Майнинг близких по смыслу языковых выражений для поисковой системы Яндекс (до 2012 года)

Алексей Сокирко, Яндекс, sokirko@yandex.ru

В данной работе описывается процесс поиска близких по смыслу слов, которыми автоматически расширялись (уточнялись) запросы, заданные пользователями поисковой системе Яндекс до 2012. В первой части мы расскажем о нашем понимании близости по смыслу и об эффективности использования поисковых расширений. Во второй части дается краткий обзор научных работ в этой области. В третьей части мы объясним, какие источники синонимии использованы и какие у них характеристики. В четвертой части дается описание системы машинного обучения, тестовой и обучающей выборок.

Введение

Общепринято, что синонимы в целом имеют общее значение (смысл), но различается в некоторых деталях. Общепринято утверждение, что полных синонимов нет. Чтобы оценить степень похожести по смыслу двух слов, нужно попробовать заменять эти слова друг на друга в разных примерах [Gauger 1972: 7-10]. Чем больше число примеров, где общее значение сохраняется, тем ближе значения, следовательно, выше синонимичность. В реальности лексикографы ограничиваются просмотром небольшого набора примеров, интуитивно обобщая эти примеры до всей генеральной совокупности. Интуитивное понятие общего смысла и степени синонимичности создает очевидные трудности. Для разных классов синонимов средняя корреляция экспертных (ручных) оценок редко достигает 0.95 (см., например, Reznik 2011]). Наши собственные эксперименты показали, что для некоторых самых широких классов синонимов эксперты в 20% случаях не могут однозначно определить степень синонимичности по двухбалльной шкале.

Можно выделить два больших класса близких по смыслу языковых выражений. Для первого класса можно найти соответствие между внутренними элементами (буквами, слогами, любыми морфологическими элементами). В первый класс включают:

1. Морфологическое словоизменение (*мама, мамой, мамами*)
2. Морфологическое словообразование (*Москва-московский, компиляция-компилирование*)
3. Аббревиатуры (МГУ-Московский государственный университет)
4. Транслиты (Майкрософт-Microsoft)
5. Слитно-раздельно (ватер-поло – ватерполо)
6. Орфоварианты (colour - color)

Во второй класс входят пары без поддержки внутренних элементов, например:

1. Переводы (стол-table)
2. Чистые синонимы (бегемот – гиппопотам)

Для любых алгоритмов в области поиска синонимов тема и жанр обрабатываемых текстов может очень сильно ограничить перебор вариантов на многих стадиях. Например, поиск синонимов для медицинских текстов[Jovic A. et al, 2011] уберет все «немедицинские» значения слов, резко упростив оценку синонимичности. В общеязыковых корпусах, каким, конечно, является корпус поисковых запросов, нет выделенной темы, все значения могут быть представлены. Однако политематичность не означает, что распределение по темам нельзя использовать. Самые распространенные темы поисковых запросов такие:

1. Кино
2. Образование
3. Музыка
4. Мебель и интерьер
5. Электроника
6. Одежда, обувь, аксессуары
7. Медицина
8. Спорт
9. Недвижимость

Видно, что эти темы примерно соответствуют страницам популярных таблоидов и составу телевизионных передач. Важно еще отметить, что средний уровень образованности пользователей поисковых систем ниже

уровня любого человека, способного составить несколько связных предложений на естественном языке. Скажем так, синонимы поисковых запросов – это синонимы, понятные даже тем, кто не может сдать ЕГЭ. Кроме этого, информационные поисковые системы освободили пользователей от любых условностей письменной речи: кавычки не нужны, регистр не нужен, порядок слов почти свободный, грамотность не необходима. Фактически мы имеем дело со свободной фонетической транскрипцией устной речи с использованием букв русского алфавита, что еще раз приводит нас к мысли о том, что без расширений запроса задачу общего информационного поиска решить нельзя.

