Г.С. Осипов, И.В. Смирнов, И.А. Тихомиров

Реляционно-ситуационный метод поиска и анализа текстов и его приложения

Аннотация. В статье кратко описан реляционно-ситуационный метод анализа текстов естественного языка, построенный на основе теории коммуникативной грамматики русского языка и теории неоднородных семантических сетей. Показано, как реляционно-ситуационный метод может быть использован для точного поиска документов в локальных и глобальных сетях и создания электронных библиотек.

Ключевые слова: семантический поиск, реляционно-ситуационная модель, значения синтаксем, семантическая сеть.

Введение

От возникновения нового знания до его публикации «на твердом носителе» проходит значительное время. Это обстоятельство приводит к задачам создания сервисов по точному поиску требуемой информации как во внешних источниках, так и в собственных электронных хранилищах, мониторинга имеющихся электронных ресурсов в тех или иных областях, обнаружения новых ресурсов, извлечения целевой информации, создания и ведения собственных электронных хранилищ. Существующие поисковые машины не обладают всеми требуемыми для решения указанных задач качествами. Кроме того, они обладают рядом хорошо известных особенностей, среди которых важнейшей является низкая релевантность поиска, т.е. высокий уровень шума. Следует также добавить, что во многих случаях отсутствует возможность «фокусировки» запроса, т.е. возможность формулировки действительных поисковых потребностей пользователя.

Решение указанных задач возможно на пути применения таких методов поиска, которые допускают запросы на естественном языке, обладают способностями к анализу семантики документов и запросов и к вычислению их семантической близости. Описанию соответствующих методов и программных средств посвящена настоящая статья.

1. Смысл и значение в естественном языке

В лингвистической семантике слову «смысл» соответствуют два близких, но различных понятия. В первом случае под семантикой понимается значение - информация, связываемая со словом конвенционально (например, в толковом словаре), во втором- смысл – совокупность ассоциаций и коннотаций, связываемых со словом в сознаниях коммуникантов. Хотя каждая из этих двух ипостасей с равным правом претендует на то, чтобы быть предметом семантики, здесь мы зафиксируем первое понимание. Тогда основная цель семантического анализа – переход от значения отдельных слов к значению высказываний.

Работы в области семантического анализа ведутся достаточно давно как отечественными, так и зарубежными специалистами. По-видимому, в числе первых следует указать работы Хомского (1957), в которых предлагалась трансформационная грамматика, целями которой было: построение глубинной синтаксической структуры; запись значений каждого предложения; обнаружение семантических аномалий.

Можно назвать и Кембриджский лингвистический кружок (1959 год), в котором был предложен некоторый язык-посредник, который должен отождествлять семантически эквива-

лентную информацию и снимать межъязыковую синонимию.

Ч. Филмор в 1966 году предложил так называемые предикатно – аргументные структуры, в которых языковым конструкциям (аргументам) приписывались роли: агент, объект, место, адресат, инструмент, источник.

Важной вехой в исследовании семантики естественного языка явилась Модель «Смысл-Текст» (Мельчук, 1974), в которой выбор значений определялся максимальной повторяемостью семантических элементов в пределах предложения.

Монтегю принадлежат попытки «логизирования» семантики высказывания; при этом значения высказывания рассматриваются как истинностные значения.

В отличие от этого, предлагаемый здесь подход опирается, с одной стороны, на коммуникативную грамматику русского языка [1,2], с другой - на теорию неоднородных семантических сетей [3].

2. Краткое описание реляционно-ситуационной модели текста

Коммуникативная грамматика русского языка [1,2] отвергает традиционное противопоставление синтаксиса семантике, которое предполагает разделение знаний о законах формирования связной речи на два уровня: знания о форме (синтаксис) и знания о значении (семантика). Главная идея коммуникативной грамматики заключается в том, что синтаксис должен изучать осмысленную речь, а синтаксические правила должны учитывать категориальные значения слов, чтобы иметь возможность определять обобщенный смысл любой синтаксической конструкции – от слова до словосочетания и простого предложения.

