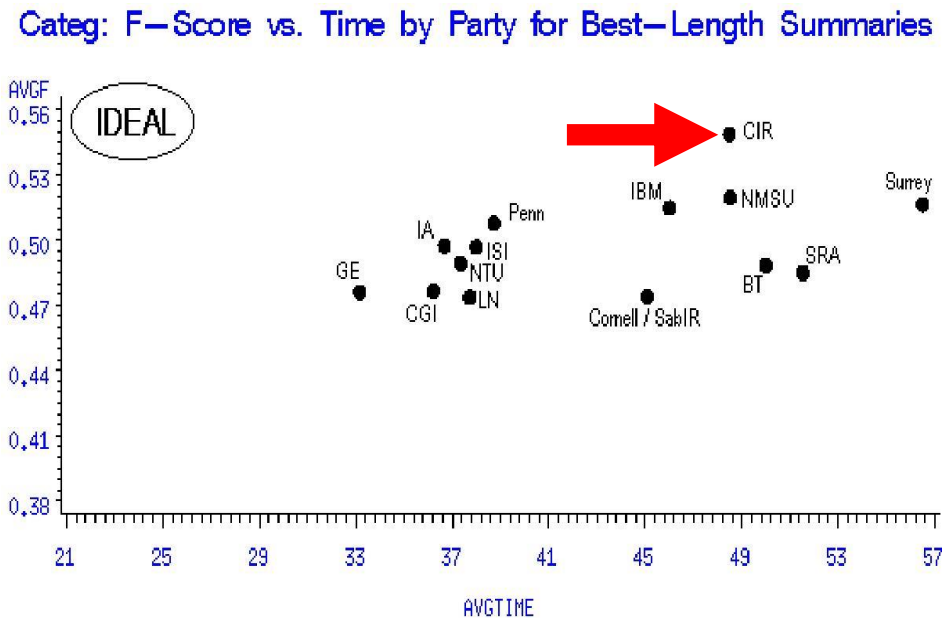


Рис. 1: Результаты SUMMAC`1998

1.2.3.2 Латентный семантический анализ (Latent Semantic Analysis, LSA)

Латентный семантический анализ ([21]) представляет собой алгоритм неявного представления семантических особенностей текста на основании статистики совместной встречаемости слов. В работе [28] предложены варианты применения результатов работы алгоритма LSA для задач аннотирования одного документа и построения обзорных рефератов коллекций документов. Результат LSA используется для определения наиболее значимых топиков, не используя при этом вспомогательные лексические ресурсы (тезаурусы, словари и т.д.).

В основе алгоритма LSA лежит представление входной коллекции документов в виде матрицы *A*: строки матрицы соответствуют словам, встречающимся во входной коллекции, а столбцы – предложениям. Каждый элемент матрицы a_{ij} соответствует весу слова *i* в предложении *j*. Если слово отсутствует в соответствующем предложении, то значение элемента равно нулю; иначе вес слова равен весу $TF \cdot IDF$. После построения матрицы к ней применяется стандартное сингулярное разложение (**S**ingular **V**alue **D**ecomposition, SVD), в результате которого может быть получено следующее представление для матрицы *A*:

$$A = U\Sigma V^T$$

В работе [28] утверждается, что строки матрицы V^T могут рассматриваться как взаимно независимые тематики, обсуждаемые во входной коллекции документов, в то время как столбцы данной матрицы продолжают соответствовать отдельным предложениям. Предлагаемый алгоритм построения аннотации является итеративным, заключающийся в обходе всех строк матрицы V^T и отбора предложений с наибольшим весом, до тех пор пока не будет достигнут лимит на длину аннотации. Применение сингулярного разложения позволяет улучшить отбор предложений для аннотации по сравнению с обычными методами, учитывающими лишь совместное появление слов. При этом алгоритмы данного класса не значительно превосходят системы аннотирования на основе $\mathrm{TF}\cdot\mathrm{IDF}$.

1.2.3.3 Байесовские тематические модели (Bayesian Topic Models)

Одной из наиболее сложных и, в то же время, активно развивающихся моделей для задачи автоматического аннотирования является использование байесовских моделей текстов ([15], [30], [20]). В данных моделях коллекция документов представляется как выборка пар документ-слово *(d, w)*, $\mathrm{d}\in\mathrm{D}$ (множество документов), $\mathrm{w}\in W_d$ (множество слов документа *d*), позволяя документу или слову относить сразу к нескольким темам с различными вероятностями. Каждая тема $\mathrm{t}\in\mathrm{T}$ описывается неизвестным распределением $p(W\mid t)$ на множестве слов $\mathrm{w}\in\mathrm{W}$. Каждый документ $\mathrm{d}\in\mathrm{D}$ описывается неизвестным распределением $p(t\mid d)$ на множестве тем $\mathrm{t}\in\mathrm{T}$.

В модели *вероятностного латентно-семантического анализа* (probabilistic latent semantic analysis, PLSA, [34]) вероятность появления пары «документ-слово» записывается тремя эквивалентными способами:

$$\mathrm{p(d,w)}=\sum_{t\in T}p(t)\,p(w\mid t)\,p(d\mid t)=\sum_{t\in T}p(d)\,p(w\mid t)\,p(t\mid d)=\sum_{t\in T}p(w)\,p(t\mid w)\,p(d\mid t)$$

где *p(t)* - неизвестное априорное распределение тем во всей коллекции; *p(d)* - априорное распределение на множестве документов; *p(w)* - априорное

распределение на множестве слов. Искомые условные распределения *p(w|t)* и *p(t|d)* выражаются через *p(t|w)* и *p(d|t)* по формуле Байеса:

$$\mathrm{p(w\,|\,t)} = \frac{p(t\,|\,w)\,p(w)}{\sum_{w'} p(t\,|\,w')\,p(w')}; \qquad \mathrm{p(t\,|\,d)} = \frac{p(d\,|\,t)\,p(t)}{\sum_{t'} p(d\,|\,t')\,p(t')}$$

Идентификация параметров тематической модели по коллекции документов происходит с использованием принципа максимума правдоподобия, который приводит к задаче минимизации с ограничениями нормировки:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \log p(d,w) \to \min_{\Phi,\Theta}$$

$$\sum_{w} p(w\,|\,t) = 1; \qquad \sum_{t} p(t\,|\,d) = 1; \qquad \sum_{t} p(t) = 1$$

где n_{dw} – число вхождений слова w в документ d; $\Phi = \|p(w\,|\,t)\|$ и $\Theta = \|p(t\,|\,d)\|$ - искомые матрицы вероятностей. Для решения данной оптимизационной задачи обычно применяется EM-алгоритм ([34]).

Проблемами модели PLSA являются линейный рост числа параметров с ростом коллекции документов и невозможность применения имеющихся формул при добавлении новых документов в коллекцию. Основные недостатки модели PLSA устранены в модели *скрытого распределения Дирихле* (Latent Dirichlet Allocation, LDA, [10], [11]), который основан на той же вероятностной модели:

$$\mathrm{p(d,w)} = \sum_{t\in T} p(d)\,p(w\,|\,t)\,p(t\,|\,d)$$

но при этом также вводятся следующие дополнительные условия:

- Векторы документов $\Theta_d = (p(t\,|\,d) : t \in T)$ порождаются одним и тем же вероятностным распределением, в качестве которого берется распределение из параметрического семейства распределений Дирихле $Dir(\Theta, \alpha), \alpha \in \Re^{|T|}$
- Векторы тем $\phi_t = (p(w\,|\,t) : w \in W)$ порождаются одним и тем же вероятностным распределением на нормированных векторах

размерности $|W|$, которое также берется из параметрического семейства распределений Дирихле $Dir(\Theta, \beta), \beta \in \Re^{|W|}$

Наиболее популярным методом идентификации параметров модели LDA по коллекции документов является сэмплирование Гиббса ([29]).

Сформированные вероятностные распределения могут сравниваться различными мерами схожести, наиболее распространенной из которых является использование ***дивергенции (расстояния) Кульбака–Лейблера*** (Kullback–Leibler divergence). Дивергенция Кульбака–Лейблера идентифицирует несогласованности в вероятностях, соответствующих одинаковым событиям распределений. В контексте автоматического аннотирования под событиями понимаются вхождения слов. Общая формула дивергенции Кульбака-Лейблера вероятностного распределения Q по отношению к распределению P для слов w определяется следующим образом:

$$\mathrm{KL}(\mathrm{P} \parallel \mathrm{Q}) = \sum_{\mathrm{w}} \mathrm{P}(\mathrm{w}) \cdot \log \frac{\mathrm{P}(\mathrm{w})}{\mathrm{Q}(\mathrm{w})}$$

Ранжирование и отбор предложений в итоговую аннотацию на основе байсовских моделей происходит с помощью итеративных локально-оптимизационных алгоритмов ([30]). В рамках каждой из итераций отбирается предложение, добавление которого приведет к наибольшему уменьшению КЛ-дивергенции между вероятностным распределением слов коллекции для аннотирования и текущим вариантом аннотации.

1.2.4 Теория графов для построения автоматических аннотаций

Использование теории графов для задачи автоматического аннотирования близко идеям известного алгоритма ссылочного ранжирования PageRank ([53]). В рамках данного подхода входная коллекция текстов представляется в виде полносвязного графа, вершины которого являются предложениями, а ребра между ними отражают веса схожести между данными предложениями. Предложения, имеющие сильные связи с большим количеством других предложений, вероятно, являются

центральными и должны иметь высокую значимость для включения в результирующую аннотацию.

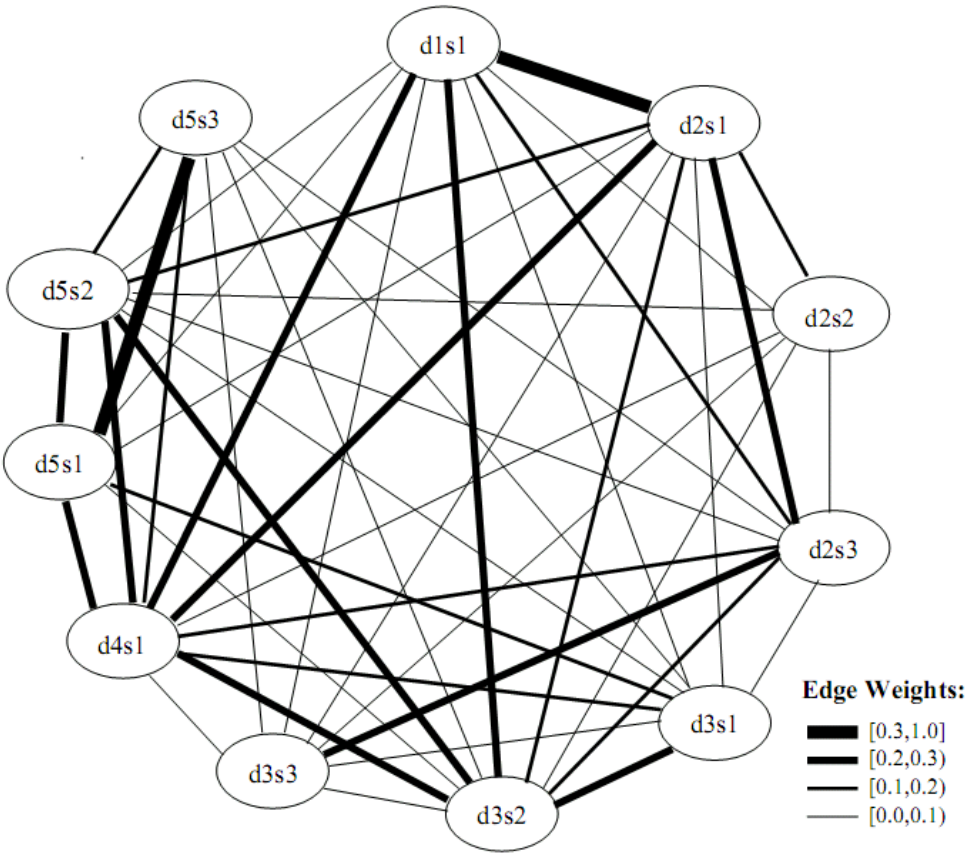


Рис. 2: Пример взвешенного графа сходства предложений

Веса ребер полученного графа могут быть нормализованы до вероятностного распределения. В этом случае исследуемый граф становится Марковской цепью, а веса ребер соответствуют вероятностям перехода из одного состояния в другое. На Рис. 2 приведен пример подобного взвешенного графа сходства предложений для некоторого кластера документов (d_i – номер документа, s_j – номер предложения в документе). На базе стандартных алгоритмов для случайных процессов могут быть получены вероятности нахождения в каждой из вершин графа. Вершины с высокой вероятностью соответствуют более значимым предложениям исходной коллекции и должны быть включены в результирующую аннотацию ([53]).

Описанный алгоритм не требует глубокой лингвистической предобработки, за исключением выделения предложений и отдельных слов, поэтому может применяться для различных языков. В то же время

добавление семантической и синтаксической информации при построении графа предложений может улучшить результаты работы алгоритма ([16]).

Дифференциация схем расчета схожести предложений принадлежащих одному и разным документам входной коллекции позволяет отделить локальные темы документов от глобальных тем, разделяемых всеми документами коллекции. В работе [64] подобная дифференциация проводится на базе графовых алгоритмов автоматического аннотирования.

1.2.5 Использование машинного обучения

Задача выбора значимых предложений в методах автоматического аннотирования на основе машинного обучения ([24], [65], [37]) представляется как задача бинарной классификации: разбиение всех предложений входной коллекции на принадлежащие и не принадлежащие итоговой аннотации. Экспертные аннотации используется для обучения статистического классификатора, который, в свою очередь, рассматривает каждое предложение как набор потенциальных характеристик для выявления значимости. Для каждого предложения s вероятность его классификации *P(s)* как предложения для аннотации *S* (данную вероятность также называют уверенностью классификатора) является весом данного предложения. При заданном наборе из *k* характеристик F_j*: j=1..k,* данная вероятность может быть представлена как:

$$P(s \in S \mid F_1, F_2, .., F_k) = \frac{P(F_1, F_2, .., F_k \mid s \in S) \cdot P(s \in S)}{P(F_1, F_2, .., F_k)}$$

С учетом статистической независимости характеристик формула может быть преобразована следующим образом:

$$P(s \in S \mid F_1, F_2, .., F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s \in S) \cdot P(s \in S)}{\prod_{j=1}^{k} P(F_j)}$$

где $P(s \in S)$ является константой, а $P(F_j \mid s \in S)$ и $P(F_j)$ могут быть рассчитаны на основе обучающей коллекции.

Выбранный классификатор играет роль оценочной функции, получая на вход промежуточное представление предложения и выдавая вес предложения в качестве результата. Предложения с наибольшими весами включаются в результирующую аннотацию, возможно, с дополнительным фильтром на схожесть с уже отобранными предложениями.

Широким является спектр применяемых характеристик для установления значимости предложений ([65], [37]), наиболее популярными из которых являются такие характеристики, как: позиция предложения в документе (например, первые предложения новостных сообщений являются обычно наиболее информативными); позиция в абзаце (первые и последние предложения обычно являются более важными); длина предложения, схожесть предложения с заголовком документа; наличие именованных сущностей или предопределенных ключевых фраз и другие.

Одной из основных проблем подходов, основанных на обучении, является необходимость подготовки размеченных данных для обучения классификатора. Привлечение экспертов для указания подходящих для аннотации предложений является дорогостоящей процедурой, и, кроме того, согласованность экспертной разметки является низкой, так как разные люди склонны выбирать разные предложения для включения в результирующую аннотацию.

В целом методы автоматического аннотирования на основе машинного обучения не показывают значительно лучших результатов в задаче составления обзорных рефератов, по сравнению с методами без обучения, основанными на единственной характеристике. При этом в задачах аннотирования одного документа и специализированного аннотирования, такого как аннотирования научных статей ([9]), машинное обучение показывает более высокие результаты ввиду возможности использования профильных характеристик значимости.

1.2.6 Стратегии отбора предложений при подготовке аннотаций

1.2.6.1 Локальная оптимизация

Последовательный отбор предложений на основании имеющейся локальной информации в рамках процедуры подготовки итоговой автоматической аннотации соответствует классу «жадных» алгоритмов или алгоритмов локальной оптимизации.

Метод Максимальной Граничной Значимости (**M**aximal **M**arginal **R**elevance, **MMR,** [14]) является наиболее популярной концепцией локальной оптимизации и представляет собой итеративный алгоритм отбора предложений для формирования автоматической аннотации. В рамках каждой итерации алгоритма происходит ранжирование предложений–кандидатов и отбор одного предложения, получившего максимальный вес. Основная идея метода MMR заключается в том, что лучшее предложение для аннотации должно быть максимально релевантно запросу для аннотирования (исходному корпусу документов в случае общего аннотирования) и максимально отлично от всех предложений, уже отобранных в итоговую аннотацию. Пусть:

- Q – запрос пользователя к системе автоматического аннотирования в случае запрос-ориентированного аннотирования / исходный корпус документов в случае общего аннотирования;
- S – множество предложений кандидатов;
- s – рассматриваемое предложение кандидат;
- E – множество предложений, уже отобранных в итоговую аннотацию.

Тогда на каждой итерации алгоритма в итоговую аннотацию будет отобрано предложение, удовлетворяющее следующему условию:

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

Итеративная процедура отбора предложений продолжается до тех пор, пока не будет достигнуто ограничение на общую длину аннотации. Для

обеспечения лучшей связности и читабельности предложения итоговой аннотации сортируются в соответствии с их порядком следования в исходных документах.

Существует ряд модификаций метод MMR для решения задач автоматического аннотирования, отличных от подготовки общих и запрос-ориентированных аннотаций. Например, в работах [12] и [13] предложена модификация алгоритма MMR для создания обновлённых аннотаций ([18]):

- *Q* – запрос пользователя к системе автоматического аннотирования;
- *s* – рассматриваемое предложение кандидат;
- *H* – рассмотренные документы (история документов, с которыми пользователь системы уже ознакомился);
- *f(H) –> 0* при увеличении *H*.

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

Возведение в степень *f(H)* необходимо для учёта информации об имеющихся знаниях пользователя - документов, которые пользователь уже прочитал, и является основным нововведением.

Одной из важных частей алгоритма MMR является оценка схожести предложений друг с другом (предложения и всего корпуса документов в случае общего аннотирования). При использовании различных вариантов алгоритма MMR для построения аннотаций (как классических, так и обновленных аннотаций) возможно использование различных мер схожести (в качестве функций Sim_1 и Sim_2), наиболее популярными из которых являются следующие меры схожести:

1. ***$Sim_1(s,Q)$:*** косинусная мера угла между векторами:

$$\cos(\Theta) = \frac{v1 \cdot v2}{\|v1\| \cdot \|v2\|}$$

2. ***$Sim_2(s,s_h)$:*** максимальная общая подпоследовательность (LCS)

$$Sim2(s, s_h) = \frac{2 \cdot Length(LCS(s, s_h))}{Length(s) + Length(s_h)}$$

Жадные алгоритмы подготовки аннотаций могут быть неэффективны при рассмотрении вопроса оптимальности порождаемых аннотаций в целом, так как включение локально-оптимальных предложений зачастую приводит к включению избыточной информации в итоговую аннотацию. Решение данной проблемы возможно с помощью алгоритмов глобальной оптимизации.

1.2.6.2 Глобальная оптимизация

Идея аннотирования на основе глобальной оптимизации заключается в построении аннотации исходя из качества результирующей аннотации в целом, а не информации о локальном качестве отдельных предложений. Автоматическая аннотация обладает рядом требований и ограничений, таких как максимизация информативности, минимизация повторов информации, ограничение на общую длину. Поиск точного решения в рамках данных ограничений является NP-трудной задачей ([44]), но приближенное решение может быть найдено с помощью алгоритмов динамического программирования ([27]). Для оценки информативности предложений в методах аннотирования на основе глобальной оптимизации используются аналогичные жадным алгоритмам характеристики, такие как частота и позиция слов в документе, $TF \cdot IDF$, схожесть с входной коллекцией и т.д.

В работе [44] проводится исследование различных вариантов разрешения оптимизационных алгоритмов и демонстрируется, что точное решение оптимизационной задачи по выбору предложений может быть найдено с помощью методов ***Целочисленного Линейного Программирования*** (**I**nteger **L**inear **P**rograming, **ILP**). В каноническом виде ILP представляет собой задачу максимизации функции линейного вида при заданном наборе ограничений – нахождение целочисленного вектора $x = (x_1, x_2, \ldots, x_n)$, такого что функция вида $z(x)=c_1x_1+ c_2x_2+ c_3x_3+ \ldots + c_nx_n \rightarrow max$ (или *min*), и удовлетворяет системе линейных неравенств:

$$\begin{cases} z(x) = \sum_{j=1}^{n} c_j x_j \to \max\ (\min) \\ \sum_{j=1}^{n} a_{ij} x_j \le b_i,\ i = \overline{1,m} \\ x_j \ge 0,\ j = \overline{1,n},\ \textit{целые} \end{cases}$$

В работе [44] предлагается модель для применения ILP для решения задачи построения автоматических аннотаций. Пусть s_i флаг включения предложения i в результирующую аннотацию, Rel_i его значимость, а Red_{ij} избыточность по отношению к предложению j. Тогда общая постановка задачи ILP имеет следующий вид:

$$\begin{cases} \sum_i \mathrm{Re}\, l_i \cdot s_i - \sum_{ij} \mathrm{Re}\, d_{ij} \cdot s_{ij} \to \max \\ \sum_j l_j \cdot s_j \le L \\ s_{ij} \le s_i \quad s_{ij} \le s_j \quad \forall i, j \\ s_i + s_j - s_{ij} \le 1 \quad \forall i, j \\ s_i \in \{0,1\} \quad \forall i \\ s_{ij} \in \{0,1\} \quad \forall i, j \end{cases}$$

где l_i и L – количество слов в предложении i и ограничение на общую длину аннотации соответственно. В данной модели заложен явный подход к устранению избыточности, основанный на сопоставлении предложений относительно друг друга (в рамках Red_{ij}), что подразумевает квадратичный рост сложности с увеличением размера входной коллекции.

В работе [27] предложен альтернативный неявный вариант устранения избыточности, на основе общего покрытия концептов входной коллекции результирующей аннотацией. Пусть c_i является флагом вхождения концепта (в качестве концептов могут рассматриваться как отдельные слова, так и более сложные конструкции, например, n-граммы слов) в итоговую аннотацию, w_i его вес, Occ_{ij} – флаг вхождения концепта i в предложение j. Тогда результирующая модель ILP приобретает следующий вид:

$$\begin{cases} \sum_i w_i \cdot c_i \to \max \\ \sum_j l_j \cdot s_j \le L \\ s_j \cdot Occ_{ij} \le c_i \quad \forall i, j \\ \sum_j s_j \cdot Occ_{ij} \ge c_i \quad \forall i \\ c_i \in \{0,1\} \quad \forall i \\ s_j \in \{0,1\} \quad \forall j \end{cases}$$

Основной проблемой методов глобальной оптимизации является их сложность. С точки зрения времени работы жадные алгоритмы являются более эффективными, в большинстве случаев отрабатывая за константное время, независимо от объема входной коллекции. Приближенные решения обычно масштабируются линейно по объему входных данных, тем самым оставаясь доступными для практического применения. Сложность же точных алгоритмов глобальной оптимизации растет экспоненциально с ростом размера входной коллекции и сложно применимо на практике ([44]), однако изменение модели требований и ограничений может давать сравнительно масштабируемые результаты ([27]).

1.3 Оценка качества автоматических аннотаций

Оценка качества автоматических аннотаций является сложной задачей, поэтому предложен спектр методов оценки, которые могут быть классифицированы по следующим ключевым параметрам:

- Степень участия человека (автоматическое, ручное, полуавтоматическое)
- Критерий оценки (содержание, читабельность и т.д., [18], [47])
- Скорость обработки
- Область применения ([17], [52], [54])

Оценка систем автоматического аннотирования является сложной задачей, ввиду высокой трудоемкости и значительной степени

несогласованности экспертов. Наиболее популярным подходом (в первую очередь ввиду минимальной трудоемкости) является использование набора автоматических мер качества ROUGE (см. Раздел 1.3.1), позволяющий производить автоматическую оценку большого количества автоматических аннотаций на базе нескольких экспертных аннотаций, составленных человеком. Метод «Пирамиды» (см. Раздел 1.3.2) также требует подготовки ручных аннотаций, но кроме того требуется дополнительная работа по выявлению ключевых фактов, которые вручную необходимо выделять и из автоматических аннотаций. Метод «Пирамиды» производит более глубокую оценку конкурсных аннотаций, но связан с большими трудозатратами. Наиболее комплексной оценкой, безусловно, является ручная оценка экспертами, но по причине своей дороговизны, а также значительной степени субъективизма, данный подход применяется значительно реже других методов оценки.

1.3.1 Автоматические меры качества ROUGE

Recall-**O**riented **U**nderstudy for **G**isting **E**valuation (ROUGE, [36]) - набор мер качества и комплекс программ для оценки систем автоматического аннотирования и машинного перевода текстов. Основная идея метода заключается в сравнении генерированной аннотации с эталонной аннотацией, сделанной экспертом. Различные способы сопоставления автоматических аннотаций с экспертными аннотациями, соответствуют различным мерам качества ROUGE, к которым относятся:

- ROUGE-N: сопоставление количества пересекающихся N–грамм слов. Наиболее распространенными являются меры качества ROUGE-1, ROUGE-2, однако также во многих работах приводятся оценки по ROUGE-3 и ROUGE-4;
- ROUGE-L: оценка длин максимальных общих подпоследовательностей (последовательность слов исходного предложения в порядке их вхождения), по отношению к общей длине предложений;

- ROUGE-W: аналог ROUGE-L, но с добавлением веса для каждой из подпоследовательностей, основанном на плотности последовательностей (среднее расстояние появления в исходном предложении);
- ROUGE-S (Skip-bigrams): анализ пересечения биграмм, находящихся на некотором расстоянии друг от друга (между первым и вторым словами биграммы могут находиться другие слова). В качестве параметра в данной мере качества выступает величина окна skip-биграммы - количество слов, которое может "вклиниваться" внутрь биграммы. Соответственно, данный параметр порождает различные варианты меры качества, такие как ROUGE-S* (нет ограничения на количество слов внутри биграммы), ROUGE-S4 (максимум 4 слова) и так далее. Стоит отметить, что стандартная мера качества ROUGE-2 является частным случаем ROUGE-S, а именно ROUGE-S0.
- ROUGE-SU: модификация ROUGE-S, добавляющая учет монограмм. Данное дополнение связано с узким местом меры качества ROUGE-S, связанной с получением нулевого веса для предложения, в котором слова находятся в обратном порядке, относительно соответствующего предложения из экспертной аннотации, так как все биграммы в данном случае будут различными.

Общая формула для мер качества ROUGE выглядит следующим образом:

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

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*.

Приведем пример расчета меры качества ROUGE-1: сравнение пересечения монограмм слов автоматической и экспертной аннотаций. Пусть автоматическая аннотация представлена следующим предложением:

- Китай и Тайвань установили авиасообщение после 60-летнего перерыва.

Эталонная аннотация, составленная экспертом:

- После почти 60-летнего перерыва открылось регулярное авиасообщение между Тайванем и материковым Китаем.

Тогда данная генерированная аннотация получит оценку, равную количеству монограмм слов, которые встречаются и в генерированной аннотации и в эталонной, по отношению к общему числу монограмм в эталонной аннотации, то есть:

$$ROUGE-1 = \frac{7}{12} = 0.58(3)$$

Метод ROUGE широко применяется для оценки генерированных аннотаций ввиду простоты и скорости его применения, однако оценки, полученные при помощи этого метода, зачастую не совпадают с человеческой точкой зрения.

1.3.2 Метод «Пирамиды» (Pyramid Evaluation)

Метод пирамидной оценки автоматических аннотаций разработан Колумбийским университетом в 2005 году и применяется на конференциях DUC и TAC. Данный метод основан на ручном выделении экспертами «информационных единиц» из эталонных аннотаций - Summary Content Units (SCUs). Каждая SCU представляет собой квант информации, которая, по мнению эксперта, должна быть также отражена в автоматической аннотации.

