А.А. Алексеев, Н.В. Лукашевич

Автоматическое извлечение сущностей на основе структуры новостного кластера

Аннотация. В данной работе рассматривается метод извлечения вариантов называния одной и той же сущности на основе структурной организации новостного кластера. Метод основан как на сравнении разного рода контекстов употребления выражений, так и на сопоставлении употребления выражений в одном и том же и соседнем предложениях.

Ключевые слова: извлечение сущностей, лексическая связность, новостные кластеры.

Введение

Современные технологии автоматической обработки новостных потоков основаны на тематической кластеризации новостных сообщений, т.е. выделении совокупностей новостей, посвященных одному и тому же событию. Именно новостные кластеры являются основными единицами представления информации в новостных сервисах таких, как yandex.news, google.news, rambler.news и др.

Важными этапами обработки полученного новостного кластера являются процедуры автоматической обработки текстов:

- из кластера удаляются дубликаты, т.е. сообщения, полностью по содержанию повторяющие сообщения первоисточников;
- кластер приписывается к той или иной рубрике новостного рубрикатора;
- создается аннотация кластера, обычно содержащая предложения из разных документов кластера;
- извлекаются разные типы информации, например, высказывания на тему кластера.

При этом само формирование кластера может представлять серьезную проблему. Особенно трудно правильно сформировать кластеры для сложных иерархических событий, имеющих некоторую длительность по времени и распределенное географическое расположение (чемпионат мира, выборы) [1, 2].

Часть проблем формирования и обработки новостных кластеров связана с тем, что в разных местах (текстах кластера) одно и то же понятие или сущность могут быть названы поразному. Подходы, основанные на лексических цепочках, частично могут преодолевать эту проблему за счет использования тезаурусных знаний [3, 4]. Однако варианты называния сущностей в том или ином кластере невозможно зафиксировать в заранее созданном ресурсе. Так, авиабаза США в Киргизии может называться в текстах одного новостного кластера, как *база Манас*, *авиабаза Манас*, *Манас*, *база в международном аэропорту Манас, база США, американская авиабаза* и др.

Частично проблему разного именования именованных сущностей снимают посредством установления кореференции имен, прежде всего, для людей и организаций (*Президент Российской Федерации Дмитрий Медведев, Президент Медведев, Дмитрий Медведев*) [5]. На конференциях TDT и ACE рассматривалась задача по извлечению и прослеживанию упоминаний сущностей, таких как организации, люди, места, по цепочкам кореференции (Entity Detection and Tracking) [6]. Однако вариативность называния в новостных кластерах относится не только к именам кон-

кретных сущностей, но и к ситуациям – *вывод авиабазы, закрытие авиабазы*.

В данной работе мы рассмотрим методы улучшения качества извлечения основных участников новостного события, описываемого в новостном кластере, что включает нахождение совокупности слов и выражений, с помощью которых тот или иной значимый участник события именовался в документах новостного кластера. Метод основан на использовании совокупности факторов, в том числе разного рода контекстов употребления слов в документах кластера.

1. Отражение участников события в документах новостного кластера

Вариативность называния участников новостного события может быть связана с такими языковыми явлениями как кореференция (т.е. отнесенность языкового выражения к одному и тому же объекту действительности) [7], перефразирование и лексическая связность. Вариативность называния участников события связана чаще с именной референцией, т.е. референцией, которая выражена именными группами.

