Информационные процессы, Том 6, № 2, 2006, стр. 218–228.

Смысловой анализ текстов

С.В. Рубаков

Московский государственный университет им. М.В. Ломоносова, Москва, Россия

Поступила в редколлегию 18.05.2006

Аннотация— На данный момент существует множество подходов к анализу текстов. В основу любого текстового анализатора обычно бывает заложен некий инструмент, позволяющий делать всевозможные лингвистические анализы – такие как семантический анализ, морфологический анализ, орфографический анализ и другие. Основной проблемой такого рода анализаторов является то, что ни один из лингвистических анализов не учитывает реального смысла текста, а лишь по косвенным признакам может выполнять какие-либо действия. В данной работе сделана попытка уйти от существующей концепции – что все языковые проблемы могут быть решены исключительно с помощью лингвистики – и взглянуть на проблему анализа текстов с несколько иной стороны. Предлагается способ анализа текста с использованием смысловой компоненты. Рассматривается подход с использованием общеупотребительной лексики, причем этот подход без ограничения общности легко переносится на любую специализированную область знания. Приводятся результаты экспериментов, поясняющих этот метод и следствия из него.

1. ВВЕДЕНИЕ

С момента появления глобальной сети Интернет существует проблема поиска информации. Существующие поисковые системы не решают проблему смыслового поиска, они лишь пытаются моделировать смысловой поиск исходя из косвенных факторов. В данной работе была сделана попытка проанализировать смысловую компоненту поиска и построить очень простой инструмент, который работает с косвенными признаками, которые более осмысленны, нежели те, которые применяются на данный момент. На данном этапе развития технологий проблема работы со смыслом становится все более актуальной, поскольку количество доступной информации превышает возможности человека в анализе этой информации.

Идея, положенная в основу этой работы очень проста. Существует большое количество различных словарей общей лексики, и человек обращается к этим словарям за информацией в случае, когда интересующее его слово ему неизвестно. Мы попытались построить систему, которая обращалась бы к словарям, как это делает человек, только система это делает всегда, поскольку не обладает механизмом памяти, и как следствие пользуется этими словарями, заменяющими ее.

2. ЧАСТЬ 1. МЕТОД.

Предпосылки

Представим мысленно такой эксперимент. Допустим, человеку было сказано некое слово, которое он не знает, также допустим, что это слово было “дератизация”.

Как в таком случае будет действовать человек, чтобы осознать **смысл** данного слова? Необходима некая информация, достаточно полно описывающая данный термин на понятном человеку языке, чтобы тот смог правильно его осознать. В большой медицинской энциклопедии

указано, что дератизация – это истребление грызунов, наносящих экономический ущерб народному хозяйству. В случае такого объяснения сразу же становится понятно, чем идет речь.

Компьютер не обладает никакими априорными знаниями об окружающем мире, а также он не обладает лексиконом. Значит, подходы к смысловому анализу текстов должны быть скрыты в:

– Информации об окружающем мире
– Способах получения новой информации, в дополнение к уже существующей.

В данной работе мы остановимся на первом пункте.

Информация об окружающем мире. Словари.

Любые словари и энциклопедии составлены для того, чтобы уточнять наши знания о мире или пополнять их. Поэтому можно утверждать, что существует прямая зависимость описания в словаре или энциклопедии от уточняемого слова. Это значит, что существует смысловое соответствие между словом и некоторой фразой, описывающей это слово, и смысл фразы гарантированно соответствует смыслу слова. Другими словами, смысловое соответствие между словом и фразой было прописано составителями словаря с максимально возможной точностью.

Значит, мы имеем некоторую структуру, в которой уже априори зашиты смысловые соответствия. Основной проблемой является то, что в словарях указаны соответствия между словом и фразой, и не указаны соответствия всех возможных комбинаций этих слов.

Постановка задачи.

Необходимо построить систему, которая помогала бы человеку искать именно ту информацию, которая человеку нужна. Таким образом есть текстовый запрос к системе, который человек вводит в некоторое поисковое поле системы, и также есть большой набор документов, которые надо отсортировать. Причем на первое место должен попасть именно тот документ, который наиболее интересен человеку, на второе – чуть менее интересный и так далее.