SCU получает вес, равный количеству эталонных аннотаций, где она встречается. Общая оценка автоматической аннотации складывается из

суммы весов SCU, которые она содержит, по отношению к общему количеству SCU для данного текста:

$$\frac{[Суммарный\_вес\_найденных\_SCU]}{[Суммарный\_вес\_всех\_SCU\_для\_данного\_топика]}$$

Пример SCU и её вхождений в текст:

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...*

Метод «Пирамиды» позволяет формализовать процедуру оценки автоматических аннотаций, что значительно облегчает человеческий труд по оценке аннотаций и позволяет повысить объективность оценки. В то же время данный метод требует значительного участия человека, так как выделение «информационных единиц» как из экспертных, так и из автоматических аннотаций производится вручную.

1.3.3 Сравнение различных методов оценки автоматических аннотаций

Развитие методов оценки автоматических аннотаций является неотъемлемой частью развития автоматического аннотирования. В настоящее время пакет автоматических мер качества ROUGE (см. главу 1.3.1) по существу является «золотым стандартом» в данной области, являясь, по сути, обязательным при представлении любых новых алгоритмов и результатов в области автоматического аннотирования. Метод пирамидной оценки автоматических аннотаций (см. Раздел 1.3.2) появился позже пакета ROUGE, но при этом быстро занял значимое место в сравнении различных методов автоматического аннотирования ([54]).

Для автоматизированных методов оценки качества автоматического аннотирования важной является корреляция с оценками экспертом. В работе [56] приводится оценка взаимной корреляции различных мер качества

ROUGE, пирамидной оценки, а также ручной оценки автоматических аннотаций (responsiveness). Сравнению подверглись следующие меры качества ROUGE: ROUGE-n (*n* = 1, 2, 3, 4), ROUGE-L, ROUGE-W-1.2, ROUGE-SU4 (см. Раздел 1.3.1). Процедура сравнения мер качества ROUGE основана на сопоставлении с ручными оценками автоматических аннотаций: для каждой пары конкурсных систем осуществляется кросс-проверка результатов, полученных на основании ручной оценки, пирамидной оценки и оценок по различным мерам качества ROUGE.

Основной характеристикой при оценке тестовых пар выступала *prediction accuracy* (*точность предсказания*), характеризующая процент согласованности данных тестовых сценариев (согласованность ручных и автоматических оценок на пространстве оценок по отдельным кластерам в разрезе пар конкурсных систем). В дополнение к характеристике accuracy были рассчитаны такие меры схожести, как precision, recall и balanced accuracy ([56]). В Табл. 1 приведены результаты оценки по всем указанным характеристикам:

	Responsiveness				Pyramid			
Metric	Acc	P	R	BA	Acc	P	R	BA
R1	0.58 (0.61)	0.24	0.64	0.57	0.62 (0.66)	0.37	0.67	0.61
R2	0.64 (0.63)	0.28	0.60	0.59	0.68 (0.69)	0.43	0.63	0.64
R3	0.70 (0.63)	0.31	0.48	0.60	0.73 (0.68)	0.49	0.53	0.66
R4	0.73 (0.64)	0.33	0.40	0.60	0.74 (0.65)	0.50	0.45	0.65
RL	0.50 (0.59)	0.20	0.56	0.54	0.54 (0.63)	0.29	0.60	0.55
R-SU4	0.61(0.62)	0.26	0.61	0.58	0.65 (0.68)	0.40	0.65	0.63
R-W-1.2	0.52(0.62)	0.21	0.54	0.55	0.57(0.64)	0.32	0.62	0.57

Табл. 1: Результаты сравнения различных мер качества ROUGE по характеристикам **A**ccuracy (**A**), **P**recision (**P**), **R**ecall (**R**) и **B**alanced **A**ccuracy (**BA**)

В данной работе также произведено исследование вариантов комбинирования различных мер качества ROUGE. В Табл. 2 представлены результаты для различных комбинированных вариантов. Наиболее значимыми результатами данного исследования являются:

I. Комбинирование различных мер качества ROUGE позволяет улучшить качество оценки автоматических аннотаций (качество моделирования автоматической оценки ручных оценок экспертов);

II. Релевантность различных мер качества ROUGE. В работе показано, что все из имеющихся мер качества ROUGE могут давать качественные результаты моделирования ручных оценок (в зависимости от специфики входных корпусов различные меры качества показывают различные результаты). Это означает, что все меры качества ROUGE являются необходимыми для вычисления при проведении комплексной оценки автоматических аннотаций.

ROUGE Combination	Acc	Prec	Rec	BA
R1_R2_R4_RBE	0.76	0.77	0.36	0.76
R1_R4_RBE	0.76	0.76	0.36	0.76
R2_R4_RBE	0.76	0.74	0.40	0.75
R4_RBE	0.76	0.73	0.41	0.75
R1_R2_R4	0.76	0.71	0.40	0.74
R1_R4	0.75	0.70	0.40	0.73
R2_R4	0.75	0.68	0.44	0.73
R1_R2_RBE	0.75	0.66	0.48	0.72
R2_RBE	0.75	0.64	0.52	0.72
R4	0.74	0.62	0.47	0.70
R1_RBE	0.74	0.62	0.49	0.70
R1_R2	0.73	0.57	0.62	0.70
RBE	0.73	0.57	0.58	0.68
R2	0.71	0.53	0.69	0.68
R1	0.62	0.43	0.69	0.63

Табл. 2: Результаты комбинирования различных мер качества ROUGE

1.4 Выводы к первой главе

В данной главе приведено описание задачи автоматического аннотирования, основных подходов, применяемых моделей и методов для её решения, а также способов оценки и сравнения полученных автоматических аннотаций. Наиболее востребованной и значимой на практике является задача подготовки обзорного реферата для коллекции документов.

Анализ предметной области показал, что большинство существующих подходов автоматического аннотирования опираются на оценку

информативности предложений коллекции для аннотирования (если речь идет об экстрактивных методах), вычисляемую на основе слов и выражений, входящих в данное предложение. Данный факт подчеркивает значительный вклад информации о лексико-семантической вариативности в общее качество результирующих аннотаций, так как без наличия данной информации подготовка полной и не избыточной аннотации становится невозможной.

Таким образом, для качественного решения задачи автоматического аннотирования алгоритмам необходима информация о вариативности именования различных сущностей в рамках входной коллекции, которая является неизбежной частью текстов на естественном языке.

2. Лексическая вариативность и ее моделирование

Данная глава посвящена исследованию природы появления вариативности в текстах на естественном языке, описываются существующие методы выявления различных типов вариативности, а также вводится формальная модель совокупности участников ситуации, описываемой в некоторой текстовой коллекции с учетом вариативности их упоминаний – тематических цепочек. Особое внимание уделяется методам смысловой группировки выражений, являющихся элементами тематических цепочек, то есть относящихся к одним и тем же участникам ситуации. Предлагается новый алгоритм построения тематических цепочек, основанный на совокупности признаков различной природы – контекстно-зависимых и контекстно-независимых признаков. Целью данной главы является анализ достоинств и недостатков существующих методов автоматического выявления вариативности при описании основных участников ситуации, различных моделей для их представления (моделей тематических цепочек), а также создание комплексного метода построения тематических цепочек с учетом проведенного анализа.

2.1 Вариативность в текстах на естественном языке

Тексты на естественном языке обладают внутренней структурой и подчиняются определенным законам устройства ([22], [76]), а также имеют ряд специфических свойств. Одним из таких свойств является наличие ***вариативности*** – использование различных языковых выражений для определения одинаковых реальных физических явлений или сущностей. Наличие подобного разнообразия выражений для описания одних и тех же аспектов является неотъемлемой частью языка и зачастую незаметно для людей, носителей данного языка, однако представляет собой значительную проблему для систем автоматической обработки текстов.

Природа возникновения данной вариативности является различной. Во-первых, выделяют различные ***цели использования*** вариативности. Одной из таких целей является ***референция*** - отнесенность языкового выражения к одному и тому же объекту действительности. Референция широко распространена в текстах на естественном языке и используется для диверсификации упоминания реальных сущностей в тексте с целью исключения повторов. Например, некоторый текст, посвященный решению правительства Киргизии о выводе авиабазы США со своей территории, содержит следующие варианты референции:

> *3 февраля* ***президент Киргизии Курманбек Бакиев*** *заявил о* ***решении правительства*** *прекратить деятельность авиабазы на территории республики…* ***Президент*** *не стал скрывать, что экономические резоны стали главной причиной побудившей правительство страны принять такое* ***решение.***

Другой причиной возникновения вариативности является ***перефразирование*** (рерайтинг) – изменение текста без изменения его смысла. Данный вид вариативности особенно распространен в контексте новостных кластеров, так как зачастую рерайтеры специально перерабатывают одни и те же новостные сообщения, чтобы новая публикация казалось новой и уникальной. Например, следующие 2 предложения обладают идентичным смыслом, но различаются на уровне языковых выражений:

- *Судьбу авиабазы США в "Манасе"* ***решит*** *парламент Киргизии.*
- *Парламент Киргизии в четверг* ***примет окончательное решение*** *о судьбе авиабазы США.*

Природа вариативности также может быть классифицирована по ***типу привязки к контексту.*** Употребление общеизвестных альтернативных вариантов именования является одним из вариантов появления

вариативности. Например, *КИРГИЗИЯ* и *КЫРГЫЗСТАН* являются равноценными названиями единственной страны, и в рамках реальных текстов могут употребляться различные варианты именования:

> *Правительство* ***Киргизии*** *передало для ратификации в законодательный орган… Парламент* ***Кыргызстана*** *в четверг примет окончательное решение о судьбе авиабазы США… стали главной причиной побудившей правительство* ***страны*** *принять такое решение.*

Также необходимо отметить, что употребление слова *СТРАНА* также в данном случае относится к общеизвестному типу вариативности, так как данный факт может быть проверен без анализа контекста (взят из предопределенных источников). В то же время может иметь место ***контекстная вариативность***, разрешение которой невозможно без анализа контекста её употребления. Данный тип вариативности является наиболее сложным для установления системами автоматической обработки текстов. Например, установление вариативности для выражений *ОХРАННИК АВИАБАЗЫ* и *АМЕРИКАНСКИЙ ВОЕННЫЙ* невозможно без глубокого смыслового анализа следующего фрагмента текста:

> В *декабре 2006 года 46-летний водитель … был расстрелян в упор* ***охранником авиабазы*** *Закари Хатфилдом …* **.** ***Американский военный*** *… также был тайно вывезен с территории страны и до сих пор не предстал перед судом.*

Таким образом, общая классификация типов вариативности в текстах на естественном языке может быть представлена в следующем виде:

I. Цель использования

a. Референция (отнесенность языкового выражения к одному и тому же объекту действительности)
b. Перефразирование (изменение текста без изменения смысла)

II. Наличие привязки к контексту

a. Общеизвестная вариативность (может быть установлена на основе предопределенных источников)
b. Контекстная вариативность (для установления необходим анализ контекста употребления вариативности)

Проблемы референции в тексте рассматриваются посредством установления ***кореференции имен*** (построения ***референциальных цепочек***). Прежде всего, данная задача решается для установления референциальных отсылок для людей и организаций (*Президент Российской Федерации Дмитрий Медведев, Президент Медведев, Дмитрий Медведев*) ([23], [79]).

Обнаружение в текстах парафраз, т.е. альтернативных способов представления одной и той же информации в текстах, основан на анализе контекстов их употребления ([8]).

В работе [19] предлагается алгоритм, акцентирующийся на качестве контекстов для извлечения контекстных синонимов (квазисинонимов). В качестве контекстов рассматриваются *N*-граммы слов вокруг слов-кандидатов в квазисинонимы. Для вычисления качества контекста *P(c)* предлагается использовать следующую формулу, основанную на количества различных контекстов, в которых употребляется слово или выражение:

$$P(c) = \frac{1}{Z} \cdot \frac{1}{W}$$

где W это количество различных контекстов, с которыми встречалось выражение c, а Z – нормализующий фактор. Вероятность квазисинонимичности для выражений w_1 и w_2 предлагается вычислять на основе совместного учета количества и качества разделяемых контекстов, по следующей формуле (C – общие контексты выражений w_1 и w_2):

$$P(w_2 \mid w_1) = \sum_{c \in C} P(w_1 \mid c) \cdot P(w_2 \mid c) \cdot P(c)$$

Близкие по смыслу выражения в текстах часто рассматриваются не попарно, а как элементы "цепочек" близких по смыслу выражений. Такие подходы будут рассмотрены более подробно.

2.2 Цепочные методы смысловой группировки языковых выражений

Группы близких по смыслу выражений в «цепочных» алгоритмах собираются в виде ***лексических цепочек***. Лексическая цепочка представляет собой последовательность семантически связанных слов (повторы, синонимы, гипонимы, гиперонимы и др.) и является известным подходом для моделирования связности текста на естественном языке ([33], [38], [81]). Алгоритмы построения лексических цепочек основаны на использовании информации о связях между словами и выражениями, описанных в некотором заранее определенном ресурсе. Например, тезаурус английского языка WordNet и тезаурус русского языка РуТез.

Во всех подходах по автоматическому моделированию лексических цепочек построение этих цепочек не является самоцелью - лексические цепочки выделяются для того, чтобы «приблизиться» к автоматическому построению тематической структуры текста, т. е. уметь выделять, что в тексте главное, что второстепенное, как текстовые сущности связаны друг с другом.

С целью выделения наиболее значимых для содержания текста лексических цепочек рассматриваются различные параметры лексических цепочек, такие как частотность ее элементов, текстовое покрытие и другие. В лексических цепочках выделяются наиболее частотные элементы цепочки в качестве наиболее важных тематических элементов текста. При этом наиболее частотные лексические цепочки определяют наиболее значимых участников ситуации, описываемой в исходном корпусе документов.

Лексические цепочки являются идейно наиболее близким из существующих подходов к предлагаемому в данной кандидатской диссертации.

2.2.1 Алгоритм построения лексических цепочек на основе тезауруса WordNet для английского языка

В работе [33] предлагается алгоритм построения лексических цепочек на основе тезауруса Wordnet для текстов на английском языке. Каждому слову в Wordnet может быть сопоставлено несколько ***синсетов*** – смысловых значений данного слова, а сами синсеты могут быть связаны различными отношениями (синонимия, часть-целое, род-вид и т.д.). Для построения лексических цепочек выделяют три типа отношений:

- ***Экстра-сильные***: повторение слова;
- ***Сильные***: наличие общего синсета или синсетов связанных горизонтальным отношением;
- ***Средне-сильные***: наличие связи между синсетами, удовлетворяющих заданному набору правил (с заданной максимальной длиной и количеством, а также типом перегибов).

Экстра-сильные и сильные отношения являются бинарными, в то время как средне-сильные отношения имеют вес, рассчитываемый по следующей формуле:

$$weight = C - path\_length - k \cdot number\_of\_direction\_changes$$

где C и k являются константами.

Алгоритм построения лексических цепочек является итеративным, в основе которого лежит последовательный просмотр текста входной коллекции. Для каждого рассматриваемого слова выполняется один следующих шагов:

1. Поиск экстра-сильных отношений с существующими цепочками. Если отношение найдено, рассматриваемое слово добавляется к данной цепочке (процедура добавления описана ниже). Если отношение не найдено – переход на следующий шаг;
2. Поиск сильных отношений (аналогично шагу 1), но с ограничением на дистанцию поиска – не более 7 предложений;

3. Поиск средне-сильных отношений (аналогично шагу 1), но с ограничением на дистанцию поиска – не более 3 предложений;
4. Если необходимых отношений найдено не было, то создается новая цепочка – рассматриваемое слово со всеми его синсетами является её единственным элементом.

При добавлении слова в цепочку производится процедура дизамбигуации слов со множественными синсетами. Синсеты добавляемого слова, связанные с синсетами слов, уже присутствующих в цепочке, сохраняются. Остальные синсеты, не имеющие связей с синсетами других элементов цепочки, удаляются.

На Рис. 3 представлен пример построения лексической цепочки (4 шага работы алгоритма).

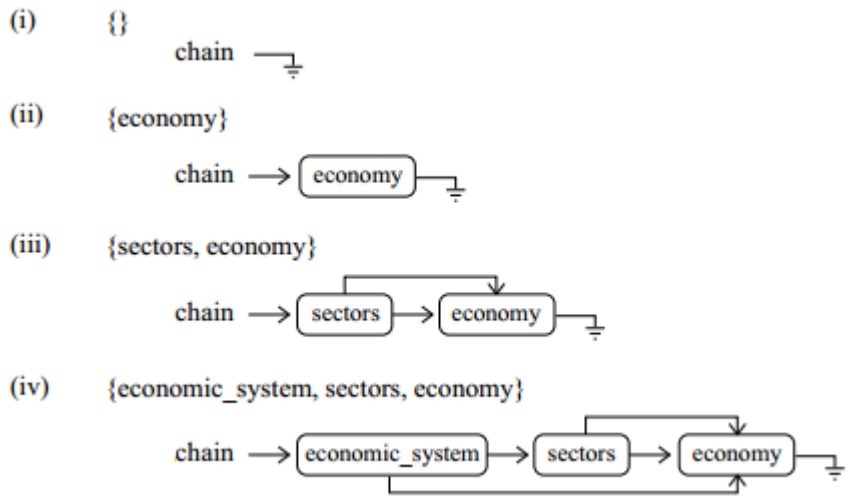


Рис. 3: Пример построения лексической цепочки

2.2.2 Алгоритм построения лексических цепочек на основе тезауруса РуТез для русского языка

В работе [81] предложен метод построения тематической структуры текста на основе знаний тезауруса русского языка РуТез. Для построения тематического представления текста сеть понятий тезауруса автоматически разбивается на совокупности близких по смыслу понятий – ***тематические узлы***. Алгоритм разбиения является жадным и состоит из следующих шагов:

1. Просмотр упомянутых понятий тезауруса, сначала по заголовку текста, далее по мере снижения частотности;

2. Если очередное понятие C_i относится к уже существующему тематическому узлу, то переходим к следующему понятию;
3. Если нет, то образуется новый тематический узел, с центральным элементом C_i. В новый тематический узел вносятся все понятия документа, которые связаны с понятием C_i по непосредственным отношениям тезауруса или по отношениям с учетом транзитивности и наследования.

Для оценки связности между понятиями в тексте вводится понятие ***текстовая связь***: понятие считается связанным по тексту с теми понятиями, которые находятся на расстоянии не более N понятий от очередного вхождения данного понятия, безотносительно к порядку следования понятий в тексте. На основе анализа суммированных текстовых связей тематические узлы разбиваются на основные и локальные тематические узлы, а все понятия делятся на пять базовых классов значимости для текста:

a. Центры основных тематических узлов;
b. Другие понятия основных тематических узлов;
c. Центры локальных тематических узлов;
d. Другие понятия локальных тематических узлов;
e. Упоминавшиеся понятия, не вошедшие в предыдущие классы.

Основные тематические узлы имеют максимальные суммарные текстовые связи друг с другом, образуя полносвязное ядро обсуждения рассматриваемого текста.

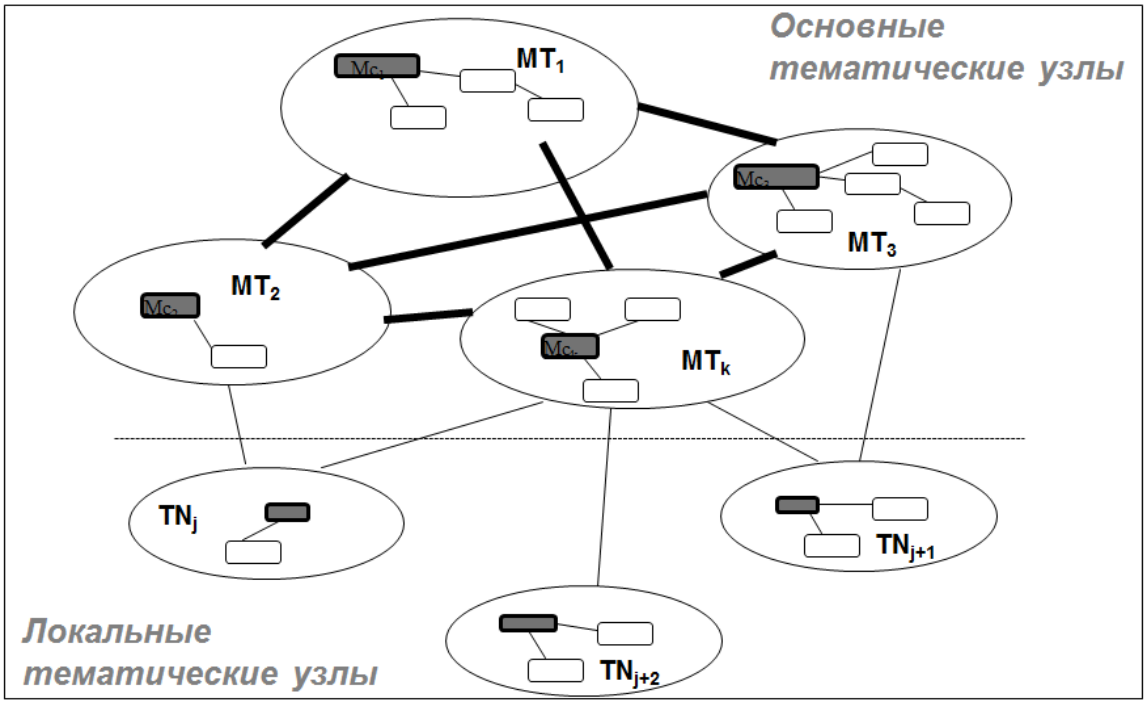


Рис. 4: Основные и локальные тематические узлы

Для использования построенного тематического представления в дальнейшей обработке каждому тематическому узлу ставится в соответствие некоторый вес. Результирующий вес понятия является линейной функцией от веса его позиции в тематическом узле и относительной частотности понятия в документе.

Описанные выше подходы к группированию близких по смыслу слов и выражений на основе лексических цепочек имеют существенный недостаток: в лексические цепочки включаются только те языковые выражения, которые заранее описаны в тезаурусе. Для подключения в лексические цепочки новых слов и выражений необходимо автоматически выделить из текстов кандидаты в лексические цепочки, установить их семантическую близость между собой и между известными (описанными в тезаурусе) лексическими единицами. Для установления такихотношений в работе предлагается использовать совокупность факторов, свидетельствующих о такой близости, а именно сходство контекстов употребления, расположение относительно друг друга, сходство по написанию. Кроме того, в работе вводится дополнительный фактор, ограничивающий вхождение лексических единиц в одни и те же цепочки.

Данный фактор основан на таких свойствах связного текста как локальная и глобальная связность.

2.3 Локальная и глобальная связность текста

Тексты на естественном языке обладают свойствами глобальной и локальной связности. Глобальная связность текста проявляется в том, что тематическая структура текста может быть представлена в виде иерархии ([22], [76]). Вершина данной иерархии представляет собой основную тему документа, в то время как нижние уровни соответствуют локальным или побочным темам текста. Локальная связность, или связность между соседними предложениями текста, часто осуществляется такими средствами, как анафорические отсылки, например, с помощью местоимений, или посредством повторения одних и тех же или близких по смыслу слов.

В работе [76] Ван Дейк и Кинч описывают тематическую структуру текста как иерархическую в том смысле что тема всего текста описывается посредством более конкретных подтем, которые в свою очередь могут быть охарактеризованы посредством еще более конкретных подтем текста и т.д.

Под темой/подтемой при этом понимается предикат $P(C_1, \ldots, C_n)$. Его атрибуты $C_{1,}\ldots,C_n$ будем называть ***тематическими элементами***. Таким образом, если текст посвящен обсуждению взаимоотношений между тематическими элементами $C_{1,}\ldots,C_n$, то в предложениях текста должны обсуждаться детали этих отношений. С формальной точки зрения иерархия тематической структуры образует отношение частичного порядка φ на множестве тем и подтем исходного документа(ов) $P = \{P_i^{level}(C_{i\_1}^{level}, \ldots, C_{i\_n}^{level})\}$, т.е. обладает свойствами:

1. Рефлективности: $P_i \varphi P_i \quad \forall P_i \in P$
2. Транзитивности: $(P_i \varphi P_j) \wedge (P_j \varphi P_k) \Rightarrow P_i \varphi P_k \quad \forall P_i, P_j, P_k \in P$
3. Антисимметричности: $(P_i \varphi P_j) \wedge (P_j \varphi P_i) \Rightarrow P_i = P_j \quad \forall P_i, P_j \in P$

Каждое предложение s связного текста посвящено раскрытию той или иной подтемы основной темы текста: $s \to \mathrm{P}_i^{level}(\mathrm{C}_{\mathrm{i\_1}}^{\mathrm{level}}, \ldots, \mathrm{C}_{\mathrm{i\_n}}^{\mathrm{level}})$, раскрывающей один из аспектов взаимоотношений тематических элементов $\mathrm{C}_{\mathrm{i\_1}}^{\mathrm{level}}, \ldots, \mathrm{C}_{\mathrm{i\_n}}^{\mathrm{level}}$. При этом отнесение к тематическим элементам $\mathrm{C}_{\mathrm{i\_1}}^{\mathrm{level}}, \ldots, \mathrm{C}_{\mathrm{i\_n}}^{\mathrm{level}}$ внутри s осуществляется с помощью конкретных языковых выражений, упомянутых в s:

$$s \leftrightarrow \mathrm{P}_i^{level}(\mathrm{C}_{\mathrm{i\_1}}^{\mathrm{level}}, \ldots, \mathrm{C}_{\mathrm{i\_n}}^{\mathrm{level}}) = P_i^{level}(\mathrm{C}_{\mathrm{i\_1}}^{\mathrm{level}} \to \{\mathrm{t}_{\mathrm{i\_1}}^{\mathrm{level}}\}, \ldots, \mathrm{C}_{\mathrm{i\_n}}^{\mathrm{level}} \to \{\mathrm{t}_{\mathrm{i\_n}}^{\mathrm{level}}\}), \quad \mathrm{t}_{\mathrm{i\_1}}^{\mathrm{level}}, \ldots, \mathrm{t}_{\mathrm{i\_n}}^{\mathrm{level}} \in s$$

Общая схема частично упорядоченного множества тем и подтем текста может быть представлена следующим образом:

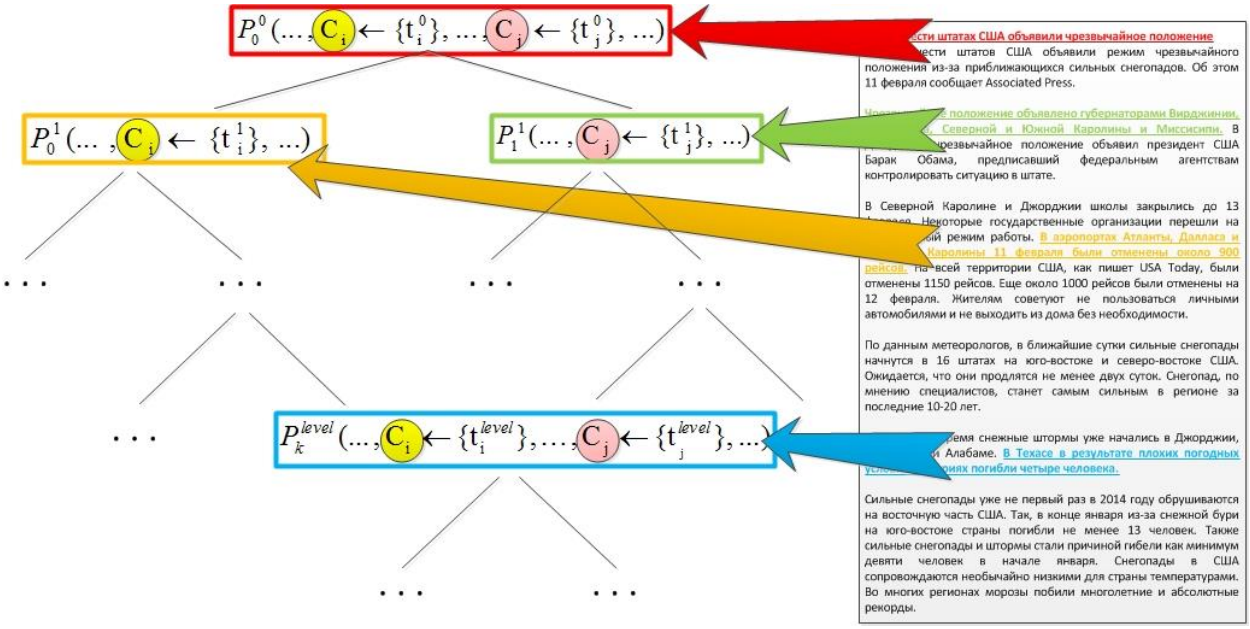


Рис. 5: Иерархия тем и подтем документа

Таким образом, для отнесения к некоторому тематическому элементу C_m используется определенный набор языковых выражений, каждое из которых используется при раскрытии определенных подтем текста:

$$\mathrm{C}_\mathrm{m} \to \{\mathrm{t}_\mathrm{m}^1, \ldots, \mathrm{t}_\mathrm{m}^\mathrm{i}, \ldots\}$$

Модель тематической структуры обладает следующими свойствами:

- Наличие отнесения как минимум к двум различным тематическим элементам C_m и C_k во всех предложениях s исходной коллекции:

$$(\forall s \in S)\,(\exists C_m, C_k) : (C_m \to \{t^i_m\}) \wedge (C_k \to \{t^j_k\}) \wedge (t^i_m \in s) \wedge (t^j_k \in s)$$

- Раскрытие деталей отношений между тематическими элементами $P_i(C_1, \ldots, C_n)$ и $P_j(C_1, \ldots, C_n)$ при выполнении отношения частичного порядка:

$$(\forall P_i, P_j \in P) : P_i \prec P_j \Rightarrow (\forall s_i \leftrightarrow P_i, s_j \leftrightarrow P_j)\,(\exists t^i_m, t^i_k \in s_i\ \exists t^j_m, t^j_k \in s_j) :$$
$$(C_m \to \{t^i_m, t^j_m\}) \wedge (C_k \to \{t^i_k, t^j_k\})$$

Из описанной модели следует важный практический вывод: если языковые выражения t_1 и t_2 часто встречаются в анализируемом тексте в одних и тех же простых предложениях, то это означает, что данный текст посвящен рассмотрению отношений между этими сущностями, т. е. t_1 и t_2 соответствуют разным тематическим элементам ([32], [38]). С другой стороны, если два языковых выражения t_1 и t_2 редко встречаются в одних и тех же предложениях текстов, но при этом часто упоминаются в соседних предложениях, то это дает возможность предположить, что они используются для осуществления локальной связности, то есть между ними имеется смысловая связь.

