Статистический анализ связности текстов по общественно-политической тематике

© В.Е. Абрамов[1], Н.Н. Абрамова[2], Е.В. Некрасова[2], Г.Н. Росс[2]

[1]ЗАО СКБ «ТЭЛКА»,
[2] ФГУП «НИЦИ при МИД России»
abramval@yandex.ru, NAbramova@mid.ru, ENekrasova@mid.ru, GRoss@mid.ru

Аннотация

В работе описаны результаты статистического анализа средств выражения межфразовых связей в русскоязычных текстах по общественно-политической тематике. Исследование проводилось с целью усовершенствования технологии реферирования текстов, что позволит разработать методы реферирования с учетом законов связности текстов.

1. Введение

В результате решения задач автоматической смысловой обработки текстовой информации, таких как реферирование или машинный перевод текстов с одного естественного языка на другой, должен быть получен связный текст, представляющий собой последовательность семантически связанных друг с другом предложений. Поэтому необходимо анализировать не только синтаксические связи внутри предложений, но и связи между предложениями - межфразовые связи. Эти связи могут выражаться различными языковыми способами. Согласно изложенной в работе [7] теории, текст подчиняется некоторым законам устройства связного текста, таким как законы повторяемости, сокращения и избыточности смысла.

Повторение элементов смысла – это необходимое условие существования теста. При этом повторение смысла в связном тексте возможно лишь при условии действия принципа сокращения. Закон сокращения требует передавать определенный смысл минимумом лексических средств. Напротив, закон избыточности позволяет увеличивать средства выражения смысла, что способствует повышению надежности восприятия текста.

Опираясь на законы связности текста, межфразовые связи можно определить через понятие замещения, т.е. повторения смысла какого-либо отрезка текста с помощью особых языковых средств. Замещение одних элементов теста другими часто называют анафорой и говорят, что между ними существует анафорическая связь. Заместители или показатели связи – это слова и словосочетания, обозначающие понятия, повторяющиеся в тексте, и указывающие на связь одних предложений с другими. Замещаемое (или антецедент) – это обозначение того же понятия в предшествующем предложении.

Знание законов распределения показателей связи дает возможность разработать методы реферирования, позволяющие автоматически создавать связные тексты рефератов.

Как совершенно справедливо отмечается в работе Ахреновой Н.А. [3], изучавшей современный политический текст на английском языке, «решение задачи автоматического нахождения анафорических связей есть часть важной и не решенной до сих пор проблемы автоматического синтаксического анализа естественных языков».

Однако в этой области достигнуты определенные успехи. В обзоре [9], посвященном разрешению анафоры, известный специалист в области компьютерной лингвистики Руслан Митков останавливается на традиционных и новых подходах к этой проблеме, описывая методы Картера, Е. Рича, Д. Карбонеллы, Р. Брауна, С. Рико Перез, Ш. Лаппина и Г. Лисса, и дает краткую характеристику известных компьютерных систем.

Для современных подходов характерно создание интегрируемой модели разрешения анафоры, использующей комбинацию традиционных лингвистических методов со статистическими методами.

Модель интегрирует модули, содержащие различные типы знаний - синтаксические, семантические, предметной области и дискурса.

Синтаксический модуль используется для проверки согласованности анафоры и антецедента в числе, роде и лице.

Семантический модуль отсеивает возможных кандидатов в антецеденты, отдавая предпочтение

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

кандидатам с такой же семантической ролью, как и у анафоры.

Модуль предметной области - это базы знаний в рассматриваемой области. Модуль дискурса определяет центральный сегмент текущего дискурса с помощью статистического метода Байеса.

Окончательный выбор антецедента происходит в модулях предметной области и дискурса.

В других работах Р. Миткова есть описание программы распознавания анафор для ряда языков: английского, французского, болгарского, польского и арабского.

Среди российских исследователей, занимающихся данной проблемой, следует отметить Ермакова А.Е. и Толпегина П. В.

Метод выявления анафоры, описанный в работе [5], предназначен для решения задачи «извлечения фактографической информации из текстовых документов особого стиля (биографий, протоколов, сводок и т.д.), назначение которых состоит в лаконичной передаче совокупности фактов о некоторых объектах».