Другими словами, необходимо построить такой алгоритм, который с максимально возможной точностью строил бы соответствия между любой комбинацией любых слов. Другими словами, этот алгоритм должен строить соответствия по смыслу для таких комбинаций:

– Слово-Слово
– Фраза-Фраза
– Слово-Фраза
– Фраза-Слово

Здесь под фразой мы понимаем неограниченный набор слов, который в русском языке также называется “Текст”, “Абзац”, “Параграф” и т.д. Будем называть все такие наборы слов для краткости “Фразами”, без уточнения их длины и ограничения общности рассуждений.

В данной работе мы будем рассматривать только соответствия типа

– Слово-Слово
– Фраза-Фраза

с возможностью дальнейшего распространения этих рассуждений на остальные случаи.

Модель.

Итак, есть многомерное пространство **всех слов**, которыми пользуется человек в рамках одного языка, или лексикон.

Введем дискретный базис, со значениями 0 и 1 , базисными векторами которого будут являться слова в словаре, причем если слово встречается в данной фразе – то координата по этому базисному вектору этой фразы будет 1, в противном случае 0. Тогда фразы в данном пространстве слов будут описываться точками, например для фраз A и B так:

$$A=(0,0,0,1,0,\ldots\ldots,0,1,1,0,0)$$

$$B=(0,0,1,1,0,\ldots\ldots,0,0,1,1,0)$$

Таким образом, каждому слову в этом пространстве соответствует точка с такими координатами в этом многомерном пространстве:

$$(0,0,0\ldots\ldots.0,0,1,0,0\ldots..0)$$

и каждому предложению в этом пространстве тоже соответствует точка, но уже с другими координатами:

$$(0,0,1\ldots\ldots.1,0,1,0,0\ldots..0)$$

Понятно, что в таком многомерном пространстве будет абсолютно недостаточно считать какие-то стандартные метрики, чтобы искать смысловое соответствие.

В математическом плане эта задача интересна тем, что в данном случае мы имеем пространство со слишком малым количеством точек в нем, в основном есть точки, лежащие на осях системы координат. И в случае добавления в пространство новых точек (например фраз) встает вопрос о разумном их сравнении друг с другом. В данной работе предлагается механизм сравнения таких точек исходя из принципа выхода в новое пространство, пространство описаний, в котором, в свою очередь, существует метод решения данной задачи.

Введем некоторые дополнительные ограничения.

Пусть каждому слову в пространстве слов соответствует некоторая другая точка. Построим соответствие таким образом: Есть слово, к каждому слову в любом словаре общей лексики есть некая фраза-описание, выражающая смысл этого слова через другие слова.

Пример. В словаре Ожегова, который был использован в эксперименте, дается такое описание слова **абразив:**