Для проверки гипотезы о том, что связанные сущности чаще встречаются в соседних предложениях, чем в одних и тех же простых предложениях текстов, был проведен эксперимент, подробно описанный автором диссертации в статье [3]. Основным результатом проведенного эксперимента является подтверждение того факта, что наиболее близкие по смыслу выражения значительно чаще встречаются в соседних предложениях, чем в одних и тех же фрагментах предложений текстов. При нарастании смысловых различий между выражениями частота встречаемости в одних и тех же предложениях нарастает, пока не стабилизируется на некотором среднем для всех выражений уровне. В частности, был получен результат, что синонимы в среднем в 5 раз чаще встречаются в соседних предложениях по отношению к вхождениям в одни и те же предложения, по сравнению с аналогичным отношением для несвязанных по тезаурусу выражений.

Гипотеза о совместной встречаемости связанных языковых выражений легла в основу ограничивающего фактора *IsNSCriterion(t_i,t_j)*, который является управляющим в предлагаемом алгоритме построения модели участников ситуации новостного кластера (см. Раздел 2.4.2):

$$count((s_k, s_m) \mid t_i \in s_k \wedge t_j \in s_m \wedge NS(s_k, s_m)) > C \cdot count(s \mid t_i \in s \wedge t_j \in s)$$

где *count(A|B)* – количество элементов *A* удовлетворяющих условию *B* (в данном случае предложений и пар предложений); $NS(s_k, s_m)$ - признак последовательного появления предложений s_k и s_m в одном из документов анализируемого новостного кластера. Необходимо отметить, что критерии, подобные критерию *IsNSCriterion(t_i,t_j)*, не использовалась ранее для решения таких задач, как установление вариантов именования основных участников ситуации, построение рядов квазисинонимов, лексических цепочек и т.п.

2.4 Предлагаемый метод построения тематических цепочек

Тематическая цепочка - способ моделирования тематических элементов в иерархической структуре тем и подтем связного текста, а именно совокупности языковых выражений, относящихся к аргументам предикатов, описывающих темы документа.

Предлагаемый метод построения тематических цепочек новостного кластера основан на комбинировании известных характеристик схожести языковых выражении, показавших свою эффективность в различных работах по данному направлению (см. Раздел 2.2), а также новой характеристики схожести, основанной на тематической структуре документов и внутреннем устройстве текстов на естественном языке в целом (см. Раздел 2.3).

Основной задачей предлагаемого алгоритма является построение структуры, содержащей информацию об имеющихся семантических связях между языковыми выражениями, упомянутыми внутри входной коллекции документов. Данную информацию предполагается использовать как дополнительное знание для работы различных алгоритмов автоматической

обработки текстов, поэтому построенная структура должна обладать достаточным уровнем качества и достоверности.

Выбор формы целевых тематических цепочек осуществлялся на основе описанных требований к качеству целевой структуры, а также предполагаемом варианте использования результатов (машинная обработка – применение результатов работы одного алгоритма как входных данных другого), и базировался на следующих принципах:

- Взаимодействие участников описывается в предложениях текста, поэтому, чем чаще слова (или многословные выражения) встречаются в одних и тех же предложениях текста, тем больше вероятность того, что эти слова (или многословные выражения) относятся к разным участникам ситуации;
- Каждому участнику в тексте соответствует группа слов и многословных выражений. Предполагается, что в тексте имеется наиболее частотное (главное название участника) и разные варианты, поэтому группа слов и выражений, относящихся к одному участнику, строится в форме тематической цепочки с выделенным центральным элементом: $tc = \{t_{main}[, t_1, \ldots, t_k]\}$, t_i – слово или многословное выражение;
- Совокупность построенных тематических цепочек является основным результатом работы предлагаемого алгоритма [32], [38]. Каждое слово (или многословное выражение), упомянутое в исходном новостном кластере, сопоставляется с одной (для однозначных языковых выражений) или несколькими (в случае неоднозначности употребления языкового выражения в рамках входного корпуса документов) тематическими цепочками.

Данные предположения основаны на внутреннем устройстве и тематической структуре текстов на естественном языке [22], [33]. Новостной кластер не является связным текстом, но посвящен одной ситуации (или совокупности связанных ситуаций) и содержит большое количество

документов, что влечет за собой усиление всех статистических особенностей. Характеристики схожести языковых выражений, описанные в разделе 2.4.2, основаны на выявлении данных статистических особенностей.

2.4.1 Формальная постановка задачи построения тематических цепочек

Пусть:

- $w \in \{1..V\}$ - множество слов (Словарь размерности *V*);
- $T = \{t_i = [w_{i1}, .., w_{in}] \mid n \geq 1, i = 1...M\}$ - множество языковых выражений (слов и многословных выражений) размерности *M*;
- $S = \{s_j = [t_{j1}, .., t_{jm}] \mid m \geq 1\}$ - множество предложений;
- $D = \{d_k = [s_{k1}, .., s_{kl}] \mid l \geq 1, k = 1..N\}$ - коллекция *N* новостных документов, посвященных одному событию (новостной кластер);
- $F_{sim} = f(t_i, t_j) \in \Re, t_i, t_j \in T$ - функция близости языковых выражений

Тогда задача построения тематических цепочек представляет собой задачу **кластеризации с ограничениями**:

- $TC = \{tc_i \in a^M\}, a_{ij} \in \{0, 1, 2\}$, где tc_i - тематическая цепочка языковых выражений с выделенным центральным элементом (кластер языковых выражений);

Целевые показатели:

- $F_0 = \frac{\sum_{i<j} [tc_i = tc_j] F_{sim}(t_i, t_j)}{\sum_{i<j} [tc_i = tc_j]} \to \min$ - минимизация среднего внутрикластерного расстояния;
- $F_1 = \frac{\sum_{i<j} [tc_i \neq tc_j] F_{sim}(t_i, t_j)}{\sum_{i<j} [tc_i \neq tc_j]} \to \max$ - максимизация среднего межкластерного расстояния;

- $F_0 / F_1 \to \min$ - минимизация отношения средних внутрикластерных и межкластерных расстояний;

Ограничения:

- $\forall i : count_{j=1..M}([tc_{ij} = 2]) = 1$ - каждая тематическая цепочка содержит один и только один центральный элемент;
- $\forall j : (sum_1, .., sum_M) = \sum_{tc_i \in TC} tc_{ij} : 1 \le sum_j \le 2$ - каждое языковое выражение является элементом не более чем 2 и не менее одной тематической цепочки, либо центром ровно одной тематической цепочки;
- $\forall tc_i : (tc_{ij} > 0 \wedge tc_{ik} > 0) \Rightarrow \text{IsNSCriterion}(tc_{ij}, tc_{ik}) = true$ - выполнено ограничивающее условие на объединение языковых выражений в тематическую цепочку (см. Раздел 2.3).

2.4.2 Характеристики схожести языковых выражений для построения тематических цепочек

2.4.2.1 Расчет контекстных параметров

Важным фактором для построения тематических цепочек являются контексты, в которых употребляются слова и выражения. Сразу несколько различных характеристик схожести основаны на контекстах употребления слов, вследствие чего первым необходимым шагом алгоритма построения тематических цепочек является предварительный расчет контекстных параметров. Для получения контекстов слов, предложения разбиваются на фрагменты между знаками препинания. Выделяются следующие типы контекстов в рамках таких фрагментов:

- соседнее прилагательное или существительное вправо или влево от исходного слова (контекстный параметр *Near*);
- во фрагментах, в которых есть глаголы, фиксируются прилагательные и существительные, между которыми и исходным словом встречается глагол (контекстный параметр *AcrossVerb*);

- прилагательные и существительные, встречающиеся во фрагментах предложений с данным словом, не разделенные глаголом и не являющиеся соседними к исходному слову (контекстный параметр *NotNear*).

Кроме того, для всех прилагательных и существительных запоминаются слова, встречающиеся в соседних предложениях (контекстный параметр *NS*). Предложения для вычисления этого показателя берутся не полностью - учитываются фрагменты предложений с начала и до фрагмента, содержащего глагол (включительно), что позволяет извлечь из соседних предложений наиболее значимые слова.

2.4.2.2 Описание характеристик схожести

Для определения семантически связанных языковых выражений и последующего построения тематических цепочек используется набор из шести основных характеристик схожести. Некоторые из данных характеристик являются контекстно-зависимыми и вычисляются непосредственно на основании рассматриваемого новостного кластера, в то время как другие определяются на основании формальной схожести выражений и информации из заранее определенных ресурсов. Каждая характеристика добавляет некоторый балл в общий вес схожести пары выражений, независимо от других характеристик схожести. В следующих секциях дается подробное описание алгоритма расчета весов схожести пар языковых выражений для контекстно-зависимых и контекстно-независимых характеристик схожести.

Контекстно-зависимые характеристики

Количество вхождений в соседние предложения (**N**eighboring **S**entence **F**eature, ***NSF***). Данная характеристика основана на гипотезе глобальной связности текстов на естественном языке [22] и её следствии о том, что элементы одной тематической цепочки чаще появляются в соседних предложениях исходных документов, чем в одних и тех же предложениях.

Характеристика *NSF* вычисляется на основе контекстных параметров *AcrossVerb*, *Near*, *NotNear* и *NS* и распределения их средних значений внутри исходного новостного кластера. Характеристика *NSF* дает численную оценку соотношения количества вхождений в соседние предложения (характеристика *NS*) по отношению к количеству вхождений в одни и те же предложения исходного корпуса (характеристики *AcrossVerb*, *Near* и *NotNear*), и основана на следующем соотношении:

$$C(t_i, t_j) = NS(t_i, t_j) - 2 \cdot \left(AcrossVerb(t_i, t_j) + Near(t_i, t_j) + NotNear(t_i, t_j) \right)$$

Общая формула вклада характеристики *NSF* в вес схожести пары выражений имеют следующую форму:

$$\underset{t_i, t_j \in T}{weight}_{NSF}(t_i, t_j) = \min(1, \frac{C(t_i, t_j)}{\underset{t_k, t_m \in T}{Avg}(C(t_k, t_m))})$$

где *AVG(C)* является средним значением *C* среди всех положительных значений в рамках всего кластера.

Характеристика *NSF* также является управляющей характеристикой, моделируя критерий лексической связности IsNSCriterion, описанный в Разделе 2.3. Это означает, что два выражения не могут быть включены в одну и ту же тематическую цепочку, если значение характеристики *NSF* имеет отрицательное значение. Подобная пара с отрицательным значением *NSF* не имеет общего веса и не рассматривается алгоритмом построения тематических цепочек.

Строгие контексты (**S**trict **C**ontext, ***SC***). Данная характеристика основана на сравнении строгих контекстов употреблений слов – текстовых шаблонов. В качестве шаблонов рассматриваются 4-граммы, два выражения влево и вправо от рассматриваемого выражения:

$$s_i = (t_{i1}, .., t_{ij-2}, t_{ij-1}, t_{ij}, t_{ij+1}, t_{ij+2}, ..)$$

где $(t_{ij-2}, t_{ij-1}, t_{ij+1}, t_{ij+2})$ является строгим контекстом выражения t_{ij} в некотором предложении s_i.

Границей рассмотрения шаблона является предложение, т.е. некоторые шаблоны (в начале и конце предложения) являются неполноценными, в связи с чем вводится весовая дифференциация шаблонов. Вес шаблона строгого контекста вычисляется на основе количества присутствующих в шаблоне слов, каждое из которых имеет вес равный *0.25*. Например, 4-грамма (*, *, *состоит, из*) имеет вес равный *0.5*, а 4-грамма (*новостной, кластер, состоит, из*) - *1.0*, что является максимальным весом полного шаблона 4-граммы.

Значение характеристики *SC* имеет вещественное значение, принадлежащее отрезку [0,1]. Вес характеристики вычисляется относительно веса пары с максимальным значением разделяемых строгих контекстов, пропорционально весу разделяемых строгих контекстов для текущей пары:

$$\underset{t_i,t_j \in T}{weight_{SC}}(t_i,t_j) = \frac{\sum_{templ \in TEMPLATES(t_i) \cap TEMPLATES(t_j)} weight(templ)}{\max\limits_{t_k,t_m \in T} \sum_{templ \in TEMPLATES(t_k) \cap TEMPLATES(t_m)} weight(templ)}$$

Схожесть контекстов употребления по внутренним характеристикам предложения (**S**calar **P**roduct **S**imilarity, ***SPS***). Каждый из контекстных параметров, описанных в разделе 2.4.2.1, представляет собой вектор частот $V=(v_1, .. , v_m)$, сопоставленных каждому слову или выражению для каждого контекстного параметра. Размерности v_i данного вектора отражают частоту совместной встречаемости рассматриваемого выражения *t* или выражения по одному из контекстных параметров с остальными словами и выражениями, упомянутыми в новостном кластере.

$$t_i \rightarrow \left| \begin{array}{l} \overrightarrow{V_i^{AcrossVerb}} = (v_{i\_1}^{AcrossVerb}, .. , v_{i\_m}^{AcrossVerb}) \\ \overrightarrow{V_i^{Near}} = (v_{i\_1}^{Near}, .. , v_{i\_m}^{Near}) \\ \overrightarrow{V_i^{NotNear}} = (v_{i\_1}^{NotNear}, .. , v_{i\_m}^{NotNear}) \\ \overrightarrow{V_i^{NS}} = (v_{i\_1}^{NS}, .. , v_{i\_m}^{NS}) \end{array} \right.$$

После построения данные контекстные вектора могут сравниваться классическими мерами схожести (например, косинусная мера угла между

векторами), характеризуя степень схожести контекстов употреблений рассматриваемых слов и выражений:

$$\underset{t_i,t_j\in T}{weight}_{SPS}(t_i,t_j)=\frac{(\overrightarrow{V_i^{Context}}1,\overrightarrow{V_j^{Context}})}{|\overrightarrow{V_i^{Context}}|\cdot|\overrightarrow{V_j^{Context}}|}$$

где *Context*={*AcrossVerb, Near, NotNear, NS*} – различные типы контекстов.

Значение характеристики *SPS* имеет вещественное значение, лежащее в пределах от 0 до 1, и вычисляется как косинусная мера схожести по всем контекстным характеристикам, ограниченная сверху значением *1.0*.

Контекстно-независимые характеристики

Формальное сходство (**B**eginning **S**imilarity, ***BS***). Рассмотрение формального сходства выражений является естественным путем обнаружения семантически-связанных объектов. В существующей реализации используется простая мера схожести – одинаковые начала слов. Данная характеристика позволяет находить сходство между такими выражениями как *Руководитель – Руководство, Президент России – Российский президент* и так далее.

Общий вес характеристики *BS* имеет вещественное значение из отрезка [0, 1] и вычисляется на основе модифицированной меры Жаккара, классический вариант которой выглядит следующим образом:

$$K=\frac{n(A\cap B)}{n(A)+n(B)-n(A\cap B)}=\frac{n(A\cap B)}{n(A\cup B)}$$

где *A* и *B* сравниваемые множества, $n(A\cap B)$ и $n(A\cup B)$ – количество элементов в пересечении и объединении множеств *A* и *B* соответственно. В нашем случае сравнению подвергаются множества слов сопоставляемых выражений, а эквивалентными считаются слова, имеющие одинаковые начала. Общий вид модифицированной меры Жаккара для расчета характеристики *BS* имеет следующий вид:

$$\underset{t_i,t_j\in T}{weight}_{BS}(t_i,t_j)=\begin{cases}\dfrac{n_{word}(t_i\cap t_j)+k}{n_{word}(t_i\cup t_j)+k}\ при\ n_{word}(t_i\cap t_j)>0,\\ 0,\ при\ n_{word}(t_i\cap t_j)=0.\end{cases}$$

где k – нормировочный коэффициент, сглаживающий значения меры для коротких многословных выражений, в том числе отдельных слов. В текущем алгоритме использовался коэффициент $k=3$.

Информация о схожести, описанная во внешнем ресурсе – тезаурусе РуТез (**T**hesaurus **S**imilarity, ***TS***). На текущий момент существует большое количество разнообразных предопределенных ресурсов, которые содержат в себе дополнительную информацию о связях слов и выражений. Данная информация может быть использована для построения тематических цепочек и сделать данное построение более стабильным и качественным. Более того, известно, что некоторые типы отношений между словами и выражениями широко используются для обеспечения связности реальных текстов (например, такие отношения как синонимия). Вычисление характеристики *TS* основано на использовании информации из тезауруса русского языка РуТез [39]. При этом в рассмотрение попадали как непосредственные связи объектов, так и «длинные» связи по транзитивным типам отношений. Рассматриваются следующие типы связей: синонимия, часть – целое, род – вид.

Значение характеристики *TS* имеет вещественное значение от 0 до 1 и убывает с ростом длины пути по отношениям между объектами в тезаурусе:

$$\underset{t_i,t_j\in T}{weight}_{TS}(t_i,t_j)=f_{path}(t_i,t_j)=f(N_{rel}(t_i,t_j),\{\mathrm{Re}\,l_{type}(t_i,t_j)\})$$

где N_{rel} – длина пути по отношениям тезауруса (количество связей), $\{Rel_{type}\}$ – информация о типах связей по данному пути. Функция $f(N_{rel},\{Rel_{type}\})$ моделирует ослабление связи между выражениями при увеличении расстояния между ними в тезаурусе, с учетом различных типов отношений. В данной работе использовалась линейная модель падение схожести для всех типов связей в размере *0.2* для каждого перехода: $TS=1-0.2\cdot N_{rel}$.

Наличие одинаковых языковых выражений (**E**mbedded **O**bjects **S**imilarity, ***EOS***). При анализе схожести тематических цепочек, включающих в себя несколько языковых выражений, важным фактором схожести является наличие общих языковых выражений. Данный фактор особенно важен на поздних итерациях работы алгоритма, когда имеется значительное количество частично сформированных тематических цепочек, и большинство характеристик схожести уже значительно проработаны. Значение характеристики *EOS* является булевым и добавляет 1 балл в общий вес схожести пары в случае наличия одинаковых языковых выражений у анализируемых тематических цепочек. Общая формула вычисления характеристики *EOS* для тематических цепочек tc_i и tc_j имеет следующий вид:

$$\underset{tc_i, tc_j \in TC}{weight_{EOS}}(tc_i, tc_j) = \begin{cases} 1 \text{ при } count(tc_i \cap tc_j) > 0, \\ 0 \text{ при } count(tc_i \cap tc_j) = 0. \end{cases}$$

2.4.3 Алгоритм построения тематических цепочек

2.4.3.1 Сборка многословных выражений

Необходимой информацией для качественного построения тематических цепочек является информация о многословных выражениях, используемых в исходном документе. По этой причине этап сборки многословных выражений рассматривается нами как необходимый предварительный этап построения целевого результата.

Важной основой извлечения многословного выражения из текста документа является частотность его встречаемости в тексте. Однако кластер представляет собой структуру, в которой многие цепочки слов повторяются многократно. Поэтому основным критерием для выделения многословных выражений является значительное превышение встречаемости слов непосредственно рядом друг с другом по сравнению с раздельной встречаемостью во фрагментах предложений [77]:

$$\text{Near} > 2 \cdot (\text{AcrossVerb} + \text{NotNear})$$

Кроме того, используются ограничения по частотности встречаемости слов рядом друг с другом.

Просмотр подходящих пар слов (выражений) для склейки производится в порядке снижения коэффициента *Near / (AcrossVerb + NotNear)*. При нахождении подходящей пары слов, они склеиваются в единый объект и все контекстные отношения пересчитываются. Процедура просмотра начинается заново и повторяется до тех пор, пока произведена хотя бы одна склейка. Псевдокод процедуры сборки многословных выражений:

Процедура: Сборка многословных выражений	
∇ Вход:	1. Новостной кластер *D* в пословном представлении: $D=\{d_1, \dots, d_n\}$, $d_i=\{s_1, \dots, s_m\}$, $s_j=\{w_1, \dots, w_t\}$, $w \in W$ *(Словарь)* 2. $AcrossVerb(a_1, a_2)$ – количество вхождений слов или выражений a_1 и a_2 в одном предложении через глагол в *D* 3. $Near(a_1, a_2)$ – количество вхождений слов или выражений a_1 и a_2 непосредственно рядом и в данной очередности в *D* 4. $NotNear(a_1, a_2)$ – количество вхождений слов или выражений a_1 и a_2 не рядом в D 5. $Frequency(a)$ – частота слова или выражения a в *D* 6. C_1, C_2, C_3 – настраиваемые параметры алгоритма (задаются в интерфейсе программного комплекса)
∇ Выход:	1. Новостной кластер *D* с выделенными многословными выражениями: $D=\{d_1, \dots, d_n\}$, $d_i=\{s_1, \dots, s_m\}$, $s_j=\{t_1, \dots, t_t\}$, $t_k=\{w_1, \dots, w_p\}$ $w \in W$ (Словарь), $t \in T$ (множество языковых выражений – слов и многословных выражений)

```
// Инициализируем множество объектов отдельными словами
T = W;

jointCount = MaxValue;

while(jointCount > 0)
   jointCount = 0;

   // Произвести расчет AcrossVerb, Near, NotNear, Frequency в соответствии с
   текущими значениями T и D
   CalculateContextParametersAndFrequencies(D, T);

   // Сформировать пары объектов
```

```
Pairs = {(t_i, t_j) | t_i, t_j ∈ T};

// Отсортировать пары по убыванию ф-ции схожести
Pairs.OrderByDescending(Near(t_i, t_j) – (AcrossVerb(t_i, t_j) + NotNear(t_i, t_j)));

foreach pair in Pairs
  if ( pair.Near > C1 · max( AcrossVerb (t_i, t_j ∈ T) ) AND pair.Near > 1
    AND pair.Near > C2 · ( pair.AcrossVerb + pair.NotNear )
    AND pair.Near > C3 · (Frequency (pair.t1) + Frequency (pair.t2)) )
      t_new = {pair.t1, pair.t2};

      // Заменить все вхождения t1 и t2 рядом в D на t_new
      ChangeAllOccurrences (D, t1, t2, t_new);

      T.Add (t_new);

      jointCount++;
  end-if;
 end-foreach;
end-while;
```

В результате данной процедуры собираются такие выражения, как *президент компании, международные экономические отношения, председатель совета директоров, контрольный пакет акций* и так далее.

Сборка многословных выражений является предварительным этапом для алгоритма построения тематических цепочек новостного кластера, поэтому качество самого алгоритма выделения многословных выражений является критичным и в значительной степени влияет на итоговый результат работы комплекса в целом.

Для тестирования предложенного метода сборки многословных выражений были взяты 10 новостных кластеров различной тематики, величиной более 20-30 документов в каждом. При тестировании качества метода проверялись два показателя, отражающие полноту и точность выделения многословных выражений в исходных новостных кластерах.

Точность сборки многословных выражений оценивалась как отношение количества синтаксически правильных групп среди всех выделенных выражений. Для оценки полноты выделения многословных

выражений был привлечен профессиональный лингвист, который для каждого тестового новостного кластера вручную выделил наиболее существенные для понимания смысла документов кластера многословные выражения (5-10 выражений для каждого кластера), упорядоченные в порядке снижения их значимости.

Так для кластера примера о закрытии авиабазы США на территории Киргизии лингвистом были сочтены значимыми следующие многословные выражения:

- *авиабаза Манас;*
- *парламент Киргизии;*
- *база Манас;*
- *киргизский парламент;*
- *денонсация соглашения;*
- *решение правительства.*

Отметим, что данная постановка задачи для лингвиста отличается от тестирования алгоритмов автоматического извлечения ключевых слов из текста [61], когда экспертов просят обозначить наиболее тематически значимые слова и выражения текста. В нашем же случае тестировалось именно выделение словосочетаний. Кроме того, в списке, создаваемом лингвистом, могли быть смысловые повторы (*парламент Киргизии – Киргизский парламент*).

В Табл. 3 представлены результаты оценки точности построения многословных выражений на наборе тестовых новостных кластеров. Данные результаты свидетельствуют о том, что точность предложенного алгоритма находится на уровне ~90%. Для оценки полноты сборки многословных выражений лингвистом были выделены 70 наиболее важных многословных выражений. При этом 72.6% из них были также автоматически выделены предлагаемым алгоритмом, что подтверждает высокую полноту покрытия выделяемых многословных выражений.

Всего выделено синтаксических групп, шт.	Правильных синтаксических групп, шт.	Правильных синтаксических групп, %	
364	312	Без учета частотности	**87,9**
		С учетом частотности	**91.4**

Табл. 3: Результаты оценки точности сбора многословных выражений

2.4.3.2 Формирование тематических цепочек

Алгоритм построения тематических цепочек является итеративным. На каждой итераций происходит объединение одной пары языковых выражений (установление семантической связи) или слияние двух тематических цепочек. Конструирование тематических цепочек из пар языковых выражений происходит в порядке убывания весов схожести пар языковых выражений. Общий вес схожести пары языковых выражений вычисляется как сумма весов по отдельным характеристикам схожести, описанным в разделе 2.4.2. Каждая пара получает вес, лежащий в пределах от 0 (отсутствие схожести) до 6 (максимальная схожесть), получаемый на основе шести характеристик схожести, лежащих в пределах от 0 до 1.

Пример ранжирования пар в соответствии с описанным алгоритмом приведен в Табл. 4 (топ-5 пар по общему весу на первой итерации работы алгоритма).