Первые исследования использования расширений для информационного поиска начались в 1970-х [Sparck Jones, 1971]. Для русского языка использование морфологического словоизменения является стандартом, начиная с 1980-х [Лахути 1988]. Расширения в поиске могут быть грубее синонимов, это могут быть любые слова, которые имеют некоторое информационное пересечение, т.е., в предельном случае - элементами онтологии на некотором фиксированном расстоянии друг от друга. Кажется естественным по запросу [романы Толстого] искать [Война и Мир], а, если в коллекции совсем нет Толстого, показать роман «Война и Мир» на первом месте по запросу [Толстой]. В теории поисковые расширения могут быть даже не иметь информационного пересечения, а быть просто словами, которые улучшают ранжирование по данным запросам на данной коллекции документов. Описание чистых ассоциативных расширений выходит за пределы данной работы, и, кажется, выходит за пределы всей компьютерной лингвистики, поскольку эти ассоциации часто нельзя понять. Отметим, что, несмотря на это, ассоциативные поисковые расширения являются основным вектором развития поисковых расширений компании Яндекс начиная с 2012 года.

Хотя расширение запросов - общепризнанная методика для улучшения поиска, оценка непосредственного вклада поисковых расширений в ранжирование является довольно сложной процедурой. Ведь формулы ранжирования используют очень много факторов, часть из которых может «подменять» факторы от поисковых расширений. Приведем два примера:

1. Одним из основных факторов ранжирования является поиск по приссылочным текстам (текст, от которого идет гиперссылка на данный сайт). Понятно, что в приссылочном тексте могут стоять слова, которых нет в самом документе. Например, в приссылочном тексте

стоит слово «Nokia», а на основном сайте используется только «Нокиа». Получается, что сайт может быть найден по Nokia, когда этого слова нет в документе, а пользователю может показаться, что было использовано поисковое расширение.

2. Для популярных поисковых систем факторы, которые учитывают пользовательское поведение, например, клики на выдачу, становятся очень важными, часто количество переходит в качество. Иногда получается, что поисковые расширения выводят хорошие сайты в топ, пользователям начинает это нравиться, после это уже трудно понять, какие факторы были причиной улучшения.

Краткий обзор литературы по поисковым расширениям

Один из первых подходов [Attar and Fraenkel, 1977] к расширению запросов относится классу так называемых локальных методов. Для поиска расширений они используют некоторое подмножество документов, полученных с помощью исходного запроса, которые рассматриваются как релевантные. Выбор такого подмножества может производиться как вручную ([Salton and Buckley, 1997]), так и автоматически, однако первый метод требует дополнительных усилий со стороны пользователя и почти не применяется на практике.

Наиболее распространённой способ автоматического получения релевантных документов является PRF (Pseudo Relevance Feedback) [Lavrenko and Croft, 2001, Buckley et al., 1995], при котором некоторое количество лучших документов, найденных поисковой системой с помощью некоторого базового метода, признаются релевантными. Тем не менее, все методы данного класса предполагают, что почти все псевдорелевантные документы в действительности релевантны запросам и, более того, посвящены исключительно теме запроса, так как в противном случае, полученные расширения будут пессимизировать релевантные документы.

Другим большим классом методов расширения запросов является класс глобальных методов. В их основе лежат статистические модели, которые заранее строятся по текстовым корпусам или поисковым логам. Первые модели использовали кластеризацию, и большинство работ использует модели статистического перевода, в которых текст документа переводится в текст запроса. В простейшем случае используются модели, основанные на вероятности отдельных слов, например, IBM Model 1 [Berger and Lafferty, 1999]. Более сложные модели при переводе документа в слово запроса учитывают контекст явно [Wang and Zhai, 2008] или с помощью введений скрытых тематик [Gao et al., 2010].

Анализ работ, основанных на различных методах [Bendersky et al., 2012, Mandala et al., 1999], показывает, что использование комбинации

нескольких источников, позволяет значимо улучшить качество поиска. Однако их компоновка выполняется либо на основании некоторых эвристических порогов, либо с помощью взвешенной суммы. В работе [Cao et al. 2008] предлагается использовать классификатор (SVM) для определения релевантных и нерелевантных поисковых расширения, который построен на ручной выборке, в которой эксперты должны были оценить, портит или помогает данное расширения данной поисковой системе. Эта работа пересекается с нашей работой в части использования методов машинного обучения, но отличается от метода получения оценок и классифицирующими факторами.

