*На правах рукописи*

**БАБИНА Ольга Ивановна**

# ПОСТРОЕНИЕ МОДЕЛИ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ТЕХНИЧЕСКИХ ТЕКСТОВ

**Специальность 10.02.21 –** Прикладная
и математическая лингвистика

**Автореферат**
диссертации на соискание ученой степени
кандидата филологических наук

**Тюмень 2006**

Работа выполнена на кафедре лингвистики и межкультурной коммуникации Южно-Уральского государственного университета.

| | |
|---|---|
| **Научный руководитель** | доктор филологических наук,<br>профессор<br>Шереметьева Светлана Олеговна |
| **Официальные оппоненты** | доктор филологических наук,<br>профессор<br>Мышкина Нелли Леонидовна |
| | кандидат технических наук,<br>доцент<br>Поляков Владимир Николаевич |
| **Ведущая организация** | ГОУ ВПО Российский государственный педагогический университет им. А.И. Герцена |

Защита состоится 28 октября 2006 года в ___ часов на заседании диссертационного совета К 212.274.05 по защите диссертаций на соискание ученой степени кандидата филологических наук при Тюменском государственном университете по адресу: 625000, г. Тюмень, ул. Семакова, 10, корпус 1, ауд. 211.

С диссертацией можно ознакомиться в читальном зале библиотеки Тюменского государственного университета по адресу: 625000, г. Тюмень, ул. Семакова, 10, корпус 1.

Автореферат разослан 22 сентября 2006 года.

Ученый секретарь
диссертационного совета
кандидат филологических наук,
доцент Т.В. Сотникова

Развитие наук сегодня идет все увеличивающимися темпами, что ведет к стремительному росту научно-технической информации, представленной в текстовых документах. Это относится, в частности, к патентным документам, так как новые технические решения и изобретения регистрируются посредством патентования.

При получении заявки на патентование изобретения перед патентными ведомствами встает задача определить новизну предлагаемого заявителем технического решения. Для этого необходимо проанализировать весь объем существующей на текущий момент патентной документации, определив, не подпадает ли новое изобретение под один из действующих патентов. В современных условиях огромных объемов этой документации такая проверка вручную становится практически неподъемной задачей. В связи с этим появляется необходимость автоматизации процесса поиска патентов, порочащих новизну изобретений.

Однако современные средства автоматического отбора информации из массивов текстов большей частью ориентированы на использование некоторого искусственного языка, представляющего собой упрощение естественного языка (ЕЯ). Семантическая сила искусственных языков в значительной степени уступает естественным. Упрощения, как правило, включают в рассмотрение далеко не все уровни ЕЯ. Синтаксический и семантический уровни часто остаются за пределами таких моделей ЕЯ. Это негативно влияет на показатели точности и полноты поиска патентной информации в информационных массивах патентных текстов.

Таким образом, с одной стороны, огромные массивы существующей информации задают необходимость автоматизировать процессы отбора информации из естественно-языковых текстов. С другой стороны, на сегодняшний момент не существует моделей, которые бы в достаточной степени учитывали особенности ЕЯ при отборе необходимой информации.

В связи с этим, **актуальность** данного исследования обусловлена необходимостью совершенствования автоматизированных средств отбора релевантной информации из информационных массивов текстов на ЕЯ (в частности, патентных документов). Такая необходимость проистекает из того, что в существующих сегодня системах используются в недостаточной степени разработанные модели извлечения информации, слабо учитывающие особенности ЕЯ.

Недостаточная степень разработки систем извлечения информации, основанных на автоматической обработке естественно-языкового текста, объясняется трудностями, возникающими при описании сложной системы ЕЯ, что обусловлено его природой. Природа естественного языка, отличительной особенностью которого является нечеткость[1], принципиально отлична от искусственных языков, которые могут быть представлены посредством исчисления. Сознание человека способно воспринимать нечеткие суждения и из контекста делать выводы о значениях, актуализованных в высказываниях на естественном языке. Машина может воспринимать только то, что эксплицитно задано в описании модели автоматиче-

[1] Заде, Л. Понятие лингвистической переменной и его применение к принятию приближенных решений / Л. Заде; пер с англ. – М.: Мир, 1976. – 165 с.; Налимов, В.В. Вероятностная модель языка / В.В. Налимов. – 2 изд. – М.: Наука, 1979. – 303 с.