В кластере могут встретиться следующие виды именной референции:

- конкретная референция: *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**;*

- неконкретная (обобщенная референция), т.е. отсылка не к конкретному объекту (ситуации), а к понятию: *Власти Киргизии не опасаются, что решение о закрытии базы может привести к **обострению взаимоотношений** с США и западноевропейскими государствами. Никаких **политических разногласий** у нас с США нет;*

- референция на один и тот же объект в разных текстах, когда один и тот же объект, явление в разных документах новостного кластера называется по-разному, например, *Киргизия и Кыргызстан; вывод авиабазы, закрытие авиабазы.*

По соотношению между кореференными именными группами могут встретиться такие случаи, как:

- кореференция выражена более короткой и более длинной именными группами: *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**."*

- кореференция выражена именными группами, частично совпадающими по своему составу: *за **вывод авиабазы** проголосовали 78 киргизских законодателей, против - один, двое воздержались...*

*Между тем, начинается скандал, который может ускорить **вывод американских войск...;***

- кореференция обозначается посредством слов, связанных между собой известными лексическими отношениями (синонимы, род-вид, часть-целое и др.), и, таким образом, возникает лексическая связность текста [8]:

- *На Сахалине в море уносит льдины со 100 **рыбаками**...Как сообщили в управлении информации МЧС, на льдинах находятся от 60 до 100 **человек**."*
- *"Судьба авиабазы — в руках **парламентариев**. **Парламент** должен проголосовать "за" или "против" вывода военных США со своей территории.";*
- ***Закон** о денонсации киргизско-американского соглашения был принят большинством парламентариев... Парламент Киргизии одобрил вывод авиабазы "Манас" с территории страны, приняв **законопроект** "О денонсации ответной ноты Министерства иностранных..."*
- *Тем временем, **Киргизия** провозгласила свое решение о дальнейшей судьбе базы "Манас" окончательным... В **Бишкеке** до последних дней демонстративно тянули с вынесением решения о выводе базы "Манас" на суд парламента."*

- кореференция осуществляется посредством выражений, связанных между собой на основе текущего контекста и требующих логического вывода на основе этого контекста. Например, в *декабре 2006 года 46-летний водитель топливозаправщика киргизской фирмы, занимающейся обслуживанием аэропорта "Манас", Александр Иванов, был расстрелян в*

*упор **охранником авиабазы** Закари Хатфилдом на КПП при въезде на перрон аэропорта"... **Американский военный**, несмотря на неоднократные требования киргизского МИДа, также был тайно вывезен с территории страны и до сих пор не предстал перед судом.*

Помимо явления референции имеется явление перефразирования (квазисинонимии) в нереферентном употреблении слов (выражений): *Судьбу авиабазы США в "Манасе" **решит** парламент Киргизии. Парламент Киргизии в четверг **примет окончательное решение** о судьбе авиабазы США.*

Таким образом, в новостных кластерах присутствует достаточно большая вариативность называния участников новостного события, которую трудно полностью описать заранее и распознавание которой в текстах не сводится к применению простых процедур анализа текста. В то же время нераспознавание этих языковых выражений как вариантов ссылки на одну и ту же сущность или понятие представляет собой проблему на разных этапах сборки и обработки новостного кластера. В частности, нераспознанная вариативность основных сущностей ведет к размыванию границ новостного кластера, при автоматическом аннотировании новостного кластера нераспознанные варианты приводят к излишней повторяемости в автоматической аннотации, снижению качества нахождения новой информации в поступающих новостных сообщениях.

Далее в тексте статьи все виды кореферентных выражений, перифраз, возникших в документах новостного кластера, мы будем называть квазисинонимами

Особенность новостного кластера как источника для извлечения квазисинонимов, перифраз, исследуется в ряде работ. Так, в работе [9] описывается процедура построения корпуса для извлечения перифраз в предметной области «терроризм». В работе [10] описывается построение корпуса похожих предложений из новостных кластеров широкой тематики как базы для последующего анализа перефразирования. Основным методом для распознавания различных способов перефразирования в данных работах является выявление похожих предложений и нацелен на накопление общелексических перифраз типа *участвововать – принимать участие*.

Наша работа сфокусирована на распознавании вариантов называния основных участников новостного события с целью применения полученной информации для улучшения качества обработки этого же кластера. Кроме того, одним из существенно новых принципов, на который можно опереться при распознавании вариативности называния основных сущностей, упоминаемых в кластере, является учет некоторых свойств связного текста, которые мы рассмотрим в следующем разделе.