Формируя и изучая связную речь, синтаксис имеет дело, прежде всего, с осмысленными единицами, несущими свой не индивидуально-лексический, а обобщенный, категориальный смысл в конструкциях разной степени сложности. В основе коммуникативной грамматики лежит понятие синтаксемы, как минимальной единицы смысла высказывания.

Синтаксемой называется минимальная синтактико-семантическая единица языка, несущая обобщенный категориальный смысл и характеризующаяся взаимодействием морфологических, семантических и функциональных признаков.

Несмотря на сложность описания, синтаксема является интуитивно понятной конструкцией для любого носителя языка и используется им повсеместно для построения различного рода высказываний. Пример: «Митрофанушка не знал, что говорит прозой». Синтаксемами имен существительных в данном предложении являются: «Митрофанушка» – личное существительное именительного падежа, принадлежащее к классу имен собственных и играющее в данном предложении роль субъекта. «Прозой» - существительное в творительном падеже, принадлежащее к классу признаковые и играющее в данном предложении роль медиатива.

Примеры *значений* синтаксем:

- Аблатив – исходная точка движения (*выйти из комнаты*).
- Агенс – производитель действия (*закон подписан президентом*)
- Адресат – лицо или реже предмет, к которому обращено информативное, донативное или эмотивное действие (*обратиться к президенту*).
- Дестинатив – назначение действия или предмета (*выступать в защиту животных; поехать на лечение*)
- Транзитив – компонент со значением пути движения (*Корейский лидер проехал почти по всей России на поезде*).
- Инструментив – орудие действия (*Что написано пером, того не вырубишь топором*).
- Каузат – объект каузирующего воздействия (*способствовать вступлению России в ВТО*).
- Каузатор – воздействующий фактор (*от ученья нравам лишь вред*).

Для понимания предложения недостаточно лишь знания значений синтаксем. Важна и их сочетаемость друг с другом. Эта сочетаемость определяется семейством бинарных отношений на множестве синтаксем [3]. Приведем несколько примеров элементов таких отношений:

отношение INS: (субъект, инструментив)
отношение TRA: (транзитив, объект)
отношение TAR: (субъект, дестинатив)
отношение DES: (объект, дестинатив)
отношение TRS: (субъект, транзитив)

Отношения на множестве синтаксем являются отражением их семантических связей, которые суть ситуативные отношения понятий в понятийной системе предметной области. Таким образом, возникает следующая

***Гипотеза**: семантика предложения определяется множеством входящих в него синтаксем и семейством отношений на них.*

Опишем сценарий определения синтаксем и множества отношений на них:

ШАГ 1. Морфологический анализ. На этапе морфологического анализа в тексте распознаются слова и разделители. Для каждого слова на основании его морфологии определяется список всех возможных грамматических форм. Словоформы, соответствующие одной нормальной форме слова и одной части речи, а также стоящие в одном числе (для изменяющихся по числам частей речи), объединяются в группы, которые далее будем условно называть «лексемами». В результате морфологического анализа формируется множество предложений, каждое из которых содержит упорядоченный список слов, у каждого слова может быть несколько вариантов омонимичных лексем.

ШАГ 2. Синтаксический анализ. Основной задачей синтаксического анализа в нашем случае является установление различных зависимостей между лексемами и выделение синтаксем. Вход синтаксического анализа представляет собой предложение, получаемое на выходе морфологического анализа. Выход – предложение в виде списка деревьев синтаксического подчинения (на основании эвристических правил выбирается наиболее вероятный вариант синтаксического разбора) и множество синтаксем. Именной синтаксемой считается именная или предложная группа – словосочетание, управляющим словом которого является имя существительное или предлог.

ШАГ 3. Реляционно-ситуационный анализ. Основной задачей реляционно - ситуационного анализа является выявление значений синтаксем и семантических связей между ними. Главную роль здесь играют глаголы, имеющие, как правило, центральное положение в семантической структуре предложения и оказывающие решающее влияние на именные словосочетания и предложения.