В работе [8] описывается алгоритм автоматизированного разрешения анафоры местоимений третьего лица на основе методов машинного обучения. Точность разрешения анафоры составляет более 70%. Автор обещает расширить область анализа «в пользу личных, возвратных, притяжательных и, в особенности, указательных местоимений».

С точки зрения выявления анафорических связей научно-технический текст изучался сотрудниками ВИНИТИ РАН под руководством проф. Г.Г. Белоногова. Однако следует отметить, что описанный в работе [1] эксперимент проводился на коротких текстах (средняя длина реферата составляла пять предложений), предварительно прошедших аналитическую обработку для публикации в реферативном журнале.

2. Выявление межфразовых связей в общественно-политических текстах

Для изучения связей между предложениями были выбраны следующие группы текстов по общественно-политической тематике:

1) сообщения информационных агентств;
2) газетные публикации;
3) брифинги официальных представителей России.

Были взяты по 35 сообщений и газетных статей за один из дней (28 марта 2011 г.) и 35 брифингов (за 28 марта – 2 апреля 2011 г.). Специального отбора текстов не проводилось, но исключались тексты, состоящие менее, чем из пяти предложений. Эту выборку можно считать репрезентативной, так как и по объему, и по характеру поступающей ежедневно информации она не отличается от всех остальных дней года.

Всего было обработано 105 текстов общим объемом 319 Кб. При этом минимальное количество предложений в текстах равнялось пяти, а максимальное - 258. В табл. 1 приводится суммарная длина текстов и средняя длина одного текста в байтах по каждой группе.

Таблица 1. Распределение различных видов общественно-политических текстов по длине

Вид текста	**Суммарная длина текстов** (в байтах)	**Средняя длина текста** (в байтах)
Сообщения информационных агентств	59913	1712
Газетные статьи	179580	5131
Брифинги	79219	2264

В процессе анализа изучались типы замещения, использующиеся в текстах. Мы выявили буквальный повтор, когда элемент текста совпадает с другим отрезком текста с точностью до словоформ, и морфо-синтаксический повтор, при котором совпадение происходит с точностью до словообразования (например, «консульство» - «консульский») или на уровне опорных слов, определители которых могут трансформироваться (например, «журналисткое расследование» – «расследование журналиста»).

Смысловая связь между предложениями выражается также с помощью синонимии, вызванной изменением состава слов и словосочетаний антецедента и заместителя («безвизовый режим» - «полная отмена виз»), или вызванной аббревиацией или сокращением слов («Содружество независимых государств» – «СНГ», «генеральное консульство» – «генконсульство»).

В текстах встречаются показатели связи, выраженные гипонимами и гиперонимами. Гипонимия – это сужение значений элементов смысла («информационная безопасность» – «международная информационная безопасность»), а гиперонимия - расширение («заместитель министра» – «руководство министерства»).

Межфразовая связь выражается иногда с помощью эллипсиса – повторения смысла с некоторыми опущенными элементами, не сводимого к отношению род-вид («Комиссия по правам человека» - «Комиссия»).

Часто используется способ замещения слова или словосочетания из предшествующего предложения местоимением (личным - «он», «она», «они»; притяжательным - «его», «ее», «их», «свой»; возвратным – «себя»; указательным – «это», «то»; относительным – «который», «где», «что») или местоименным наречием («куда», «там», «туда»).

Такой вид межфразовой связи называют местоименной анафорой.

Кроме перечисленных способов замещения, в текстах встречается выражение связи с помощью вводных слов, наречий и союзов («таким образом», «в связи с изложенным», «выше», «далее», «поэтому»).

Пример выявления показателей межфразовой связи приведен на рис. 1. Все предложения текста перенумерованы в порядке их следования в тексте и дается их полный текст, в котором показатели связи заключены в косые скобки. Между предложениями обозначен тип связи.