ской обработки текста, используемой в системе. Причем многозначность языковых единиц значительно снижает качество работы систем автоматической обработки текстов, так как ставит проблему выбора из множества альтернатив, что не доступно «пониманию» машины.

В связи с этим повышение качества систем отбора информации возможно, с одной стороны, посредством ограничения обрабатываемой в системе информации до подъязыка конкретной предметной области[2] (ПО), что позволяет за счет сужения контекста максимально сократить число возможных актуализаций значений в конкретных высказываниях. С другой стороны, посредством максимально полного и эксплицитного представления знаний[3] об особенностях выбранного подъязыка в модели извлечения информации из поискового массива текстов, что дает возможность в оставшихся случаях неоднозначности принимать по возможности верные решения.

В своем исследовании мы ограничиваем предметную область патентами на способ в фармакологии. Наиболее важной частью патентного документа является формула изобретения. Именно она имеет «решающее значение для оценки органов, осуществляющих государственную научно-техническую экспертизу изобретений, новизны и существенных отличий, а также положительного эффекта заявляемого объекта»[4]. Поэтому целесообразным представляется производить поиск релевантных для экспертизы патентных документов на основании анализа именно текста формулы изобретения. Руководствуясь вышеизложенными соображениями, мы определили объект и предмет исследования.

**Объектом** исследования является семантико-синтаксическая структура формулы изобретения на способ (ФИС) патентов по фармакологии.

**Предметом** исследования является разработка процедуры автоматизированного отбора релевантной информации из информационного массива ограниченной ПО, использующей предикатно-аргументную конструкцию в качестве единицы поиска.

**Материалом** для исследования послужил корпус текстов, включающий ФИС 295 патентов США по фармакологии объемом около 210 тыс. словоупотреблений.

**Целью** нашего исследования является построение модели отбора информации из патентных текстов в узкой ПО, использующей модуль автоматической обработки текста на естественном языке для максимально полного представления знаний.

Для достижения поставленной цели реализуется ряд задач теоретического и практического характера:

---

[2] Городецкий, Б.Ю. Методы семантического исследования ограниченного подъязыка / Б.Ю. Городецкий, В.В. Раскин. – М.: Изд-во Моск. ун-та, 1971. – 414 с.; Kittredge, K. Synthesizing Whether Forecasts from Formatted data / K. Kittredge, A. Polguere, E. Goldberg // *Proceedings of the 11th International Conference on Computational Linguistics (COLING-86)*. Bonn, Germany. 1986. Pp. 563-565.

[3] Мельников, Г.П. Системология и языковые аспекты кибернетики / Г.П. Мельников; под ред. Ю.Г. Косарева. – М.: Сов. радио, 1978. – 368 с.

[4] Изобретателям и рационализаторам: Сб. офиц. материалов / сост. В.И. Божинский. – М.: Профиздат, 1980. – 256 с.

1. Исследовать существующие подходы к построению информационно-поисковых систем (ИПС) и границы использования модулей автоматической обработки текста в этих системах;

2. Определить лингвистические особенности организации ФИС;

3. Провести сравнительный анализ отличий лексики и грамматики в формулах изобретения патентов на устройство (ФИУ) и на способ;

4. Модифицировать процедуру анализа текста ФИУ, настроив на обработку текстов ФИС;

5. Расширить процедуру автоматического анализа текста для решения задач индексирования патентных документов посредством представления семантико-синтаксической структуры ФИС;

6. Определить возможности переиспользования интерфейса системы автоматического синтеза формулы изобретения для определения запроса к системе автоматического поиска патентных текстов;

7. Разработать критерии оценки сходства образа запроса и документа для текстов формул изобретения патентов на способ в фармакологии.