Сведения о синтаксической сочетаемости каждого глагола с синтаксемами заносятся лингвистами в словарь предикатных слов. Кроме того, в словаре имеются указания на то, как могут быть связаны между собой именные синтаксемы. Набор заносимых в словарь бинарных связей на множестве ролей также является специфичным для каждого типа предикатных слов и определяется априорно.

ШАГ 3.1. Определение предикатного слова. Принципы выделения предикатных слов заключаются в следующем:

- поиск глаголов, девербативов и предикативных наречий выполняется по данным морфологии и словаря предикатных слов;
- сочетания [неполнознаменательный глагол в личной форме / предик. наречие + инфинитив] рассматриваются в комплексе как одно предикатное слово;
- если в одном фрагменте встретились несколько предикатных слов (например, глагол + инфинитив, глагол + девербатив), то синтаксемы для них ищутся в ограниченных контекстах.

ШАГ 3.2. Установление значений синтаксем и отношений на них. После того, как выделены предикатное слово и синтаксемы, связанные с ним, определяются значения синтаксем с использованием указанных лингвистических словарей. В статье словаря для каждого предикатного слова ищутся вхождения морфологических признаков (предлог, падеж, категориально-семантический класс) связанных с ним синтаксем. Если морфологические признаки синтаксемы найдены в статье, то ей приписывается соответствующее значение. После того, как установлены значения синтаксем при предикатном слове, определяются устойчивые отношения между ними: в той же статье словаря ищутся пары значений синтаксем, образующие элемент семантического отношения.

В безглагольных предложениях значение синтаксемы зависит от контекста – других синтаксем и слов определенных частей речи. Такая зависимость записывается в виде правила, посылка которого содержит характеристики контекста синтаксемы, а следствие указывает значение, которое необходимо установить рассматриваемой синтаксеме. Правила установления значений синтаксем были получены с помощью индуктивного метода машинного обучения, по примерам из электронной версии синтаксического словаря Г.А. Золотовой [3]. Приведем примеры правил для синтаксем родительного падежа:

Правило: Если встречается синтаксема в падеже <родительный> с предлогом <для>, имеющая категориальный класс <личное>, а до неё встречается синтаксема в падеже <именительный>, имеющая категориальный класс <предметное>, то полагается, что первая синтаксема имеет значение <дестинатив>.

Правило: Если встречается синтаксема в падеже <родительный> с предлогом <из, изо>, имеющая категориальный класс <локативное>, а рядом с ней встречается синтаксема в падеже <именительный>, имеющая категориальный класс <личное>, то полагается, что первая синтаксема имеет значение <аблатив>.

Таким образом, результатом работы описанных процедур анализа являются структуры, описывающие семантическую информацию, передаваемую текстом, в виде реляционно-ситуационной модели. С формальной точки зрения они являются неоднородными семантическими сетями и называются семантическими образами текстов запросов и документов. Семантические образы текстов запросов и документов используются в вычислении степени семантической близости запроса и документов при поиске и ранжировании (Рис.1).

3. Взаимодействие реляционно-ситуационной модели и статистических механизмов

При использовании реляционно-ситуационной модели текста в задачах поиска не обязательно отказываться от традиционных статистических методов. Напротив, стандартные словарные методы помогают выполнить предварительный отбор документов по статистическим критериям соответствия слов документа словам запроса, тем самым отбросить документы с незначимыми, случайными появлениями слов запроса в них. Для этого используются известные показатели TF*IDF и подобные им. Для более «тонкого» ранжирования полученных документов необходимо использовать семантические значения слов, содержащихся в документах и запросе. Документ, содержащий слова запроса в тех же семантических значениях, в которых они входят в данный запрос, имеет больший вес по сравнению с документами, в которых слова имеют другие семантические значения или не имеют их совсем. Аналогичным способом можно учитывать и семантические связи между словами предложений. Таким образом, после сортировки документы более близкие по смыслу запросу выводятся пользователю раньше.