Источники синонимии (базовые майнеры)

Как уже было сказано выше, в данное работе мы пытаемся обнаружить только такие поисковые расширения, которые являются близкими по смыслу выражениями (синонимами). На самом первом этапе нам необходимо получить список возможных гипотез из **источников синонимии**:

1. Выравнивание параллельных текстов;
2. Линковая база;
3. Скобочные написания;
4. Открытые словари, например, Википедия;
5. Переформулировки запросов;
6. Кликовые данные.

Каждый источник выдает данные разные по качеству, количеству и характеру. Для нас принципиально, что возникновение синонимии в этих источниках является не случайным процессом, а вполне объяснимым – во всех случаях это основано на разном назывании одного и того же объекта разными выражениями. Понятно, что список источников является открытым. Рассмотрим подробнее каждый из них.

Выравнивание предполагает существование большой базы пар выравненных словосочетаний[Yandex MT], типа:

киотский протокол - kyoto protocol 20

киотские соглашения - kyoto treaty 10

киотские соглашения - kyoto protocol 11

киотский протокол - kyoto treaty 40

киотский протокол - kyoto agreement 1

В правой колонке стоит частота, с которой данная пара встречается в параллельных текстах. Два русских выражения можно объявить гипотезами синонимов, когда они переводятся в одно и то же английское выражение. Общих английских выражений может быть много, чем больше, тем лучше. С другой стороны, нужно учитывать степень однозначности английского выражения: чем больше у него переводов, тем хуже. В конечном итоге мы приходим к проблеме бикластеризации двудольных графов, для которых было предложено много решений, см., например [Dhillon I. S 2001]. Поскольку выравнивание – всего лишь один из источников, на текущем этапе не нужно решать проблему кластеризации в окончательном виде, нужно лишь только породить набор значимых сигналов (факторов) для последующего машинного обучения. Из исходной таблицы в 107 млн. выравненных пар, после обрезания по некоторому порогу, можно получить 21 млн. пар гипотез. По нашим сведениям, впервые для поисковых расширений результаты выравнивания использовались в работе [L van der Plas et al 2008], хотя в этой работе улучшения качества они не дали.

Линковая база – это набор приссылочных текстов на гиперлинках, которые ведут на один сайт. Этот источник очень сильно зашумлен, например, часто приссылочный текст содержит только что-то подобное:

смотри подробнее здесь

В исходной таблице линков, ведущих на один и тот же сайт, содержится 8 млрд. пар. Использование приссылочных текстов для расширений – один из самых распространенных методов, например [Kraft R et al 2004]

Скобочное написание – это набор n-gram, которые встречаются с текстах рунета в контексте скобок, например:

Московский государственный университет (МГУ)
Владимир Путин (Vladimir Putin)

При работе с этим источником не определена левая граница, поэтому приходится использовать статистические методы вычленения границы словосочетания. Исходная таблица скобочного написания содержит 5 млрд. пар, после определения границ словосочетаний получается 5 млн пар. Анализ скобочных конструкций является стандартным методом для вопросно-ответных систем, см., например, [Soubbotin 2001].

Открытые словари. Наша система использует Википедию и другие открытые словари для формирования множества гипотез. Русская Википедия содержит около миллиона жестких редиректов, типа:

Абрикос сибирский *Даурсат*
Авачинская бухта *Авачинская губа*

Проблема со словарями и любыми энциклопедиями в чрезмерной академичности, и все же они часто содержат множество высокоточных синонимов, но стараются исключать оценочные и просторечные переходы: типа *вконтакте-вконташа*. Однако словарный метод, основанный на Wordnet, является обычным base line для многих систем поисковых расширений.

Переформулировки запросов - это пары запросов, которые часто возникают внутри поисковых сессий. Например:
[Апокалипсис смотреть] –> [Апокалипсис фильм]. Если в запросах есть что-то общее и они были заданы одним пользователем и на протяжении короткого времени, можно предположить, что отличающиеся части имеют общий смысл. Понятно, что нас интересует частые переформулировки, т.е., идущие от многих запросов, конечно, важна не только частота запросов, но и их разнообразие. В исходной таблице содержится примерно 2 млрд. пар запросов, из которых, используя несколько слабых эвристических порогов, можно получить около 80 млн пар.