Занимаясь алгоритмизацией автоматического определения межфразовых связей, важно учитывать распределение частот встречаемости различных типов связей. Так, по результатам анализа научно-технических текстов рефератов в ВИНИТИ РАН [4], среди самых распространенных способов выражения связей отмечены лексический повтор, синонимия и эллипсис, составляющие 78% от всех встречающихся связей.

Мы анализировали неформализованные тексты. Была выявлена зависимость распределения частот от вида текста. В табл. 2 приводится распределение частот встречаемости различных типов связей для исследуемых групп текстов, а на рис.2 это распределение представлено в виде диаграммы.

Оказалось, что для всех видов текстов по общественно-политической тематике наиболее распространенным типом межфразовой связи является лексический повтор, чаще всего использующийся в газетных публикациях (около 43%).

Распределения остальных типов связи близки для текстов сообщений информационных агентств и брифингов: почти одинаково популярны способы выражения межфразовых связей с помощью синонимии, гипонимии и местоименной анафоры.

В текстах газет на втором месте по использованию находится местоименная анафора, затем эллипсис и синонимия, а гипонимия является самой редко встречающейся связью.

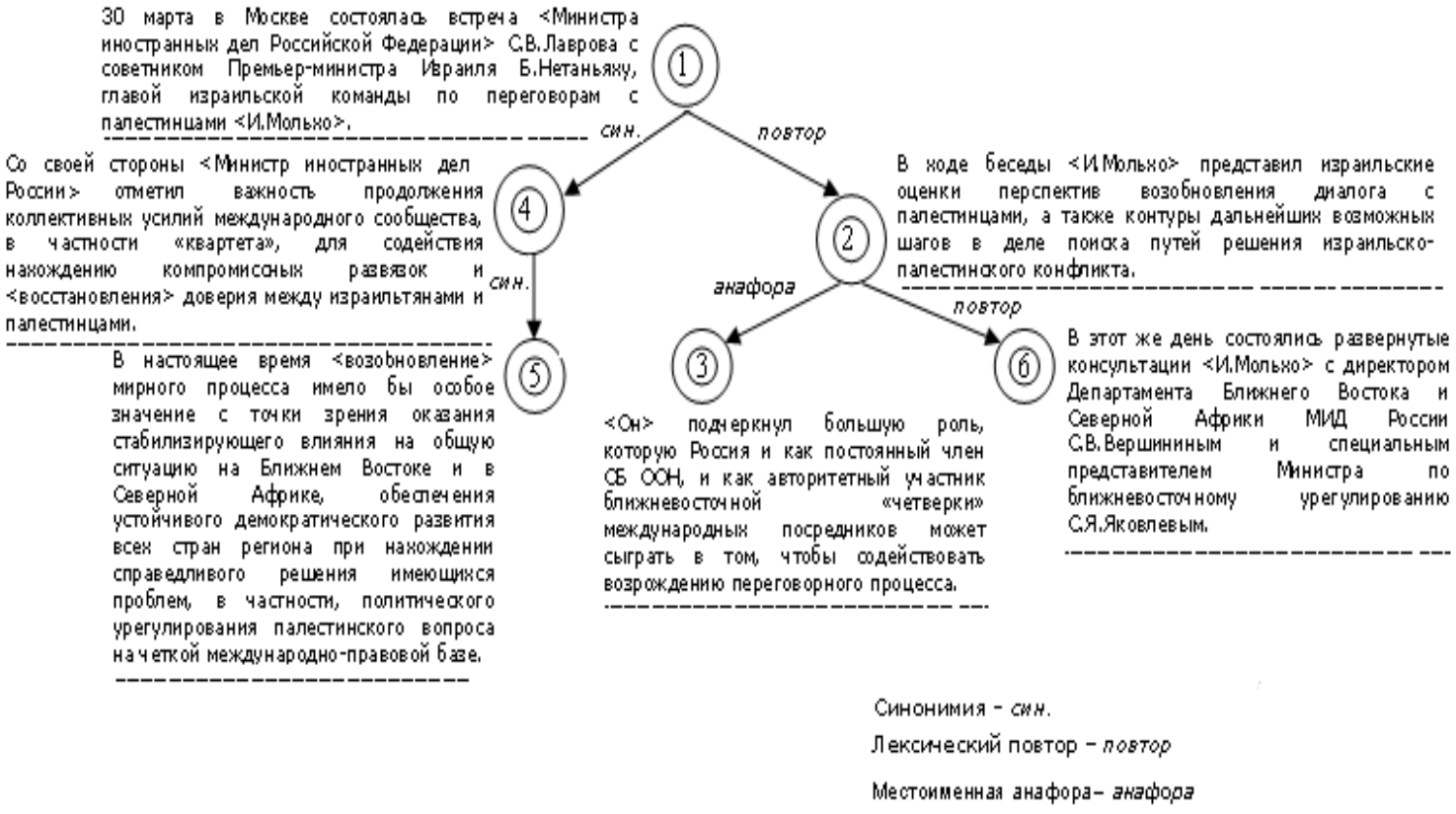

Рис.1. Установление межфразовых связей

Таблица 2. Распределение частот встречаемости различных типов межфразовой связи в общественно-политических текстах

Тип связи / Вид текста	Лексический повтор	Синонимия	Гипонимия, гиперонимия	Эллипсис	Местоим. анафора	Другие связи	Кол-во связей на док-т
Сообщения информагентств	0,28	0,2	0,16	0,09	0,15	0,12	9,8
Газеты	0,43	0,13	0,04	0,15	0,17	0,08	27,5
Брифинги	0,36	0,2	0,15	0,07	0,16	0,06	12,3

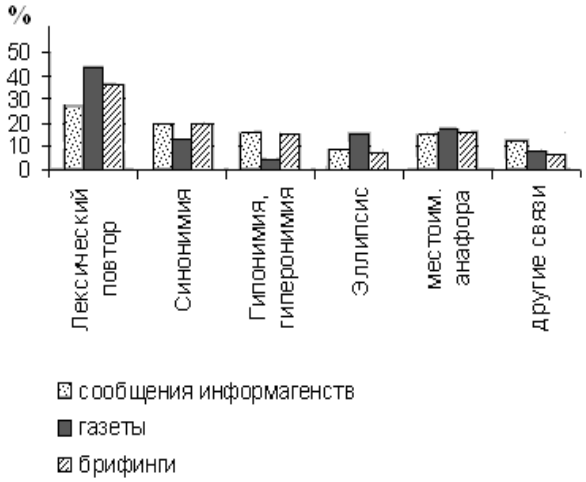

Рис.2. Частота встречаемости различных типов межфразовой связи