2. Принципы обработки текстов новостного кластера

Как известно, текст обладает такими свойствами как глобальная и локальная связность. Глобальная связность текста проявляется в том, что содержание текста может быть представлено в виде иерархической структуры пропозиций [11]. 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). Лексическая связность моделируется посредством таких структур, как лексические цепочки [12].

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

Отсюда следует практический вывод: если даже очень близкие по смыслу лексические сущности C_1 и C_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. C_1 и C_2 соответствуют

разным тематическим элементам основной темы или подтемы текста, соответствуют разным обсуждаемым в тексте сущностям [13, 14]. С другой стороны, если две лексические сущности C_1 и C_2 редко встречаются в одних и тех же предложениях текстов, но часто в соседних предложениях, то это дает возможность предположить, что они используются для осуществления локальной связности, т.е. между ними имеется какая-то смысловая связь.

Новостной кластер не является единым связным текстом, но тексты кластера посвящены одной теме, и поэтому статистическое проявление свойств связного текста в новостном кластере многократно усиливается, что мы и пытаемся использовать для автоматического извлечения из кластера неизвестной заранее информации.

Для проверки вышеуказанной мысли о том, что квазисинонимы чаще встречаются в соседних предложениях, чем в одних и тех же простых предложениях текста, был предпринят следующий эксперимент. Десять новостных кластеров, каждый с количеством документов более 40, были сопоставлены с Общественно-политическим тезаурусом [4], и были выделены пары слов (выражений), между которыми в тезаурусе установлены типы отношений, которые чаще всего могут соответствовать квазисинонимам:

- синонимы-существительные (*Киргизия – Киргизстан*);
- дериваты прилагательные-существительные (*Киргизия – Киргизский*);
- отношения род-вид существительных (*депутат – представитель*);
- отношение род-вид прилагательное-существительное (*национальный – Россия*);
- отношение часть-целое существительные (*парламентарий – парламент*);
- отношение часть-целое прилагательное-существительное (*американский – Вашингтон*);
- отношение ассоциации существительные (*нефть – баррель*);
- отношение ассоциации существительное-прилагательное (*нефтяной – баррель*).

Для всех таких типов пар слов, употреблявшихся в текстах кластера с частотностью более четверти числа документов кластера, было подсчитано отношение встречаемости в пределах сегментов без запятых F_{segm} и встречаемости в соседних предложениях F_{sent}. В Табл. 1 отражены результаты данного подсчета.

Табл. 1. Соотношение частотностей встречаемости близких по смыслу пар выражений внутри сегментов предложений и в соседних предложениях

Тип отношения	Соотношение частотностей встречаемости внутри сегмента предложения и в соседних предложениях F_{segm}/F_{sent}	Количество пар
Синонимы-существительные	0,309	31
Дериваты прилагательные-существительные	0,491	53
Отношение род-вид между существительными	1,130	88
Отношение род-вид прилагательное-существительное	1,471	28
Отношение часть-целое существительные	0,779	58
Отношение часть-целое прилагательное-существительное	1,580	29
Отношение ассоциации существительные	1,248	37
Отношение ассоциации существительное-прилагательное	0,898	18
Выражения, между которыми не установлены перечисленные типы отношений	1,440	21483

Из таблицы видно, что самые близкие по смыслу выражения (синонимы, дериваты) значительно чаще встречаются в соседних предложениях, чем в одних и тех же фрагментах предложений. Далее с нарастанием смысловых различий между выражениями частота встречаемости в одних и тех же предложениях нарастает, пока не стабилизируется на некотором уровне.

Еще можно заметить, что пары существительное-существительное и существительное-прилагательное имеют существенно отличные величины соотношений. На наш взгляд это связано с тем, что в большинстве случаев прилагательные входят в состав именных групп, которые и выступают в роли наименований тех или иных участников описываемого события. Таким образом, прежде чем использовать выявленное соотношение по расположению близких

по смыслу или по денотату языковых выражений, необходимо автоматически выделить многословные выражения, служащие обозначениями тех или иных участников события.