Характеристики / Пары	Контекстно-независимые		Контекстно-зависимые			SCORE
	BS	TS	NSF	SC	SPS	
Президент России – Президент РФ	0.66	1.00	0.00	0.50	0.68	**2.84**
Инвестгруппа – Инвестиционная группа	0.80	1.00	0.40	0.00	0.63	**2.83**
ГМК Норильский никель – Норильский никель	0.83	1.00	0.40	0.00	0.21	**2.44**
Российская Федерация – Россия	0.80	1.00	0.00	0.00	0.51	**2.31**
Отставка – Отставка с должности	0.80	1.00	0.40	0.00	0.00	**2.20**

Табл. 4: Пример ранжирования пар-кандидатов на основании характеристик схожести

Желаемые тематические цепочки должны максимально близко моделировать структуру отношений между языковыми выражениями в текстах на естественном языке. В основе конструируемых тематических цепочек лежит гипотеза глобальной связности и её прямые следствия ([22], [76]), а также концепции подхода лексических цепочек ([38], [33], [81]).

Предлагаемая структура тематической цепочки обладает следующими свойствами:

- текстовое выражение может принадлежать одной или двум тематическим цепочкам; разрешение множественной принадлежности обеспечивает возможность представления различных аспектов исходного текстового выражения, а также его лексической многозначности.
- каждая тематическая цепочка имеет главный элемент – центр тематической цепочки, который может принадлежать только одной цепочке. Центр тематической цепочки является наиболее частотным элементом среди всех элементов цепочки.

Построение тематических цепочек состоит из следующих шагов:

- Рассматривается пара текстовых выражений с наибольшим весом схожести среди всех пар-кандидатов;
- Более частотный элемент пары поглощает менее частотный элемент вместе со всеми его текстовыми вхождениями и контекстными характеристиками, и становится представителем данной пары текстовых выражений – центром новой тематической цепочки;
- Менее частотный элемент рассматриваемой пары может в дальнейшем аналогичным образом присоединиться к другой тематической цепочке;
- Объединение тематических цепочек, состоящих из нескольких элементов, происходит аналогично объединению одиночных

текстовых выражений. Центр более частотной тематической цепочки становится центром новой объединенной цепочки.

В целом каждая итерация алгоритма состоит из трех основных шагов:

1. Ранжирование пар-кандидатов (на основании характеристик схожести)
2. Выбор пары для объединения (наибольший вес + удовлетворение ограничений)
3. Процедура объединения (установление семантической связи)

Итеративный процесс продолжается до тех пор, пока есть пары-кандидаты для объединения с весом схожести выше установленного порога. Псевдокод алгоритма построения тематических цепочек:

Процедура: Построение тематических цепочек	
∇ Вход:	1. Новостной кластер *D* с выделенными многословными выражениями: $D=\{d_1, \dots, d_n\}$, $d_i=\{s_1, \dots, s_m\}$, $s_j=\{t_1, \dots, t_t\}$, $t_k=\{w_1, \dots, w_p\}$ $w \in W$ (Словарь), $t \in T$ (множество языковых выражений – слов и многословных выражений) 2. *Similarity_Score*(tc_1, tc_2) – общий вес по характеристикам схожести для тематических цепочек tc_1 и tc_2 3. *IsNSCriterion*(tc_1, tc_2) – признак выполнения ограничивающего фактора (см. Раздел 2.3) 4. C_1, C_2, C_3 – настраиваемые параметры алгоритма (задаются в интерфейсе программного комплекса)
∇ Выход:	1. Набор тематических цепочек *TC* новостного кластера *D* с выделенными центральными элементами – разметка для объектов $t \in T$: $t \to \{tc_1, \dots, tc_n\},\ 1 \le n \le 2,\ tc_i \in TC,\ tc_i = \{t_{main}[, t_1, \dots, t_k]\}$

```
// Инициализируем множество тематических цепочек отдельными языковыми выражениями
TC = T;
joinFlag = true;
while(joinFlag)
   joinFlag = false;

   // Сформировать пары цепочек, удовлетворяющих ограничению
   Pairs = {(tc_i, tc_j) | IsNSCriterion(tc_i, tc_j)=true, tc_i, tc_j ∈ TC};

   // Отсортировать пары по убыванию схожести
```

```
   Pairs.OrderByDescending(Similarity_Score(tc_i, tc_j) );

   // Выбрать пару для объединения
   { tc_i, tc_j } = Pairs[0];

   // Объединение в случае достаточной схожести
   if ( Similarity_Score(tc_i, tc_j) > C)
      if ( Frequency(tc_i) > Frequency(tc_j) )
         tc_new={t_main=t_main_i, t_i1, ... , t_in, t_j1, ... , t_jm};
         TC.Remove(tc_i);
      else
         tc_new={t_main=t_main_j, t_i1, ... , t_in, t_j1, ... , t_jm};
         TC.Remove(tc_j);
      end-if;

      // Произвести расчет характеристик для новой пары tc_new
      CalculateParameters (D, TC, tc_new);

      TC.Add ( tc_new );
      joinFlag = true;
   end-if;
end-while;
```

Например, тематическая цепочка с центральным элементом *пост* проходит следующие этапы в процессе построения (показаны пары с максимальным весом схожести на разных итерациях; более частотный элемент пары является первым элементом):

Итерация 7: (***Отставка)*** ← ***(Отставка с должности)***

Итерация 33: (***Отставка,*** *Отставка с должности)* ← ***(Уход в отставку)***

Итерация 44: (***Отставка,*** *Отставка с должности, Уход в отставку)* ← ***(Отставка президента)***

Итерация 61: (***Уход с поста)*** ← ***(Уход в отставку)***

Итерация 62: (***Отставка,*** *Отставка с должности, Уход в отставку, Отставка президента)* ← ***(Уход с поста,*** *Уход в отставку)*

Итерация 102: (***Отставка,*** *Отставка с должности, Уход в отставку, Отставка президента, Уход с поста)* ← ***(Пост)***

Итерация 103: (***Пост,*** *Отставка, Отставка с должности, Уход в отставку, Отставка президента, Уход с поста)* ← ***(Должность)***

Итерация 104: (***Пост,*** *Отставка, Отставка с должности, Уход в отставку, Отставка президента, Уход с поста, Должность)* ← ***(Уход)***

Следующие тематические цепочки были получены в результате работы описанного алгоритма для кластера примера. Представлены 5 наиболее частотных тематических цепочек, в порядке убывания их частоты. Данные цепочки не подвергались какой-либо постобработке, центры тематических цепочек выделены жирным шрифтом:

Пост: *уход с поста; должность; уход; отставка; отставка с должности; уход в отставку; отставка президента*

Алроса: *президент Алроса; АК Алроса*

Компания: *акция компании; владелец компании; объединение компаний; акция; акционер компании; владелец; пакет акций; состав владельцев; контрольный пакет акций; контрольный пакет; владение*

Ничипорук*: Александр Ничипорук*

Якутия*: президент Якутии; якутский; якутский президент*

Общая схема работы алгоритма построения тематических цепочек представлена на Рис. 6.

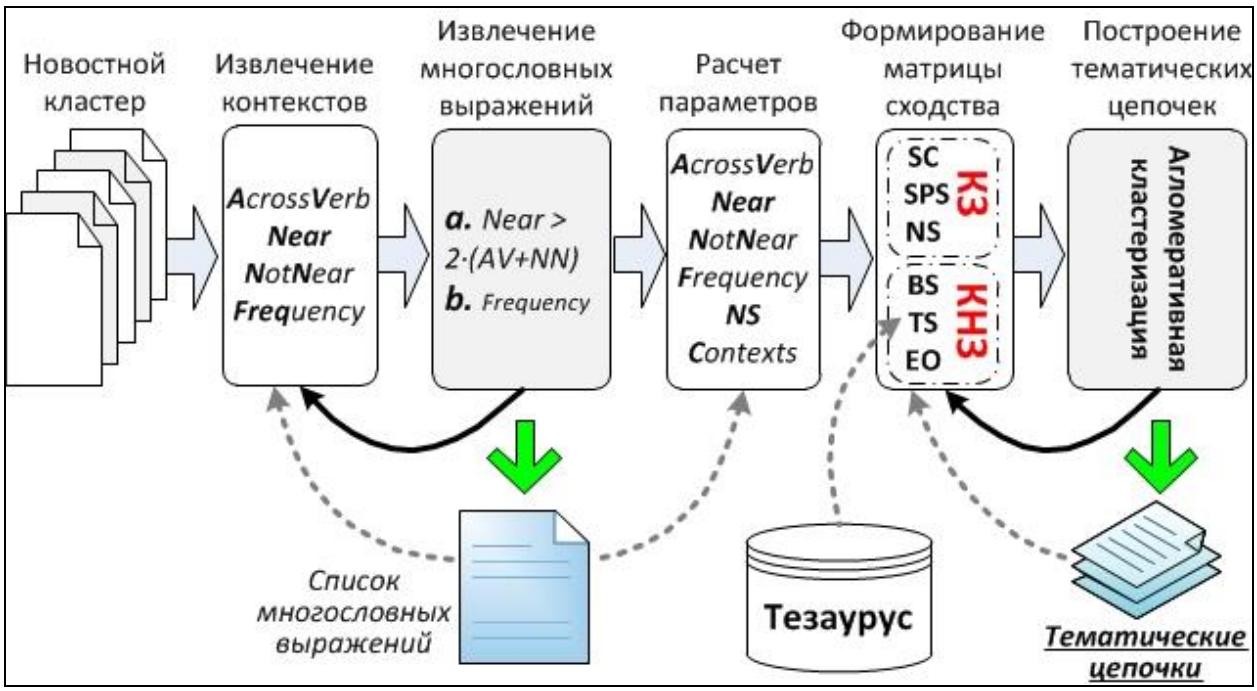


Рис. 6: Общая схема алгоритма построения тематических цепочек

2.5 Алгоритмическая сложность и производительность алгоритма построения тематических цепочек

Описанный алгоритм построения тематических цепочек состоит из двух основных этапов:

1. Формирование матрицы сходства
2. Построение тематических цепочек (итерация алгоритма)

Каждый из этапов обладает различной алгоритмической сложностью, основанной на количестве языковых выражений n, находящихся в рассмотрении (содержащихся в новостном кластере). При этом необходимо отметить, что рост количества языковых выражений и количество документов в новостном кластере связаны, но не обязательно данная зависимость является линейной. В реальных новостных кластерах одни и те же языковые выражения многократно повторяются, добавление новых документов в некоторых случаях может не приводить к увеличению количества языковых выражений.

Первый этап алгоритма предполагает попарное рассмотрение всех языковых выражений, т.е. в худшем случае обладает квадратичной сложностью $O(n^2)$. В практических задачах отмечено, что матрица сходства

является сильно разреженной (вес схожести между большинством пар равен нулю), в связи с чем рассматривается только часть пар языковых выражений.

В рамках второго этапа происходит построение тематической цепочки из двух частей, для чего необходимо рассмотрение связей данных частей с другими языковыми выражениями. Таким образом, в худшем случае необходимо рассмотреть *2·n* связей, а алгоритмическая сложность второго этапа, соответственно, является линейной - *O(2·n)*. Практические результаты скорости работы модуля построения тематических цепочек (см. Раздел 4.2) для различных новостных кластеров подтверждают теоретическую оценку сложности:

Новостной кластер	Количество языковых выражений, шт.	Количество документов, шт.	Формирование матрицы сходства, мс	Среднее время итерации, мс
Алроса	**631**	**12**	**332**	**37**
Снег	**460**	**8**	**109**	**26**
Авиабаза	**2132**	**195**	**5696**	**112**
Грипп	**142**	**5**	**17**	**8**
Талибы	**358**	**20**	**106**	**22**

Табл. 5: Результаты скорости работы модуля построения тематических цепочек

Общая сложность предложенного алгоритма построения тематических цепочек новостного кластера с *m* итерациями работы имеет следующий вид:

$$O(n^2) + m \cdot O(2 \cdot n), \quad где\, n - количество\, языковых\, выражений$$

2.6 Влияние лексической вариативности на установление схожести

Вариативность естественного языка оказывает значительное влияние на решение базовых задач автоматической обработки текстов. Одной из таких задач является установление схожести фрагментов текстов, в частности установление схожести предложений. Классическим подходом для решения данной задачи является сравнение векторных представлений предложений по косинусной мере угла между векторами:

$$\cos(\Theta) = \frac{\vec{s_1} \cdot \vec{s_2}}{\left\|\vec{s_1}\right\| \cdot \left\|\vec{s_2}\right\|}$$

Наиболее простым и широко употребляемым вариантом является использование пословного представления, т.е. размерности векторов предложений соответствуют отдельным словам предложений:

$$\vec{s_1} = (w_1^1, \ldots, w_n^1), \quad \vec{s_2} = (w_1^2, \ldots, w_m^2), \quad w_1^1, \ldots, w_n^1, w_1^2, \ldots, w_m^2 \in W$$

Рассмотрим, как ведет себя мера схожести между предложениями в базовом случае и в случае усложнения пословного представления с помощью применения следующих операций:

- Добавление многословного выражения $w_i^1 w_j^1$ в предложение s_1 - операция $f_{MWE}(w_i^1, w_j^1, \vec{s_1})$:

  $$f_{MWE}(w_i^1, w_j^1, \vec{s_1}) : \vec{s_1} = (w_1^1, \ldots, w_i^1, \ldots, w_j^1, \ldots, w_n^1) \rightarrow (w_1^1, \ldots, w_i^1 w_j^1, \ldots, w_n^1)$$

- Установление сходства между выражениями w_i^1 и w_j^2 в предложениях s_1 и s_2 - операция $f_{TC}(w_i^1, w_j^2, tc, \vec{s_1}, \vec{s_2})$:

  $$f_{TC}(w_i^1, w_j^2, tc, \vec{s_1}, \vec{s_2}) : \begin{cases} \vec{s_1} = (w_1^1, \ldots, w_i^1, \ldots, w_n^1) \rightarrow (w_1^1, \ldots, tc, \ldots, w_n^1) \\ \vec{s_2} = (w_1^2, \ldots, w_j^2, \ldots, w_m^2) \rightarrow (w_1^2, \ldots, tc, \ldots, w_m^2) \end{cases}$$

- Последовательное применение операций добавления многословного выражения $w_i^1 w_j^1$ в предложение s_1 и установления сходства выражений w_r^1 и w_m^2 из предложений s_1 и s_2 соответственно: $f_{TC}(w_r^1, w_m^2, tc, f_{MWE}(w_i^1, w_j^1, \vec{s_1}), \vec{s_2})$.

Пусть предложения s_1 и s_2 имеют длины L_1 и L_2 соответственно, бинарные веса *{0, 1}* размерностей (для простоты расчетов), и содержат k одинаковых слов $w_1^1 = w_1^2, \ldots, w_k^1 = w_k^2$. Тогда вес схожести будет вычисляться следующим образом:

$$sim_{words}(\vec{s_1}, \vec{s_2}) = \cos_{words}(\Theta) = \frac{\vec{s_1} \cdot \vec{s_2}}{\left\|\vec{s_1}\right\| \cdot \left\|\vec{s_2}\right\|} = \frac{k}{\sqrt{L_1} \cdot \sqrt{L_2}}$$

При добавлении информации о многословных выражениях возможно как уменьшение, так и увеличении меры схожести между предложениями. Рассмотрим подробно варианты объединения слов w_i^1 и w_j^1 в многословное выражение (применение операции $f_{MWE}(w_i^1, w_j^1, \vec{s_1})$). Возможно четыре варианта:

1. $(i <= k) \wedge (j <= k) \wedge (f_{MWE}(w_i^1, w_j^1, \vec{s_1}) <> f_{MWE}(w_i^2, w_j^2, \vec{s_2})) : sim_{MWE}(\vec{s_1}, \vec{s_2}) = \frac{k-2}{\sqrt{L_1 - 1} \cdot \sqrt{L_2}}$
2. $(i <= k) \wedge (j <= k) \wedge (f_{MWE}(w_i^1, w_j^1, \vec{s_1}) = f_{MWE}(w_i^2, w_j^2, \vec{s_2})) : sim_{MWE}(\vec{s_1}, \vec{s_2}) = \frac{k-1}{\sqrt{L_1 - 1} \cdot \sqrt{L_2 - 1}}$
3. $(i <= k) \wedge (j > k) : sim_{MWE}(\vec{s_1}, \vec{s_2}) = \frac{k-1}{\sqrt{L_1 - 1} \cdot \sqrt{L_2}}$
4. $(i > k) \wedge (j > k) : sim_{MWE}(\vec{s_1}, \vec{s_2}) = \frac{k}{\sqrt{L_1 - 1} \cdot \sqrt{L_2}}$

Таким образом, добавление информации о многословных выражениях может приводить как к уменьшению, так и к увеличению меры схожести между предложениями, в зависимости от соотношений между параметрами *i, j, k, L_1, L_2*.

При установлении схожести между словами w_i^1 и w_j^2 (применение операции $f_{TC}(w_i^1, w_j^2, tc, \vec{s_1}, \vec{s_2})$ или операции $f_{TC}(t_i^1, t_j^2, tc, \vec{s_1}, \vec{s_2})$ в случае перехода к пространству многословных выражений) происходит сокращение размерности общего пространства предложений s_1 и s_2, приводящее к увеличению скалярного произведения между ними:

$$sim_{TC}(\vec{s_1}, \vec{s_2}) = \frac{k+1}{\sqrt{L_1} \cdot \sqrt{L_2}}$$

Интеграция тематических цепочек, представленных в Разделе 2.4, в пословное представление текстов предполагает комбинирование операций добавления многословных выражений $f_{MWE}(w_i^1, w_j^1, \vec{s_1})$ и установления

схожести $f_{TC}(w_i^1, w_j^2, tc, \overrightarrow{s_1}, \overrightarrow{s_2})$. В связи с чем чрезвычайно значимым становится утверждение следующей леммы:

Лемма 1: Последовательное применение операций добавления многословного выражения $f_{MWE}(w_i^1, w_j^1, \overrightarrow{s_1})$ и установления схожести $f_{TC}(w_r^1, w_m^2, tc, \overrightarrow{s_1}, \overrightarrow{s_2})$ при выполнении условия на установление схожести для одной из частей многословного выражения (*) [1] приводит к ***увеличению*** косинусной меры схожести между предложениями.

$$f_{MWE}(w_i^1, w_j^1, \overrightarrow{s_1}) \wedge (w_i^1 \in s_2) \wedge (w_j^1 \in s_2) \Rightarrow \exists tc`: (w_i^1 w_j^1 \in tc`) \wedge ((w_i^1 \in tc`) \vee (w_j^1 \in tc`)) \ (*)$$

Доказательство: Необходимо отметить, что добавление многословного выражения предполагает наличие как минимум двух его исходных частей, что влечет за собой ограничение на минимальную длину предложения s_1: $L_1>1$. Косинусная мера схожести между предложениями до применения операций добавления многословного выражения и установления схожести имеет следующий вид:

$$sim_{\text{before}}(\overrightarrow{s_1}, \overrightarrow{s_2}) = \cos_{before}(\Theta) = \frac{\overrightarrow{s_1} \cdot \overrightarrow{s_2}}{\left\|\overrightarrow{s_1}\right\| \cdot \left\|\overrightarrow{s_2}\right\|} = \frac{k}{\sqrt{L_1} \cdot \sqrt{L_2}}$$

Для доказательства леммы отдельно рассмотрим три возможных варианта:

1. Ни одна из частей добавляемого многословного выражения $f_{MWE}(w_i^1, w_j^1, \overrightarrow{s_1})$ не встречалась в предложении s_2: $(i > k) \wedge (j > k)$. Мера схожести между предложениями s_1 и s_2 имеет следующий вид:

$$sim_{after}(\overrightarrow{s_1}, \overrightarrow{s_2}) = \frac{k+1}{\sqrt{L_1 - 1} \cdot \sqrt{L_2}} \quad > \quad \frac{\mathrm{k}}{\sqrt{\mathrm{L}_1} \cdot \sqrt{\mathrm{L}_2}} = sim_{\text{before}}(\overrightarrow{s_1}, \overrightarrow{s_2}), \forall k >= 0, L_1 > 1, L_2 > 0$$

[1] Добавление многословного выражения $w_i^1 w_j^1$ в предложение s_1 в случае вхождения компонентов данного выражения w_i^1 и w_j^1 в предложение s_2 требует установления дополнительной связи нового выражения $w_i^1 w_j^1$ с одним из его компонентов

2. Только одна из частей добавляемого многословного выражения $f_{MWE}(w_i^1, w_j^1, \vec{s_1})$ встречалась в предложении s_2: $(i <= k) \wedge (j > k)$. Данный вариант подразумевает наличие как минимум одного одинакового элемента в предложениях s_1 и s_2, т.е. $k>0$. Мера схожести между предложениями s_1 и s_2 имеет следующий вид:

$$sim_{after}(\vec{s_1}, \vec{s_2}) = \frac{k}{\sqrt{L_1 - 1} \cdot \sqrt{L_2}} > \frac{\mathrm{k}}{\sqrt{\mathrm{L}_1} \cdot \sqrt{\mathrm{L}_2}} = sim_{\mathrm{before}}(\vec{s_1}, \vec{s_2}), \forall k > 0, L_1 > 1, L_2 > 0$$

3. Обе части добавляемого многословного выражения $f_{MWE}(w_i^1, w_j^1, \vec{s_1})$ встречались в предложении s_2 непосредственно рядом: $(i <= k) \wedge (j <= k) \wedge f_{MWE}(w_i^1, w_j^1, \vec{s_2})$. Данный вариант подразумевает наличие как минимум двух одинаковых элементов в предложениях s_1 и s_2, т.е. $k>1$. Мера схожести между предложениями s_1 и s_2 в данном случае принимает следующий вид (новое многословное выражение $w_i^1 w_j^1$ встречается в обоих предложениях s_1 и s_2 и учитывается при вычислении косинусной меры):

$$sim_{after}(\vec{s_1}, \vec{s_2}) >= \frac{k}{\sqrt{L_1 - 1} \cdot \sqrt{L_2 - 1}} > \frac{\mathrm{k}}{\sqrt{\mathrm{L}_1} \cdot \sqrt{\mathrm{L}_2}} = sim_{\mathrm{before}}(\vec{s_1}, \vec{s_2}), \forall k > 1, L_1 > 1, L_2 > 0$$

4. Обе части добавляемого многословного выражения $f_{MWE}(w_i^1, w_j^1, \vec{s_1})$ встречались в предложении s_2 раздельно: $(i <= k) \wedge (j <= k) \wedge NOT(f_{MWE}(w_i^1, w_j^1, \vec{s_2}))$. Данный вариант является наиболее сложным, и подразумевает наличие как минимум двух одинаковых элементов в предложениях s_1 и s_2, т.е. $k>1$. Кроме того, по условию леммы в случае вхождения обеих составных частей конструируемого многословного выражения (выполнения условия $(w_i^1 = w_i^2 \in s_2) \wedge (w_j^1 = w_j^2 \in s_2)$), в дополнение к установлению схожести $f_{TC}(w_r^1, w_m^2, tc, \vec{s_1}, \vec{s_2})$ также будет установлена одна из связей с частями выражения $w_i^1 w_j^1$: $\exists tc : (w_i^1 w_j^1 \in tc) \wedge ((w_i^1 \in tc) \vee (w_j^1 \in tc))$. Мера схожести между предложениями s_1 и s_2 в данном случае принимает вид:

$$sim_{after}(\vec{s_1},\vec{s_2}) >= \frac{k}{\sqrt{L_1-1}\cdot\sqrt{L_2}} \quad > \quad \frac{\mathrm{k}}{\sqrt{\mathrm{L}_1}\cdot\sqrt{\mathrm{L}_2}} = sim_{\mathrm{before}}(\vec{s_1},\vec{s_2}),\ \forall k>1,\ L_1>1,\ L_2>0$$

Таким образом, рассмотрены все варианты последовательного применения операций добавления многословного выражения $f_{MWE}(w_i^1, w_j^1, \vec{s_1})$ и установления схожести $f_{TC}(w_r^1, w_m^2, tc, \vec{s_1}, \vec{s_2})$, доказано увеличение меры схожести между предложениями s_1 и s_2, *что и требовалось доказать*:

$\forall k,\ L_1,\ L_2 \in Z,\ k >= 0,\ \ L_1 > 1,\ L_2 > 0:$

$$\min(sim_{\mathrm{after}}(\vec{s_1},\vec{s_2})) = \frac{k}{\sqrt{L_1-1}\cdot\sqrt{L_2}} \quad > \quad \frac{\mathrm{k}}{\sqrt{\mathrm{L}_1}\cdot\sqrt{\mathrm{L}_2}} = sim_{before}(\vec{s_1},\vec{s_2})$$

Одним из условий Леммы 1 является установление схожести для одной из частей многословного выражения (*), которое при первичном анализе видится довольно искусственным:

$$f_{MWE}(w_i^1, w_j^1, \vec{s_1}) \wedge (w_i^1 \in s_2) \wedge (w_j^1 \in s_2) \Rightarrow \exists tc : (w_i^1 w_j^1 \in tc) \wedge ((w_i^1 \in tc) \vee (w_j^1 \in tc)) \quad (*)$$

В то же время данное условие имеет практическое обоснование, которое изначально заложено в одной из характеристик схожести предложенного алгоритма построения тематических цепочек – характеристике ***BS*** (см. Раздел 2.4.2), использующей естественную идею обнаружения семантически-связанных объектов, а именно использования формального сходства анализируемых выражений. В случае выполнения условия (*) выражения кандидаты будут гарантированно иметь близкое к максимальному значение по характеристике схожести ***BS***.

2.7 Выводы ко второй главе

В данной главе приведен обзор природы появления вариативности в текстах на естественном языке, описаны существующие подходы к выявлению вариативности, а также разработана формальная модель описания участников ситуации новостного кластера с учетом вариативности их именования.

На основании разработанной модели был предложен алгоритм построения тематических цепочек – структур, соответствующих основным участникам новостного кластера. В Главе 3 производится интеграция построенных тематических цепочек в методы автоматического аннотирования.

3. Интеграция тематических цепочек в методы автоматического аннотирования

Сконструированные тематические цепочки, являющиеся моделью описания совокупности участников ситуации, обсуждаемой в рамках исследуемого новостного кластера, сами по себе не является практически полезными (по крайней мере, варианты практического использования не были найдены на момент написания данной работы). Но при этом построенные структуры несут в себе дополнительную информацию о внутреннем устройстве новостного кластера. Гипотеза проведенного исследования заключалась в том, что полученные структуры могут повысить качество решения других задач автоматической обработки текстов, имеющих практическое значение.

Одной из таких важных прикладных задач является автоматическое аннотирование (см. Главу 1), качество решения которой в значительной степени зависит от наличия информации о лексико-семантической вариативности, содержащейся в анализируемой текстовой коллекции. Построенные тематические цепочки содержат в себе данную информацию, поэтому предполагается, что интеграция тематических цепочек в методы автоматического аннотирования должна улучшать общее качество полученных автоматических аннотаций.

Проверка описанной гипотезы производится двумя способами. Во-первых, в Разделе 3.1 предлагается алгоритм интеграции построенных тематических цепочек в известные метода автоматического аннотирования, такие как Maximal Marginal Relevance (MMR, см. Раздел 1.2.6.1) и SumBasic (см. Раздел 1.2.2.1). Предполагается, что интеграция построенных тематических цепочек должна улучшить общее качество исходных методов. Во-вторых, в Разделе 3.2 предлагается два новых метода автоматического аннотирования, опирающихся исключительно на предоставленные тематические цепочки. В работах [80] и [81] предлагается алгоритм

построения тематического представления на основе единственной характеристики схожести – информация о связи по тезаурусу РуТез (см. Раздел 1.2.3.1), а также алгоритм аннотирования на основе данного тематического представления. Предложенная в рамках данной кандидатской диссертации модель тематических цепочек строится на основе объединения нескольких разнородных характеристик схожести (см. Раздел 2.4.2), и предполагает более высокое качество построенной модели. По причине чего предполагается, что методы автоматического аннотирования на основе тематических цепочек, обогащенных новыми характеристиками схожести, также будут показывать лучшие результаты.