3. Автоматическое разрешение анафоры

В рамках создания системы автоматического реферирования [2] проводился эксперимент по распознаванию местоименных анафорических связей.

Перед нами стояла сугубо прагматическая цель: достичь того, чтобы в текстах рефератов не появлялись местоимения, смысл которых не был бы ясен из контекста.

Алгоритм распознавания межфразовых анафорических связей разработан нами с учетом специфики общественно-политических текстов. Принципы работы алгоритма устанавливались эмпирически на основе анализа текстов со следующими ограничениями:

1) между анафорой и антецедентом имеется явная кореферентность;
2) случаи катафоры не анализируются;
3) предполагается, что в тексте нет референциальных конфликтов, т.е. для каждого референта существует только один антецедент.

Мы исходили из того, что предсказуемость антецедента зависит от “референциального расстояния”, т.е. расстояния от него до ближайшего местоимения, поэтому кореферентным может быть признан только ближайший по тексту антецедент.

Местоименные анафоры определялись с помощью алгоритма, в котором использовался частичный анализ синтаксической структуры предложения. Анализ проводился на основе морфологической информации о словах, являющейся входной для работы алгоритма.

Синтаксический анализ начинался с выявления слов, грамматически не связанных с членами предложения, т.е. вводных слов, вставных предложений и оборотов. Для этих целей использовались два словаря: словарь вводных слов и словарь оборотов. Отличие между словарем оборотов и словарем вводных слов состоит в том, что вводные слова целиком задаются списком, а в словаре оборотов имеется только начальная часть фразы (например, "Как сообщается…", "Как стало известно…"), с помощью которой происходит распознавание оборота в тексте. Границей оборота является запятая.