Основным **методом исследования** является метод моделирования[5], посредством которого определяется структура подъязыка и на этой основе осуществляется построение процедуры отбора документов из информационного массива патентных текстов. Моделирование процесса извлечения релевантной информации строится на базе использования также следующих вспомогательных методов исследования:

- метод сплошной выборки при отборе документов, составивших корпус текстов;
- валентный анализ языкового материала;
- классификационно-типологический подход при анализе языкового материала;
- метод компонентного анализа лексики;
- метод статистического анализа для определения особенностей функционирования лингвистических единиц в тексте ФИС;
- метод дистрибутивно-статистического анализа при настройке процедуры автоматического анализа текста;
- метод экспериментальной проверки модели, воспроизводящей процедуру отбора релевантных текстов документного массива;
- аппарат теории множеств, математической логики, а также теории представления знаний и теории алгоритмов при описании основных положений модели отбора информации.

Диссертационное исследование опирается на работы по изучению семантики предикатов (А.А. Уфимцева, Е.В. Падучева, У. Чейф, Б. Левин), семантических и синтаксических отношений в предикатной структуре предложения (Ч. Филмор,

[5] Лосев, А.Ф. Введение в общую теорию языковых моделей: Уч. пособие / А.Ф. Лосев; под ред. И.А. Василенко. – М.: Изд-во Моск. гос. пед. инст., 1968. – 296 с.; Степанов, Ю.С. Методы и принципы современной лингвистики / Ю.С. Степанов. – 2-е изд. – М.: Эдиториал УРСС, 2001. – 312 с.; Ревзин, И.И. Современная структурная лингвистика: Проблемы и методы / И.И. Ревзин; отв. ред. Вяч. Вс. Иванов. – М.: Изд-во «Наука», 1977. – 263 с.

Л. Теньер, М. Минский, И.М. Богуславский, И.М. Мельчук, Н.Н. Леонтьева и др.), семиотические исследования природы и структуры языка (Р.Г. Пиотровский, В.В. Налимов, Л. Заде), а также на работы отечественных и зарубежных ученых по созданию прикладных систем автоматической обработки текста (С.О. Шереметьева, Е.А. Шингарева, К. Киттридж, А. Джоши и др.).

**Научная новизна** работы определяется тем, что данный языковой материал впервые исследуется с применением указанной совокупности современных лингвистических методов, что определяет новизну полученных результатов. Существенной новизной отличается разработанный метод отбора информации, основанный на использовании предикатно-аргументной структуры текста формулы изобретения в качестве единицы поиска при сопоставлении образов документа и запроса. Впервые разработаны формальные правила сопоставления патентного документа и запроса, использующие лингвистические особенности структуры формулы изобретения.

**Теоретическая значимость** исследования заключается в формальном описании одной из обособленных языковых подсистем (подъязыка ФИС), а также в моделировании системы отбора информации на основе использования в качестве образа документов в информационном массиве результата применения к ФИС процедуры автоматического лингвистического анализа текста. Полученные результаты вносят определенный вклад в разработку общей таксономии подъязыков науки и техники. Предложенный способ отбора информации дает основания расширить теорию информационного поиска, включив в область ее рассмотрения модели, использующие в качестве единицы поиска не только номинативные элементы, но и ситуативные (предикативные) единицы.

**Практическая значимость** исследования заключается в возможности создания на базе разработанных правил системы автоматического отбора информации из массива патентных документов, с помощью которой решается задача автоматизации патентной экспертизы в ходе рассмотрения заявки на вновь патентуемые объекты. Тем самым облегчается труд и значительно уменьшаются затраты времени работников патентных ведомств. Результаты исследования подъязыка ФИС могут быть использованы также при разработке других приложений автоматической обработки текста: систем автоматического перевода, аннотирования и реферирования текстов, а также при чтении курсов по прикладной лингвистике. Описанная модель в дальнейшем может быть модифицирована для автоматизации не только этапа поиска, но и всей процедуры патентной экспертизы.