3. Этапы обработки текстов новостного кластера

Обработка новостного кластера состоит из трех основных этапов. На первом этапе накапливаются контексты о существительных и прилагательных, упоминаемых в текстах новостного кластера. На втором этапе производится сбор многословных выражений, обозначающих значимую сущность, из отдельных слов. На третьем этапе производится поиск вариантов называния сущностей в текстах кластера. Общая схема обработки новостного кластера изображена на рисунке.

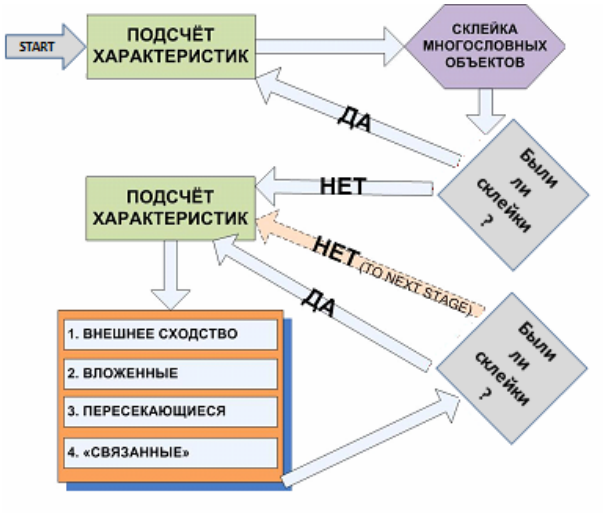
Рассмотрим эти этапы обработки новостных кластеров более подробно. В качестве примера будем использовать новостной кластер от 19.02.2009, посвященный денонсации соглашения между Киргизией и США по поводу авиабазы США, расположенной в международном аэропорту Манас. Кластер собран на основе алгоритма, описанного в [2] и содержит 195 новостных документов.

3.1. Извлечение контекстов употребления слов

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 (Av);
- 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



Общая схема обработки новостного кластера

предложений с начала и до фрагмента, содержащего глагол (включительно), что позволяет извлечь из соседних предложений наиболее значимые слова.

3.2. Распознавание многословных выражений, обозначающих отдельную сущность

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 [15]:

$$Near > 2 * (Av + NotN) \qquad (1)$$

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 / (Av + 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 *парламент Киргизии,*

американский военный, денонсация соглашений с США, президент Киргизии Курманбек Бакиев (кластер относится к 2009 году).

3.3. Нахождение близких по смыслу выражений

На третьем этапе производится поиск вариантов называния сущностей в текстах кластера. В этой процедуре для выдвижения и проверки предположения о семантической связи между выражениями U_1 и U_2 используются такие факторы, как

- внешнее сходство U_1 и U_2 (например, слова с одинаковым началом);
- фактор употребления U_1 и U_2 в соседних предложениях по сравнению с их употреблением внутри фрагментов одного предложения, здесь используются соотношения, выявленные в эксперименте, описанном в разделе 2;
- сходство U_1 и U_2 по векторам Near, Av, NotN, Ns, что определяется вычислением скалярного произведения соответствующих векторов (NearScalProd, AverbScalProd, NotNearScalProd, NsentScalProd). Пороги сходства основываются на предварительно вычисленных скалярных произведениях для тезаурусных квазисинонимов, подобно процедуре описанной в разделе 2.

Отметим, что если процедура проверки сходства контекстов слов на основе вычисления скалярного произведения между их контекстами является стандартной процедурой выявления близких по смыслу слов [16], то использование в этом качестве такого фактора, как расположение слов в соседних предложениях, в литературе не описано.