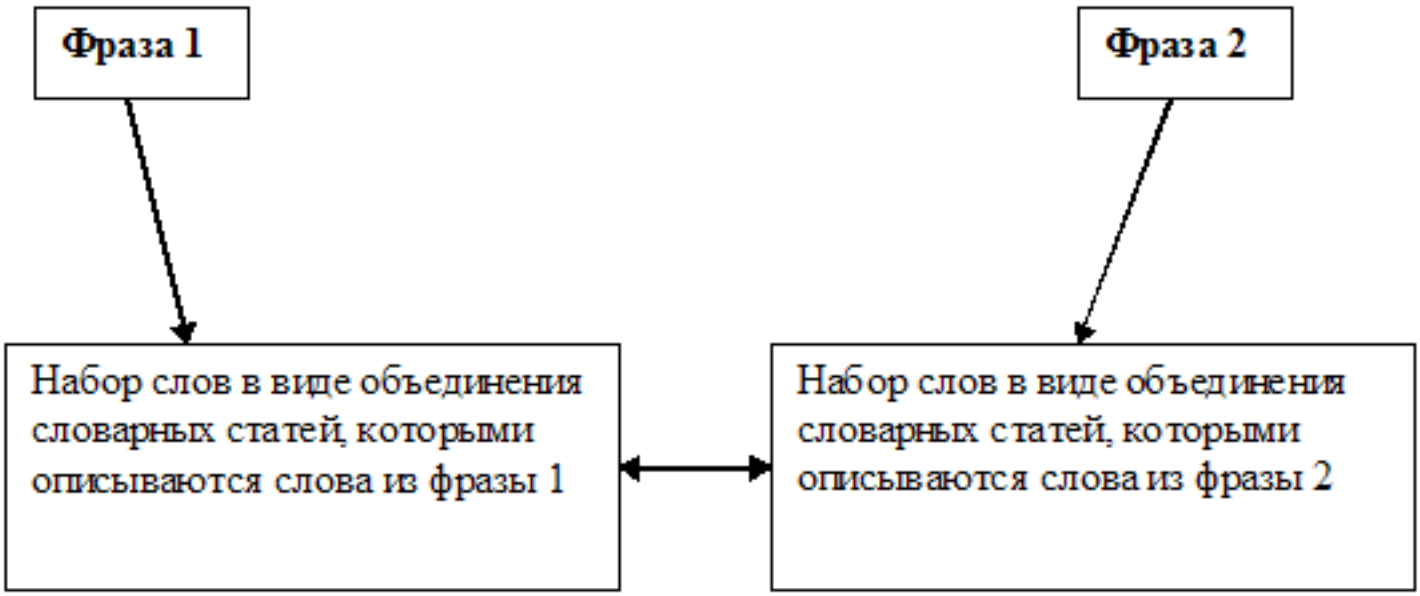
Абразив - твердое мелкозернистое или порошкообразное вещество (кремень наждак, корунд, карборунд, пемза, гранат) применяемое для шлифовки, полировки, заточки.

Еще раз подчеркнем, словарь содержит в себе **уже сформированные человеком** связи между различными словами, причем связи **смысловые**.

Воспользуемся этим фактом.

Из эмпирических соображений понятно, что если два слова описываются похоже в словарях, то скорее всего они должны быть очень близки по смыслу, причем тем ближе по смыслу, чем большим количеством одинаковых слов они описываются.

В терминах многомерного пространства, каждой точке-слову соответствует другая точка-фраза, причем для двух различных слов точки фразы будут тем ближе по мере близости, которую мы введем чуть ниже, чем большее количество одинаковых координат они имеют.



Процедура сравнения двух фраз, которая предлагается в данной статье выглядит следующим образом:

Каждой фразе сопоставляется набор слов в виде объединения словарных статей, которыми описываются слова из фразы, после чего производится сравнение между собой этих наборов по определенному правилу.

Какими свойствами должна обладать мера близости для удобного анализа?

1. При выяснении степени близости слова с самим собой значение ее должно быть 1
2. Для двух абсолютно различных слов значение ее 0
3. Мера близости учитывает количество совпадений слов в фразах-описаниях.

Изначально была предложена следующая мера близости:

Пусть P – число одинаковых координат в двух точках-фразах

Пусть F – полное число ненулевых координат в двух фразах

Тогда мера будет такой:

$$M = 2\frac{P}{F}$$

Таким образом, если одинаковых слов в двух фразах нет, то $P = 0$ и $M = 0$.

Если все слова совпадают, то $2P = F$ и $M = 1$

После проведения эксперимента, описанного ниже, стало понятно, что при использовании такой метрики достаточно сильный шум в этот метод вносят слова “как”, “что”, “либо”, “и” и так далее.

В процессе исследования метрика была изменена, после чего:

1. Из описаний были удалены все повторы слов
2. Были оставлены в словаре только существительные, прилагательные, наречия.
3. Был произведен пересчет для всех слов на предмет частоты встречаемости слова в описании, и в метрике были учтены эти частоты.