Эти соображения находят ещё одно применение – поиск в документах ответов на вопросы. Принцип поиска ответа на вопрос заключается в том, что в предложении документа ищутся слова в том же семантическом значении, что и вопросительное слово запроса, при наличии в предложении остальных слов запроса. Например, если имеется вопросительный запрос «К чему приводит окисление наночастиц?», то ответом может быть предложение «Окисление наночастиц приводит к формированию структуры наночастицы – металлическое ядро-оксидная оболочка», т.к. именная группа «к формированию структуры наночастицы – металлическое ядро-оксидная оболочка» имеет семантическое значение «результатив», как и синтаксема «к чему» в запросе. На Рис. 1 приводятся семантические сети для запроса и ответа. Сплошной стрелкой показано подчинение именных синтаксем предикатному слову, подпись к стрелке обозначает значение синтаксемы, пунктирной стрелкой показана семантическая связь между синтаксемами. Значение «результатив» применяется к компонентам предложения, выражающим результат некоторого действия, «каузатив» обозначает производителя действия.

Для быстрого поиска необходимо создание семантического индекса, который содержит список вхождений слов в документы, статистический вес значимости слова в документе и значение соответствующей синтаксемы. Предварительное индексирование текстов документов позволяет при поиске выполнять быструю выборку вхождений слов запроса в документы для дальнейшего ранжирования результатов по описанным выше методам.

Следует заметить, что при использовании системы для поиска в хранилищах эффективность поиска можно увеличить, выполнив адаптацию реляционно-ситуационной модели.

Указанная адаптация к предметной области заключается в пополнении словаря предикатных слов, необходимых для установления значения синтаксем, глаголами и отглагольными существительными, характерными для данной предметной области. Это возможно сделать на

Рис. 1. Семантические сети запроса и предложения, являющегося ответом

основе полуавтоматического анализа текстов предметной области. Важную роль для определения значений синтаксем играют категориальные классы существительных, поэтому необходимо введение классов, характерных для данной предметной области. Составление списка элементов класса также может выполняться в полуавтоматическом режиме на основе текстовых документов по рассматриваемой теме.

4. Создание электронных библиотек в различных предметных областях

Одним из возможных применений изложенных соображений является создание и использование электронных библиотек.

Электронная библиотека (ЭБ) – это совокупность электронных текстовых документов, сервисов быстрого доступа к ним и сервисов наполнения библиотеки новыми документами. Сервисы доступа разделяются на поиск документов по заданным параметрам и просмотр содержимого электронной библиотеки, удовлетворяющего некоторым критериям.

Электронные документы содержатся в библиотеке в структурированном виде.

Реляционно-ситуационный поиск позволяет автоматически загружать документы в библиотеку из разнородных источников. Источниками могут быть, например, сайты журналов по требуемой тематике. Наполнение ЭБ может происходить в автоматическом режиме по расписанию. В процессе наполнения опрашиваются внешние источники (сайты журналов), автоматически извлекается нужная информация и помещается в ЭБ. Модули наполнения в полуавтоматическом режиме с помощью методов машинного обучения настраиваются на извлечение из информационных ресурсов только необходимой информации, например, текстов статей, но не сопутствующей информации на сайте журнала.

Электронный каталог – это систематизированное описание хранящихся в библиотеке документов. Обычно каталог представляет собой иерархическую или сетевую структуру рубрик, каждая из которых содержит описания документов, соответствующих данной рубрике, и ссылки непосредственно на сами документы.

Рубрики каталога могут составляться по полям документа, например по авторам, или по названиям издательств, но возможно и создание тематических каталогов, где рубрики соответствуют темам в предметной области. Тематические каталоги разбивают все документы электронной библиотеки на множества (возможно, пересекающиеся) тематически близких документов, что может использоваться при поиске и навигации.