Кликовые данные - это пары разных запросов, с которых пользователи кликнули по одному и тому же сайту на поисковой выдаче. С одной стороны, это довольно хороший источник синонимии, с другой стороны, для поисковой системы он не является самым полезным, поскольку поисковая система уже некоторым образом уравняла эти два запроса, показав по ним кликнутый сайт. Более подробно об использовании переформулировок смотри [Cui H. et al. 2002]

Машинное обучение

Результаты работы всех майнеров объединяются в одну таблицу (примерно 200 млн. гипотез). Эта таблица **нормализуется** и **фильтруется**.
Нормализация – это приведение гипотез к нормальной форме (как они должны быть в любом ручном словаре):
Государственной Думы
Государственную Думу -> *Государственная Дума*

Нормальные формы заимствуются из поисковых запросов, например, для словоформы *одноклассниках* нормальной считается форма *одноклассники*, а не *одноклассник* (из-за популярности соответствующей социальной сети)

Фильтрация гипотез для левой части гипотезы осуществляется по логу запросов, а правой части – по текстам рунета.

После фильтрации и нормализации получается порядка 150 млн. пар, что уже подается на вход модели машинного обучения, в нашем случае - это Матрикснет [Гулин 2010]. Метод Матрикснет показал лучшие результаты, хотя методы типа RandomForest не так уж сильно от него отстают. Основной проблемой на этом этапе является не метод машинного обучения, а обучающая выборка. По многим нашим экспериментам выборка должна быть достаточно большой. Идеально было бы размечать выборку из входного множества гипотез, но это представляет собой серьезные трудности, поскольку, по примерным оценкам, во входном множестве гипотез возможных синонимов меньше 10 процентов. С учетом разногласий между разметчиками, качество разметки очень низкое. Выборку необходимо перепроверять много раз. Для русского языка было собрано порядка 40 тысяч пар.

Факторы машинного обучения во многом пересекаются с источниками синонимии, можно сказать даже жестче, все источники синонимии являются одновременно факторами обучения. Кроме этого, используются следующие сигналы:

- FactorAnd – как часто два выражения стоят рядом в тексте, это фактор «антисинонимии», поскольку два синонима обычно не стоят вместе в тексте.
- FactorCtxt – насколько часто два выражения встречаются в похожих контекстах. Мы не используем этот фактор в качестве источника, поскольку он очень грязный, например, когипонимы, типа «январь»-«февраль» будут по этому фактору максимально близки.
- ExtTypes - тип поискового расширения (транслит, аббревиатура и т.д.)
- Leven, Translit – близость по левенштейну, транслитности
- Частоты обеих частей по запросам и текстам

Все факторы разделены на несколько подфакторов с небольшими , но иногда существенными статистическими вариациями (где-то берется среднее арифметическое, где-то медиана, где-то используются разные пороги отсечения). Если пытаться обобщить значимость факторов по их вкладу в качество классификации, получается примерно такой результат: ExtTypes 25%; Dict 24%; Частоты 12%; Reform 10%; Clicks 10%; Align 11%; FactorAnd 8%.

В результате применения модели для каждой входной гипотезы получается число (Predict), которое до некоторой степени говорит о степени синонимичности этих двух выражений. Мы выбираем первые несколько миллионов самых близких гипотез и объявляем их **бесконтекстным словарем** поисковых расширений. Этот словарь будет использован для **контекстного** расширения в пределах одного запроса при обработке запроса поисковой системы Яндекс. Контекстное расширение имеет некоторое пересечение с бесконтекстными механизмами, но в целом его описание выходит за пределы данной работы.

Аналитика

Полностью понять, как работает машинное обучение с сотнями решающих деревьев, очень сложно. Интересно взглянуть на случаи, когда система сильно ошибается или когда не находит очевидные синонимы. В следующей таблице приведены примеры синонимов:

		Hum	ML	clicks	reform	parenth	align	dicts	and	RevFreq
яник	yanik	1	1	22	3	521	5	0	382	5372135
топограф	землемер	0	1	181	10	4918	24	2	100702	3962614
обанкро-титься	разоряться	1	1	2185	25	64	406	6	15234	1372981
утвержде-ние	апробация	0	1	0	34	221	788	2	21565	34989
яндэкс видео	яндекс видео	1	1	31	63	0	0	0	0	1200
ушкино	пушкино	0	1	44	65	0	0	0	296	1334080 28
москва	moskwa	1	0	1522	105	52347	3	0	251657	11568
кормящая	содержать	0	1	4940	130	3608	59	8	464996	64908
гадать	гадалка	1	0	2688	576	269	0	0	139020	399027
гоголь	гоголевски й	1	0	4229	650	3636	633	0	104104	128421
сайт	веб страница	1	0	20406	991	2977	1985 8	0	57622	303
племена	роды	0	1	12685	1602	32278	2749	8	3917014	66223
хара-ктерные	отличите-льный	1	1	9366	2561	852	5051	4	617187	58277
становить ся	вставать	0	1	19770	3111	11207	4284	3	1856933	2020
освещени	света	0	1	212979	6200	325343	2369 3	6	1157679 6	24540

е										
голые	раздетые	1	1	128426	24225	2998	284	4	1266094	32955
забере-менеть	бере-менность	1	0	146710	24651	7116	741	0	4529567	464297
шины	шинный	1	0	158473	26532	7165	2772	2	4759250	6037
деревянные	из дерева	1	0	4488191	29929	295122	4671	0	1721138	19958
шины	резина	1	0	2522069	54960	489180	2649	2	21958046	6037
курсовые	реферат	1	0	1270261	55446	13728	111	0	26917782	18584

Таблица 1. Сложные случаи синонимии.

В первых двух колонках стоят гипотезы синонимов. В колонке 3 – ручная экспертная оценка, которая говорит о наличии синонимической связи. В колонке 4 - результат работы машинного обучения (“0” – ниже порога, “1” – выше порога). Дальше идут сырые данные для факторов машинного обучения. Последняя колонка – обратная частота слова в первой колонке в рунете. Данная таблица не является случайной выборкой из результирующего словаря, мы специально искали больше плохих примеров для наглядности. Точность этих данных по колонке 3 – около 57%, а в случайной выборке – больше 90%.

Если принять, что нулевая гипотеза – это отсутствие синонимической связи между словами, в таблице 1 есть шесть ошибок первого рода: <*утверждение, апробация*>, <*ушкино, пушкино*>, <*кормящая, содержать*>, <*племена, роды*>, <*освещение, света*>, <*топограф, землемер*> и девять ошибок второго рода <*москва, moskwa*>, <*гадать, гадалка*>, <*гоголь, гоголевский*>, <*сайт, веб страница*>, <*забеременеть, беременность*>, <*шины, шинный*>, <*деревянные, из дерева*>, <*шины, резина*>, <*курсовые, реферат*>. Сразу бросается в глаза, что в ошибках первого рода часто встречается омонимия (*роды, света),* разные значения (*кормящая),* искусственные синонимы, свойственные только письменной речи (*утверждение, топограф).* Ошибки второго рода гораздо сложнее объяснить, стоило ожидать, что для них будет недостаточно статистики (*гоголь-гоголевский*), но видно, что и

очень частые примеры (*курсовые-реферат*) не проходят. Если пытаться построить разные модели машинного обучения по похожим выборкам, а начать даже с изучения корреляции между столбцами, видно, что даже самыми важными факторами часто являются Dict и Reform, они же входят в топ важности большой модели. Нормализация значений факторов с помощью колонки RevFreq приводит к более аккуратному предсказанию, но слабо меняет топ значимых факторов. Значимость фактора Dict приводит нас к мысли о важности конкретных случаев, непохожести одного случая на другой. Выглядит так, будто мы одним методом пытаемся решить все задачи из школьного курса математики (геометрия, арифметика, алгебра). Интересно, что вывод о слабой силе обобщения прямого моделирования синонимии методами машинного обучения был до некоторой степени сделан раньше в [Сокирко 2010].

Заключение

В данной работе мы объяснили, как работал сбор гипотез синонимов (расширений запросов) для поисковой системы Яндекс. Попытки объединить разные сигналы синонимичности в единую платформу предпринимались уже много раз, однако на таком уровне это сделано впервые для русского языка. Несмотря на то, что система проработала довольно успешно, с теоретической стороны она довольно сложна. Размеры обучающих выборок сопоставимы с размерами небольших опубликованных словарей синонимов. Значимость словарных факторов обучения говорит о том, что обучение не находит достаточного сигнала в других факторах. **Ручные непосредственные** сигналы слишком важны для машинного обучения, кажется, эти сигналы не поддаются обобщению. Вместе с тем, трудно себе представить, чтобы поисковая система вообще отказалась от автоматизированной оценки степени синонимичности, поэтому работа в этом направлении будет продолжена.

