Тематическое представление новостного кластера как основа для автоматического аннотирования

© А.А. Алексеев
Московский государственный университет имени М.В. Ломоносова,
г. Москва
a.a.alekseevv@gmail.com

Аннотация

В работе предложен метод извлечения цепочек семантически близких слов и выражений, описывающих различных участников сюжета – тематических узлов. Метод основан на объединении различных факторов схожести, включающих структурную организацию новостных кластеров, анализ контекстов вхождения языковых выражений, а также информацию из предопределенных ресурсов. Контексты слов используются в качестве базиса для извлечения многословных выражений и построения тематических узлов. Оценка предложенного алгоритма произведена в задаче построения обзорных рефератов новостных кластеров.

1 Введение

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 [17].

Кластер документов должен соответствовать ситуации или совокупности связанных ситуаций (основная тема кластера) [2, 17]. 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.

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, значит, повысить качество различных операций с новостными кластерами, например, автоматическое аннотирование, определение новизны информации и др.

В данной работе предлагается модель представления содержания новостного кластера, описывающая основных участников ситуации с учетом вариативности их именования – тематическое представление новостного кластера. Мы рассмотрим 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. Метод 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.

Статья организована следующим образом: после обзора существующих подходов, приведенного в разделе 2, в разделе 3 обсуждается теоретическая основа предлагаемого алгоритма, в частности, модель связанного текста. Подробное описание предлагаемого алгоритма и интеграция результатов работы алгоритма в методы автоматического аннотирования представлены в разделах 4 и 5 соответственно. Оценка полученных результатов произведена в разделе 6. Все примеры взяты из новостного кластера, посвященного смене руководства алмазодобывающей компании «Алроса», содержащего 12 новостных документов.

2 Обзор существующих методов

Проблема определения вариативности именования в текстах является актуальной для различных задач автоматической обработки естественного языка. С формальной точки зрения данная проблема является задачей группирования набора языковых выражений входной текстовой коллекции на тематические группы, относящиеся к одинаковым сущностям. Существует ряд различных подходов со схожими постановками задач, наиболее близкими из которых являются:

- построение лексических цепочек;
- построение референциальных цепочек;

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

- вероятностные тематические модели (в частности, LDA).

Лексическая цепочка представляет собой последовательность семантически связанных слов (повторы, синонимы, гипонимы, гиперонимы и др.) и является известным подходом к моделированию связности текста на естественном языке [10, 14, 21]. Алгоритмы построения лексических цепочек основаны на использовании информации о связях между словами и выражениями, описанных в некотором заранее определенном ресурсе, например, тезаурусе английского языка WordNet и тезаурусе русского языка РуТез.

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.

Использование лексических цепочек является из существующих подходов идейно наиболее близким к предлагаемому в данной статье. Основные отличия предлагаемого подхода заключаются в расширении рассмотрения с одного документа на коллекцию документов, а также в использовании совокупности различных факторов для группирования семантически близких слов и выражений (а не только информации, описанной в предопределенном ресурсе).

Частично проблему разного именования именованных сущностей снимают посредством ***установления кореференции имен*** (***референциальных цепочек***), прежде всего, для людей и организаций (*Президент Российской Федерации Дмитрий Медведев, Президент Медведев, Дмитрий Медведев*) [19].

На конференциях TDT и ACE рассматривалась задача по извлечению и прослеживанию упоминаний сущностей по цепочкам кореференции (Entity Detection and Tracking) [6]. Специфика данной задачи заключается в обнаружении ограниченного набора типов сущностей, существующих в реальном мире.

Вероятностные тематические модели, такие, как Latent Dirichlet Allocation [2, 3, 7], основаны на предположении, что документы на естественном языке являются комбинацией различных тем (топиков), в то время как каждая тема (топик) является вероятностным распределением над словами. В подобных моделях обычно рассматриваются два вероятностных распределения:

- Темы (Топики) vs Документы (распределение тем (топиков) по документам коллекции);
- Слова vs Темы (Топики) (распределение слов по темам (топикам)).