Процедура поиска сходных по смыслу выражений состоит из нескольких шагов. На каждом шаге применятся свой набор критериев. Просмотр осуществляется по мере снижения частотности выражений: для каждого выражения U_1 рассматриваются все выражения U_2, частотность которых ниже данного. Если все условия очередного этапа оказываются выполненными, то менее частотное выражение U_2 вносится в ряд синонимов выражения U_1, все контексты U_2 перемещаются в контексты U_1, выражения как бы склеиваются.

Таким образом, формируется ряд условных синонимов, т.е. языковых выражений, которые предполагаются эквивалентными относительно содержания кластера. Помещенные в такой синонимический ряд выражения U_1 и U_2 связаны между собой по смыслу, или референты U_1 и U_2 тесно связаны между собой в рамках тематики кластера так, что U_2 не представляет отдельной тематической значимости по сравнению с U_1. Например, в общем контексте такие слова, как *парламент* и *парламентарий* имеют между собой тесную смысловую связь, но при этом не являются синонимами. Но в рамках конкретного кластера, например, в котором обсуждается процесс принятия решения в парламенте, эти слова могут быть отнесены к классу условной синонимии.

Для обнаружения синонимичных вариантов названия сущностей осуществляются следующие автоматические шаги.

На первом шаге (4.1) ищется смысловое сходство между выражениями, состоящими из похожих слов, например, *Киргизия – Киргизский; Парламент Киргизии – Киргизский парламент.*

Для соединения слов с одинаковым началом в синонимический ряд требуется выполнение следующих условий: встречаемость в соседних предложениях значительно выше встречаемости в одном предложении (2, 3), оба выражения должны быть достаточно частотны в кластере. Процедура является итерационной:

- $Ns > 2 * (Av + Near + NotN)$; (2)
- $Ns > 1$. (3)

В тех случаях, когда выражения редко располагались в соседних предложениях ($Ns<2$), дополнительно требуется сходство по скалярным произведениям контекстов:

- NearScalProd + NotNearScalProd + AVerbScalProd + NSentScalProd > 0.4 (4)

На втором шаге (4.2) ищется смысловое сходство между выражениями, одно из которых включается в другое, например, *парламент - парламент Киргизии, авиабаза – авиабаза Манас*. Суть этого шага заключается в том, что в кластере могло и не упоминаться других парламентов, кроме парламента Киргизии, т.е. в обоих случаях идет ссылка на один и тот же объект. Здесь используется условие на сходство контекстов встречаемости выражения (5):

- $NearScalProd > 0.1$. (5)

На третьем шаге (4.3) ищется смысловое сходство между выражениями одной длины, в состав которых входит хотя бы одно одинако-

вое слово, например, *база Манас – авиабаза Манас, американский военный – американская сторона.* Дополнительным условием здесь является высокая встречаемость в соседних предложениях (6, 7):

- $NS > 2 * (Av + Near + NotN);$ (6)
- $NS > 1.$ (7)

Наконец, на последнем шаге (4.4) производится поиск смыслового сходства между различными языковыми выражениями, упомянутыми в тексте, например, *США – американский, Киргизия – Бишкек.*

Предположение о смысловом сходстве непохожих друг на друга выражений требует максимального числа проверок: превышения частотности встречаемости выражений по сравнению с встречаемости в тех же предложениях (8, 9), высокие ограничения по частотности, сходство по контекстам употребления:

- $NS > 2 * (Av + Near + NotN);$ (10)
- $NS > 0.1 * MaxAv.$ (11)

В результате проведенных этапов для кластера примера автоматически собрались следующие синонимические группы (жирным шрифтом выделен полученный автоматически заглавный синоним группы):

База Манас*: база, авиабаза Манас, авиабаза, Манас;*

США: *американский, американец, Америка;*

Киргизия: *Киргизстан, киргизский, киргизско-американский, Бишкек;*

Парламент Киргизии: *Киргизский парламент, парламент, парламентский, парламентарий;*

Международный аэропорт Манас: *аэропорт Манас, аэропорт;*

Законопроект*: закон, законодательство, законодательный, законный* и др.