**Положения, выносимые на защиту**:

1. Использование лингвистической базы знаний, определяемой предложенной методологией извлечения информации, обеспечивает более полное и глубокое представление поисковых образов документа и запроса, учитывающее семантические отношения между участниками описываемых в текстах ситуаций;

2. Использование унифицированной формы для представления поискового образа полнотекстового документа и запроса с помощью набора фреймоподобных предикатно-аргументных структур расширяет возможность сравнивать образы на семантическом уровне;

3. Разработанные правила и предложенные метрики для сличения образов запроса и документа позволяют проранжировать результаты в зависимости от степени релевантности запросу отобранных документов.

4. Переиспользование некоторых алгоритмов и правил автоматического анализа текста, настроенных для использования в другой предметной области, повышает эффективность разработки новых приложений на новом материале, уменьшая затраты труда и времени.

**Апробация материалов исследования**. По теме диссертации были сделаны доклады на Международной научно-практической конференции «Теория и методика преподавания языков в вузе» (Челябинск, 15-17 декабря 2003 г) и на Второй международной конференции по модели «Смысл ⇔ Текст» (Москва, 23-25 июня 2005 г). Отдельные этапы исследования обсуждались на научных семинарах кафедры лингвистики и межкультурной коммуникации Южно-Уральского государственного университета. По теме диссертационной работы опубликовано 7 работ общим объемом 2 п.л.

**Объем и структура исследования**. Структура работы соответствует целям и задачам исследования. Работа состоит из введения, трех глав, заключения, списка литературы, включающем наименования на русском, английском, французском и немецком языках, и 9 приложений. Общий объем диссертационной работы составляет 235 страниц печатного текста.

Во **введении** обосновывается актуальность темы исследования, научная новизна, теоретическая и практическая значимость работы, определяется объект и предмет исследования, его основная гипотеза, формулируется цель, задачи и выносимые на защиту положения, дается описание материала и методики исследования. Введение также содержит данные об апробации результатов, структуре и объеме диссертационной работы.

В **первой главе** «Модели и средства извлечения информации» рассматриваются информационно-поисковые системы и модели извлечения информации из массива текстов. Особое внимание уделяется лингвистическому компоненту, являющимся ключевым при отборе текстов на естественном языке. Рассматривается место поиска при проведении патентных исследований.

Во **второй главе** «Подъязык формул изобретения патентов на способ в фармакологии» представляется результат лингвистического анализа языкового материала подъязыка ограниченной предметной области. Акцент делается на синтактико-семантической структуре исследуемых текстов, в частности на особенностях предикатно-аргументной структуры текстов формул изобретения патентов на способ. Детально исследуется семантика предикатов подъязыка ФИС.

В **третьей главе** «Модель извлечения информации из поискового массива формул изобретения патентов на способ» описывается модель извлечения информации из корпуса текстов формул изобретения. Описывается методика переиспользования и применения процедур автоматической обработки текста для представления поисковых образов патентных документов в информационном массиве. Показывается способ формирования образа запроса на основе использования интерфейса системы формального синтеза ФИС. Определяются принципы и правила

сопоставления образов запроса и патентных документов с целью отбора релевантных текстов. Приводится пример применения описанной модели для извлечения из патентной базы патентных документов.

В **заключении** подводятся общие итоги работы, намечаются направления для дальнейших исследований, обозначаются перспективы для применения и совершенствования описанных в работе правил и процедур.

## Основное содержание работы

В **первой главе** дается представление о средствах, используемых для отбора информации из массива текстов, рассматривается лингвистическая составляющая процедур извлечения информации из поисковых массивов, а также определяется роль средств автоматического извлечения информации при проведении патентного поиска.

В качестве современных средств отбора информации выступают информационно-поисковые системы, каждая из которых использует индивидуальную модель извлечения информации из поискового массива. Основными составляющими, определяющими существо модели извлечения информации, являются: 1) структура информационного массива; 2) лингвистический компонент, лежащий в основе процедуры отбора информации; 3) правила и процедуры, с помощью которых осуществляется отбор информации непосредственно.

По способу организации информационного массива среди поисковых систем выделяют: 1) документальные; 2) фактографические; 3) документально-фактографические (смешанные). В документальных массивах информация представляется в форме текстов, каждый из которых представляет собой единицу информации. В фактографических системах в качестве единицы информации выступает факт/событие/ситуация с описанием значений его основных признаков/участников. В смешанных системах описание каждого факта соотносится с документами, в которых имеется информация о нем. Для нашего исследования интерес представляют документальные ИПС.