Восстановление информации об исходном распределении тем (топиков) основано на итеративном применении статистических методов (например, Семплинга Гиббса [7]), использующих информацию о совместном появлении слов в документах исследуемой коллекции. При этом подобный статистический вывод обычно не учитывает информацию о существующих лексических отношениях между словами и внутреннем устройстве текстов на естественном языке. Результаты работы алгоритмов, основанных на подобных моделях, имеют вероятностный результат и трудно интерпретируемы.

3 Тематическое представление

Как известно, текст обладает такими свойствами, как глобальная и локальная связности. Глобальная связность текста проявляется в том, что его содержание может быть представлено в виде иерархической структуры пропозиций [5]. Самая верхняя пропозиция представляет собой основную тему документа, а пропозиции нижних уровней представляют собой локальные или побочные темы документа.

Локальная связность, т. е. 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 (лексическая связность).

Пропозиция основной темы документа, т. е. взаимоотношения участников основной темы, должна находить свое отражение в конкретных предложениях текста, которые должны раскрывать и уточнять взаимоотношения между тематическими элементами. Если текст посвящен обсуждению взаимоотношений между тематическими элементами C_1, , C_n, то в предложениях текста должны обсуждаться детали этих отношений. Это проявляется в том, что сами тематические элементы C_1, , C_n или их лексические представители должны встречаться как разные актанты одних и тех же предикатов в конкретных предложениях текста.

Исходя из данных идей, для выявления участников ситуации, описываемой в исходном новостном кластере, мы сделали ряд следующих предположений:

1) взаимодействие участников описывается в предложениях текста, поэтому чем чаще слова (или выражения) встречаются в одних и тех же предложениях текста, тем больше вероятность того, что эти слова (или выражения) относятся к разным участникам ситуации;

2) каждому участнику в тексте соответствует группа слов и выражений; предполагается, что в тексте имеются наиболее частотное (главное название участника) и разные варианты, поэтому группа слов и выражений, относящихся к одному участнику, строится в форме узла, т. е. главное выражение и относящиеся к нему выражения – ***тематический узел***;

3) тематическое представление в предлагаемом подходе представляет собой совокупность выявленных тематических узлов и отношений между ними [9, 14].

Данные предположения основаны на внутреннем устройстве и тематической структуре текстов на естественном языке [5, 10]. Более подробная ин-

формация о сделанных предположениях, а также описание проведенных экспериментов по проверке сделанных гипотез, представлены в работе [1]. 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.

4 Алгоритм построения тематического представления

4.1 Контексты употребления слов

Важным фактором для построения тематического представления являются контексты, в которых употребляются слова и выражения. 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 (Near);
- во фрагментах, в которых есть глаголы, фиксируются прилагательные и существительные, между которыми и исходным словом встречается глагол (AcrossVerb);
- прилагательные и существительные, встречающиеся во фрагментах предложений с данным словом, не разделенные глаголом и не являющиеся соседними к исходному слову (NotN).

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

4.2 Сборка многословных выражений

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 [18]:

$$Near > 2 * (AcrossVerb + NotN). \quad (1)$$

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 + NotN). 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.

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

4.3 Характеристики для определения семантических связей

Для определения семантически связанных выражений и последующего построения тематических узлов 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 следующей секции будет дано подробное описание алгоритма расчета весов схожести пар выражений.

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 [5] и её следствии о том, что элементы одного тематического узла 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:

$$C = NS - 2 * (\mathrm{AcrossVerb} + Near + NotNear). \quad (2)$$

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

$$NSF = Min\left[0.5, \frac{C}{Avg(C)}\right], \quad (3)$$

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

NSF также является управляющей характеристикой. Это означает, что два выражения не могут быть включены в один и тот же тематический узел, если значение характеристики NSF имеет отрицательное значение. Подобная пара с отрицательным значением NSF не имеет общего веса и не рассматривается алгоритмом построения тематических узлов. Стоит

отметить, что подобная характеристика не использовалась раньше для подобных задач, таких, как определение вариантов именования основных участников ситуации, построение рядов квазисинонимов, а также лексических 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: по два слова справа и слева от рассматриваемого выражения. Чем больше одинаковых шаблонов разделяет пара-кандидат, тем больше схожесть по данной характеристике. Контексты с недостающей информацией (или неполные 4-граммы контекстов, в начале и конце предложений) получают меньший вес, чем целые шаблоны.