Оценка качества всех полученных автоматических аннотаций производится с помощью автоматических мер качества ROUGE (см. Раздел 1.3.1). Для подтверждения полученных результатов лучшие методы дополнительно подвергались оценке методом «Пирамид» (см. Раздел 1.3.2).

3.1 Интеграция в существующие методы аннотирования

Большинство существующих методов автоматического аннотирования работают на основе пословной модели (bag-of-words model) представления. В рамках данной модели входная коллекция (документы, предложения) представляются векторами, размерности которых соответствуют отдельным словам. Веса для данных размерностей вычисляются на основе значимости или информативности соответствующих им слов. На основе данной модели работают и рассматриваемые методы MMR и SumBasic. Описанная пословная модель не подразумевает учета различных вариантов именования одних и тех же сущностей в рамках входной коллекции документов, вследствие чего алгоритмы автоматического аннотирования интерпретируют все слова как несвязанные сущности.

Интеграция построенных тематических цепочек в методы автоматического аннотирования предполагает уход от оперирования словом как атомарной единицей, анализируемой при вычислении информативности

предложений. Вместо слов предлагается использовать ***тематические цепочки*** (thematic chain, ***tc***), каждая из которых является описанием некоторого участника ситуации или сущности входной коллекции, агрегируя в себе все варианты её именования, используемые в рамках данной коллекции. Таким образом, предлагаемая модель заключается в совершении двухступенчатого перехода от слова к тематической цепочке, как атомарной единицы рассмотрения:

Слова → *Объекты (слова + мног.выр.)* → *Тематические цепочки*

I. Замена слов на многословные выражения. Добавление информации о многословных выражениях – переход от слова к объекту (отдельные слова или многословные выражения);

II. Добавление информации о принадлежности объектов тематическим цепочкам. В рамках предлагаемой модели тематических цепочек каждый объект может принадлежать к одной или двум цепочкам.

Каждая тематическая цепочка имеет вес, равный сумме частот его элементов (объектов):

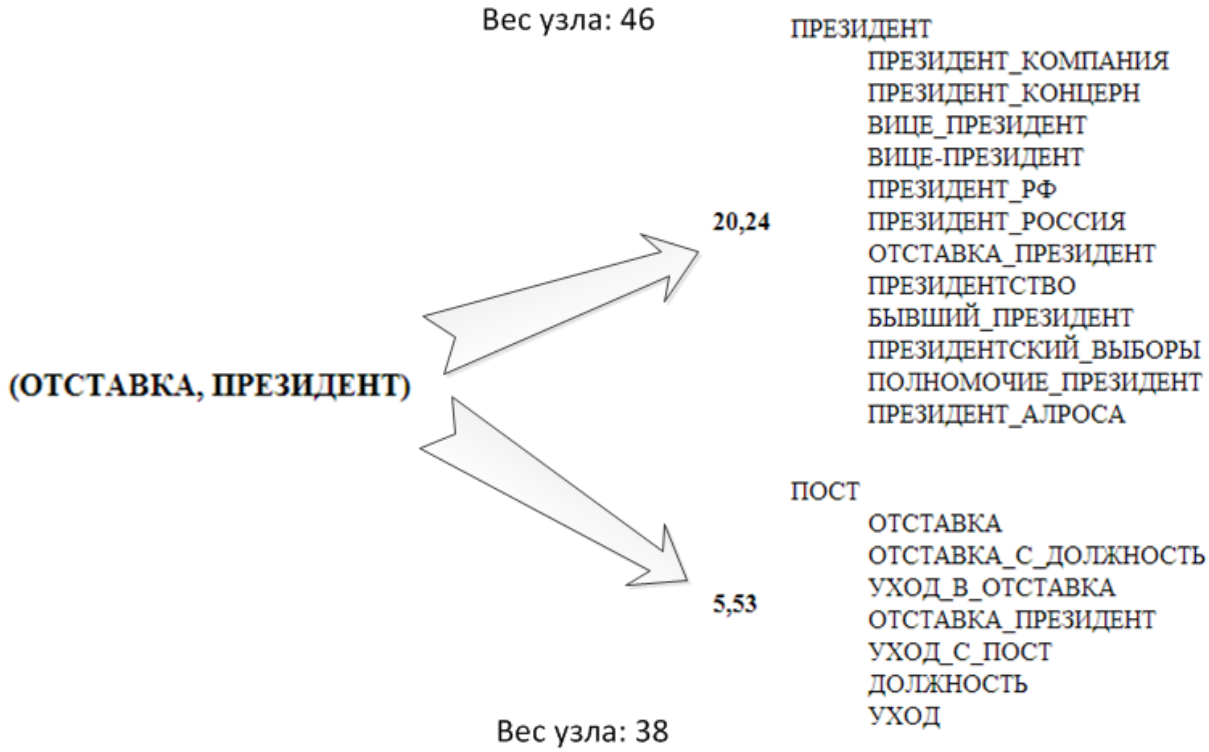
$$weight(tc) = \sum_{tc_{elem\_i} \in tc} freq(tc_{elem\_i})$$

Предложение представляет собой последовательность объектов, каждый из которых является либо элементом одной или двух тематических цепочек, либо является центром какой-либо тематической цепочки (см. Раздел 2.4.3.2). Элементы цепочек имеют вес схожести с центральным элементом, равный отношению набранного суммарного балла по характеристикам схожести (при построении данной тематической цепочки) к максимально возможному баллу схожести:

$$weight(tc_{elem}) = \frac{similarity(tc_{elem}, tc_{center})}{\max\limits_{tc_{elem\_i} \in tc} (similarity(tc_{elem\_i}, tc_{center}))} \cdot weight(tc)$$

Вес центрального элемента цепочки равен весу всей тематической цепочки. Таким образом, веса объектов разбиваются на веса в

соответствующих им тематических цепочках, пропорционально силе связи с центрами цепочек. Например, для кластера, посвященного смене руководства алмазодобывающей компании «Алроса», объект *(отставка президента)* был сопоставлен двум тематическим цепочкам:



Данная информация о принадлежности объектов тематическим цепочкам, с учетом указанной логики вычисления весов, может быть использована для интеграции построенных тематических цепочек в методы автоматического аннотирования.

3.1.1 Учет TF·IDF для многословных выражений

Важной частью многих методов автоматического аннотирования является вычисление информативности для отдельных слов. Наиболее популярной является стратегия TF·IDF (см. Раздел 1.2.2.1), расчет которой связан с анализом вхождений слова в некоторую вспомогательную коллекцию. Интеграция тематических цепочек в методы автоматического аннотирования и, в частности, переход от слов к единому рассмотрению слов и многословных выражений (объектам) подразумевает изменение схемы расчета информативности. При этом непосредственный переход от

пословных алгоритмов расчета информативности к учету многословных выражений не является очевидный. В рамках данной работы был разработан вариант вычисления IDF для многословных выражений по следующей формуле:

$$MWE\_IDF = \ln\left(\frac{doc\_count}{[\prod_{w_i \in MWE} freq(w_i)]/[doc\_count]^{N-1}}\right)$$

где *MWE* - многословное выражение (**M**ulti-**W**ord **E**xpression), *doc_count* – число документов во вспомогательной коллекции, $freq(w_i)$ – документная частота слова w_i во вспомогательной коллекции.

3.1.2 Интеграция в метод MMR

В основе алгоритма MMR лежит сопоставление пословных векторов предложений, содержащих информацию о значимости отдельных слов, друг с другом и с общим вектором входной текстовой коллекции. Интеграцию построенных тематических цепочек предлагается делать путем замены размерностей отдельных слов и многословных выражений на соответствующие им размерности тематических цепочек. Таким образом, на основании принципа, изложенного в начале данного раздела, базовые вектора предложений MMR на основе пословного представления, заменяются векторами по тематическим цепочкам. Например, для языкового выражения *ОСТАВКА ПРЕЗИДЕНТА*, упомянутого выше и сопоставленного двум тематическим цепочкам [*ПРЕЗИДЕНТ*; *ПРЕЗИДЕНТ КОМПАНИИ*; …] и [*ПОСТ*; *ОТСТАВКА*; …] преобразования векторов предложений будет происходить следующим образом:

$$(\ldots, ОТСТАВКА\ \ ПРЕЗИДЕНТА, \ldots) \rightarrow (\ldots, 4.0, \ldots)$$

$$\Downarrow$$

$$(\ldots, [\text{ПРЕЗИДЕНТ};\ldots], [ПОСТ;\ldots], \ldots) \rightarrow (\ldots, 20.24, 5.53, \ldots)$$

Общий вектор для входной коллекции при этом заменяется вектором, агрегирующим все тематические цепочки построенной модели. Все остальные этапы работы метода MMR остаются без изменений.

3.1.3 Интеграция в метод SumBasic

Базовый вариант работы алгоритма SumBasic (см. Раздел 1.2.2.1) на основе пословного представления предполагает жадный итеративный процесс отбора предложений *s* в результирующую аннотацию, обладающих наибольшей средней вероятностью слов в рамках текущей итерации:

$$weight\ (s) = \sum w_i \in s \frac{p(w_i)}{|\ \{\ w \mid w \in s_j\ \}\ |}, \quad \mathrm{p(w_i)} = \frac{\mathrm{n}}{\mathrm{N}}$$

где вероятность слова $p(w_i)$ определяется отношением числа появлений слова w_i в исходной коллекции *n* к общему числу слов в данной коллекции *N*.

При интеграции построенных тематических цепочек в метод SumBasic и осуществлении перехода от рассмотрения отдельных слов к рассмотрению тематических цепочек, отбор предложений начинает осуществляться на основании средней вероятности объектов (слов или многословных выражений). Каждый из объектов *O* сопоставляется одной или двум тематическим цепочкам tc_1 и tc_2, имея вес (и, соответственно, вероятность) в каждой из данных цепочек:

$$p(O) = \{\ sim(O, tc_1) \cdot p(tc_1)\ \ [\ , sim(O, tc_2) \cdot p(tc_2)]\ \}$$

Итоговой вероятностью объекта при расчете средней вероятности выступает максимальная из вероятностей по соответствующим тематическим цепочкам для данного объекта:

$$p_{final}(O) = \max(\ p(O))$$

Вес предложения *S* в рамках данной модели вычисляется следующим образом:

$$score\ (S) = \sum_{O \in S} p_{final}(O) \Big/ count\ (o \in S)$$

Важной частью алгоритма SumBasic является пересчет вероятностей при добавлении предложения в итоговую аннотацию. В классическом варианте вероятности всех слов w, входящих в выбранное предложение, квадратично уменьшаются:

$$p_{new}(w) = p_{old}(w) \cdot p_{old}(w)$$

При отборе предложений на основании объектов и соответствующих им тематическим цепочкам данная модель усложняется:

- Уменьшаются вероятности для всех элементов тематических цепочек, которые соответствуют объектам отобранного предложения;
- Вводится дифференциация при уменьшении вероятностей для центральных и обычных элементов тематических цепочек.

Таким образом, общая схема понижения вероятностей для отобранного в итоговую аннотацию предложения S выглядит следующим образом:

$$\begin{aligned}
&\textit{foreach}\ (O \in S)\\
&\quad \textit{if}\ (\textit{IsMain}\ (O))\\
&\qquad \textit{foreach}\ (O_{tc} \in \{tc \mid O \in tc\})\\
&\qquad\quad p_{new}(O_{tc}) = p_{old}(O_{tc}) \cdot p_{old}(O_{tc})\\
&\quad \textit{else}\\
&\qquad p_{new}(O) = p_{old}(O) \cdot p_{old}(O)\\
&\qquad \textit{foreach}\ (O_{tc} \in \{tc \mid O \in tc\})\\
&\qquad\quad p_{new}(O_{tc}) = p_{old}(O_{tc}) \cdot (1 - \textit{Sim}\ (O, tc))
\end{aligned}$$

Остальная логика работы алгоритма SumBasic используется без изменений.

3.2 Новые методы аннотирования на основе построенных тематических цепочек

В работе [80] предложен метод автоматического аннотирования новостных кластеров на основе тематического представления на базе тезауруса русского языка РуТез (см. Раздел 1.2.3.1). Данное тематическое представление строится на основе единственной характеристики схожести – наличие связи в предопределенном ресурсе (тезаурус РуТез). Разработанная в рамках данной кандидатской диссертации модель тематических цепочек комбинирует несколько разнородных характеристик схожести.

Предполагается, что это позволит сформировать более полную и комплексную структуру, отражающую максимальное число взаимосвязей, присутствующих во входной коллекции документов.

Для сравнения качества построенных тематических цепочек и тематического представления на основе тезауруса РуТез были разработаны два новых метода автоматического аннотирования, базирующихся на информации из построенных тематических цепочек:

- Отбор предложений на основе участников ситуации (по тематическим цепочкам)
- Отбор предложений на основе взаимоотношений участников ситуации (по связям тематических цепочек)

Предлагаемые алгоритмы используют аналогичные работе [80] идеи, а именно учет особенностей устройства текстов на естественном языке и избыточности, в них имеющейся ([22]).

3.2.1 Построение аннотации по тематическим цепочкам

Классической моделью выделения значимой информации является анализ наиболее частотных сущностей – то, что часто повторяется, с большой вероятностью является важным. Сформированные тематические цепочки характеризуют основных участников ситуации и обладают весами, рассчитываемыми на основе частотных характеристик элементов данной цепочки. Эти веса используются как основа метода автоматического аннотирования, который представляет собой жадный алгоритм отбора предложений, включающих в себя наиболее значимые цепочки. Для избегания повышенного веса длинных предложений, содержащих большое количество различных сущностей, в рассмотрение попадают только три наиболее весомые тематические цепочки предложения. В рамках каждой итерации i отбирается по одному предложению s_i, содержащему 3 тематических цепочки с наибольшим весом и ещё не упомянутых в предложениях, отобранных в итоговую аннотацию на более ранних этапах работы алгоритма:

$$s_i \Rightarrow \max\left(\sum_{tc_{new\_j} \in s_i,\ j=1..3}^{desc\ weight(tc_{new})} weight\ (tc_{new\_j})\right)$$

где *tc* – тематическая цепочка, tc_{new} – новая тематическая цепочка, не упомянутая в уже отобранных предложениях. Итеративный процесс отбора предложений завершается при достижении ограничения на длину итоговой аннотации.

3.2.2 Построение аннотации по связям тематических цепочек

Гипотеза Ван Дика об иерархическом устройстве тематической структуры текстов на естественном языке ([76]) является основой второго предлагаемого метода автоматического аннотирования, на основе связей тематических цепочек. В соответствии с данной гипотезой, отдельные темы документа представляют собой отношения между основными участниками ситуации, которые раскрываются в отдельных предложениях текста. При рассмотрении новостного кластера это означает, что сущности, наиболее часто употребляемые в одних и тех же предложениях исходной коллекции, являются также наиболее значимыми для общей тематики исходной коллекции. Предлагаемый алгоритм построения автоматической аннотации по связям тематических цепочек является итеративным, в рамках каждой итерации отбирается предложение, такое что, если:

- $tc_{rel} = \{tc_1, tc_2\}$ – пара тематических цепочек;
- $weight(tc_{rel})$ – число вхождений пары в одни и те же предложения кластера;
- tc_{rel_new} – новая пара тематических цепочек, не упомянутая в одних и тех же предложениях, уже отобранных в аннотацию.

то обязательным условием включения предложения s_i в итоговую аннотацию является наличие в нем неупомянутой пары тематических цепочек, обладающей наибольшим весом:

$$s_i \supset \max_{tc_{rel\_new} \in Cluster} \left(weight(tc_{rel\_new})\right)$$

Предложений, удовлетворяющих данному условию, может быть несколько, из которых дополнительным фильтром выбирается одно, обладающее наибольшим суммарным весом неупомянутых пар тематических цепочек:

$$s_i \Rightarrow \max\left(\sum_{tc_{rel\_new\_j} \in s_i} weight(tc_{rel\_new\_j}) \right)$$

Итеративный процесс отбора предложений завершается при достижении ограничения на длину итоговой аннотации или отсутствия предложений, удовлетворяющих критериям отбора.

3.3 Оценка автоматических аннотаций и основные результаты

Оценка качества полученных автоматических аннотаций, а именно сравнение модификаций методов с интеграцией и без интеграции построенных тематических цепочек является мерой качества самих тематических цепочек. Для проведения оценки были подготовлены 11 новостных кластеров различной тематики (спорт, политика, происшествия), построенных на основе пословной модели одним из известных алгоритмов кластеризации ([78]). К каждому из данных кластеров профессиональными лингвистами были подготовлены от 2 до 4 ручных аннотаций. Оценке подверглись следующие методы и их модификации (всего 10 модификаций):

I. Модификации метода Maximal Marginal Relevance (MMR)

1. Классический MMR («Classic MMR»)
2. MMR с интегрированными тематическими цепочками без учета IDF («MMR + Groups»)
3. MMR с интегрированными тематическими цепочками и учетом IDF («MMR With IDF + Groups»)

II. Модификации метода SumBasic

1. Классический SumBasic («SumBasic»)

2. SumBasic с интегрированными тематическими цепочками («SumBasic + Groups»)

III. Методы аннотирования на основе тематических моделей

1. Аннотирование на основе тематического представления на базе тезауруса РуТез («ThematicLines»)
2. Аннотирование на основе построенных тематических цепочек, без учета IDF («OurSummary (Nodes)»)
3. Аннотирование на основе построенных тематических цепочек, с учетом IDF («OurSummary (Nodes) With IDF»)
4. Аннотирование на основе связей построенных тематических цепочек, без учета IDF («OurSummary (Relations)»)
5. Аннотирование на основе связей построенных тематических цепочек, с учетом IDF («OurSummary (Relations) With IDF»)

Процедура оценки состояла из двух этапов. Сначала все модификации методов были оценены автоматическими мерами качества официального пакета ROUGE (см. Раздел 1.3.1). В Табл. 6 представлены результаты оценки данным пакетом по всем основным мерам качества. По причине значимости различных мер качества для разных типов задач и входных данных ([56]), в качестве основы используется средняя позиция по всем мерам качества ROUGE (колонка **Avg**).

Метод	1	2	L	S	SU	Avg
1. MMR + Groups	**0,62499 (1)**	**0,41633 (1)**	**0,6021 (1)**	**0,35529 (1)**	**0,36649 (1)**	**1,0**
2. OurSummary (Nodes)	0,58652 (2)	0,36154 (2)	0,56450 (2)	0,32113 (2)	0,33203 (2)	2,0
3. OurSummary (Nodes) with IDF	0,58497 (3)	0,33918 (4)	0,55745 (3)	0,30124 (3)	0,31283 (3)	3,2
4. Classic MMR	0.55917 (4)	0.34539 (3)	0.54012 (4)	0.29428 (4)	0.30519 (4)	3,8
5. ThematicLines	0,53416 (5)	0,33364 (5)	0,51238 (5)	0,27130 (5)	0,28243 (5)	5,0
6. OurSummary (Relations)	0,53141 (6)	0,28920 (6)	0,50422 (6)	0,25382 (6)	0,26509 (6)	6,0
7. SumBasic + Groups	0,52255 (7)	0,22881 (9)	0,49300 (8)	0,24356 (7)	0,25525 (7)	7,6
8. SumBasic	0,51847 (8)	0,24735 (8)	0,49786 (7)	0,23064 (8)	0,24257 (8)	7,8
9. OurSummary (Relations) with IDF	0,45494 (9)	0,24856 (7)	0,43768 (9)	0,19419 (10)	0,20492 (10)	9,0
10. MMR with IDF + Groups	0,44475 (10)	0,22238 (10)	0,42318 (10)	0,20627 (9)	0,21648 (9)	9,6

Табл. 6: Результаты оценки методом ROUGE

Наиболее значимыми являются следующие результаты:

I. Интеграция построенных тематических цепочек в классические методы автоматического аннотирования MMR и SumBasic улучшает качество исходных методов;
II. Методы аннотирования на основе обогащенной модели тематических цепочек показывают лучшее качество по сравнению с методом ThematicLines, основанным на единственной характеристике схожести.

Для подтверждения результатов оценки методом ROUGE, лучшие и наиболее приоритетные модификации методов были дополнительно оценены методом «Пирамиды» (см. Раздел 1.3.2). В Табл. 7 представлены результаты оценки данным методом.

Метод	**Score**
MMR + Groups	**0,645 (1)**
OurSummary (Nodes)	0,602 (2)
Classic MMR	0,578 (3)
SumBasic + Groups	0,575 (4)
SumBasic	0,567 (5)

Табл. 7: Результаты оценки методом «Пирамиды»

Результаты оценки методом «Пирамиды» подтверждают факты, установленные при оценке методом ROUGE, а именно улучшение качества методов автоматического аннотирования при интеграции в них построенных тематических цепочек на основе совокупности разнородных факторов.

3.4 Выводы к третьей главе

В данной главе предлагается алгоритм интеграции разработанной в рамках данной кандидатской диссертации модели основных участников ситуации новостного кластера (тематических цепочек) в существующие методы автоматического аннотирования, на примере методов MMR и SumBasic, а также предлагается два новых метода автоматического

аннотирования на основе построенной модели. Произведена оценка полученных автоматических аннотаций методами ROUGE и «Пирамиды».

Интеграция тематических цепочек в методы автоматического аннотирования позволило улучшить качество исходных методов по обоим методам оценки. Новые методы автоматического аннотирования также показали улучшение качества по сравнению с существующими методами аналогичного класса, что подтверждает полезность построенной модели.

4. Система автоматического аннотирования на основе тематических цепочек

В результате данного диссертационного исследования был разработан программный комплекс по автоматической обработке новостных кластеров, включающий в себя следующие условно-независимые модули:

- Модуль построения тематических цепочек новостного кластера на основе разработанного алгоритма (см. раздел 4.2)
- Модуль автоматического аннотирования, реализующий более 10 различных методов аннотирования (см. раздел 4.3)
- Модуль автоматической оценки аннотаций новостного кластера на основе метода ROUGE (см. раздел 4.4)

Данные модули объединены в единое приложение и могут взаимодействовать друг с другом по принципу конвейера (результаты работы одного модуля передаются как входные данные другому модулю) в указанной последовательности, обеспечивая замкнутый цикл обработки новостного кластера всеми функциональными блоками.

4.1 Общее описание программного комплекса

4.1.1 Архитектурная схема

На Рис. 7 представлена общая архитектурная схема разработанного программного комплекса. В качестве входных данных для программного комплекса выступает новостной кластер (или набор новостных кластеров - разработанный программный комплекс поддерживает пакетную обработку), который проходит предварительную обработку внешним модулем - морфологическим анализатором (подробное описание данного модуля, а также формат входных данных, представлены в Разделе 4.1.2).

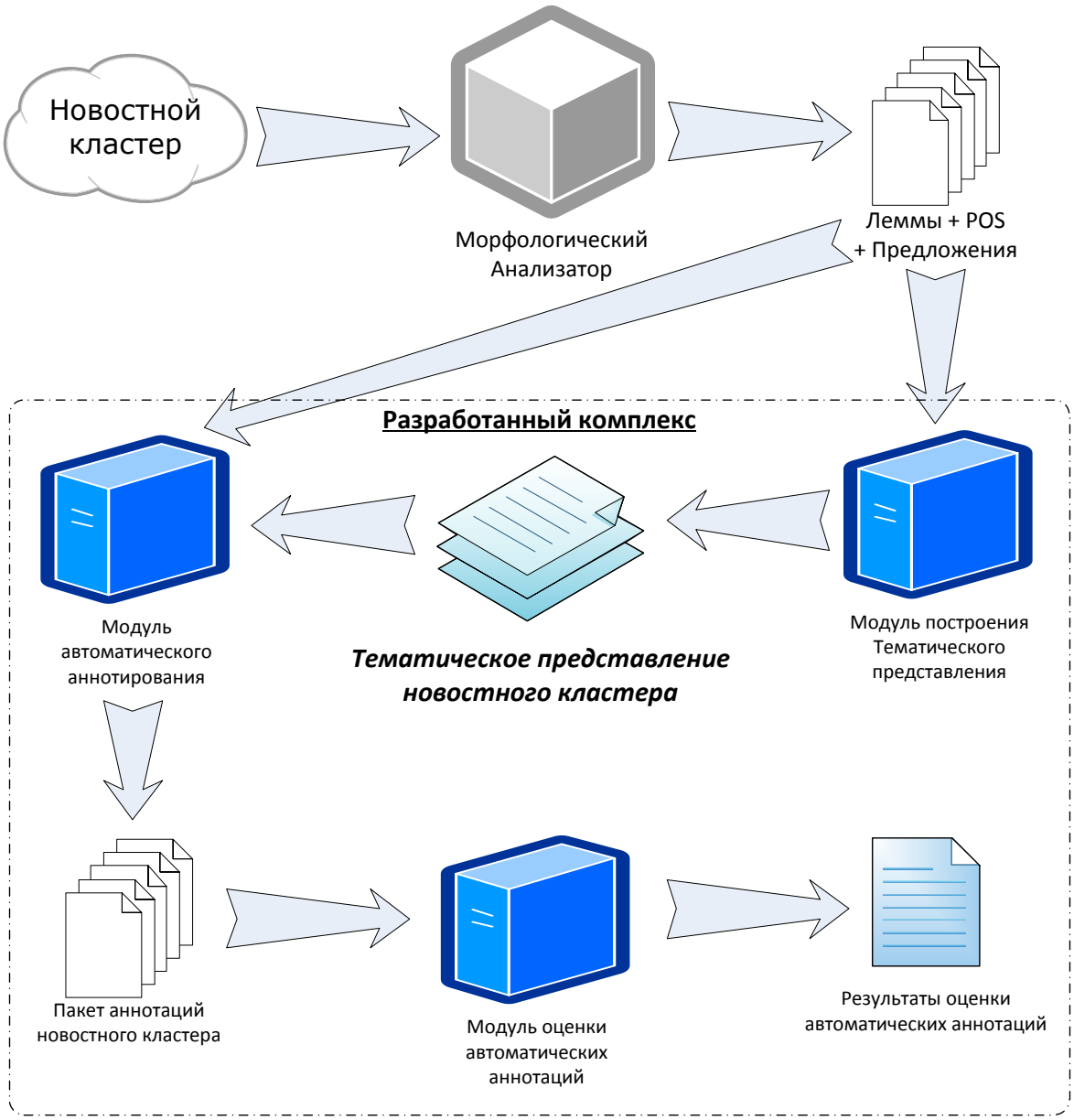


Рис. 7: Архитектура разработанного программного комплекса

Результаты работы морфологического анализатора (леммы слов, информация о частях речи, разбиение текстов по предложениям) передаются на вход стартовому модулю системы – модулю построения тематических цепочек. Кроме того, на вход программному комплексу передается следующая дополнительная информация, необходимая для построения тематических цепочек разработанным алгоритмом (вычисления характеристик схожести), а также реализации методов автоматического аннотирования:

- Информация о концептах тезауруса (опционально, в случае отсутствия данной информации соответствующая

характеристика алгоритма построения тематических цепочек всегда будет равна нулю, см. раздел 2.4.2.2);

- Информация о частотных характеристиках слов по коллекции документов (1 миллион новостных сообщений, данная информация необходима для расчета весов **I**nverse **D**ocument **F**requency (IDF) слов, используемых методами аннотирования);
- Список стоп-слов, исключаемых из рассмотрения при работе алгоритмов построения тематических цепочек, автоматического аннотирования, а также оценки автоматических аннотаций.

Более подробная информация о специфике работы отдельных модулей будет представлена в разделах 4.2 - 4.4.

В качестве среды разработки использовалась Microsoft Visual Studio 2012. Графический интерфейс пользователя программного комплекса построен на платформе Microsoft .NET Framework 4.0, с использованием модуля Windows Forms (высокоуровневый интерфейс Win32 API в ОС Windows).

4.1.2 Входные данные: Структура и предварительная обработка

4.1.2.1 Описание модуля морфологического анализатора

Необходимым предварительным этапом обработки исходного новостного кластера является проведение морфологического анализа. В результате работы модуля морфологического анализатора новостной кластер представляется в строго формализованном виде, ключевыми аспектами которого являются:

- Токенизация исходной текстовой коллекции. Входной поток разбивается на отдельные токены, которые должны соответствовать словоформам естественного языка;
- Лемматизация потока токенов. Каждому токену сопоставляется лемма – словарная форма слова. Дальнейшая обработка ведется именно на уровне лемм слов, так как естественные языки содержат

большое количество словоформ одних и тех же слов, которые, в свою очередь, должны единообразно интерпретироваться алгоритмами автоматической обработки текстов. Именно механизм лемматизации решает данную проблему.