Литература

[Gauger 1972] H.-M.Gauger. Zum Problem der Synonymie/Hrsg.v.Gunter Narr.- Tübinger Beiträge zur Linguistik, Band 9.- Tübingen 1972.

[Reznik 2011] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language //arXiv preprint arXiv:1105.5444. – 2011.

[Jovic A.,, 2011] Jovic A., Prcela M., Gamberger D. Ontologies in medical knowledge representation //Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on. – IEEE, 2007. – С. 535-540.

[Sparck Jones, 1971] Automatic Keyword Classification for Information Retrieval. . Butterworth,London.

[Лахути 1988] Автоматизированные документально-фактографические информацион-но-поисковые системы/ Д.Г. Лахути// Итоги науки и техники. Сер. Информатика. Т.12. – М., 1988

[Cui H. et al. 2002] Cui H. et al. Probabilistic query expansion using query logs //Proceedings of the 11th international conference on World Wide Web. – ACM, 2002. – С. 325-332.

[Attar and Fraenkel, 1977] Attar, R. and Fraenkel, A. S. (1977). Local feedback in full-text retrieval systems.Journal of the ACM (JACM), 24(3):397–417.

[Bendersky et al., 2012] Bendersky, M., Metzler, D., and Croft, W. B. (2012). Effective query formulation with multiple information sources. In Proceedings of the fifth ACM international conference on Web search and data mining - WSDM ’12, page 443, New York, New York, USA. ACM Press.

[Berger and Lafferty, 1999] Berger, A. and Lafferty, J. (1999). Information retrieval as statistical translation. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 222–229. ACM.

[Buckley et al., 1995] Buckley, C., Salton, G., Allan, J., and Singhal, A. (1995). Automatic query expansion using smart: Trec 3. NIST SPECIAL PUBLICATION SP, pages 69–69.

[Cui et al., 2002] Cui, H., Wen, J.-R., Nie, J.-Y., and Ma, W.-Y. (2002). Probabilistic query expansion using query logs. In Proceedings of the 11th international conference on World Wide Web, pages 325–332. ACM.

[Salton and Buckley, 1997] Salton, G. and Buckley, C. (1997). Improving retrieval performance by relevance feedback. Readings in information retrieval, 24:5.

[Voorhees, 1994] Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pages 61–69. Springer-Verlag New York, Inc.

[Mandala et al 1999] Mandala R., Tokunaga T., Tanaka H. Combining multiple evidence from different types of thesaurus for query expansion //Proceedings of the

22nd annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 1999. – С. 191-197.
[Cao 2008] CAO, G., GAO, J., NIE, J.-Y., AND ROBERTSON, S. 2008. Selecting good expansion terms for pseudorelevance feedback. In *Proceedings of the 31st Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 243–250.
[Yandex MT] Yandex's Machine Translation Technology http://company.yandex.com/technologies/translation.xml]
[Dhillon I. S 2001] Dhillon I. S. Co-clustering documents and words using bipartite spectral graph partitioning //Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2001. – С. 269-274.
[L van der Plas et al 2008] L van der Plas L., Tiedemann J. Using lexico-semantic information for query expansion in passage retrieval for question answering //Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering. – Association for Computational Linguistics, 2008. – С. 50-57.
[Kraft R et al 2004] Kraft R., Zien J. Mining anchor text for query refinement //Proceedings of the 13th international conference on World Wide Web. – ACM, 2004. – С. 666-674.
[Soubbotin 2001] Soubbotin M. M., Soubbotin S. M. Patterns of potential answer expressions as clues to the right answers //Proceedings of the Tenth Text REtrieval Conference (TREC 2001). – 2001.
[Гулин 2010] Матрикснет http://download.yandex.ru/company/experience/searchconf/Searchconf_Algoritm_MatrixNet_Gulin.pdf
[Сокиро 2010] Быстрословарь: предсказание морфологии русских слов с использованием больших лингвистических ресурсов//Конференция Диалог. - 2010