Вес шаблона строгого контекста рассчитывается следующим образом: каждое слово n-граммы шаблона контекста имеет вес, равный 0.25. Например, n-грамма (*, *, *состоит*, *из*) будет иметь вес 0.5, а n-грамма (*новостной*, *кластер*, *состоит*, *из*) будет иметь вес, равный 1.0, что является максимальным весом полного шаблона n-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.

Схожесть контекстов употребления по внутренним характеристикам предложения (**S**calar **P**roduct **S**imilarity, **SPS**). Каждый из контекстных параметров, описанных в разделе 4.1, представляет собой вектор частот, сопоставленных с каждым словом или выражением. Размерности данного вектора отражают частоту совместной встречаемости рассматриваемого слова или выражения со всеми остальными словами и выражениями, упомянутыми в новостном кластере. После построения данных контекстных векторов они могут быть сопоставлены классическими метриками схожести, например, такой, как косинусная мера угла между векторами. Характеристика SPS может быть рассмотрена как более сглаженная и гибкая характеристика по отношению к характеристике SC, так как обе данных характеристики основаны на контекстах употребления слов и выражений.

Значение характеристики SPS имеет вещественное значение, лежащее в пределах от 0 до 0.5 (половинный вес характеристики), и вычисляется как косинусная мера схожести по всем контекстным характеристикам (AcrossVerb, Near и NotNear), ограниченная сверху значением 0.5.

Контекстно-независимые характеристики:

Формальное сходство (**B**eginning **S**imilarity, **BS**). Рассмотрение формального сходства выражений является естественным путем обнаружения семантически связанных объектов. На текущий момент используется простая метрика схожести – одинаковые начала слов. Данная характеристика позволяет находить сходство между такими выражениями, как *Руководитель – Руководство, Президент России – Российский президент* и т. д.

Общий вес характеристики BS имеет вещественное значение из отрезка [0,1] и вычисляется по следующей формуле (в случае, если есть слова с одинаковыми началами, иначе вес равен нулю):

$$BS = 1.0 - 0.1 * N_{diff},$$

где N_{diff} – число слов с различными началами.

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 узлов 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 [13]. 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 вычисляется обратно-пропорционально расстоянию между объектами в тезаурусе:

$$TS = 1.0 - 0.2 * N_{rel},$$

где N_{rel} – длина пути по отношениям тезауруса (количество связей).

Наличие одинаковых языковых выражений (**E**mbedded **O**bjects **S**imilarity, **EOS**). При анализе схожести комплексных тематических узлов, включающих несколько языковых выражений, важным фактором схожести является наличие общих языковых выражений у двух различных узлов. Данный фактор особенно важен на поздних итерациях работы алгоритма, когда имеется значительное количество сформированных тематических узлов и остальные характеристики схожести уже проработаны. Значение данной характеристики является булевым и может добавлять 1 балл в общий вес схожести пары в случае наличия одинаковых языковых выражений.

Общий вес схожести пары рассматриваемых объектов вычисляется как сумма весов по отдельным характеристикам схожести, описанным выше. Таким образом, каждая пара получает вес, лежащий в пределах от 0 (отсутствие схожести) до 5 (максимальная схожесть), получаемый на основе шести характеристик (3 контекстно-зависимых и 3 контекстно-независимых), лежащих в пределах от 0 до 1 (SC, BS, TS, EOS) и от 0 до 0.5 (NSF, SPS). Пример ранжирования пар в соответствии с описанным алгоритмом приведен в Таблица 1 (топ-5 пар по общему весу на первой итерации работы алгоритма, характеристика EOS равна нулю для всех пар на первой итерации).