4. Тестирование метода

Для тестирования предложенного метода мы взяли 10 новостных кластеров различной тематики величиной более 20-30 документов.

При тестировании качества сборки многословных выражений проверялись два показателя. Во-первых, мы выясняли процент синтаксически правильных групп среди выделенных выражений. Во-вторых, мы привлекли профессионального лингвиста и попросили для каждого кластера выделить наиболее существенные для понимания смысла документов кластера многословные выражения (5-10), упорядоченные в порядке снижения их значимости.

Так для кластера примера лингвистом были сочтены важными следующие 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 [17], когда экспертов просят обозначить наиболее тематически значимые слова и выражения текста. В нашем же случае мы тестировали именно выделение словосочетаний. Кроме того, в списке, создаваемом лингвистом, могли быть смысловые повторы (*парламент Киргизии – Киргизский парламент*).

В результате тестирования было получено, что автоматически было выделено 364 многословных выражений, из них 312 являются правильными синтаксическими группами, что составляет 87,9%. С учетом частотности правильные синтаксические выражения составляют 91,4%. Лингвист выделил для кластеров 70 наиболее важных многословных выражений, 72,6 % из них было автоматически собрано системой.

При тестировании автоматического извлечения синонимичных выражений кластера просматривались все вхождения выражений, вошедших в синонимичные ряды, и тестировалось, имеет ли каждое вхождение отношение к заголовочному выражению синонимического ряда. Если для большей части таких вхождений было установлено существование отношения, то склейка для выражения в целом считалась правильной. Табл. 2 содержит информацию о качестве произведенных синонимических склеек по 10 кластерам, исчисляемых как в количестве отдельных выражений, так и в частотном выражении.

Для оценки вклада информации о встречаемости выражений в соседних предложениях мы провели подробное тестирование склейки выражений с одинаковым началом (шаг 4.1) для кластера примера (Табл. 3). Из таблицы видно, что добавление условий на показатель Ns, как это сделано в условиях шага 4.1, улучшает точность и полноту принятых решений о синони-

Табл. 2. Результаты тестирования автоматического соединения слов и выражений новостного кластера в синонимические ряды

Этап	Количество склеек по отдельным выражениям	Количество склеек по частотности	Процент правильных склеек для выражений	Процент правильных склеек по частотности
4.1. Склейка выражений с одинаковым началом	155	4383	87,9%	91,4%
4.2. Склейка вложенных выражений	99	9131	91,4%	92,9%
4.3. Склейка выражений, с пересекающимися компонентами	8	677	85,7%	80,8%
4.4. Склейка остальных видов выражений	38	4822	62,5%	62,4%

Табл. 3. Результаты тестирование синонимических склеек выражений с одинаковыми началами

Метод	Число склеенных выражений	Общая частота склеек	Частота правильных склеек	Точность по частоте, %	Полнота по частоте, %
Совпадение начал слов (BasicLine)	383	2266	1472	65	100
Совпадение начал слов + скалярные произведения (порог 0.1)	38	996	834	83,7	56,7
Совпадение начал слов + скалярные произведения (порог 0.4)	36	976	814	83,4	55,3
Условия шага 4.1	36	965	873	90,5	59,3

мических склейках, по сравнению с более традиционными условиями по близости линейных контекстов.

Заключение

В статье был представлен эксперимент, в котором для новостного кластера извлекаются многословные выражения, и производится формирование синонимических рядов выражений, близких по смыслу, по употреблению в данном кластере. Часто такие выражения представляют собой альтернативные наименования одной и той же сущности. Для нахождения таких выражений мы, помимо известного метода сопоставления контекстов употребления выражений, используем еще и информацию о встречаемости выражений в соседних предложениях. Было проведено тестирование реализованного метода и показан вклад предложенного фактора.