Поиск документа в электронной библиотеке может проводиться как по отдельным полям документа, так и по его тексту. Параметром поиска по текстовым полям документа является запрос на естественном языке или вопросительное предложение. На основе реляционно-ситуационной модели создаются семантические индексы текстовых полей документов, которые позволяют значительно повысить скорость поиска в электронной библиотеке. Метод семантического поиска по индексам значительно повышает точность поиска. Реляционно-ситуационная модель используется также для повышения точности автоматической каталогизации.

Навигация по электронной библиотеке обеспечивается электронными каталогами, позволяющими просматривать описание документов. Например, возможен просмотр всех выпусков некоторого журнала за заданный год, или просмотр всех публикаций некоторого автора в заданном издательстве.

Навигация по тематическому каталогу позволяет просматривать описания документов, относящихся к заданной теме. Тематический каталог может также использоваться при поиске в электронной библиотеке для фильтрации результатов поиска по заданной теме.

Основу тематического каталога электронной библиотеки могут составлять стандартные библиотечные классификаторы УДК, ББК. Возможно использование нескольких тематических классификаторов. Автоматическая каталогизация документов заключается в автоматическом определении принадлежности нового документа рубрикам классификаторов тематического каталога. Методы автоматической каталогизации распределяют документы по рубрикам классификаторов каталога на основе содержания документов.

Средства уточнения запроса позволяют выявлять действительные потребности пользователя при поиске в ЭБ, предлагая ему расширение первоначального запроса связанными словами, или перевод терминов запроса на другой поддерживаемый язык. Автоматическое формирование и пополнение для каждого из пользователей ЭБ собственной «виртуальной библиотеки» может основываться на анализе его предпочтений. Предпочтения пользователя определяются исходя из истории работы пользователя с ЭБ. Накопление такой истории формирует информационный портрет пользователя, позволяющий определить круг его научных интересов и потребностей, что может быть использовано для адаптации всех параметров поиска под индивидуальные особенности пользователя. Это позволит пользователю быстрее найти нужный ответ на поисковый запрос.

Информационный портрет пользователя позволит при поиске автоматически подобрать соответствующие его специализации коллекции и подборки документов, предложить наиболее правильный путь уточнения запроса.

5. Экспериментальная оценка поисковых алгоритмов системы EXACTUS

Экспериментальная оценка проводилась в рамках российского семинара по оценке методов информационного поиска в 2007 году [6]. Участникам семинара раздавались коллекции документов и запросов (несколько миллионов документов и несколько тысяч запросов). Коллекции документов индексировались, после чего по ним в автоматическом режиме прогонялись запросы. Результаты поиска помещались в файл, который затем обрабатывали независимые эксперты-оценщики, определяя степень релевантности запросов и документов. Методология оценки основывалась на следующих принципах:

1. Эксперт оценивает соответствие документов исходному запросу на основе расширенного описания информационной потребности (к каждому запросу прилагается краткое описание того, что должно быть по нему найдено).

2. Используется метод оценки типа "общего котла" (pooling) с глубиной пула 50 [6].

3. Используются следующие шкалы оценки релевантности:

- точно релевантно;
- возможно релевантно;
- вероятно релевантно;
- не релевантно;
- невозможно оценить.

4. Результат считается релевантным, если он получил оценку по одному из первых двух пунктов шкалы.

5. Для выставления оценки результата используются два способа:

- Строгая оценка AND – документ получает оценку релевантен или нерелевантен, если все оценщики выставили соответствующую оценку.
- Нестрогая оценка OR - результат получает оценку релевантен, если хотя бы один оценщик выставил соответствующую оценку.

Для оценки используются метрики точности и полноты, а также 11-точечный график TREC, который отображает совмещенные показатели точности и полноты при разных показателях точности [6].

На Рис. 2 и Рис. 3 приводятся 11-точечные графики TREC для оценок AND и OR для системы Exactus и других участников соревнования.