- Выделение частей речи. Помимо сопоставления лемм для токенов, используемый модуль морфологического анализатора также определяет часть речи для каждой словоформы. Информация о частях речи используется алгоритмом построения тематических цепочек, а также алгоритмами автоматического аннотирования.
- Определение границ предложений. Для входного потока токенов определяются знаки пунктуации, разбивающие данный поток на отдельные предложения исходного текста.

На вход разработанному программному комплексу подается структурированное представление новостного кластера (вектор лемм слов с указанием частей речи и разбивкой по предложениям, модель bag-of-words в международной терминологии).

В качестве модуля морфологического анализатора использовался внешний модуль, разработанный в Научно-Исследовательском Центре Московского Государственного Университета им. М.В. Ломоносова (НИВЦ МГУ). Исследование и разработка данного модуля не является предметом данной кандидатской диссертации. Пример обработки текста модулем морфологического анализатора представлен на Рис. 8.

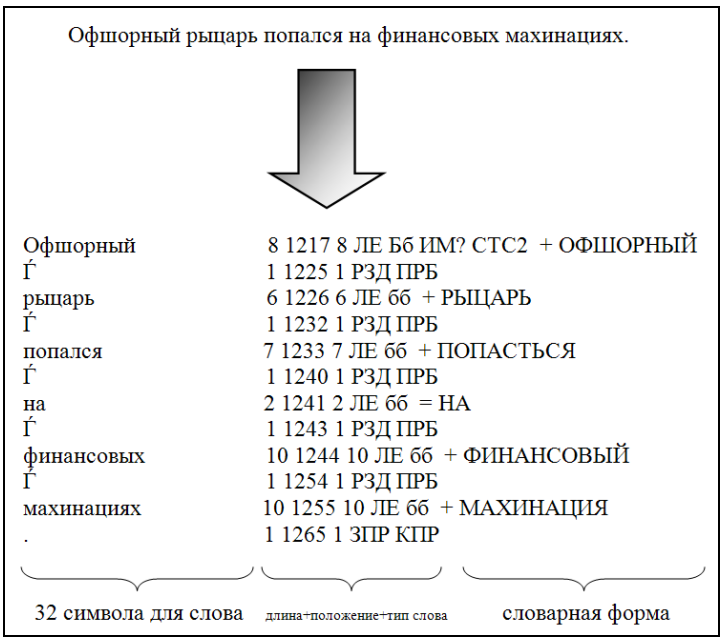


Рис. 8: Пример работы модуля морфологического анализатора

4.2 Модуль построения тематических цепочек

Модуль построения тематических цепочек реализует алгоритм, описанный в Главе 2.4. На вход данному модулю поступает новостной кластер (или несколько новостных кластеров в случае пакетной обработки), прошедший предварительную обработку модулем морфологического анализатора.

Одной из характеристик, используемых алгоритмом построения тематических цепочек, является наличие связи между рассматриваемыми сущностями в некотором предопределенном ресурсе. Для русского языка таким ресурсом выступает тезаурус РуТез. В связи с этим, в дополнение к результату обработки новостного кластера модулем морфологического анализатора, модуль построения тематических цепочек дополнительно получает результат сопоставления входного новостного кластера с ресурсом, содержащим описание семантических связей (тезаурус русского язык РуТез) в структурированном виде (в случае отсутствия данной информации на входе соответствующая характеристика схожести не будет функционировать).

Также на вход модулю построения тематических цепочек подаются следующие файлы с необходимой дополнительной информацией:

- Информация о семантических связях между объектами в используемом предопределенном ресурсе (информация о связях в тезаурусе РуТез для русского языка);
- Информация о стоп-словах, исключаемых из рассмотрения алгоритмом построения тематических цепочек.

Возможен выбор следующих параметров алгоритма:

- Порог схожести Score для объединения сущностей в тематическую цепочку в случае отсутствия между ними формального сходства (характеристика BS) и связи по тезаурусу (характеристика TS).
- Порог схожести Score для объединения сущностей в тематическую цепочку в случае наличия между ними формального сходства (характеристика BS) или связи по тезаурусу (характеристика TS).
- Порог на частоту сущностей-кандидатов на объединение в тематическую цепочку (***FreqThreshold***). Данный порог задается относительно количества документов в исходном новостном кластере. Частоты сущностей $Frequency_1$ и $Frequency_2$ должны удовлетворять следующему условию (*doc_count* - число документов в новостном кластере):

$$Min(Frequency_1, Frequency_2) > \mathrm{FreqThreshold} \cdot doc\_count$$

  В случае невыполнения данного условия пара сущностей не может быть объединена в одну тематическую цепочку.
- Порог на количество рассматриваемых сущностей (в порядке убывания их частоты) при формировании пар-кандидатов для объединения в тематические цепочки. Наиболее важными для рассмотрения являются сущности с высокой частотой упоминаний в рамках исходного новостного кластера – основные тематические цепочки. Сущности с низкой частотой в меньшей степени влияют на основную тему коллекции и поэтому не принципиальны для алгоритма построения тематических цепочек. Параметр имеет

целое значение: количество объектов для рассмотрения на каждой итерации алгоритма.

- Параметр для склейки многословных выражений: порог на значение характеристики Near относительно максимального значения характеристики AcrossVerb (***NearToMaxAcrossVerb***). Выполнение следующего условия является необходимым для объединения двух выражений в единое многословное выражение (MaxAcrossVerb – максимальное значение характеристики AcrossVerb на первой итерации алгоритма среди всех пар-кандидатов):

$$Near > NearToMaxAcrossVerb \cdot MaxAcrossVerb$$

- Параметр для склейки многословных выражений: порог на значение характеристики Near относительно суммы характеристик AcrossVerb и NotNear (***NearToAcrossVerbPlusNotNear***). Выполнение следующего условия является необходимым для объединения двух выражений в единое многословное выражение:

$$Near > \mathrm{NearToAcrossVerbPlusNotNear} \cdot (AcrossVerb + NotNear)$$

- Параметр для склейки многословных выражений: порог на значение характеристики Near относительно частотностей $Freq_1$ и $Freq_2$ (***NearToMinFreq$_1$AndFreq$_2$***). Выполнение следующего условия является необходимым для объединения двух выражений в единое многословное выражение:

$$Near > \mathrm{NearToMinFreq_1AndFreq_2} \cdot (Freq_1 + Freq_2)$$

- Понижающие коэффициенты на характеристику схожести по тезаурусу в случае многозначности одного (двух) концепта(ов) тезауруса, сопоставленных сущностям рассматриваемой пары-кандидатов для объединения в тематическую цепочку. Некоторые языковые выражения не могут быть однозначно сопоставлены конкретному концепту тезауруса, а могут быть отнесены сразу к

нескольким. Например, языковое выражение *пост* может относиться к следующим концептам тезауруса:

- *ДОЛЖНОСТЬ*
- *ПУНКТ ОХРАНЫ*
- *ПОСТИТЬСЯ*
- *ПОСТНЫЕ ДНИ*

При вычислении схожести по тезаурусу для сущностей, сопоставленных нескольким концептам тезауруса, вводятся соответствующие понижающие коэффициенты для связей многозначных концептов, так как используемое в тексте значение может не коррелировать с анализируемой связью по тезаурусу, вследствие чего подобные связи тезауруса для многозначных концептов не могут быть учтены наравне со связями однозначных концептов. Понижающие коэффициенты позволяют регулировать вес связей для многозначных объектов: в случае многозначности концепта тезауруса для одного или двух сущностей-кандидатов на объединение в тематический узел, вес схожести по характеристике TS умножается на соответствующий понижающий коэффициент.

Результатом работы модуля построения тематических цепочек является:

1. Информация о многословных выражениях, присутствующих в исходной текстовой коллекции;
2. Разбиение всех языковых выражений (слов и многословных выражений) на тематические цепочки, с указанием центральных элементов цепочек и весов схожести с центральными элементами цепочек для вложенных языковых выражений.

Пример результатов работы модуля построения тематических цепочек (топ-7 тематических цепочек по частоте, с указанием весов схожести с главным элементом) приведен на Рис. 9.

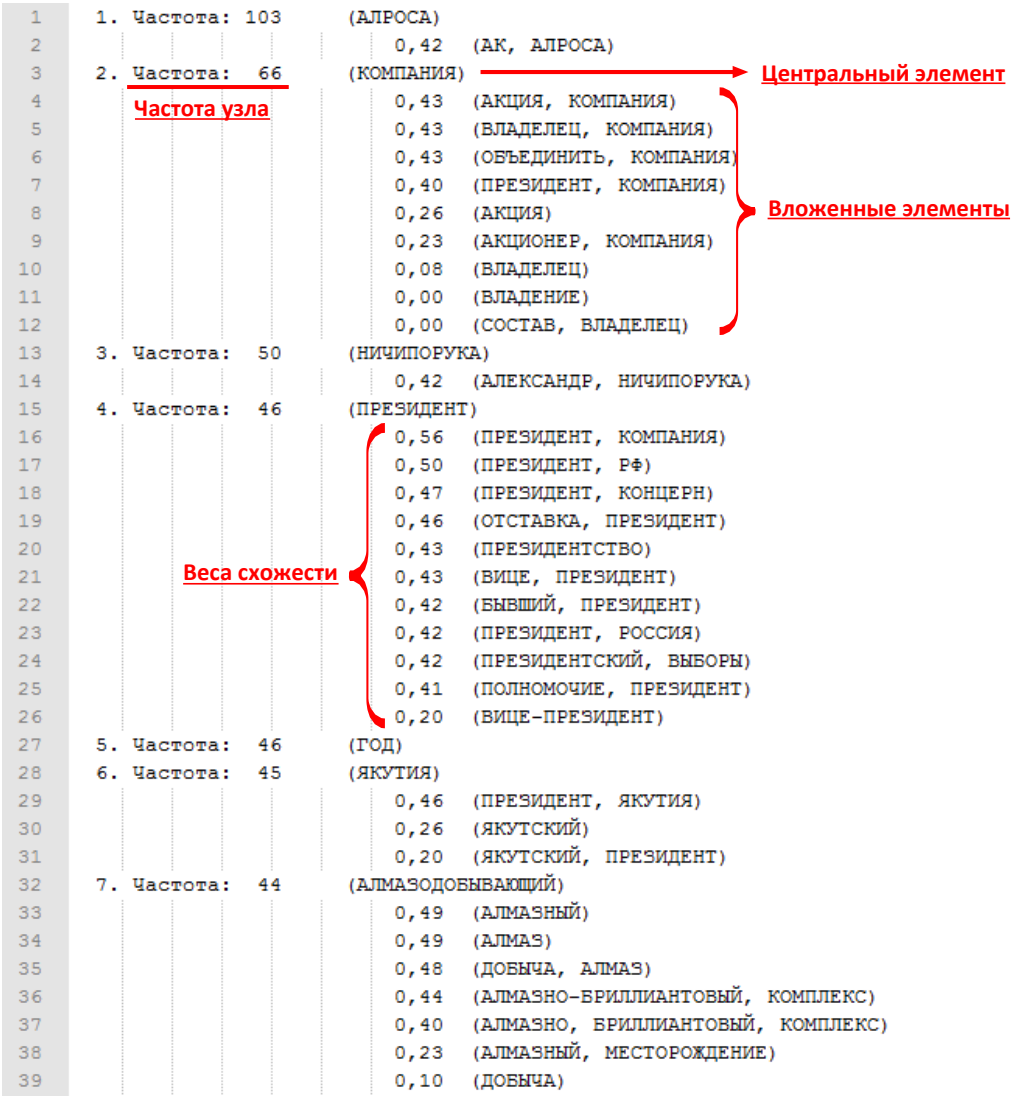


Рис. 9: Результаты работы модуля построения тематических цепочек

Полные результаты работы модуля построения тематических цепочек для кластера-примера, посвященного смене руководства алмазодобывающей корпорации «Алроса», представлены в Приложение 1.

4.3 Модуль автоматического аннотирования

Основной задачей модуля аннотирования является реализация различных алгоритмов автоматического аннотирования (см. Главу «Автоматическое аннотирование»). Помимо известных и широко-применяемых методов аннотирования, таких как Maximal Marginal Relevance, SumBasic и так далее, в модуле реализован ряд новых методов аннотирования, в том числе модификаций существующих методов,

связанных с интеграцией построенных в рамках работы первого модуля системы тематических цепочек (см. Главу 3).

В качестве входных данных модуль автоматического аннотирования получает следующую информацию:

- Результат обработки исходного новостного кластера модулем морфологического анализатора (работа большинства классических методов аннотирования основана на модели bag-of-words, предоставляемых морфологическим анализатором, см. Главу 4.1.2);
- Тематические цепочки новостного кластера (см. Главы 2.4, 4.2);
- Информация о документной частотности слов по новостной коллекции размером 1 миллион новостных документов (характеристика **D**ocument **F**requency, используемая рядом классических методов аннотирования);
- Информация о концептах тезауруса русского языка РуТез, упомянутых в исходном новостном кластере, а также тематическое представление, построенное на основе данного тезауруса (используется методом аннотирования на основе тематического представления РуТез, см. Раздел 1.2.3.1).

Результатом работы модуля автоматического аннотирования является пакет автоматических аннотаций, соответствующих различных методам аннотирования, а также дополнительная вспомогательная информация (информация об упомянутых в предложениях сущностях, их веса, веса предложений и так далее). Всего в модуле реализовано 11 различных методов аннотирования:

- Классический метод Maximal Marginal Relevance (MMR, см. Раздел 1.2.6);
- Классический метод Maximal Marginal Relevance без учета характеристики IDF (см. см. Раздел 1.2.6);

- Классический метод Maximal Marginal Relevance с интеграцией построенных тематических цепочек, без учета характеристики IDF (см. Раздел 3.1.2);
- Классический метод Maximal Marginal Relevance с интеграцией построенных тематических цепочек, с учетом характеристики IDF (см. Раздел 3.1.2);
- Классический метод SumBasic (см. Раздел 1.2.2);
- Классический метод SumBasic с интеграцией построенных тематических цепочек (см. Раздел 3.1.3);
- Метод аннотирования на основе тематического представления построенного на основе тезауруса РуТез (см. Раздел 1.2.3.1);
- Новый метод аннотирования на основе связей построенных тематических цепочек, без учета IDF (см. Раздел 3.2.1);
- Новый метод аннотирования на основе связей построенных тематических цепочек, с учетом IDF (см. Раздел 3.2.2);
- Новый метод аннотирования на основе упоминаний построенных тематических цепочек, без учета IDF (см. Раздел 3.2.1);
- Новый метод аннотирования на основе упоминаний построенных тематических цепочек, с учетом IDF (см. Раздел 3.2.2);

На Рис. 10 приведен пример результирующего файла работы модуля автоматического аннотирования для классического метода MMR с учетом IDF, для кластера-примера, посвященного смене руководства алмазодобывающей корпорации «Алроса».

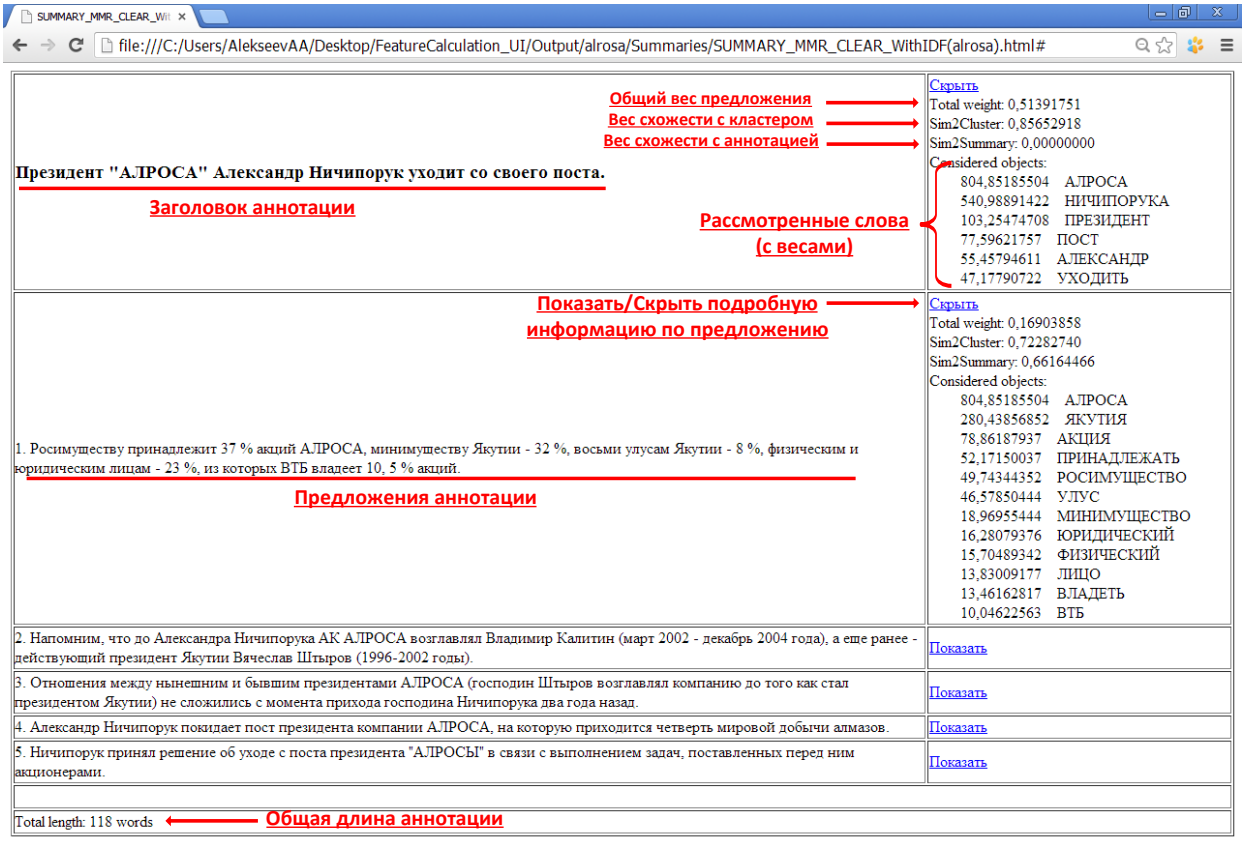


Рис. 10: Пример результата работы модуля автоматического аннотирования (метода MMR)

4.4 Модуль оценки автоматических аннотаций

Функционал модуля оценки автоматических аннотаций обеспечивает массовую оперативную оценку набора конкурсных аннотаций методом ROUGE (см. Раздел «Автоматические меры качества ROUGE»). В качестве ядра алгоритма оценки автоматических аннотаций выступает официальный пакет ROUGE [36], представляющий собой набор скриптов на языке PERL. В разработанном модуле осуществляется:

1. Подготовка автоматических аннотаций, полученных в результате работы модуля автоматического аннотирования, для обработки официальным пакетом ROUGE;
2. Подготовка экспертных аннотаций для обработки официальным пакетом ROUGE;

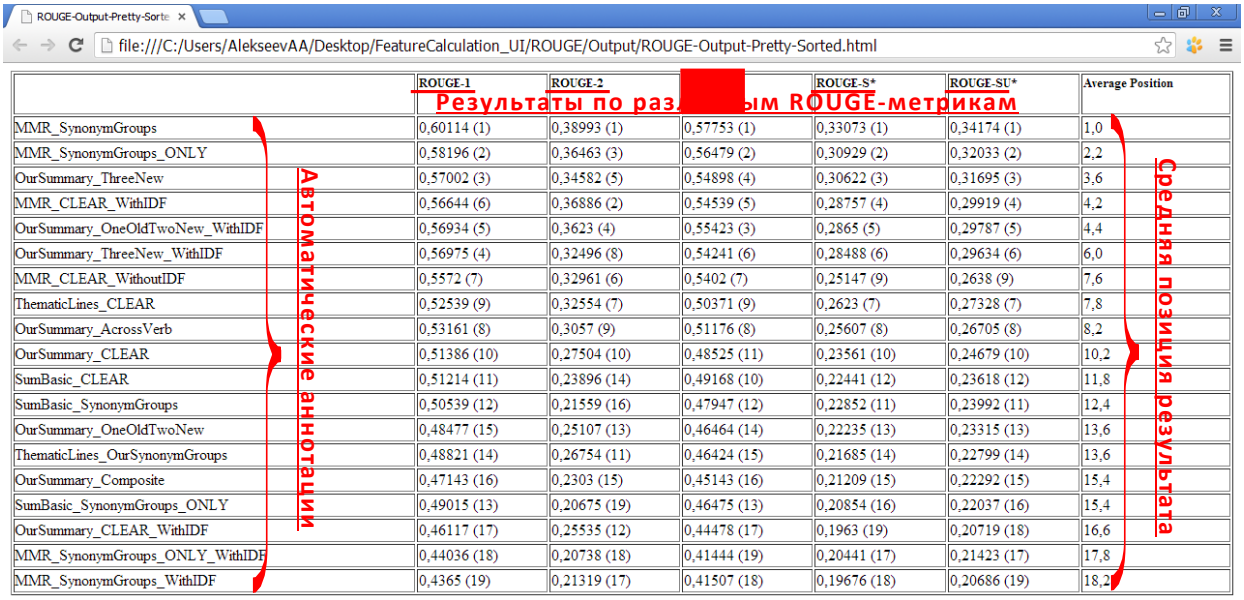
3. Интеграция с официальным пакетом ROUGE: автоматический вызов интерпретатора языка PERL с необходимыми параметрами и входными данными;
4. Считывание, обработка и запись в структурированном виде результатов работы официального пакета ROUGE.

Входными данными для модуля оценки автоматических аннотаций являются:

1. Пакет экспертных аннотаций, сформированных экспертами для оцениваемых новостных кластеров;
2. Пакет автоматических аннотаций, сформированных модулем автоматического аннотирования (см. 4.3).

Официальный пакет ROUGE создан для оценки автоматических аннотаций на английском языке и не поддерживает в качестве входных данных аннотации на других языках (по причине использования однобайтной таблицы кодировки). Для решения данной проблемы все автоматические аннотации, подаваемые на вход модулю оценки автоматических аннотаций, проходят процедуру транслитерации: каждому символу кириллицы сопоставляется уникальная строка латиницы, в соответствии со стандартом ISO 9-95.

Результатом работы модуля оценки автоматических аннотаций является HTML-файл с агрегированной информацией об оценках входных автоматических аннотаций выбранными мерами качества ROUGE на заданном пакете новостных кластеров. На Рис. 11 приведен пример подобного результирующего файла, с описанием значащих полей.



	ROUGE-1	ROUGE-2		ROUGE-S*	ROUGE-SU*	Average Position
MMR_SynonymGroups	0,60114 (1)	0,38993 (1)	0,57753 (1)	0,33073 (1)	0,34174 (1)	1,0
MMR_SynonymGroups_ONLY	0,58196 (2)	0,36463 (3)	0,56479 (2)	0,30929 (2)	0,32033 (2)	2,2
OurSummary_ThreeNew	0,57002 (3)	0,34582 (5)	0,54898 (4)	0,30622 (3)	0,31695 (3)	3,6
MMR_CLEAR_WithIDF	0,56644 (6)	0,36886 (2)	0,54539 (5)	0,28757 (4)	0,29919 (4)	4,2
OurSummary_OneOldTwoNew_WithIDF	0,56934 (5)	0,3623 (4)	0,55423 (3)	0,2865 (5)	0,29787 (5)	4,4
OurSummary_ThreeNew_WithIDF	0,56975 (4)	0,32496 (8)	0,54241 (6)	0,28488 (6)	0,29634 (6)	6,0
MMR_CLEAR_WithoutIDF	0,5572 (7)	0,32961 (6)	0,5402 (7)	0,25147 (9)	0,2638 (9)	7,6
ThematicLines_CLEAR	0,52539 (9)	0,32554 (7)	0,50371 (9)	0,2623 (7)	0,27328 (7)	7,8
OurSummary_AcrossVerb	0,53161 (8)	0,3057 (9)	0,51176 (8)	0,25607 (8)	0,26705 (8)	8,2
OurSummary_CLEAR	0,51386 (10)	0,27504 (10)	0,48525 (11)	0,23561 (10)	0,24679 (10)	10,2
SumBasic_CLEAR	0,51214 (11)	0,23896 (14)	0,49168 (10)	0,22441 (12)	0,23618 (12)	11,8
SumBasic_SynonymGroups	0,50539 (12)	0,21559 (16)	0,47947 (12)	0,22852 (11)	0,23992 (11)	12,4
OurSummary_OneOldTwoNew	0,48477 (15)	0,25107 (13)	0,46464 (14)	0,22235 (13)	0,23315 (13)	13,6
ThematicLines_OurSynonymGroups	0,48821 (14)	0,26754 (11)	0,46424 (15)	0,21685 (14)	0,22799 (14)	13,6
OurSummary_Composite	0,47143 (16)	0,2303 (15)	0,45143 (16)	0,21209 (15)	0,22292 (15)	15,4
SumBasic_SynonymGroups_ONLY	0,49015 (13)	0,20675 (19)	0,46475 (13)	0,20854 (16)	0,22037 (16)	15,4
OurSummary_CLEAR_WithIDF	0,46117 (17)	0,25535 (12)	0,44478 (17)	0,1963 (19)	0,20719 (18)	16,6
MMR_SynonymGroups_ONLY_WithIDF	0,44036 (18)	0,20738 (18)	0,41444 (19)	0,20441 (17)	0,21423 (17)	17,8
MMR_SynonymGroups_WithIDF	0,4365 (19)	0,21319 (17)	0,41507 (18)	0,19676 (18)	0,20686 (19)	18,2

Рис. 11: Пример результатов работы модуля оценки автоматических аннотаций

Итоговая сортировка методов аннотирования в результирующем файле производится по значению колонки “Average Position” – среднее значение позиции результата по всем мерам качества ROUGE.

4.5 Выводы к четвертой главе

В данной главе приведено описание разработанного программного комплекса, реализующего модели и алгоритмы, предложенные в рамках данной диссертационной работы, а именно:

- Модель и алгоритм построения тематических цепочек новостного кластера
- Метод интеграции построенных тематических цепочек в существующие методы автоматического аннотирования
- Новые методы автоматического аннотирования на основе построенных тематических цепочек

Также в разработанном программном комплексе реализован модуль для автоматической оценки аннотаций методом ROUGE.

Заключение

В ходе диссертационной работы были получены следующие результаты:

1. Предложена модель, позволяющая с помощью тематических цепочек новостного кластера описывать основных участников этого кластера с учетом вариативности их именования и специфики внутреннего устройства текстов на естественном языке;
2. Предложен и реализован новый метод автоматического построения тематических цепочек новостного кластера, основанный на комбинировании разнородных признаков схожести;
3. Предложен и реализован метод применения построенной модели в существующие методы автоматического аннотирования, а также два новых метода автоматического аннотирования на основе тематических цепочек. Показано улучшение качества работы алгоритмов аннотирования на основе построенной модели.

Список литературы

[1] Alekseev A.A., Loukachevitch N.V Use of Multiple Features for Extracting Topics from News Clusters // Proceedings of the Spring Researchers Colloquium on Databases and Information Systems. – 2012. – P. 3-11. URL: http://ceur-ws.org/Vol-899/paper2.pdf

[2] Alekseev A.A., Loukachevitch N.V. Automatic detection of near-synonyms in news clusters // Труды международной конференции «Диалог». – 2011. – С. 32-41. URL: http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/5.pdf

[3] Alekseev A.A., Loukachevitch N.V. Automatic Entity Detection Based on News Cluster Structure // Proceedings of the International Workshop on Concept Discovery in Unstructured Data. – 2011. – P. 1-10. URL: http://ceur-ws.org/Vol-757/paper_1.pdf

[4] Alekseev A.A., Loukachevitch N.V. The automatic retrieval of news entities based on the structure of a news cluster // Scientific and Technical Information Processing. – 2012. – Vol. 39, № 6. – P. 303-309. URL: http://link.springer.com/article/10.3103%2FS0147688212060019

[5] Allan J.: Introduction to Topic Detection and Tracking // Topic detection and tracking, Kluwer Academic Publishers Norwell. – USA, 2002. – P. 1-16.