Таким образом, в процессе исследований метрика преобразовалась в следующую:

Пусть P – число одинаковых координат в двух точках-фразах

Пусть F – полное число ненулевых координат в двух фразах

$$M = \frac{\sum\limits_{i=1}^{P} \frac{1}{\sqrt{K_i+1}}}{\sum\limits_{i=1}^{F} \frac{1}{\sqrt{K_i+1}}}$$

где K – частота встречаемости этого слова во всех описаниях в словаре.

Как нетрудно увидеть, эта метрика также удовлетворяет поставленным требованиям.

3. ЧАСТЬ 2. ЭКСПЕРИМЕНТЫ

Анализатор

Для подтверждения или опровержения предлагаемой модели был сделан простой анализатор, реализующий следующие действия:

- Поиск слова в словаре
- Выделение описания данного слова
- Поиск пересечений в описаниях двух слов
- Построение степени близости двух слов в относительной шкале

Для корректного сравнения двух слов в описаниях их нужно было привести к одной форме.

В начале был использован т.н. “стеммер” – программа усечения окончаний в словах.

Достаточно быстро стало очевидно, что с помощью стеммера не удается решать задачу приведения двух слов к единой форме. Так слово “шел” ни коем образом не может быть приведено к слову “ идти” с помощью стеммера, хотя это одно слово в разных формах.

После этого был применен анализатор “Лемматайзер”, бесплатно предоставленный сайтом http://www.aot.ru. Этот инструмент, немного измененный под наши нужды, полностью решил проблему приведения двух слов к единой форме. В качестве исходного словаря был взят словарь Ожегова в электронном виде. Для быстрого поиска в нем была использована СУБД MySQL, анализатор был написан на языке PHP. Результаты работы инструмента можно посмотреть по следующему адресу: http://search.k69.ru

3.1. Постановка эксперимента. Поиск.

В данном эксперименте уточнялось качество полученной модели построения степени близости для двух фраз.

Нескольким людям предлагался небольшой отрывок из текста, и предлагалось пересказать его своими словами и желательно одной фразой.

Отрывок

Когда здешние богатейшие рыбные ловли попадут в руки капиталистов, то, по всей вероятности, будут сделаны солидные попытки к очистке и углублению фарватера реки; быть может, даже по берегу до устья пройдет железная дорога, и, нет сомнения, река с лихвою окупит все затраты. Но это в далеком будущем. В настоящем же, при существующих средствах, когда приходится иметь в виду лишь ближайшие цели, богатства Тыми почти призрачны.

Ссыльному, населению она дает до обидного мало. По крайней мере, тымовский поселенец живет так же впроголодь, как и александровский.

Полученные пересказы

1. *жить в Тыми невозможно.*
2. *Богатства реки Тымь достаточно скудны и помочь её развитию могут капиталисты, очистив ее, углубив фарватер и построив железную дорогу!*
3. *улучшить жизнь тымовских поселенцев можно за счёт капитальных вложений в реку*
4. *Край: в случае прихода к власти капиталистов, экономика и хозяйство Тымской области может сильно вырасти, в несколько раз, но только в долгосрочном периоде, а в ближайшие годы обычный народ как жил впроголодь, так и будет жить!*
5. *Когда в Тымь доберутся капиталисты, то богатейшие рыбные ловли окупят все затраты на модернизацию, но сейчас каждому поселенцу приходится жить впроголодь.*
6. *После того, как у местных рыболовных угодий появятся частные владельцы, они, видимо, начнут активные действия по обустройству территории. Очистка и углубление фарватера реки, вероятно также - железная дорога вдоль берега до устья - все это несомненно повысит эффективность использования ресурсов реки и затраты окупятся. Однако произвести это в короткие сроки не представляется возможным. При имеющихся на настоящий момент средствах, возможно концентрироваться на достижении лишь ближайших целей, поэтому эффективность использования Тыми очень низка. Ссыльное население получает мало выгоды от близости реки и живет на грани нищенства, так же, как и александровские поселенцы.*

3.2. Постановка эксперимента. 2 часть.