Features / Pairs	Context-independent		Context-dependent			SCORE
	BS	TS	NSF	SC	SPS	
Президент России – Президент РФ	0.90	1.00	0.00	0.50	0.34	**2.74**
Инвестгруппа– Инвестиционная группа	0.90	1.00	0.20	0.00	0.32	**2.42**
ГМК Норильский никель – Норильский никель	1.00	1.00	0.20	0.00	0.11	**2.31**
Российская Федерация – Россия	0.90	1.00	0.00	0.00	0.25	**2.15**
Отставка – Отставка с должности	0.90	1.00	0.20	0.00	0.00	**2.10**

Таблица 1: Пример ранжирования пар-кандидатов

4.4 Алгоритм построение тематического представления на основе совокупности факторов

Алгоритм построения тематического представления конструирует тематические узлы из пар выражений в порядке убывания их схожести. Предлагаемая структура тематического узла обладает следующими свойствами:

- текстовое выражение может принадлежать к одному или двум тематическим узлам; разрешение множественной принадлежности обеспечивает возможность представления различных аспектов исходного текстового выражения, а также его лексической многозначности;
- каждый тематический узел имеет главный элемент – центр тематического узла, который может принадлежать только к одному тематическому узлу; центр тематического узла является наиболее частотным элементом среди всех элементов тематического узла.

Построение тематического представления 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 нового тематического узла;
- менее частотный элемент рассматриваемой пары может в дальнейшем аналогичным образом присоединиться к другому тематическому узлу;
- объединение тематических узлов, состоящих из нескольких элементов, происходит аналогично объединению одиночных текстовых выражений; центр более частотного тематического узла становится центром нового, объединенного тематического узла.

В целом каждая итерация алгоритма состоит из трех основных шагов:

1. ранжирование пар-кандидатов;
2. выбор пары для объединения (наибольший вес + удовлетворение ограничений);
3. процедура объединения.

Итеративный процесс продолжается до тех пор, пока есть пары-кандидаты для объединения с весом схожести выше заданного порога. Например, тематический узел 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 узлы были получены в результате работы описанного алгоритма для кластера примера. Представлены 5 наиболее частотных тематических узлов в порядке убывания частоты. Данные узлы не подвергались какой-либо постобработке, центры тематических узлов 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*

5 Порождение аннотаций на основе тематического представления

Тематическое представление содержит в себе дополнительную информацию о внутреннем устройстве исходного новостного кластера, которая может быть использована для улучшения автоматических операций над текстовыми данными. Одной из таких задач является задача автоматического аннотирования, т. е. подготовки краткого изложения содержания исходного документа(ов). Решение данной задачи в значительной степени связано с наличием информации о различном именовании одних и тех же участников ситуации, описанной во входных документах, так как практически невозможно построить полную и не избыточную аннотацию без учета вариативности упоминаний наиболее значимых сущностей. В данном разделе описаны как существующие известные методы аннотирования (MMR, SumBasic), так и новые подходы, основанные на тематическом представлении. Все представленные подходы используются для оценки качества построенного тематического представления путем его интеграции в исходную структуру данных алгоритмов аннотирования.

5.1 Метод Maximal Marginal Relevance (MMR)

Метод Maximal Marginal Relevance для задачи многодокументного аннотирования является классическим не порождающим (выбирающим для аннотации целые предложения из исходных документов) методом аннотирования, который основан на концепции Maximal Marginal Relevance для информационного поиска [4]. В оригинале данный алгоритм является запрос-ориентированным, но существует также вариант и для общего аннотирования, когда в качестве запроса для аннотирования выступает исходных корпус документов.

Критерий MMR заключается в том, что лучшее предложение для аннотации должно быть максимально релевантным исходному набору документов и максимально отличным от всех предложений, уже отобранных в итоговую аннотацию.

Аннотация строится итеративно на основании списка ранжированных предложений. Предложение с максимальным значением MMR выбирается на каждой итерации алгоритма:

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

где S – множество предложений-кандидатов в аннотацию, E – множество отобранных в аннотацию; λ – представляет собой интерполяционный коэффициент между релевантностью отбираемых предложений и их не избыточностью; Sim_1 – метрика схожести между предложением и запросом для аннотирования (например, косинусная мера угла между векторами, широко применяемая в информационном поиске); Sim_2 может быть такой же, как Sim_1, или другой метрикой схожести. В нашей работе в качестве метрик Sim_1 и Sim_2 использовалась косинусная мера угла между векторами.