[6] Barzilay R., Elhadad M. Text summarizations with lexical chains / Inderjeet Mani and Mark Maybury // Advances in Automatic Text Summarization. – MIT Press, 1999. – P. 111–121.

[7] Barzilay R., Elhadad M. Using Lexical Chains for Text Summarization // Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization. – 1997. – P. 10-17. URL: http://acl.ldc.upenn.edu/W/W97/W97-0703.pdf

[8] Barzilay R., McKeown K. Extracting Paraphrases from a Parallel Corpus // Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. – 2001. – P. 50-57. URL: http://acl.ldc.upenn.edu/acl2001/MAIN/BARZILAY.PDF

[9] Biadsy F., Hirschberg J., Filatova E. An unsupervised approach to biography production using wikipedia // Proceedings of the Annual Meeting of the Association for Computational Linguistics. – 2008. – P. 807–815. URL: http://www.cs.columbia.edu/nlp/papers/2008/fadi_al_08a.pdf

[10] Blei D., Griffiths T., Jordan M., Tenenbaum J. Hierarchical topic models and the nested Chinese restaurant process // Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference. – MIT Press, 2003.

URL: http://www.cs.princeton.edu/~blei/papers/BleiGriffithsJordanTenenbaum2003.pdf

[11] Blei D., Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. – 2003. – P. 993-1022. URL: http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf

[12] Boudin F., El-Beze M., Torres-Moreno J.-M. A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization // Proceedings of the 22nd International Conference on Computational Linguistics. – 2008. – P. 23–26. URL: http://www.aclweb.org/anthology-new/C/C08/C08-2006.pdf

[13] Boudin F., El-Beze M., Torres-Moreno J.-M. The LIA Update Summarization Systems at TAC-2008 // Proceedings of the Text Analyze Conference. – USA: Gaithersburg, 2008. URL: http://www.nist.gov/tac/publications/2008/participant.papers/LIA.proceedings.pdf

[14] Carbonell J., Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries // Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. – Australia: Melbourne, 1998. – P. 335–336. URL: http://www.cs.cmu.edu/~jgc/publication/The_Use_MMR_Diversity_Based_LTMIR_1998.pdf

[15] Celikyilmaz A., Hakkani-Tur D. A hybrid hierarchical model for multi-document summarization // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. – 2010. – P. 815–824. URL: https://www.aclweb.org/anthology-new/P/P10/P10-1084.pdf

[16] Chali Y., Joty S. Improving the performance of the random walk model for answering complex questions // Proceedings of Annual Meeting of the Association for Computational Linguistics. – 2008. – P. 9–12. URL: http://www-devel.cs.ubc.ca/~rjoty/paper/acl08Joty.pdf

[17] Dang H.T. Overview of DUC 2006 // Proceedings of the Document Understanding Conferences. – USA: New York, 2006. URL: http://duc.nist.gov/pubs/2006papers/duc2006.pdf

[18] Dang H.T., Owczarzak K. Overview of the TAC 2008 Update Summarization Task // Proceedings of the Text Analyze Conference. – USA: Gaithersburg, 2008. URL: http://www.nist.gov/tac/publications/2008/additional.papers/update_summ_overview08.proceedings.pdf

[19] Dang V., Xue X., Croft B. Context-based Quasi-Synonym Extraction // Massachusetts Center for Intelligent Information Retrieval Technical Report. – 2009. URL: http://maroo.cs.umass.edu/getpdf.php?id=882

[20] Daum´e H., Marcu D. Bayesian query-focused summarization // Proceedings of the International Conference on Computational Linguistics and the annual meeting of the Association for Computational Linguistics. – 2006. – P. 305–312. URL: http://acl.ldc.upenn.edu/P/P06/P06-1039.pdf

[21] Deerwester S., Dumais S., Furnas G., Landauer T., Harshman R. Indexing by latent semantic analysis // Journal of the American Society for Information Science. – 1990. – P. 391–407. URL: http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf

[22] Dijk van T. Semantic Discourse Analysis / Teun A. van Dijk // Handbook of Discourse Analysis. – London: Academic Press, 1985. – V. 2. – P. 103-136. URL: http://www.discourses.org/OldArticles/Semantic%20discourse%20analysis.pdf

[23] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weishedel, R.: The Automatic Content Extraction (ACE): Task, Data, Evaluation // Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC`2004). – 2004. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.78.8442&rep=rep1&type=pdf

[24] Edmundson H.P. New methods in automatic extracting // Journal of the ACM. – 1969. – Vol. 16, № 2. – P. 264–285. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.5638&rep=rep1&type=pdf

[25] Erkan G., Radev D. Lexrank: Graph-based centrality as salience in text summarization // Journal of Artificial Intelligence Research. – 2004. URL: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/jair/OldFiles/OldFiles/pub/volume22/erkan04a.pdf

[26] Galley M., McKeown K. Improving word sense disambiguation in lexical chaining // Proceedings of the international joint conference on Artificial intelligence. – 2003. – P. 1486–1488. URL: http://www.cs.columbia.edu/nlp/papers/2003/galley_mckeown_03.pdf

[27] Gillick D., Favre B. A scalable global model for summarization // Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing. – 2009. – P. 10-18. URL: http://www.aclweb.org/anthology/W/W09/W09-18.pdf#page=20

[28] Gong Y., Liu X. Generic text summarization using relevance measure and latent semantic analysis // Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2001. – P. 19–25. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.109.5097&rep=rep1&type=pdf

[29] Griffiths T., Steyvers M. Finding scientific topics // Proceedings of the National Academy of Sciences of the United States of America. – 2004. – Vol. 101, № 1. – P. 5228-5235. URL: http://people.csail.mit.edu/brussell/research/words/ICCV05/GS04.pdf

[30] Haghighi A., Vanderwende L. Exploring content models for multi-document summarization // Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. – 2009. – P. 362–370. URL: http://www.aclweb.org/anthology-new/N/N09/N09-1041.pdf

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

[32] Hasan R. Coherence and Cohesive harmony / J. Flood // Understanding reading comprehension. – DE: IRA, 1984. – P. 181-219.

[33] Hirst G., St-Onge D. Lexical Chains as representation of context for the detection and correction malapropisms / C. Fellbaum // WordNet: An electronic lexical database and some of its applications. – MA: The MIT Press, 1998. URL: http://www.cs.swarthmore.edu/~richardw/cs65-f08/litreview/meggie-malcolm.pdf

[34] Hofmann T. Probabilistic Latent Semantic Analysis // Proceedings of the Uncertainty in Artificial Intelligence (UAI'99). – Stockholm, 1999. – P. 289-296. URL: http://cs.brown.edu/~th/papers/Hofmann-UAI99.pdf

[35] Li J., Sun L., Kit C., Webster J. A Query-Focused Multi-Document Summarizer Based on Lexical Chains // Proceedings of the Document Understanding Conference. – 2007. URL: http://www-nlpir.nist.gov/projects/duc/pubs/2007papers/cas-uhongkong.final.pdf

[36] Lin C.-Y. ROUGE: a Package for Automatic Evaluation of Summaries // Proceedings of the Workshop on Text Summarization Branches Out (ACL'2004). – Spain: Barcelona, 2004. – P. 74-81. URL: http://acl.ldc.upenn.edu/acl2004/textsummarization/pdf/Lin.pdf

[37] Louis A., Joshi A., Nenkova A. Discourse indicators for content selection in summarization // Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue. – 2010. – P. 147–156. URL: https://www.aclweb.org/anthology/W/W10/W10-4327.pdf

[38] Loukachevitch N. Multigraph representation for lexical chaining // Proceedings of SENSE workshop. – 2009. – P. 67-76. URL: http://ceur-ws.org/Vol-476/paper8.pdf

[39] Loukachevitch N.V., Dobrov B.V. Evaluation of Thesaurus on Sociopolitical Life as Information Retrieval Tool // Proceedings of Third International Conference on Language Resources and Evaluation (LREC`2002). – 2002. – P. 115-121. URL: http://www.lrec-conf.org/proceedings/lrec2002/pdf/188.pdf

[40] Luhn H.P. The automatic creation of literature abstracts // IBM Journal of Research and Development. – 1958. – Vol. 2, № 2. – P. 159–165. URL: https://text-analysis.googlecode.com/files/luhn58.pdf

[41] Mani I. Automatic Summarization // John Benjamins Publishing Co. – Netherlands: Amsterdam, 2001. URL: http://benjamins.com/#catalog/books/nlp.3/main

[42] Mani I., Firmin T., Sundheim B. The TIPSTER SUMMAC Text Summarization Evaluation // Proceedings of Annual Meeting of the Association for Computational Linguistics: European Chapter. – 1999. – P. 77-85. URL: http://acl.ldc.upenn.edu/E/E99/E99-1011.pdf

[43] Mani I., Klein G., House D., Hirschman L., Firmin T., Sundheim B. SUMMAC: A text summarization evaluation // Natural Language Engineering. – 2002. – Vol. 8 № 1. – P. 43–68. URL: http://www1.cs.columbia.edu/~smaskey/candidacy/cand_papers/mani_summac.pdf

[44] McDonald R. A study of global inference algorithms in multi-document summarization // Proceedings of the European Conference on IR Research. – 2007. – P. 557–564. URL: http://ryanmcd.com/papers/globsumm.pdf

[45] McKeown K., Barzilay R., Evans D., Hatzivassiloglou V., Klavans J., Nenkova A., Sable C., Schiffman B., Sigelman S.. Tracking and summarizing news on a daily basis with Columbia's Newsblaster // Proceedings of the International Conference on Human Language Technology Research. – 2002. URL: http://www1.cs.columbia.edu/~bschiff/papers/hlt02-blast.pdf

[46] McKeown K., Passonneau R. J., Elson D. K., Nenkova A., Hirschberg J. Do summaries help? A task-based evaluation of multi-document summarization // Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2005. – P. 210–217. URL: http://www.cs.columbia.edu/~ani/papers/f98-mckeown.pdf

[47] Nenkova A. Automatic text summarization of newswire: lessons learned from the document understanding conference // Proceedings of the National Conference on Artificial Intelligence. – 2005. – P. 1436–1441. URL: http://www1.cs.columbia.edu/~ani/papers/AAAI051NenkovaA.pdf

[48] Nenkova A., McKeown K. A Survey of Text Summarization Techniques // Mining Text Data Book. – US: Springer, 2012. – P. 43-76. URL: http://vahabonline.com/wp-content/uploads/2013/06/Survey33.pdf

[49] Nenkova A., McKeown K. Automatic Summarization // Foundations and Trends in Information Retrieval. – 2011. – Vol. 5, № 2-3. – P. 103-233. URL: http://www.cis.upenn.edu/~nenkova/1500000015-Nenkova.pdf

[50] Nenkova A., Vanderwende L., McKeown K. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization // Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2006. – P. 573–580. URL: http://www.cis.upenn.edu/~nenkova/papers/fp285-nenkova.pdf

[51] Nenkova, A. and L. Vanderwende. The impact of frequency on summarization // Microsoft Research Technical Report, MSR-TR-2005-101. – 2005. URL: http://www.cs.bgu.ac.il/~elhadad/nlp09/sumbasic.pdf

[52] Over P., Dang H., Harman D. DUC in context // Information Processing and Management. – 2007. – Vol. 43, № 6. – P. 1506–1520. URL: http://dl.acm.org/citation.cfm?id=1285157

[53] Page L., Brin S., Motwani R., Winograd T. The PageRank Citation Ranking: Bringing Order to the Web // Technical Report, Stanford InfoLab. – 1999. URL: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

[54] Passonneau R., Nenkova A., McKeown K., Sigelman S. Applying the Pyramid Method in DUC 2005 // Proceedings of the Document Understanding Conferences. – Canada: Vancouver, 2005. URL: http://duc.nist.gov/pubs/2005papers/columbiau.passonneau2.pdf

[55] Radev D., Hovy E., McKeown, K. Introduction to the special issue on summarization // Computational Linguistics Journal – Summarization. – 2002. – Vol. 28, № 4. – P. 399-408. URL: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671927

[56] Rankel P., Conroy J., Dang H., Nenkova A. A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art // Proceedings of The 51st Annual Meeting of the Association for Computational Linguistics. – 2013. – P. 131-136. URL: http://aclweb.org/anthology/P/P13/P13-2024.pdf

[57] Rankel P., Dang H., Conroy J., Nenkova A. A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. – 2013. – P. 131–136. URL: http://newdesign.aclweb.org/anthology/P/P13/P13-2024.pdf

[58] Salton G., Buckley C. Term-weighting approaches in automatic text retrieval. // Information Processing and Management. – 1988. – P. 513–523. URL: http://comminfo.rutgers.edu/~muresan/IR/Docs/Articles/ipmSalton1988.pdf

[59] Schiffman B., McKeown K. Columbia University in the Novelty Track at TREC 2004 // Proceedings of the Thirteenth Text Retrieval Conference (TREC'2004). – 2004. URL: http://trec.nist.gov/pubs/trec13/papers/columbiau.novelty.pdf

[60] Soboroff I. Overview of the TREC 2004 Novelty Track // Proceedings of the Thirteenth Text Retrieval Conference (TREC'2004). – 2004. URL: http://trec.nist.gov/pubs/trec13/papers/NOVELTY.OVERVIEW.pdf

[61] Su Nam Kim S., Medelyan O., Min-Yen Kan, Baldwin T. SemEval-2010 Task-5. Automatic Keyphrase Extraction from Scientific Articles // Proceedings of the 5-th International Workshop on Semantic Evaluation (ACL`2010). – 2010. – P. 21-26. URL: http://www.aclweb.org/anthology-new/S/S10/S10-1004.pdf

[62] Vanderwende L., Suzuki H., Brockett C. Microsoft Research at DUC2006: Task-Focused Summarization with Sentence Simplification and Lexical Expansion // Proceedings of the Document Understanding Conference. – 2007. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.2486&rep=rep1&type=pdf

[63] Vanderwende L., Suzuki H., Brockett C., Nenkova A. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion // Information Processing and Management Journal. – 2007. – Vol. 43, № 6. – P. 1606-1618. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.105.9491&rep=rep1&type=pdf

[64] Wan X., Yang J. Improved affinity graph based multi-document summarization // Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. – 2006. – P. 181-184. URL: http://acl.ldc.upenn.edu/N/N06/N06-2046.pdf

[65] Wong K., Wu M., Li W. Extractive summarization using supervised and semi-supervised learning // Proceedings of the 22nd International Conference on Computational Linguistics (Coling`2008). – 2008. – P. 985–992. URL: http://speech.ee.ntu.edu.tw/~aaron/acl/www.aclweb.org/anthology-new/C/C08/C08-1124.pdf

[66] Алексеев А.А. Определение новизны информации в новостном кластере // Сборник трудов 17-ой Международной конференции "МАТЕМАТИКА. КОМПЬЮТЕР. ОБРАЗОВАНИЕ". – 2010. – С. 78. URL: http://www.mce.biophys.msu.ru/archive/doc62241/doc.pdf

[67] Алексеев А.А. Определение новизны информации в новостном кластере // Сборник трудов 13-ой Международной телекоммуникационной конференции студентов и молодых ученых "Молодежь и наука". – 2010. – С.

77-78. URL: http://library.mephi.ru/data/scientific-sessions/2010/confmin/ch2/0-1-32.doc

[68] Алексеев А.А. Тематический анализ новостного кластера как основа для автоматического аннотирования // Программная инженерия. – 2014. – № 3. – С. 41-48. URL: http://novtex.ru/prin/pi314_web.pdf

[69] Алексеев А.А. Тематическое представление новостного кластера как основа для автоматического аннотирования // Труды 15ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL). – 2013. – С. 359-369. URL: http://rcdl2013.uniyar.ac.ru/doc/full_text/ds_1.pdf

[70] Алексеев А.А., Добров Б.В., Лукашевич Н.В. Лингвистическая онтология – тезаурус РуТез // Труды международной конференции «Открытые семантические технологии проектирования интеллектуальных систем». – 2013. – С. 153-158. URL: http://conf.ostis.net/images/7/70/%D0%98%D0%B7%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5_%D0%BC%D0%B0%D1%82%D0%B5%D1%80%D0%B8%D0%B0%D0%BB%D1%8B_OSTIS-2013.pdf

[71] Алексеев А.А., Лукашевич Н.В. Автоматическое выделение близких по смыслу выражений из новостных кластеров // Труды конференции «Системный анализ и семиотическое моделирование». – 2011. – С. 150-154.

[72] Алексеев А.А., Лукашевич Н.В. Автоматическое извлечение сущностей на основе структуры новостного кластера // Искусственный интеллект и принятие решений. – 2011. – № 4. – С. 51-59. URL: http://aidt.ru/images/documents/2011-04/51_59.pdf

[73] Алексеев А.А., Лукашевич Н.В. Автоматическое порождение обновления к аннотации новостного кластера // Труды 12ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL). – 2010. – С. 84-91. URL: http://rcdl.ru/doc/2010/084-91.pdf

[74] Алексеев А.А., Лукашевич Н.В. Комбинирование признаков для извлечения тематических цепочек в новостном кластере // Труды Института системного программирования РАН. – 2012. – Т. 23. – С. 257-276. URL: http://www.ispras.ru/ru/proceedings/docs/2012/23/isp_23_2012_257.pdf

[75] Алексеев А.А., Мальковский М.Г. Автоматическое аннотирование новостного кластера на основе тематического анализа // Тезисы докладов конференции «Тихоновские чтения». – М: МГУ, 2013. – С. 55.

[76] Дейк В., Кинч В. Стратегии понимания связного текста // Новое в зарубежной лингвистике. – 1988. – В. 23. – С. 153-211.

[77] Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции "Электронные библиотеки: Перспективные методы и технологии, электронные коллекции. – 2003. – С. 201-210. URL: http://lvk.cs.msu.su/~bruzz/articles/knowledge_engineering/F2.pdf

[78] Добров Б.В., Павлов А.М. Исследование качества базовых методов кластеризации новостного потока в суточном временном окне // Труды 12ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL). – 2010. URL: http://rcdl.ru/doc/2010/287-295.pdf

[79] Ермаков А.Е. Референция обозначений персон и организаций в русскоязычных текстах СМИ: эмпирические закономерности для компьютерного анализа // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции «Диалог». – 2005. URL: http://www.dialog-21.ru/Archive/2005/Ermakov%20A/ErmakovAE.pdf

[80] Лукашевич Н.В., Добров Б.В. Автоматическое аннотирование новостного кластера на основе тематического представления // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции «Диалог». – 2009. – С. 299-305. URL: http://www.dialog-21.ru/dialog2009/materials/html/46.htm

[81] Лукашевич Н.В., Добров Б.В. Исследование тематической структуры текста на основе большого лингвистического ресурса // По материалам ежегодной Международной конференции «Диалог». – 2000. – С. 252-258. URL: http://www.cir.ru/docs/ips/publications/2000_dialog_text_analisys.pdf

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

Приложение 1

Пример результатов работы разработанного программного комплекса для кластера-примера, посвященного смене руководства алмазодобывающей корпорации «Алроса».

Примеры исходных документов новостного кластера.

2007.02.03 17:57
ИА "Росбалт"
http://www.rosbalt.ru/2007/02/03/284935.html
UN07020309111
*rus-feb07-000001-100000.txt *** 49460*

"АЛРОСА" перейдет к подземной разработке алмазных месторождений

МОСКВА, 3 февраля. "АЛРОСА" планирует прейти на подземный способ разработки алмазных месторождений в Якутии. Это позволит стабилизировать, а в дальнейшем и увеличить добычу алмазов, сообщили ИА "Росбалт-Север" в пресс- службе компании.

Активизация геологоразведки, начало реализации самостоятельной сбытовой политики, значительно расширившиеся международные контакты и укрепление позиций в Республике Ангола будут способствовать дальнейшему увеличению прибыли компании, отмечают в "АЛРОСА".

Напомним, в 2005 и 2006 годах компания получила рекордную чистую прибыль. Объемы реализации основной продукции "АЛРОСА" впервые превысили $3 млрд.

"АЛРОСА" является одной из крупнейших алмазодобывающих компаний мира, на долю которой приходится 25% мировой добычи. Акционерами "АЛРОСА" являются Росимущество - 37% акций, Министерство по управлению госимуществом Якутии - 32%, физические и юридические лица - 23%. Восьми улусам Якутии принадлежит 8% акций компании.

Инвестгруппа "АЛРОСА" управляет проектом освоения месторождения алмазов имени М.В.Ломоносова в Архангельской области. Ей также раньше принадлежали три золотодобывающих актива в Якутии, которые были проданы ОАО "Полюс золото", которое, в свою очередь, было выделено в независимую компанию из ГМК "Норильский никель".

2007.02.03 02:48
"Полит.Ру"
http://www.polit.ru/news/2007/02/03/_alros.popup.html
UN07020301128
*rus-feb07-000001-100000.txt *** 45454*

Президент "АЛРОСА" подал в отставку

Александр Ничипорук подал в отставку с должности президента алмазодобывающей АК "АЛРОСА". Причиной этого может быть конфликт с президентом Якутии Вячеславом Штыровым. Отставка Ничипорука еще не утверждена, однако уже полагают, что на его место придет гендиректор инвестгруппы "АЛРОСА" Сергей Выборнов, пишет "Коммерсантъ".

Недавно Штыров был утвержден на второй срок на посту президента Якутии. До того, как придти на должность главы республики, он был руководителем "АЛРОСА". Пришедший ему на смену Ничипорук должен был провести федерализацию "АЛРОСА", в ходе которой пакет РФ должен увеличиться с нынешних 37% до 50% плюс одна акция. В конце прошлого года федеральному центру удалось оспорить у Якутии права на большую часть имущества ПНО "Якуталмаз", на мощностях которого создавалась "АЛРОСА".

2007.02.03 00:29
Издательский дом "Коммерсантъ"
http://www.kommersant.ru/doc.html?DocID=739530&IssueId=36192
UN07020300581
*rus-feb07-000001-100000.txt *** 45134*

Алмазный фронт

Президент Якутии выжил своего преемника из АЛРОСА

Фото: Дмитрий Костюков / Коммерсантъ

Александр Ничипорук покидает пост президента компании АЛРОСА, на которую приходится четверть мировой добычи алмазов. Господин Ничипорук выполнил поставленную федеральным центром задачу - вернул государству контроль над компанией. Но испортил отношения с президентом Якутии Вячеславом Штыровым. Что, скорее всего, и стало причиной отсавки.

"Президент АК АЛРОСА Александр Ничипорук уходит со своего поста, он принял это решение в связи с выполнением задач, поставленных перед ним акционерами",- говорится в официальном сообщении компании. Кто может занять место господина Ничипорука, в компании не сообщают. По информации Ъ, один из наиболее вероятных претендентов на этот пост - президент инвестиционной группы АЛРОСА Сергей Выборнов. Сам он вчера вечером не отвечал на звонки по мобильному телефону.

Такое развитие событий не стало неожиданностью для участников алмазного рынка. После утверждения Вячеслава Штырова президентом Якутии на второй срок (его инаугурация прошла в минувшие выходные) уход из АЛРОСА его давнего оппонента Александра Ничипорука являлся лишь вопросом времени.

Отношения между нынешним и бывшим президентами АЛРОСА (господин Штыров возглавлял компанию до того как стал президентом Якутии) не сложились с момента прихода господина Ничипорука два года назад. Последний был поставлен во главе алмазного монополиста с вполне определенной целью - провести федерализацию АЛРОСА, в ходе которой пакет РФ должен увеличиться с нынешних 37% до 50% плюс одна акция. Господин Штыров до последнего стремился сохранить влияние республиканских властей в компании, и это стало причиной его конфликта с господином Ничепоруком. Тем не менее в конце прошлого года АЛРОСА фактически перешла под контроль федерального центра: РФ удалось оспорить у Якутии права на большую часть имущества ПНО "Якуталмаз", на мощностях которого создавалась АЛРОСА. В Якутии считают, что назначение на новый срок Вячеслав Штыров получил в обмен на уступки федеральным властям.

Как утверждают в Минфине (глава ведомства Алексей Кудрин возглавляет наблюдательный совет АЛРОСА), отставка господина Ничипорука пока не утверждена. "Этот вопрос входит в компетенцию наблюдательного совета",- сообщил Ъ представитель Минфина, заявив, что совет пройдет в среду.