Как видно из графиков, алгоритм Exactus показал результаты, которые не хуже, а в некоторых точках лучше результатов остальных участников соревнования. Следует отметить, что остальными участниками являются такие известные поисковые машины, как Mail.Ru, KM.RU, поисковая машина УИС РОССИЯ, свободно распространяемая поисковая машина

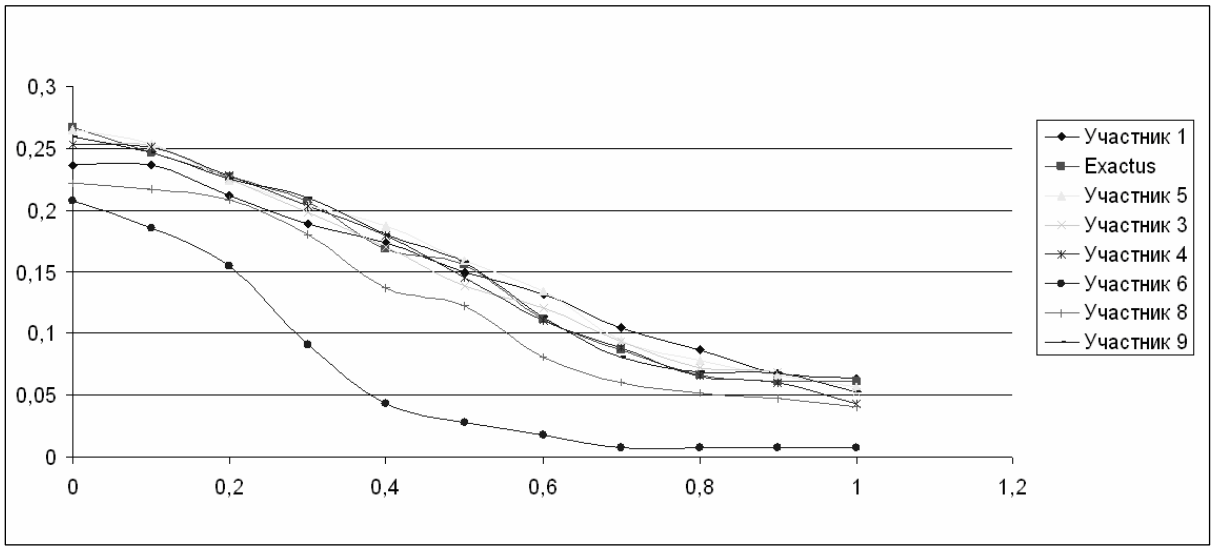

Рис. 2. График TREC: AND-оценка

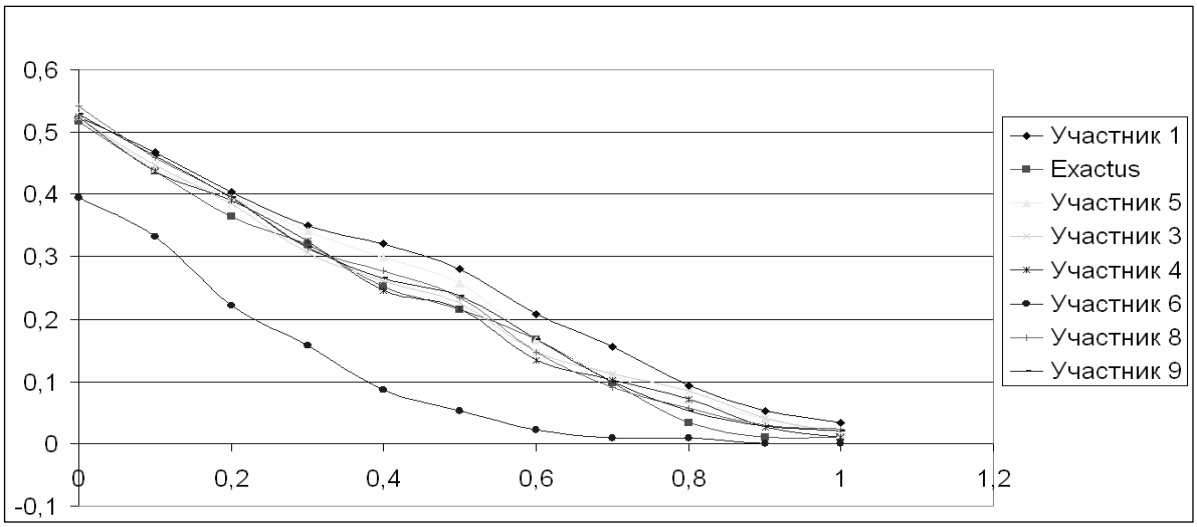

Рис. 3. График TREC: OR-оценка

Lucene и другие. Exactus показал лучшие результаты на строгой оценке AND, при высокой точности, показав высокие показатели полноты. Строгая оценка AND отражает мнения всех экспертов, т.е. более близка к независимой, объективной оценке.

Некоторое отставание от лидеров в оценке OR можно объяснить тем, что при поиске выполнялись алгоритмы семантического поиска, увеличивающие приоритет смысловой составляющей текста по сравнению с лексической. Поэтому многие эксперты могли счесть результаты не соответствующими *словам* запроса, т.е. недостаточно релевантными.

Заключение

Применение реляционно-ситуационной модели текста, дополненной статистическими методами поиска информации, позволяет повысить качество поиска текстовых документов. Повышение точности поиска, демонстрируемое системой EXACTUS особенно заметно на запросах достаточно большой длины, в то время как другие машины демонстрируют ухудшение результатов при увеличении длины запроса. Однако, именно запросы достаточно большой длины, написанные на естественном языке, и являются тем средством, которое позволяет точно сфокусировать поисковые потребности пользователя и, взаимодействуя с механизмами, реализованными в системе, существенно сократить время на получение действительно полезной информации.

Литература

1. Золотова Г.А., Онипенко Н. К., Сидорова М. Ю. Коммуникативная грамматика русского языка. Институт русского языка РАН им. В. В. Виноградова, М. 2004 – 544 с.