В дальнейшем предполагается изучение влияния выделенных внутри кластера рядов условной синонимии на качество выполнения различных операций с кластерами (например, уточнение границ кластера, автоматическое аннотирование кластера, выявление новизны в предложениях кластера и др.). Также мы предполагаем изучение способов комбинирования автоматически выделенных синонимов с информацией, описанной в тезаурусе.

Литература

1. Лукашевич Н.В., Добров Б.В., Штернов С.В. Обработка потоков новостей на основе больших лингвистических ресурсов // Интернет-математика-2005. (http://download.yandex.ru/company/grant/2005/10_Loukachevitch_103030.pdf)
2. Добров Б.В., Павлов А.М. Исследование качества базовых методов кластеризации новостного потока в суточном временном окне // Труды конференции RCDL-2010. 2010.
3. Li J., Sun L., Kit C., Webster J. A Query-Focused Multi-Document Summarizer Based on Lexical Chains // Proc. of the Document Understanding Conference DUC-2007. 2007.
4. Лукашевич Н.В., Добров Б.В. Автоматическое аннотирование новостного кластера на основе тематического представления // Компьютерная лингвистика и интеллектуальные технологии по материалам ежегодной Международной конференции «Диалог 2009». Вып. 8 (15), 2009. С. 299-305.
5. Ермаков А.Е. Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2007. – М.: Наука. 2007.
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, LREC 2004 (2004)
7. Арутюнова Н.Д. Предложение и его смысл. // М. 1976.
8. Halliday M., Hasan R. 1976. Cohesion in English. - Longman, London.
9. Barzilay R., Lee L. Learning to Paraphrase: an Unsupervised Approach Using Multiple Sequance Alignment // In Proceedings of HLT/NAACL. 2003.
10. Dolan B., Quirk Ch., Brockett Ch. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources // In Proceedings of COLING-2004 2004.
11. Dijk van T. Semantic Discourse Analysis // Handbook of Discourse Analysis / Teun A. van Dijk, (Ed.), vol. 2. London: Academic Press. 1985. pp. 103-136.
12. 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.
13. Hasan R. Coherence and Cohesive harmony // Understanding reading comprehension / J. Flood, editor, Newark, DE: IRA. 1984. pp. 181-219.
14. Loukachevitch N. Multigraph representation for lexical chaining // Proc. of SENSE workshop, 2009. pp. 67-76.
15. 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.
16. Yang H., Callan J. A metric-based framefork for automatic taxonomy induction // Proc. of ACL-2009. 2009.Singapore.
17. Su Nam Kim, Medelyan O., Min-Yen Kan, Baldwin T. SemEval-2010 Task-5. Automatic Keyphrase Extraction from Scientific Articles // Proc. of the 5-th International Workshop on Semantic Evaluation ACL -2010, 2010. pp. 21-26.
18. Добров Б.В., Лукашевич Н.В. Тезаурус РуТез как ресурс для решения задач информационного поиска. Труды Всероссийской конференции ЗНАНИЯ-ОНТОЛОГИИ-ТЕОРИИ. Том.1, С. 250-259.

Алексеев Алексей Александрович. Аспирант факультета вычислительной математики и кибернетики МГУ им. М.В. Ломоносова. Окончил МГУ имени М.В. Ломоносова в 2010 году. Автор пяти печатных работ. Область научных интересов: искусственный интеллект, компьютерная лингвистика, интеллектуальный анализ данных, извлечение информации. E-mail: a.a.alekseevv@gmail.com.

Лукашевич Наталья Валентиновна. Ведущий научный сотрудник НИВЦ МГУ имени М.В. Ломоносова. Окончила МГУ имени М.В. Ломоносова в 1986 году. Кандидат физико-математических наук. Автор 130 печатных работ и монографий. Область научных интересов: искусственный интеллект, компьютерная лингвистика, интеллектуальный анализ данных, извлечение информации. E-mail: louk_nat@mail.ru.