Вводные слова и обороты могут быть средством межфразовой связи, если включают в себя местоимения или уточняющие прилагательные (например, «по его словам», «в связи с вышеизложенным»). Поэтому для выявления этой связи в выделенных из текста вводных словах и оборотах проводился поиск местоимений и уточняющих прилагательных, задаваемых списком.

Вводные слова и обороты, не содержащие местоимений и уточняющих прилагательных, не учитывались при дальнейшем анализе.

Следующий шаг синтаксического анализа был связан с выяснением вида предложения: простое, сложное или предложение с прямой речью.

Предложение считается простым при наличии в нем только одного глагола (слова, относящегося к одному из грамматических классов: «глагол в личной форме», «глагол прошедшего времени», «инфинитив», «модальный глагол») или глагола и группы рядом стоящих глаголов, не разделенных

знаками препинания и союзами, а также при отсутствии глаголов.

Если слева и справа от знаков препинания и союзов имеются слова, относящиеся к классу глаголов, то предложение считается сложным. Внутри сложных предложений выделялись части, являющиеся простыми предложениями. Членение на простые предложения проводилось по знакам препинания и союзам, непосредственно предшествующим глаголу так, чтобы слева от этого знака был бы хотя бы один глагол.

Предложения с прямой речью определялись на основе формальных признаков (знаков препинания и использования больших или маленьких букв в первом слове автора) в зависимости от расположения прямой речи относительно авторских слов. Рассматривались случаи, когда прямая речь стоит после и перед словами автора, или прерывается словами автора, или стоит внутри авторских слов.

При выявлении анафор в предложениях с прямой речью анализу подлежали только слова автора. Если в них было обнаружено личное местоимение 3-его лица, то устанавливалась связь входящего в прямую речь блока предложений с предшествующим прямой речи предложением. Если местоимение было найдено в предложении, которое следовало за прямой речью, то также устанавливалась связь этого предложения с предшествующим прямой речи предложением.

В простом предложении поиск анафор проводился среди всех слов с признаками местоименности до границы предложения. В сложных бессоюзных предложениях или в предложениях с сочинительными союзами они искались только в первой части предложения, а в сложноподчиненных предложениях – в главном и придаточном предложениях. Причем, в предложении с несколькими придаточными анализировалось только предложение, ближайшее к главному.

При выявлении связей, основанных на местоименной анафоре, нужно учитывать, что одно и то же местоимение в одних случаях может быть выражением межфразовой связи, а в других нет. Так, слово «это», перед которым стоит тире (например, «БРИКС – это площадка диалога») не является местоименной анафорой.

Если местоимение входит в именную группу, состоящую из однородных членов предложения, связанных союзами «и»/»или» (например, «министры и их постоянные представители», «ООН и ее работники»), то это также указывает на отсутствие межфразовой связи.

В алгоритме не заложено распознавание местоименных анафор, использующихся для характеристики времени, потому что, как правило, их антецедент не находится в предыдущем предложении, и нужно проводить поиск по дополнительным признакам.

Такие анафоры выражаются с помощью указательных местоимений «этот» или «тот» (иногда с частицей «же»), согласованных в роде, числе и падеже с рядом стоящим существительным из следующего списка: «год», «век», «столетие», «месяц», «неделя», «день», «время», «час», «минута», «секунда», «период» (например, «этой неделе», «тот же год», «этот месяц» и т.д.).

Найденная временная анафора пропускалась, и выбиралось следующее за ней слово с признаком местоименности, которое сравнивалось с заданным списком личных, относительных, притяжательных и указательных местоимений. В этот список не включались возвратные и взаимные местоимения, так как согласно принципу связности текста [8] связь указанных видов местоимений с их антецедентом не может выходить за границу предложения, т.е. она не является межфразовой связью.

Поскольку часто межфразовая связь выражается с помощью уточняющих прилагательных и причастий, согласованных с рядом стоящими существительными в роде, числе и падеже (например, «данный случай», «последнее рассуждение», «вышеуказанный пример», «указанный способ» и т.д.), они отыскивались в тексте, и проводилась проверка на согласование. Уточняющие слова задавались списком.