В подавляющем большинстве случаев информационный массив в современных документальных поисковых системах представляет собой набор текстов на естественном языке. Это обусловлено тем, что: 1) естественный язык обладает наибольшей семантической силой, и поэтому является наиболее «эффективным» (с точки зрения человека) средством представления информации в терминах смыслоразличения и смыслоотождествления; 2) естественный язык является наиболее типичным средством экспликации смыслов и передачи информации, что является причиной того, что знания в современном мире, чаще всего, представляются в форме текстов на естественном языке, в частности, письменных текстов. Представление информации в каком-либо другом формате требует дополнительных усилий по преобразованию информационных сообщений на естественном языке в сообщения, представленные через знаковую систему иного рода. Видимо, подобные преобразования должны проводиться вручную, что является практиче-

ски неподъемной задачей для человека и даже для группы людей в условиях высокого роста информации, в частности, научно-технической информации.

Способ представления информации на естественном языке в поисковых массивах предопределяет ключевую роль лингвистического компонента в моделях извлечения информации. Лингвистический компонент системы поиска включает: 1) информационно-поисковый язык (ИПЯ), являющийся, как правило, ограничением естественного языка; степень и виды ограничения в ИПЯ определяются процедурами отбора информации, применяемыми в данной ИПС; 2) словарная база, включающая используемые в процедурах отбора информации лексиконы, тезаурусы, онтологии.

Среди ИПЯ выделяют предкоординируемые и посткоординируемые ИПЯ. Предкоординируемые ИПЯ строятся, как правило, в форме иерархии, а поиск с их использованием включает продвижение по ветвям иерархии с последовательным сужением области релевантных для поиска единиц информации. Язык, представляющий собой иерархию терминов, является закрытой системой с жесткой структурой. Посткоординируемые ИПЯ более свободны по своей структуре. Они обладают вокабуляром и грамматикой. Термины посткоординируемых языков связаны между собой парадигматическими и синтагматическими отношениями. Степень семантической силы таких языков тем выше, чем более полно вокабуляр и грамматика ИПЯ соответствует естественному языку.

Вокабуляр ИПЯ инвентаризуется в автоматических словарях (лексиконах). Лексиконы могут составляться автоматически или вручную. В последнем случае в лексиконе часто представляется информация о морфологических характеристиках каждого вхождения.

Для отражения парадигматических отношений между лексическими единицами ИПЯ используются информационно-поисковые тезаурусы. В тезаурусе показываются отношения меронимии, синонимии, антонимии и т.д. между терминами ИПЯ. Таким образом, в тезаурусе отражается не только языковая информация о лексической единице, но также ее место в структуре терминов предметной области, представленной в тезаурусе.

Структурно аналогичны тезаурусам онтологии, но в качестве вхождения в последних используются не термины ИПЯ, а понятия (концепты), которые связаны между собой парадигматическими отношениями. Каждому концепту может соответствовать список терминов ИПЯ.

Лингвистический компонент составляет основу моделей поиска релевантной информации в поисковом массиве. Исторически первыми и наиболее распространенными в настоящее время являются статистические модели поиска, среди которых выделяют: 1) теоретико-множественные; 2) векторные; 3) вероятностные модели. Фундамент статистических моделей составляет лексический состав ИПЯ. Поиск осуществляется по фразам, состоящим из ключевых слов (терминов ИПЯ), связанных набором допустимых в ИПЯ операторов: морфологических, логических, операторов фрагментирования поискового образа документа, дополнительных контекстных операторов, операторов поиска по числовым параметрам. Степень релевантности документа определяется по соответствию фразы запроса до-

кументу на основании: 1) наличия/отсутствия указанных в запросе ключевых слов в документе; 2) значений векторных коэффициентов, определяющих степень сходства векторов, репрезентирующих запрос и документ; 3) значений вероятностных коэффициентов, учитывающих степень важности каждого ключевого слова, указанного в запросе, для характеристики данного документа.