Далее нескольким людям предлагалось отсортировать пересказы по убыванию по степени их близости к исходному отрывку текста.

3.3. Результаты и анализ

В табл. 1 указаны варианты сортировки двадцати людей

Практически все люди отсортировали пересказы по-своему.

В таблице 2 указана некоторая дополнительная информация по сортировкам – такая как среднее значение по всем пересказам и стандартное отклонение. Такой расчет позволил понять в среднем как сортируются данные пересказы людьми, для того чтобы впоследствии сравнивать это среднее со значениями, полученными с инструмента.

Далее в таблице 2 указаны числа, полученные с помощью разрабатываемой системы при разных начальных данных – с использованием стеммера с метрикой 1, лемматайзера с метрикой 1 и лемматайзера с метрикой 2. В следующих строчках указаны порядковые номера, соответствующие разным начальным данным.

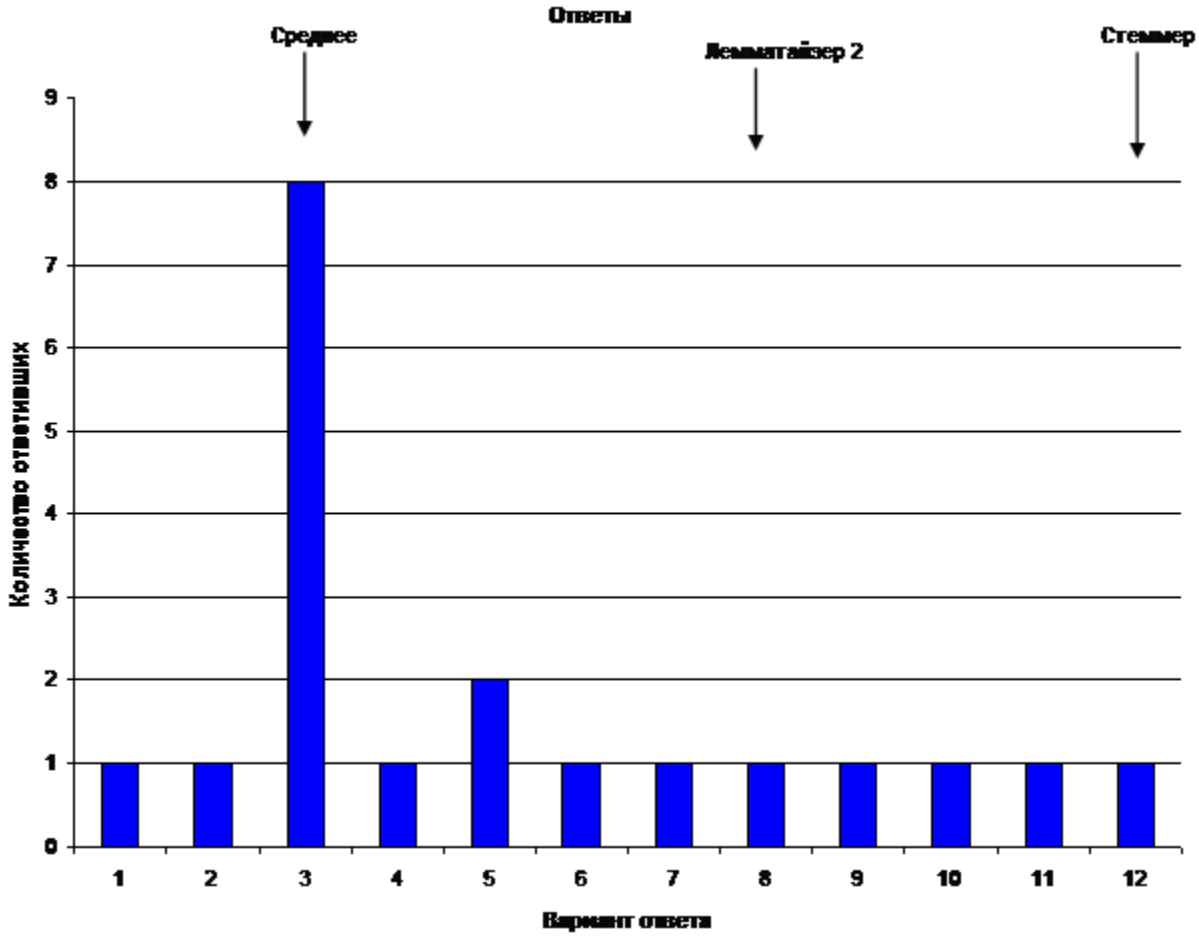
В таблице 3 представлены результаты статистического анализа с использованием ранговой корреляции Спирмана. Для каждого пересказа было получено значение корреляции этого пересказа со средним, стеммером, лемматайзером с метрикой 1 и лемматайзером с метрикой 2, после этого для всех строк было найдено среднее по столбцу, что позволило показать степень правильности результатов инструмента с различными начальными данными. Так, видно, что инструмент со стеммером справляется с задачей хуже всех, лемматайзер с метрикой 1 справляется лучше и лемматайзер с метрикой 2 справляется с задачей лучше всех предыдущих, но все равно существенно отстает от среднего, что показывает, что используемая метрика (2) все равно еще не достаточна для корректной работы инструмента. Работы в направлении усовершенствования метрики уже осуществляются.