При выявлении анафор в сложноподчиненных предложениях сначала выполнялся поиск личного местоимения 3-его лица («он», «оно», «она», «они») в придаточном предложении.

Если оно было найдено, то в главном предложении искалось подлежащее (существительное в именительном или винительном падеже).

Если подлежащее отсутствовало или было выражено дейктическим местоимением («я», «мы», «ты», «вы»), или не было согласовано в роде и числе с местоимением, то устанавливалась связь этого предложения с предыдущим предложением.

В приведенном ниже примере местоимение «они» явно указывает на межфразовую связь:

Думаю, что ***они*** *должны активно подключиться к выработке оптимального курса, по которому пойдет БРИКС.*

Если в главном предложении было подлежащее, согласованное в роде и числе с местоимением из придаточного предложения, то связь местоимения с его антецедентом не считалась межфразовой связью.

Это положение иллюстрируется следующим примером:

Президент сообщил, что он издал Указ о …

При отсутствии в придаточном предложении личных местоимений 3-его лица искалось указательное местоимение («этот», «тот», «такой»). Наличие местоимения указывает на существование межфразовой связи, что показано в приведенных ниже примерах:

*Думаю, что **это** окажется полезным для политиков и дипломатов, причастных к принятию коллективных решений в рамках БРИК.*

*Маркин сообщил, что **эти** решения Генпрокуратуры будут обжалованы….*

Входной информацией для работы алгоритма выявления местоименных анафор (см. табл. 3) является исходный текст, каждая словоформа которого сопровождается:

- сведениями о ее месторасположении;
- признаком буквы, с которой она начинается;
- сведениями о длине окончания и о принадлежности к части речи;
- набором грамматической информации (род, число, падеж и т.д.).

В первой и второй позиции записи через разделитель "#" указывается номер предложения, в котором находится словоформа, и ее порядковый номер в этом предложении; в третьей позиции – признак буквы (1 – большая буква, 0 – маленькая буква); затем сама словоформа и ее признаки, разделяемые косой линией (/):

- количество букв в окончании;
- номер флективного класса;
- дополнительные признаки грамматического класса (сюда входит и признак местоименности);
- набор грамматической информации в закодированном виде.

Каждый элемент набора состоит из одной или двух цифр. Например, для класса существительных первая цифра означает число (1 – единственное, 2 – множественное), вторая – падеж слова (1 – именительный, 2 – родительный и т.д.).

Таблица 3. Фрагмент входных данных для работы алгоритма распознавания анафор

5#1#1#ранение 01/073/01/1114
5#2#0#получил 00/125/10/1
5#3#0#российский 02/106/01/1114
5#4#0#гражданин 00/037/01/11
5#5#1#п 00/145/01
5#6#0#. 00/2000/01
5#7#1#о 00/164/46
5#8#0#. 00/2000/01
5#9#1#ершов 00/042/01/11
5#10#0#, 00/2000/01
5#11#1#работающий 02/105/10/1114
5#12#0#в 00/164/046
5#13#0#миссии 01/061/01/1213162124
5#14#1#оон 00/146/01
5#15#0#в 00/164/046
5#16#1#афганистане 01/001/01/16
5#17#0#. 00/2000/01
6#1#1#сейчас 00/152/01
6#2#0#он 00/145/02
6#3#0#находится 04/124/10/3
6#4#0#в 00/164/046
6#5#0#больнице 01/067/01/1316
6#6#0#. 00/2000/01
7#7#1#российские 02/106/01/4144
7#8#0#дипломаты 01/021/01/21
7#9#0#выехали 01/125/10/4
7#10#0#на 00/164/046
7#11#0#место 01/070/01/1114
7#12#0#происшествия 01/073/01/122124
7#13#0#для 00/155/2
7#14#0#оказания 01/073/10/122124
7#15#0#помощи 01/054/01/1213162124
7#16#1#пострадавшему 05/105/10/13
7#17#0#. 00/2000/01

В результате работы алгоритма формируется список пар из номеров связанных предложений: в первом предложении пары содержится антецедент, а во втором – анафора. При отсутствии анафорической связи первым элементом будет номер предложения, а вместо второго элемента пары ставится ноль. Для фрагмента текста, приведенного в табл.3, на выходе были получены следующие данные: (5,6), (7,0).