5.2 Метод SumBasic

SumBasic – алгоритм для общего многодокументного аннотирования [14, 15]. В его основе лежит эмпирическое наблюдение о том, что более частотные слова рассматриваемого кластера документов с большей вероятностью попадают в экспертные аннотации, нежели слова с низкой частотностью.

Алгоритм SumBasic строится на базе частотного распределения слов в исходном документе и состоит из пяти шагов. На первом шаге происходит расчет вероятностей слов исходного кластера $p(w_i)$:

$$p(w_i) = n/N,$$

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:

$$weight(S_j) = \sum w_i \in S_j \frac{p(w_i)}{|\{w_i \mid w_i \in S_j\}|}.$$

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:

$$p_{new}(w_i) = p_{old}(w_i) * p_{old}(w_i).$$

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

5.3 Аннотирование на основе тематического представления РуТез

В работе [20] предложен метод аннотирования на основе тематического представления, построенного на базе тезауруса русского языка РуТез. Одной из ключевых задач данного алгоритма является обеспечение одинаково высокого качества как полноты изложения информации, представленной в автоматической аннотации, так и её связности.

Алгоритм аннотирования на основе тематического представления на базе РуТез является итеративным, на каждой итерации отбирается по одному предложению. Поставленные цели по комбинации полноты и связности конечной аннотации решаются за счет наложения ограничений на этапе отбора предложений, а именно:

- для обеспечения полноты изложения информации предложение-кандидат должно содержать новый (ещё не упомянутый в отобранных предложениях) тематический узел;
- для обеспечения связности предложение-кандидат должно содержать уже упомянутый в отобранных предложениях тематический узел.

Из всех предложений, удовлетворяющих данным условиям, отбирается предложение с наибольшим весом тематических узлов – отражение основной темы исходного новостного кластера.

Алгоритм аннотирования на основе тематического представления на базе РуТез интересен нам в контексте его основы – тематического представления. Оно имеет схожую структуру с тематическим представлением, описанным в данной работе. При этом оно построено с использованием только одной характеристики – информации о наличии связей в тезаурусе РуТез. В данной работе добавлен ряд новых факторов, которые призваны обогатить полученное тематическое представление, а также повысить его качество.

5.4 Собственные методы аннотирования на основе тематического представления

Документы новостного кластера содержат в себе описание некоторого события (ситуации) или ряда связанных событий (ситуаций). Основная цель автоматического аннотирования заключается в наиболее полном отражении значимых фактов, относящихся к данному событию (ситуации) и описанных в исходной коллекции документов. Событие (ситуация), в свою очередь, характеризуется набором её участников. Под «фактом» в данном случае понимаются описание взаимоотношений между некоторыми участниками события (ситуации) или же детализация описания отдельного её участника.

Тематический узел предлагаемого тематического представления является воплощением некоторого участника события (ситуации) и в идеале должен содержать всевозможные варианты именования данного участника в рамках исходного новостного кластера. Таким образом, в основе предлагаемых методов аннотирования лежит учет наиболее значимых взаимоотношений тематических узлов построенного тематического представления – учет взаимоотношений основных участников события (ситуации). Предлагается два итеративных метода аннотирования, отличающихся стратегией учета значимости отношений между тематическими узлами. На каждой итерации в обоих подходах отбирается по одному предложению.

В качестве основы для расчета значимости отношений между тематическими узлами в первом алгоритме (OurSummary_Nodes) выступает значимость самих тематических узлов. Каждый тематический узел имеет вес, равный суммарной частоте его элементов. На каждой итерации в итоговую аннотацию отбирается предложение, содержащее три наиболее значимых и ещё не упомянутых тематических узла (ТУ_NEW):

$$s_i \Rightarrow \max\left(\sum_{TV\_NEW_j \in s_i,\, i=1..3}^{desc\,weight(TV\_NEW_j)} weight(TV\_NEW_j)\right).$$

В рамках второго предлагаемого алгоритма аннотирования (OurSummary_Relations) критерием для отбора предложения выступает наличие наиболее обсуждаемой и ещё не упомянутой пары тематических узлов. Для каждой пары тематических узлов предварительно рассчитывается количество вхождений в одни и те же предложения исходного новостного кластера – «обсуждаемость» пары. На каждой итерации отбирается предложение, содержащее наиболее обсуждаемую и неупомянутую в отобранных предложениях пару, а также обладающее наибольшим общим весом тематических узлов:

$$s_i \Rightarrow \max\left(\sum_{TV\_REL\_NEW_j \in s_i} weight(TV\_REL\_NEW_j)\right).$$

Итеративный процесс отбора предложений для аннотации в обоих алгоритмах продолжается до тех пор, пока не будет превышен заданный порог по количеству слов. Во всех генерируемых аннотациях данный порог равен 100 словам – стандартный размер аннотации для подобных задач на соревнованиях мирового уровня (DUC, TAC).