Таблица 1. Варианты смысловых сортировок людьми.

Номер уникального ответа	Жить	Богатства	Улучшить	Край	Когда	После
1	6	2	1	3	4	5
2	6	3	4	1	5	2
3	6	4	5	2	3	1
4	6	2	5	4	3	1
3	6	4	5	2	3	1
5	6	5	4	2	3	1
3	6	4	5	2	3	1
6	6	4	5	1	3	2
7	6	4	5	3	2	1
3	6	4	5	2	3	1
8	6	3	5	4	2	1
3	6	4	5	2	3	1
9	6	4	3	2	3	1
3	6	4	5	2	3	1
3	6	4	5	2	3	1
10	6	5	4	2	1	3
11	6	5	4	1	3	2
12	5	6	4	2	3	1
5	6	5	4	2	3	1
3	6	4	5	2	3	1

Таблица 2. Дополнительные расчеты

Среднее	5,95	4,00	4,40	2,15	2,95	1,45
Стандартное отклонение	0,22	0,97	0,99	0,81	0,76	1,00
Стеммер	0.013	0.005	0.015	0.17366	0.17369	0.08
Лемматайзер (метрика 1)	0.01	0.513	0.265	0.37	0.269	0.543
Лемматайзер (метрика 2)	0.038	0.139	0.053	0.097	0.142	0.36
Порядковая сортировка (люди)	6	4	5	2	3	1
Порядковая сортировка (стеммер)	5	6	4	2	1	3
Порядковая сортировка(Лемматайзер, метрика 1)	6	2	5	3	4	1
Порядковая сортировка (Лемматайзер, метрика 2)	6	3	5	4	2	1


Ответы
Среднее
Лемматайзер 2
Стеммер
Количество ответивших
0
1
2
3
4
5
6
7
8
9
1
2
3
4
5
6
7
8
9
10
11
12
Вариант ответа