2. Золотова Г.А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. – М.: Наука, 1988 – 440 с.
3. Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука, Физматлит, 1997.
4. Осипов Г.С., Завьялова О.С., Климовский А.А., Кузнецов И.А., Смирнов И.В., Тихомиров И.А. Проблемы обеспечения точности и полноты поиска: Пути решения в интеллектуальной метапоисковой системе "Сириус". //Труды международной конференции Диалог'2005, Москва, Наука, 2005, с. 390-395.
5. Осипов Г.С., Тихомиров И.А., Смирнов И.В. Exactus – система интеллектуального метапоиска в сети Интернет. // Труды десятой национальной конференции по искусственному интеллекту с международным участием КИИ-2006. М: Физматлит, 2006. т. 3. – С. 859-866.
6. Russian Information Retrieval Evaluation Seminar // http://romip.ru.

Осипов Геннадий Семенович. Заместитель директора по научной работе Института системного анализа РАН, окончил в 1972 г. Ростовский государственный университет, д.ф.- м.н., профессор. Имеет 120 печатных работ, 5 монографий. Область научных интересов: представление знаний, приобретение знаний интеллектуальными системами, динамические интеллектуальные системы, семантический поиск.

Смирнов Иван Валентинович. Инженер-исследователь лаборатории интеллектуальных динамических систем Института системного анализа РАН. Окончил Российский университет дружбы народов в 2003 году. Автор 14 научных работ. Область научных интересов: искусственный интеллект, компьютерная лингвистика.

Тихомиров Илья Александрович. Научный сотрудник лаборатории интеллектуальных динамических систем Института системного анализа РАН. Окончил Рыбинскую государственную авиационную технологическую академию в 2002 году. Кандидат технических наук (2006). Автор 17 научных работ. Область научных интересов: искусственный интеллект, компьютерная лингвистика.