6 Оценка качества аннотаций

6.1 Методы оценки автоматических аннотаций ROUGE и Пирамид

Оценка качества порождаемых аннотаций является достаточно сложной процедурой. Несомненно, наиболее правдоподобные оценки можно получить при помощи ручной оценки путём привлечения большого количества экспертов. Но данный метод является очень дорогим и трудоёмким. Поэтому используются автоматические методы оценки качества аннотаций ROUGE [12] и формализованный метод Пирамид.

Метод ROUGE основан на автоматическом сравнении порожденной аннотации с эталонными аннотациями, созданными экспертами. Существуют различные модификации алгоритма, связанные с различными способами сравнения: сравнение n-грамм (ROUGE-N); сравнение максимальных общих последовательностей (ROUGE-L и ROUGE-W); сравнение пропусков монограмм и биграмм (ROUGE-S и ROUGE-SU). В статье [12] показано, что все основные ROUGE-метрики являются значимыми, так как в зависимости от специфики конкретной задачи каждая из метрик может иметь наилучшую корреляцию с ручными аннотациями.

В основе метода Пирамид также лежит сравнение автоматических аннотаций с эталонными аннотациями. Но в отличие от метода ROUGE данное сравнение происходит не в автоматическом режиме, а в ручном, на основании формализованного алгоритма сравнения. Эксперты выделяют из эталонных аннотаций все «информационные единицы» (**S**ummary **C**ontent **U**nits, **SCU**) – факты, описанные в аннотации. Каждая информационная единица получает вес пропорционально количеству упоминаний в экспертных аннотациях. Далее полученные информационные единицы вручную ищутся в автоматических аннотациях. Итоговая оценка аннотации равна общему весу упомянутых информационных единиц по отношению к суммарному весу информационных единиц, извлеченных для данного новостного кластера.

В данной работе оценка качества аннотаций построена на оценке методом ROUGE и дополнительной оценке методом Пирамид.

6.2 Автоматические аннотации и их оценка

Для оценки качества автоматических аннотаций были подготовлены 11 новостных кластеров по различным тематикам, собранные на основе пословной модели представления данных [17]. Независимые эксперты-лингвисты подготовили от 2 до 4 ручных аннотаций для каждого из данных кластеров. Все автоматические аннотации прошли единообразную обработку для автоматической оценки программным пакетом ROUGE [12]:

- ограничение аннотаций длиной свыше 100 слов (рассмотрение только первых 100 слов);
- приведение слов к соответствующим леммам;
- исключение стоп-слов и незначащих частей речи;
- транслитерация всех русскоязычных слов (пакет ROUGE работает только с латинскими символами);
- преобразование автоматических аннотаций в необходимый для пакета ROUGE входной формат (см. документацию пакета).

Всего в оценке участвовали 11 различных модификаций алгоритмов:

- классический пословный MMR в модификациях с учетом и без учета IDF (MMR_WithIDF и MMR_WithoutIDF соответственно);
- MMR с добавлением информации из построенного тематического представления (модификации с и без учета IDF, MMR_WithIDF+Groups и MMR+Groups соответственно);
- классический пословный SumBasic;
- SumBasic с добавлением информации из построенного тематического представления (SumBasic+Groups);
- аннотирование на основе тематического представления на базе тезауруса РуТез (ThematicLines);
- собственный алгоритм аннотирования на основе построенного тематического представления, по тематическим узлам (модификации с и без учета IDF, OurSummary_Nodes_WithIDF и OurSummary_Nodes соответственно);
- собственный алгоритм аннотирования на основе построенного тематического представления, по связям тематических узлов (модификации с и без учета IDF, OurSummary_Relations_WithIDF и OurSummary_Relations соответственно)

6.3 Результаты

В результате работы программного пакета ROUGE каждая автоматическая аннотация получает набор результатов по различным метрикам сопоставления автоматических аннотаций с аннотациями, составленными экспертами. По причине значимости различных ROUGE-метрик для задач с различной спецификой (см. раздел 6.1) в качестве основного параметра для сравнения автоматических аннотаций была взята средняя позиция в результатах по всем основным ROUGE-метрикам.