Ниже приводятся примеры работы алгоритма.

Правильно распознанные анафоры:

*Ранение получил российский гражданин П.О. Ершов, работающий в миссии ООН в Афганистане. Сейчас **он** находится в больнице.*

*30 марта в рамках продолжающейся эвакуации российских граждан, изъявивших желание покинуть Ливию, автоколонной Посольства России из Триполи на тунисскую территорию вывезены 107 человек. **Эту** гуманитарную операцию обеспечивают Посольства России в Триполи и Тунисе.*

Неправильно распознанные анафоры:

*Полагаю, что БРИК не может быть альтернативой традиционным приоритетам внешней политики каждой из четырех стран. Однако взаимопонимание, которое может быть достигнуто в формате БРИК, способно позитивно влиять на определение курса **их** дальнейшего развития.*

Работа алгоритма оценивалась на случайной выборке из 50 текстов. Было получено около 70% правильно разрешенных анафор.

4. Заключение

Проводя выявление межфразовых связей текстов по общественно-политической тематике на сравнительно небольшом по объему материале, авторы попытались найти зависимость частоты встречаемости различных типов связей от вида и длины текста. Для двух видов – «Сообщения информационных агентств» и «Брифинги официальных представителей России», были получены близкие распределения и по средней

длине текста документа, и по частоте встречаемости типов связи. Тексты из газет оказались в среднем в 2,5 раза длиннее, чем тексты «Брифингов», и имеют отличия в способах выражения связей.

Проведенный нами анализ полезен при разработке алгоритмов автоматического выявления межфразовых связей.

Такой приближенный алгоритм был разработан нами для системы автоматического реферирования русских текстов. Он основан на частичном синтаксическом анализе предложения.

Направление дальнейших работ мы видим в разработке программных средств для распознавания межфразовых связей, создании лингвистических словарей (синонимов, гипонимов и гиперонимов) и баз данных названий объектов и персон.

Литература

[1] Абрамова Н.Н., Матвеева Е.Г., Новоселова Л.Н., Панова Н.С., Рыжова Е.Ю. Синтаксическая структура текстов рефератов. - Материалы Х1У Всесоюзного научного семинара "Системные исследования ГАСНТИ", г. Кишинев, 21-23 июня 1983 г.

[2] Абрамов В.Е. Автоматическое рубрицирование и реферирование текстов (в том числе на иностранных языках): Автореферат дис. ….. канд. технич. наук: М., 2008, 27 с.

[3] Ахренова Н.А. Нахождение анафорических связей при автоматическом анализе текста (на материале английского языка) : Дис. ... канд. филол. наук : М., 2003, 219 с.

[4] Белоногов Г.Г., Ю.П.Калинин, Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии. – М.: Русский мир, 2004, 248 с.

[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] Кибрик А.А. Современная американская лингвистика: Фундаментальные направления. Изд. 2-е, испр. и доп. - М.: Едиториал УРСС, 2002. - 480 с.

[7] Откупщикова М.И. Синтаксис связного текста. – Л., 1982.

[8] Толпегин П. В. Новые методы и алгоритмы автоматического разрешения референции местоимений третьего лица русскоязычных текстов. – М.: Комкнига, 2006. – 88с.

[9] Mitkov R. Anaphora resolution: the state of the art. Working paper (based on the COLING'98/ACL'98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton. 1999. clg.wlv.ac.uk›papers/mitkov-99a.pdf

Statistical Analysis of the Coherence of Texts on Social and Political Issues

© V. Abramov, N. Abramova, E. Nekrasova, G. Ross

This paper describes the results of statistical analysis of the expression connection between the phrases in Russian texts on the social and political issues. The study was made to improve technology of text summarization, which allows to develop methods of summarization with the help of coherence texts laws.