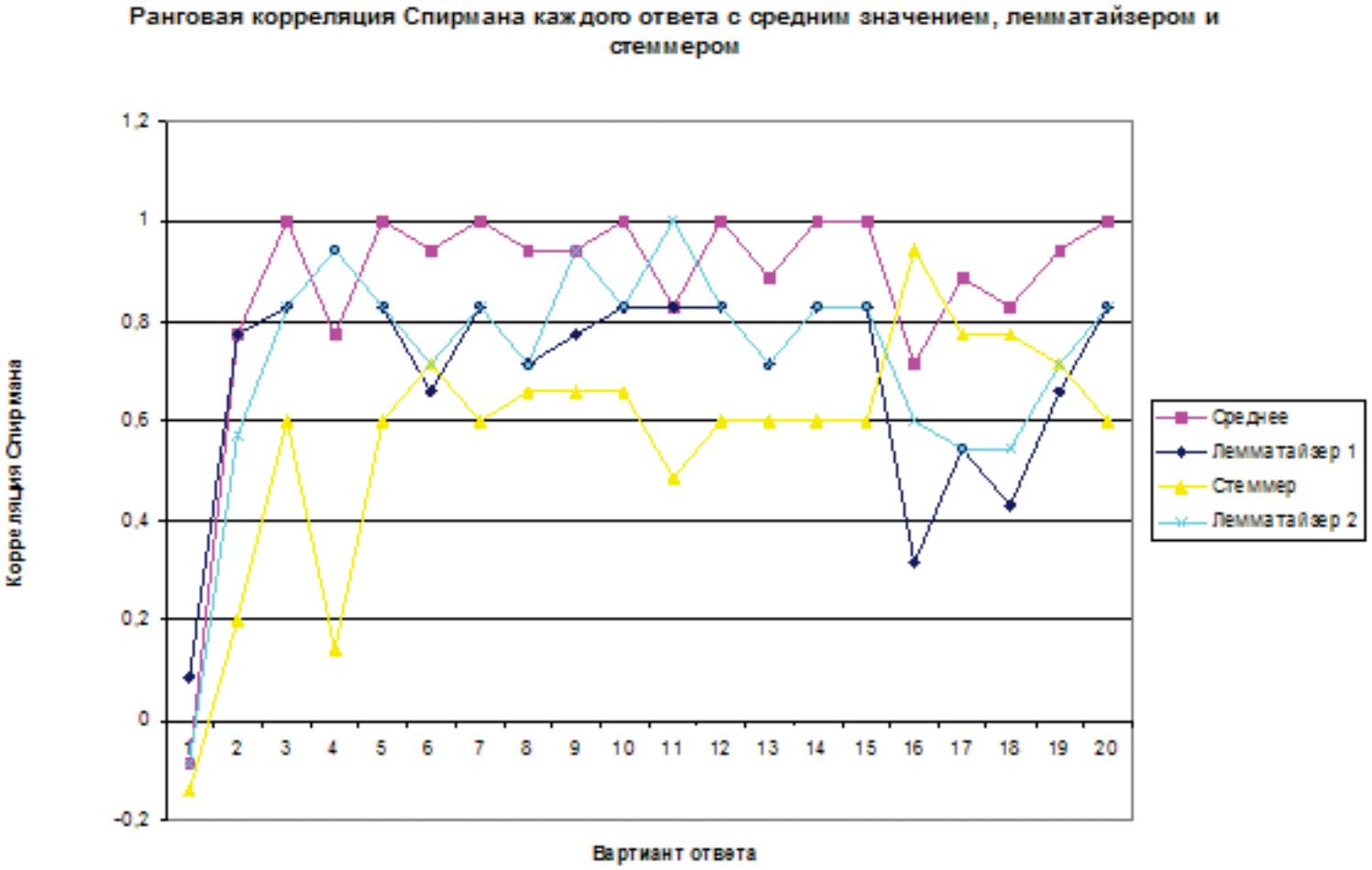

Ранговая корреляция Спирмана каждого ответа с средним значением, лемматайзером и стеммером
Корреляция Спирмана
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Вартиант ответа
Среднее
Лемматайзер 1
Стеммер
Лемматайзер 2