Метод	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-S	ROUGE-SU	Avg
MMR + Groups	**0,62499 (1)**	**0,41633 (1)**	**0,6021 (1)**	**0,35529 (1)**	**0,36649 (1)**	**1,0**
OurSummary_Nodes	0,58652 (2)	0,36154 (3)	0,5645 (2)	0,32113 (2)	0,33203 (2)	2,2
OurSummary_Nodes_WithIDF	0,58497 (3)	0,33918 (5)	0,55745 (3)	0,30124 (3)	0,31283 (3)	3,4
MMR_WithIDF	0,57623 (4)	0,38116 (2)	0,55503 (4)	0,29792 (4)	0,30971 (4)	3,6
MMR_WithoutIDF	0,56784 (5)	0,34595 (4)	0,55124 (5)	0,26092 (6)	0,27349 (6)	5,2
ThematicLines	0,53416 (6)	0,33364 (6)	0,51238 (6)	0,2713 (5)	0,28243 (5)	5,6
OurSummary_Relations	0,53141 (7)	0,2892 (7)	0,50422 (7)	0,25382 (7)	0,26509 (7)	7,0
SumBasic + Groups	0,52255 (8)	0,22881 (10)	0,493 (9)	0,24356 (8)	0,25525 (8)	8,6
SumBasic	0,51847 (9)	0,24735 (9)	0,49786 (8)	0,23064 (9)	0,24257 (9)	8,8
OurSummary_Relations_WithIDF	0,45494 (10)	0,24856 (8)	0,43768 (10)	0,19419 (11)	0,20492 (11)	10,0
MMR_WithIDF + Groups	0,44475 (11)	0,22238 (11)	0,42318 (11)	0,20627 (10)	0,21648 (10)	10,6

Таблица 2: Результаты оценки автоматических аннотаций методом ROUGE

В Таблица 2 приведены итоговые результаты оценки всех исследуемых модификаций алгоритмов по всем основным ROUGE-метрикам, а также агрегирующая оценка, по которой выполнена сортировка.

Для проведения дополнительной оценки качества автоматических аннотаций лучших и наиболее значимых для нас методов (с и без интеграции построенного тематического представления) была проведена альтернативная оценка полученных аннотаций методом Пирамид [8]. Результаты данной оценки представлены в Таблице 3.

Метод	Оценка по Пирамидам
MMR + Groups	**0,645 (1)**
MMR_WithIDF	0,617 (2)
OurSummary_Nodes	0,602 (3)
SumBasic + Groups	0,575 (4)
SumBasic	0,567 (5)

Таблица 3: Результаты оценки методом Пирамид

На основе результатов оценки полученных аннотаций методами ROUGE и Пирамид необходимо отметить, что:

- наилучший результат показал метод аннотирования, основанный на построенном тематическом представлении;
- добавление тематических узлов к обоим базовым методам улучшило результаты исходных методов;
- предлагаемое тематическое представление показало более высокий результат в аннотировании, чем тематическое представление только на основе тезауруса.

7 Заключение

В статье предложен алгоритм выявления семантически связанных слов и выражений, описывающих различных участников ситуации новостного кластера – тематических узлов. Предложенный алгоритм основан на совместном использовании характеристик схожести различной природы. В дополнение к известным контекстным характеристикам схожести, таким, как анализ жестких контекстов (шаблонов) употребления слов и выражений, используется характеристика, основанная на внутреннем устройстве текстов на естественном языке – анализ встречаемости в соседних предложениях кластера по отношению к встречаемости в одних и тех же предложениях. В едином алгоритме объединены характеристики следующих различных типов:

- формальное сходство слов и выражений;
- информация из предопределенных ресурсов (тезаурус русского языка РуТез [13]);
- контекстные характеристики схожести.