Другой класс моделей поиска включает лингвистические модели, в которых предпринимается попытка учесть при отборе релевантной информации особенности вокабуляра, а также синтаксическую и семантическую сторону естественного языка. Соответственно можно выделить: 1) синтаксические; 2) семантические модели поиска.

В синтаксических моделях в качестве единиц поиска рассматриваются словосочетания (чаще именные группы) или клаузы. Посредством лингвистического процессора осуществляется полный или частичный синтаксический анализ текста запроса и документа. Отбор релевантной информации осуществляется в результате сопоставления деревьев/сетей, полученных в результате синтаксического разбора предложений/словосочетаний в запросе и документе. Это дает возможность увеличить точность поиска, так как сходство выявляется не только на уровне лексических единиц, но и на уровне синтагматических отношений между ними.

В семантических моделях предпринимается попытка учесть лексико-семантические варианты слов с целью улучшения показателей полноты поиска. Ключевую роль в таких моделях играют тезаурусы и онтологии. Учет лексико-семантических вариантов посредством словарных средств осуществляется двумя способами: 1) расширение запроса терминами, связанными определенными семантическими отношениями с терминами запроса; 2) избыточное индексирование документов.

Поиск патентной документации является одним из основных этапов патентных исследований. Автоматизация процесса поиска посредством использования ИПС, осуществляющих поиск на патентных базах данных, является неизбежной необходимостью в свете неумолимого роста объемов научно-технической информации, в частности, патентной.

Основным разделом патента является формула изобретения, обладающая технической, экономической и юридической силой. Патентная формула представляет собой специфичный текст на естественном языке, в котором описывается изобретение с его существенными признаками[6]. Например, один из пунктов формулы изобретения патента US 6,485,910 имеет вид:

*A method for using a cDNA to detect the differential expression of a nucleic acid in a sample comprising:*

*a) hybridizing the probe of claim 4 to the nucleic acids, thereby forming hybridization complexes; and*

*b) comparing hybridization complex formation with a standard, wherein the comparison indicates the differential expression of the cDNA in the sample.*

[6] Киселева, Т.С. Экспертиза объектов техники на патентную чистоту: Уч. пособие / Т.С. Киселева. – М.: ВНИИПИ, 1991. – 116 с.; Фейгельсон, В.М. Методика и практика экспертизы объектов техники на патентную чистоту / В.М. Фейгельсон. – М.: ИНИЦ Роспатента, 2001. – 343 с.

В нашем исследовании именно этой части патента уделяется внимание при моделировании извлечения информации из массива патентных текстов, так как она является основной для проведения патентной экспертизы. Причем при поиске представляется наиболее целесообразным использование лингвистических методов отбора информации с целью более «тонкого» учета семантики текста патентной формулы. В нашей работе это подразумевает использование лингвистического процессора для представления образов запроса и документа и разработку процедур сопоставления составленных таким путем образов и принятия решения о релевантности документов.

Для определения эффективных правил и процедур представления и извлечения информации из массива естественно-языковых текстов патентных формул необходимо рассмотреть структуру и функционирование подъязыка формул изобретения. В связи с этим во **второй главе** представляется описание подъязыка формул изобретения. В исследовании мы вводим следующие ограничения на материал: 1) рассматривается только ПО Фармакология; 2) в качестве объекта изобретения рассматриваются способы.

Описание подъязыка проводится на основе анализа корпуса текстов патентных формул на способ в фармакологии, включающего формулы 295 патентов общим объемом около 210 тыс. словоупотреблений.

Лексический состав подъязыка ФИС по фармакологии можно условно разделить на три группы: 1) предикаты; 2) знаменательная лексика аргументов предикатов; 3) служебные слова.

Под *предикатом* понимается элемент пропозиции, который обозначает ситуацию, имеющую некоторое число обязательных участников, выполняющих определенные роли[7].

Предикаты в ФИС выражены: 1) глаголами; 2) существительными; 3) прилагательными. Наиболее представительна группа глаголов. В подъязыке патентных формул предикаты представлены ограниченной семантикой и морфологией по сравнению с общеупотребительным языком.