ЕЛЕНА Ъ-КИСЕЛЕВА

Пример результатов работы модуля построения тематических цепочек

Тематические цепочки с указанием общей частоты и схожести вложенных элементов с центральными элементами цепочек (представлены цепочки с частотой не менее 5):

1. Частота: 103 (АЛРОСА)
   0,42 (АК, АЛРОСА)
2. Частота: 66 (КОМПАНИЯ)
   0,43 (АКЦИЯ, КОМПАНИЯ)
   0,43 (ВЛАДЕЛЕЦ, КОМПАНИЯ)
   0,43 (ОБЪЕДИНИТЬ, КОМПАНИЯ)
   0,40 (ПРЕЗИДЕНТ, КОМПАНИЯ)
   0,26 (АКЦИЯ)
   0,23 (АКЦИОНЕР, КОМПАНИЯ)
   0,08 (ВЛАДЕЛЕЦ)
   0,00 (ВЛАДЕНИЕ)
   0,00 (СОСТАВ, ВЛАДЕЛЕЦ)
3. Частота: 50 (НИЧИПОРУКА)
   0,42 (АЛЕКСАНДР, НИЧИПОРУКА)
4. Частота: 46 (ПРЕЗИДЕНТ)

0,56 (ПРЕЗИДЕНТ, КОМПАНИЯ)
0,50 (ПРЕЗИДЕНТ, РФ)
0,47 (ПРЕЗИДЕНТ, КОНЦЕРН)
0,46 (ОТСТАВКА, ПРЕЗИДЕНТ)
0,43 (ПРЕЗИДЕНТСТВО)
0,43 (ВИЦЕ, ПРЕЗИДЕНТ)
0,42 (БЫВШИЙ, ПРЕЗИДЕНТ)
0,42 (ПРЕЗИДЕНТ, РОССИЯ)
0,42 (ПРЕЗИДЕНТСКИЙ, ВЫБОРЫ)
0,41 (ПОЛНОМОЧИЕ, ПРЕЗИДЕНТ)
0,20 (ВИЦЕ-ПРЕЗИДЕНТ)

5. Частота: 46 (ГОД)
6. Частота: 45 (ЯКУТИЯ)
0,46 (ПРЕЗИДЕНТ, ЯКУТИЯ)
0,26 (ЯКУТСКИЙ)
0,20 (ЯКУТСКИЙ, ПРЕЗИДЕНТ)
7. Частота: 44 (АЛМАЗОДОБЫВАЮЩИЙ)
0,49 (АЛМАЗНЫЙ)
0,49 (АЛМАЗ)
0,48 (ДОБЫЧА, АЛМАЗ)
0,44 (АЛМАЗНО-БРИЛЛИАНТОВЫЙ, КОМПЛЕКС)
0,40 (АЛМАЗНО, БРИЛЛИАНТОВЫЙ, КОМПЛЕКС)
0,23 (АЛМАЗНЫЙ, МЕСТОРОЖДЕНИЕ)
0,10 (ДОБЫЧА)
8. Частота: 38 (ПОСТ)
0,44 (УХОД, С, ПОСТ)
0,42 (ДОЛЖНОСТЬ)
0,32 (ОТСТАВКА)
0,31 (УХОД)
0,24 (ОТСТАВКА, С, ДОЛЖНОСТЬ)
0,15 (УХОД, В, ОТСТАВКА)
0,13 (ОТСТАВКА, ПРЕЗИДЕНТ)
9. Частота: 29 (НОРИЛЬСКИЙ, НИКЕЛЬ)
0,59 (ГМК, НОРИЛЬСКИЙ, НИКЕЛЬ)
0,53 (НОРИЛЬСКИЙ)
0,50 (РАО, НОРИЛЬСКИЙ, НИКЕЛЬ)
0,35 (НОРНИКЕЛЬ)
10. Частота: 28 (АЛЕКСАНДРА)
0,28 (АЛЕКСАНДР, НИЧИПОРУКА)
0,25 (АЛЕКСАНДР)
11. Частота: 24 (ПАКЕТ, АКЦИЯ)
0,43 (КОНТРОЛЬНЫЙ, ПАКЕТ, АКЦИЯ)
0,43 (АКЦИЯ)
0,40 (КОНТРОЛЬНЫЙ, ПАКЕТ)
0,36 (ПАКЕТ)
12. Частота: 22 (РОССИЙСКИЙ, ФЕДЕРАЦИЯ)
0,54 (РОССИЯ)
0,48 (РОССИЙСКИЙ)
0,41 (ФЕДЕРАЛЬНЫЙ, ЦЕНТР)
0,28 (РФ)
0,23 (РОССИЙСКО-ФРАНЦУЗСКИЙ)
0,23 (ФЕДЕРАЛИЗАЦИЯ)
13. Частота: 21 (АКЦИОНЕР)
0,48 (АКЦИОНЕР, КОМПАНИЯ)
0,44 (МАЖОРИТАРНЫЙ, АКЦИОНЕР)
0,44 (МИНОРИТАРНЫЙ, АКЦИОНЕР)
0,22 (МИНОРИТАРИЕ)
14. Частота: 19 (ПОТАНИН)
0,27 (ВЛАДИМИР, ПОТАНИН)
15. Частота: 18 (РЕСПУБЛИКА)
0,51 (РЕСПУБЛИКА, САХА, ЯКУТИЯ)
0,45 (РЕСПУБЛИКАНСКИЙ)
0,41 (РЕСПУБЛИКА, САХА)

0,35 (ГОСУДАРСТВЕННЫЙ, СОБРАНИЕ, РЕСПУБЛИКА, САХА)
0,35 (ГЛАВА, РЕСПУБЛИКА)
0,03 (ГЛАВА)
0,01 (ВЕДОМСТВО)
0,00 (ГЛАВА, ВЕДОМСТВО)
16. Частота: 17 (ИНВЕСТГРУППА)
0,63 (ИНВЕСТИЦИОННЫЙ, ГРУППА)
0,40 (ИНВЕСТОР)
0,38 (ИНВЕСТИЦИЯ)
17. Частота: 17 (ГОСУДАРСТВЕННЫЙ, СОБСТВЕННОСТЬ)
0,42 (СОБСТВЕННОСТЬ)
0,40 (ГОСУДАРСТВЕННЫЙ, КОМПАНИЯ)
0,40 (ФЕДЕРАЛЬНЫЙ, СОБСТВЕННОСТЬ)
0,35 (ГОСУДАРСТВЕННЫЙ, КОРПОРАЦИЯ)
0,23 (ГОСУДАРСТВО)
0,20 (ГОСУДАРСТВЕННЫЙ, СТРУКТУРА)
0,00 (КОРПОРАЦИЯ)
18. Частота: 16 (ИНТЕРРОС)
0,27 (ИНТЕРЕС)
0,25 (ИНТЕРФАКС)
0,25 (ИНТЕРРОСОВСКИЙ)
0,23 (КОНФЛИКТ, ИНТЕРЕС)
0,00 (КОНФЛИКТ)
19. Частота: 15 (ОСВОЕНИЕ, МЕСТОРОЖДЕНИЕ)
0,45 (РАЗРАБОТКА, МЕСТОРОЖДЕНИЕ)
0,30 (АЛМАЗНЫЙ, МЕСТОРОЖДЕНИЕ)
0,24 (МЕСТО)
20. Частота: 14 (ШТЫРОВ)
0,37 (ШТЫРОВА)
21. Частота: 14 (РЫНОК)
0,41 (МИРОВОЙ, РЫНОК)
0,16 (ЭКОНОМИЧЕСКИЙ)
0,00 (МЕЖДУНАРОДНЫЙ, ЭКОНОМИЧЕСКИЙ, ОТНОШЕНИЕ)
0,00 (РЕФОРМИРОВАНИЕ, ЭКОНОМИКА)
0,00 (МИРОВОЙ, ЭКОНОМИКА)
22. Частота: 13 (МИНИСТЕРСТВО)
0,43 (КОЛЛЕГИЯ, МИНИСТЕРСТВО)
0,38 (МИНИСТР, ФИНАНСЫ)
0,00 (ФИНАНСОВЫЙ, ИНСТИТУТ)
0,00 (ФИНАНСЫ)
0,00 (ФИНАНСОВО-ЭКОНОМИЧЕСКИЙ)
0,00 (ФИНАНСОВЫЙ, ДЕЯТЕЛЬНОСТЬ)
23. Частота: 13 (ЗОЛОТО)
0,43 (ЗОЛОТОДОБЫВАЮЩИЙ)
24. Частота: 13 (ПРЕДСТАВИТЕЛЬ)
0,29 (ПРЕДСТОЯЩИЙ)
0,28 (ПРЕДСЕДАТЕЛЬ)
0,21 (ПРЕДСЕДАТЕЛЬ, СОВЕТ, ДИРЕКТОР)
25. Частота: 13 (ПРОХОРОВ)
26. Частота: 12 (ЗАДАЧА)
27. Частота: 12 (ГЕНДИРЕКТОР)
0,33 (ГЕНЕРАЛЬНЫЙ, ДИРЕКТОР)
0,22 (ДИРЕКТОР)
28. Частота: 10 (ВЫПОЛНЕНИЕ)
29. Частота: 10 (ВЯЧЕСЛАВ)
30. Частота: 10 (СЕРГЕЙ, ВЫБОРНОВ)
31. Частота: 10 (АКТИВ)
0,25 (АКТИВИЗАЦИЯ)
32. Частота: 9 (ГОСПОДИН)
33. Частота: 9 (ДОХОД, БЮДЖЕТ)
0,41 (ДОХОДНЫЙ)
0,41 (БЮДЖЕТ)
0,25 (ФЕДЕРАЛЬНЫЙ, БЮДЖЕТ)

0,20 (БЮДЖЕТНЫЙ, КОДЕКС)
34. Частота: 9 (МЕЖДУНАРОДНЫЙ, ОТНОШЕНИЕ)
0,48 (МЕЖДУНАРОДНЫЙ)
0,40 (МЕЖДУНАРОДНЫЙ, КОНТАКТ)
0,24 (ОТНОШЕНИЕ)
35. Частота: 9 (РУКОВОДСТВО)
0,51 (СМЕНА, РУКОВОДСТВО)
0,45 (РУКОВОДИТЕЛЬ)
36. Частота: 9 (ИМУЩЕСТВЕННЫЙ, СПОР)
0,42 (СПОР)
0,38 (ИМУЩЕСТВО)
0,20 (ИМУЩЕСТВЕННЫЙ, КОМПЛЕКС)
37. Частота: 8 (НАБЛЮДАТЕЛЬНЫЙ, СОВЕТ)
0,41 (СОВЕТ)
0,23 (СОВЕТНИК)
38. Частота: 8 (МОСКВА)
0,34 (МОСКОВСКИЙ)
39. Частота: 8 (РЕАЛИЗАЦИЯ)
40. Частота: 8 (ЧИСТЫЙ, ПРИБЫЛЬ)
0,40 (ПРИБЫЛЬ, КОМПАНИЯ)
41. Частота: 7 (ПОЛЮС)
42. Частота: 7 (ОАО)
43. Частота: 7 (РАЗВИТИЕ)
44. Частота: 7 (ГРУППА)
45. Частота: 7 (ДОЛЯ)
46. Частота: 7 (ВОПРОС)
47. Частота: 7 (УПРАВЛЕНИЕ)
0,43 (УПРАВЛЕНЕЦ)
48. Частота: 7 (КОНТРОЛЬ)
0,53 (КОНТРОЛЬНЫЙ)
49. Частота: 7 (МИРОВОЙ)
0,25 (МИРО)
50. Частота: 7 (РОСИМУЩЕСТВО)
51. Частота: 7 (КОМПЕНСАЦИЯ)
0,27 (КОМПЕТЕНЦИЯ)
52. Частота: 7 (НОВЫЙ)
53. Частота: 7 (МЛРД)
54. Частота: 7 (ОСНОВНОЙ)
55. Частота: 7 (ОЧЕРЕДЬ)
56. Частота: 7 (ИГО)
57. Частота: 6 (ПАРТНЕР)
0,41 (УПРАВЛЯЮЩИЙ, ПАРТНЕР)
58. Частота: 6 (РЕШЕНИЕ)
59. Частота: 6 (ПРИЧИНА)
0,25 (ПРИЧИННО-СЛЕДСТВЕННЫЙ)
60. Частота: 6 (СУЩЕСТВЕННЫЙ)
0,26 (СУЩЕСТВОВАНИЕ)
61. Частота: 6 (ОКОНЧАТЕЛЬНЫЙ)
0,25 (ОКОНЧАНИЕ)
62. Частота: 6 (ПРОШЛЫЙ)
63. Частота: 6 (ФЕВРАЛЬ)
64. Частота: 6 (КОНЕЦ)
65. Частота: 6 (МИРОВОЙ, СОГЛАШЕНИЕ)
0,47 (СОГЛАШЕНИЕ)
66. Частота: 6 (ПОСЛЕДНИЙ)
67. Частота: 6 (АРХАНГЕЛЬСКИЙ, ОБЛАСТЬ)
68. Частота: 6 (ЛОМОНОСОВ)
69. Частота: 5 (ВЫКУП)
70. Частота: 5 (МЕХАНИЗМ)
0,45 (МЕХАНИЗМА)
71. Частота: 5 (УЧАСТНИК, РЫНОК)
0,43 (УЧАСТНИК)
0,23 (УЧАСТИЕ)

72. Частота: 5 (ВЛАДИМИР)
0,24 (ВЛАДИМИР, ПУТИН)
73. Частота: 5 (НЫНЕШНИЙ)
74. Частота: 5 (МИХАИЛ)
75. Частота: 5 (СТАТЬ)
76. Частота: 5 (КОММЕНТАРИЙ)
0,25 (КОММЕРСАНТЪ)
77. Частота: 5 (БИЗНЕС)
0,25 (БИЗНЕС-КРУГ)
78. Частота: 5 (ОБЪЕМ)
79. Частота: 5 (ПРЕСС-СЛУЖБА)
0,25 (ПРЕСС-РЕЛИЗ)
0,25 (ПРЕСС-КОНФЕРЕНЦИЯ)
80. Частота: 5 (ДЕКАБРЬ)
81. Частота: 5 (ПРОДУКЦИЯ)
82. Частота: 5 (УЛУС)
83. Частота: 5 (ГОСКОМПАНИЯ)
0,45 (ГОСКОРПОРАЦИЯ)
0,25 (ГОСКОНЦЕРН)
84. Частота: 5 (РЕКОРДНЫЙ)
85. Частота: 5 (ДАЛЬНЕЙШИЙ)

Пример результатов работы модуля автоматического аннотирования.

Классический метод MMR (см. Раздел 4.3):

Президент "АЛРОСА" Александр Ничипорук уходит со своего поста.	Показать
1. Росимуществу принадлежит 37 % акций АЛРОСА, минимуществу Якутии - 32 %, восьми улусам Якутии - 8 %, физическим и юридическим лицам - 23 %, из которых ВТБ владеет 10, 5 % акций.	Показать
2. Напомним, что до Александра Ничипорука АК АЛРОСА возглавлял Владимир Калитин (март 2002 - декабрь 2004 года), а еще ранее - действующий президент Якутии Вячеслав Штыров (1996-2002 годы).	Показать
3. Отношения между нынешним и бывшим президентами АЛРОСА (господин Штыров возглавлял компанию до того как стал президентом Якутии) не сложились с момента прихода господина Ничипорука два года назад.	Показать
4. Александр Ничипорук покидает пост президента компании АЛРОСА, на которую приходится четверть мировой добычи алмазов.	Показать
5. Ничипорук принял решение об уходе с поста президента "АЛРОСЫ" в связи с выполнением задач, поставленных перед ним акционерами.	Показать
Total length: 118 words	

Классический метод MMR с интеграцией тематических цепочек:

Президент "АЛРОСА" Александр Ничипорук уходит со своего поста.	Показать
1. По мнению участников рынка, причиной отставки стал передел контроля над алмазным монополистом, 37 % акций которого принадлежит Росимуществу, а 32 % - Министерству по управлению госимуществом Якутии.	Показать
2. Отношения между нынешним и бывшим президентами АЛРОСА (господин Штыров возглавлял компанию до того как стал президентом Якутии) не сложились с момента прихода господина Ничипорука два года назад.	Показать
3. Как отмечается, Александр Ничипорук выполнил основную задачу - защитил госсобственность в алмазно-бриллиантовом комплексе, включая выкуп существенного пакета акций у миноритарных акционеров.	Показать
4. Через год был назначен президентом концерна, занимающего 25 % мирового рынка алмазов.	Показать
5. Добиться комментариев руководства "Норильского никеля" не удалось.	Показать
6. Странно на первый взгляд : именно на прошлой неделе, 31 января, группа "Интеррос" объявила, что Прохоров и Потанин решили весь свой бизнес разделить.	Показать
Total length: 119 words	

Классический метод SumBasic:

Александр Ничипорук уходит из АЛРОСА.	Показать
1. Восьми улусам Якутии принадлежит 8 % акций компании.	Показать
2. Ничипорук был президентом "АЛРОСА" с декабря 2004 года.	Показать
3. Не понарошку ли "разводятся" Владимир Потанин и Михаил Прохоров.	Показать
4. Добиться комментариев руководства "Норильского никеля" не удалось.	Показать
5. Ничипорук принял решение об уходе с поста президента "АЛРОСЫ" в связи с выполнением задач, поставленных перед ним акционерами.	Показать
6. К тому же ИГ АЛРОСА управляет проектом освоения месторождения алмазов им.	Показать
7. Такое развитие событий не стало неожиданностью для участников алмазного рынка.	Показать
8. Новым президентом госкомпании может стать генеральный директор инвестиционной группы "АЛРОСА" Сергей Выборнов.	Показать
9. М. В. Ломоносова в Архангельской области.	Показать
10. МОСКВА, 3 февраля.	Показать
Total length: 100 words	

Классический метод SumBasic с интеграцией тематических цепочек:

Александр Ничипорук уходит из АЛРОСА.	Показать
1. Таким образом, у компании будет уже третья смена руководства за десять лет.	Показать
2. Президент АК "АЛРОСА" Александр Ничипорук покидает свой пост.	Показать
3. Восьми улусам Якутии принадлежит 8 % акций компании.	Показать
4. Алмазодобывающей компании принадлежит контрольный пакет акций инвестгруппы.	Показать
5. Ничипорук принял решение об уходе с поста президента "АЛРОСЫ" в связи с выполнением задач, поставленных перед ним акционерами.	Показать
6. Таким образом, контроль над алмазным гигантом получил федеральный центр.	Показать
7. Владимир Калитин был связан с президентом Якутии Вячеславом Штыровым.	Показать
8. Не понарошку ли "разводятся" Владимир Потанин и Михаил Прохоров.	Показать
9. Добиться комментариев руководства "Норильского никеля" не удалось.	Показать
10. МОСКВА, 3 февраля.	Показать
11. К тому же ИГ АЛРОСА управляет проектом освоения месторождения алмазов им.	Показать
Total length: 109 words	

Приложение 2

Интерфейсы разработанного программного комплекса

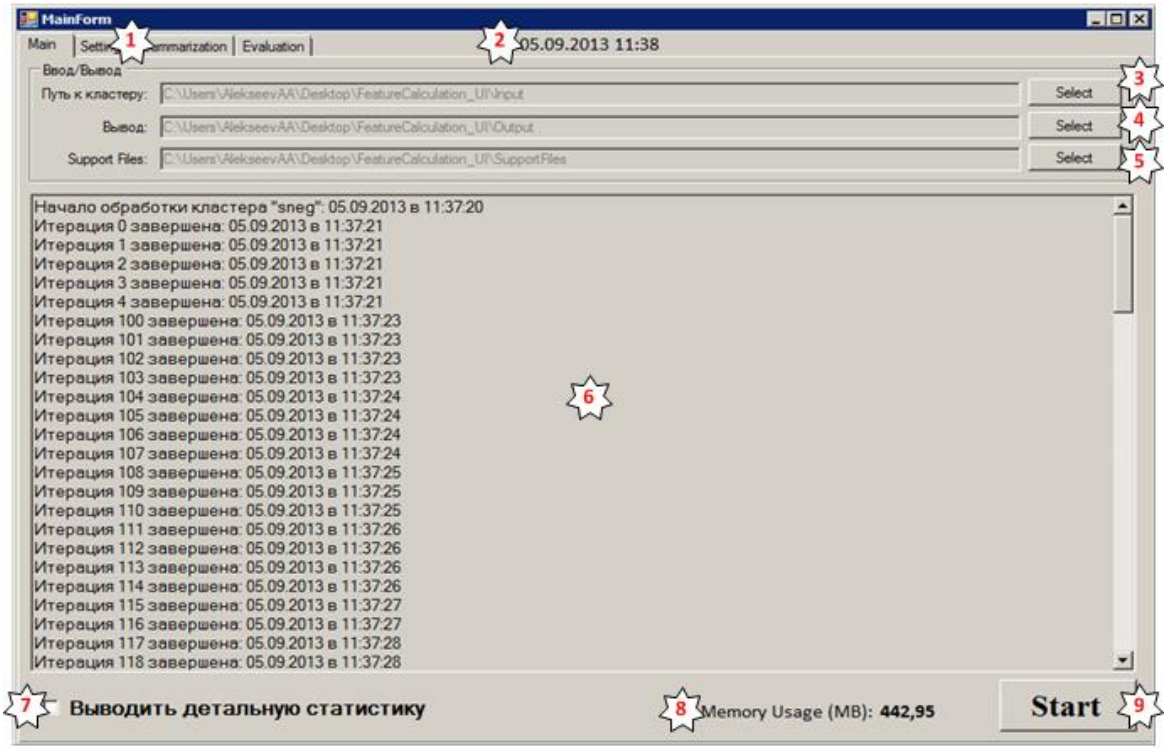


Рис. 12: Интерфейс разработанного программного комплекса – основная форма Main

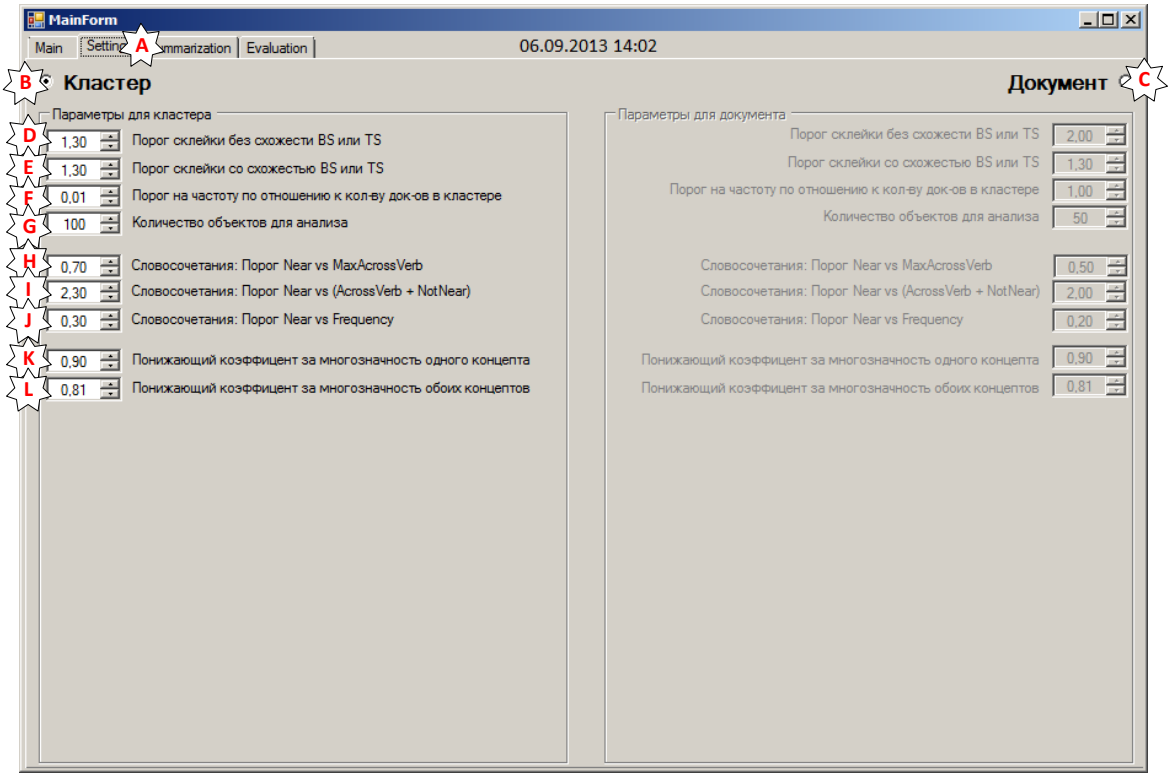


Рис. 13: Форма ввода настроечных параметров алгоритма построения тематических цепочек

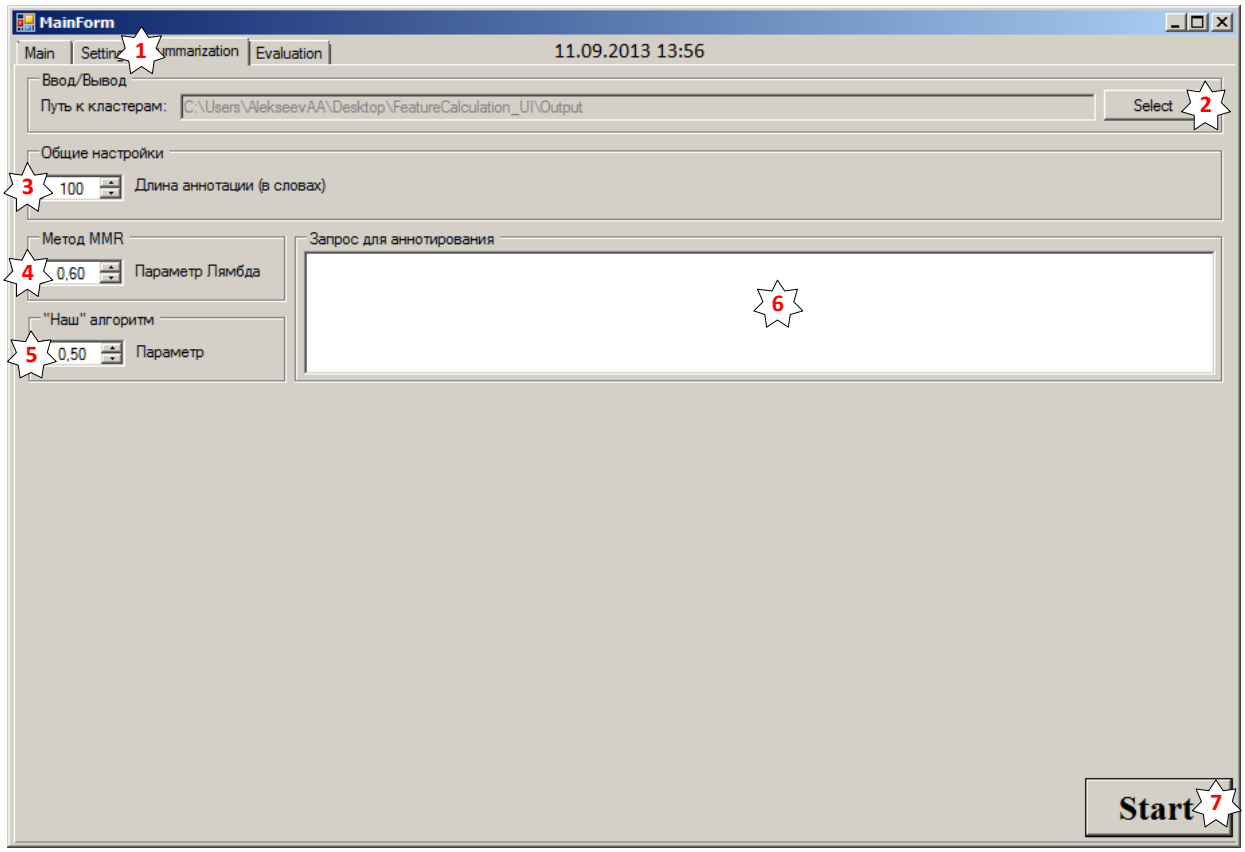


Рис. 14: Форма ввода настроечных параметров модуля автоматического аннотирования

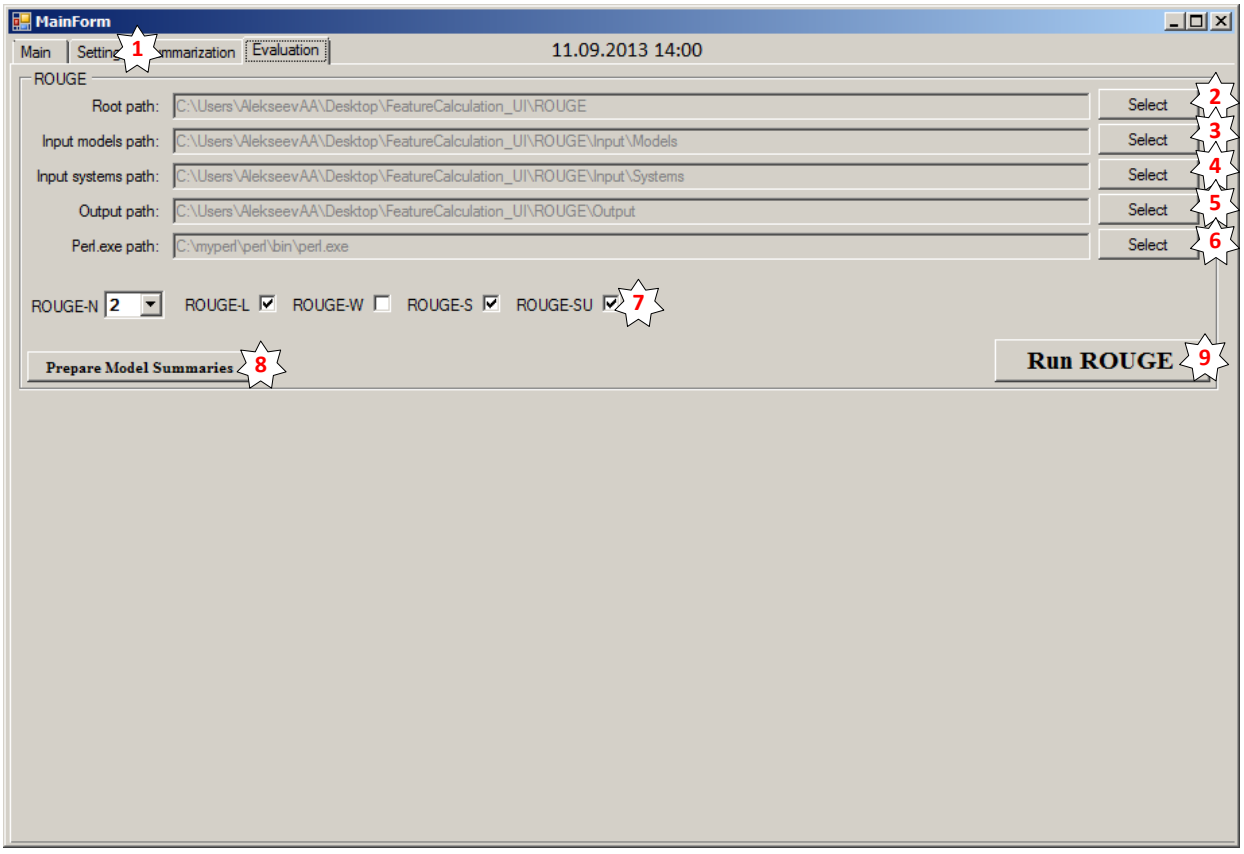


Рис. 15: Форма ввода настроечных параметров для модуля оценки автоматических аннотаций