Оценка предложенного алгоритма производилась в контексте применения полученного тематического представления к задаче автоматического аннотирования. Полученные результаты подтверждают, что информация, заложенная в построенных тематических узлах, позволяет улучшать качество алгоритмов много документного аннотирования.

Литература

[1] Alekseev A., Loukachevitch N. *Use of Multiple Features for Extracting Topics from News Clusters* // Труды конференции SYRCODIS'2012, 2012. pp. 3-11.

[2] Allan J. *Introduction to Topic Detection and Tracking* // Topic detection and tracking, Kluwer Academic Publishers Norwell, MA, USA, 2002. pp. 1-16.

[3] Blei D., Ng A., Jordan M. *Latent Dirichlet Allocation* // Journal of Machine Learning Research, 3, 2003. pp. 993-1022.

[4] 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, Melbourne, Australia, 1998. pp. 335-336.

[5] Dijk van T. *Semantic Discourse Analysis* // Handbook of Discourse Analysis / Teun A. van Dijk, (Ed.), vol. 2. London: Academic Press, 1985. pp. 103-136.

[6] 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, 2004.

[7] Griffiths T., Steyvers M. *Finding scientific topics* // Proceedings of the National Academy of Sciences of the United States of America, Vol. 101, No. Suppl. 1, 2004. pp. 5228-5235.

[8] Harnly A., Nenkova A., Passonneau R., Ram-bow O. *Automation of summary evaluation by the pyramid method* // Proceedings of the International Conference on Recent Advances in Natural Language Processing, 2005.

[9] Hasan R. *Coherence and Cohesive harmony* // Understanding reading comprehension / J. Flood, editor, Newark, 1984. pp. 181-219.

[10] Hirst G., St-Onge D. *Lexical Chains as representation of context for the detection and correction malapropisms* // WordNet: An electronic lexical database and some of its applications / C. Fellbaum, editor. Cambrige, MA: The MIT Press, 1998.

[11] 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.

[12] Lin C.-Y. *ROUGE: a package for automatic evaluation of summaries* // Proceedings of the Workshop on Text Summarization Branches Out (ACL'2004), Barcelona, Spain, 2004. pp. 74-81.

[13] Loukachevitch N., Dobrov B. *Evaluation of Thesaurus on Sociopolitical Life as Information Retrieval Tool* // Proceedings of Third International Conference on Language Resources and Evaluation, Vol.1, 2002. pp. 115-121.

[14] Loukachevitch N. *Multigraph representation for lexical chaining* // Proceedings of SENSE workshop, 2009. pp. 67-76.

[15] Nenkova A., Vanderwende L. *The impact of frequency on summarization* // Microsoft Research Technical Report, MSR-TR-2005-101, 2005.

[16] Vanderwende L., Suzuki H., Brockett C., Nenkova A. *Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion* // Information Processing and Management Journal, Volume 43 Issue 6, November, 2007. pp. 1606-1618.

[17] Добров Б.В., Павлов А.М. *Исследование качества базовых методов кластеризации новостного потока в суточном временном окне* // Труды конференции RCDL'2010, 2010.

[18] 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.

[19] 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, 2005.

[20] Лукашевич Н.В., Добров Б.В. *Автоматическое аннотирование новостного кластера на основе тематического представления* // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2009, Вып. 8 (15), 2009. с. 299-305.

[21] Лукашевич Н.В., Добров Б.В. *Исследование тематической структуры текста на основе большого лингвистического ресурса* // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2000, 2000. с. 252-258.

Thematic representation of a news cluster as a basis for summarization

Aleksey A. Alekseev

In this paper we consider a method for extraction of various references of a concept or a named entity mentioned in a news cluster. The method is based on the structural organization of news clusters and exploits comparison of various word contexts. The word contexts are used as a basis for multiword expression extraction and main entity detection. At the end of cluster processing we obtain groups of thematically-related elements, in which the main element of a group is determined. Evaluation of the proposed algorithm is performed in news cluster summarization task.