Таблица 3. Ранговая корреляция Спирмана

Номер уникального ответа	№	Среднее	Лемматайзер, метрика 1	Стеммер	Лемматайзер, метрика 2
1	1	-0,08571	0,085714	-0,14286	-0,0857143
2	2	0,771429	0,771429	0,2	0,5714286
3	3	1	0,828571	0,6	0,8285714
4	4	0,771429	0,942857	0,142857	0,9428571
3	5	1	0,828571	0,6	0,8285714
5	6	0,942857	0,657143	0,714286	0,7142857
3	7	1	0,828571	0,6	0,8285714
6	8	0,942857	0,714286	0,657143	0,7142857
7	9	0,942857	0,771429	0,657143	0,9428571
3	10	1	0,828571	0,657143	0,8285714
8	11	0,828571	0,828571	0,485714	1
3	12	1	0,828571	0,6	0,8285714
9	13	0,885714	0,714286	0,6	0,7142857
3	14	1	0,828571	0,6	0,8285714
3	15	1	0,828571	0,6	0,8285714
10	16	0,714286	0,314286	0,942857	0,6
11	17	0,885714	0,542857	0,771429	0,5428571
12	18	0,828571	0,428571	0,771429	0,5428571
5	19	0,942857	0,657143	0,714286	0,7142857
3	20	1	0,828571	0,6	0,8285714
Среднее значение		0,868571	0,702857	0,568571	0,7271429

4. ПОСТАНОВКА ЭКСПЕРИМЕНТА. РЕФЕРИРОВАНИЕ.

С помощью данного инструмента становится возможным смысловое реферирование. Для этого необходимо провести следующее преобразование анализатора.

Как мы предположили ранее, два слова тем ближе, чем ближе их описания по указанной метрике в многомерном пространстве. А значит, что для любого произвольного текста может быть найдено такое описание в словаре, которое будет наиболее близким к исходному тексту по той же метрике. А значит, найдя такое описание, мы сразу можем сказать, что данное слово наиболее точно описывает исходный текст.

Для анализа можно взять не одно слово, а несколько, и тогда мы получим некоторый набор ключевых слов, каждое слово в котором, вообще говоря, может не соответствовать ни одному слову из исходного текста, однако должно описывать текст в целом. Если из большого текста взять некоторое количество слов, и с помощью некоего преобразователя связать в небольшой текст, то получится текст, хорошо соответствующий исходному, причем более короткий, и как следствие являющийся рефератом, если под рефератом понимать более короткий текст в целом описывающий исходный.

4.1. Эксперименты.

Для проведения экспериментов был построен другой анализатор, который осуществлял реферирование текстов произвольной длины. В процессе экспериментов стало понятно следующее. В словаре Ожегова представлены описания, очень кратко описывающие данное слово, и из-за этого не происходила необходимая смысловая фокусировка, и поэтому анализатор не давал результатов, соответствующих нашему представлению о корректном реферировании. По этой причине результаты данного эксперимента не приводятся здесь, однако в данном направлении работы ведутся, и в дальнейшем мы надеемся представить результаты корректного автоматического реферирования.

5. ЗАКЛЮЧЕНИЕ.

В целом есть множество предпосылок того, что направление исследований верное, и в самое ближайшее время удастся достигнуть поставленных целей – смыслового поиска и автоматического реферирования. Этот метод достаточно гибкий, и как следствие может быть распространен на смысловой автоматический перевод с русского языка на другие, или же какие-то другие варианты применения смысловой компоненты в работе с текстами. В работе указываются лишь методы и направление исследований, которые продолжаются. Результаты, которые получены к данному моменту, пока не позволяют говорить об успешном завершении работ, но из данной работы становится понятно, что исследования могут позволить качественно выйти на другой уровень в использовании смысловой компоненты в интернет-поиске, реферировании, возможно переводе и в других областях знания, где смысловая компонента играет важнейшую роль.

СПИСОК ЛИТЕРАТУРЫ

1. П. Линсдей, Д. Норман. Переработка информации у человека. – Мир, 1974
2. М.М. Бонгард. Проблема узнавания. – Наука, 1967
3. Бонгард М.М., Лосев И.С., Смирнов М.С. Проект модели организации поведения – “Животное” // Моделирование обучения и поведения. М.: Наука, 1975. С.152-171.
4. М.Н.Вайнцвайг, М.П. Полякова. Архитектура мыслящей системы и нейронные сети, Сборник РАН “ Интеллектуальные процессы и их моделирование. Информационные сети”. – М. 1994
5. Rudi Cilibrasi, Paul M. B. Vitanyi, Automatic Meaning Discovery Using Google. – arxiv.org, cs.CL/0412098