Морфология предикатов-существительных и предикатов-прилагательных представлена одной формой: единственное число существительных и положительная степень прилагательных соответственно. Эти формы мы рассматриваем как начальные для соответствующих предикатов. Предикаты-глаголы более разнообразны в морфологическом отношении, хотя их морфология значительно беднее, чем в общеупотребительном языке. В подавляющем большинстве случаев используются формы глаголов в изъявительном наклонении, причем 99,79% словоупотреблений предикатов приходится на следующие формы:

1) Present Simple Active (*represent*);

2) Present Participle Simple Active (*suffering*);

---

[7] См. Теньер, Л. Основы структурного синтаксиса / Л. Теньер; пер. с франц. И.М. Богуславского, Л.И. Лухт, Б.П. Нарумова, С.Л. Сахно. – М.: Прогресс, 1988. – 653 с.; Филмор, Ч. Дело о падеже / Ч. Филмор // Зарубежная лингвистика. III / общ. ред. В.Ю. Розенцвейга, В.А. Звегинцева, Б.Ю. Городецкого. – М.: Изд. группа «Прогресс», 2002. – С. 127-258; Мельчук, И.А. Опыт теории лингвистических моделей «Смысл⇔Текст» / И.А. Мельчук. – М.: Школа «Языки русской культуры», 1999. –XXII, 345 с.; Богуславский, И.М. Исследования по синтаксической семантике: сферы действия логических слов. – М.: Наука, 1985. – 176 с.

3) Present Simple Passive (*is selected*);
4) Past Participle (*connected*);
5) Gerund Simple Active (*obtaining*);
6) Infinitive Simple Active (*to inhibit*);
7) Present Participle Simple Passive (*being created*);
8) Infinitive Simple Passive (*to be treated*).

Начальной формой предикатов-глаголов мы считаем формы причастия (Present Participle Simple Active, Past Participle), рассматривая в дальнейшем один и тот же глагол в форме пассивного и активного причастий как два различных предиката. Производные от причастий формы образуют морфологическую парадигму соответствующего предиката.

В семантической структуре предикатов выделены семантические отношения (валентности), инвентарь которых для рассмотренного корпуса включает:

1. Субъект (S, subject)
2. Объект (O, dir-obj)
3. Косвенный объект (IO, indir-obj)
4. Место (Pl, place)
5. Время (T, time)
6. Образ действия (M, manner)
7. Средство (Ms, means)
8. Цель (Pr, purp)
9. Результат (R, result)
10. Условие (Cond, cond)
11. Количество (Qu, quantity)
12. Эталон (E, equal)
13. Источник (Sr, source)
14. Конечная точка (D, destination)

Каждый предикат содержит в своей логической структуре одну или более валентностей из представленного инвентаря. Для многих из них синтаксические способы заполнения одноименных валентностей у различных предикатов в большинстве случаев совпадают.

На основании анализа логической структуры предикатов, а также лексической семантики предикатных слов, предикаты были разбиты на семантические классы:

1. Меронимические отношения (*having, including*);
2. Соединение между объектами (*adjacent, linked*);
3. Структурные особенности (*isolated, covered*);
4. Причинные отношения (*associated, causing*);
5. Целевые отношения (*resulting, giving rise*);
6. Перемещение (*collecting, circulating*);
7. Свойства (*effective, sensitive*);
8. Сравнение (*relative to, comparing*);
9. Изменение состояния (*treating, reduced*);
10. Динамическое взаимодействие (*associating, combined*);

11. Получение/появление нового объекта (*obtaining, prepared*);
12. Выявление объектов или явлений (*detecting, identified*);
13. Воздействие одних объектов на другие (*affecting, inhibited*);
14. Другие (*rendering, elicited*).

Синонимичные предикатные слова в пределах одного семантического класса могут быть объединены в один класс условной эквивалентности.

Сравнив морфологические и семантические особенности функционирования предикатов в патентах на устройства (описанные ранее в работах С.О. Шереметьевой[8]) и на способы, мы отметили, что: 1) качественный состав преобладающих морфологических форм предикатов практически идентичен в патентах на различные объекты изобретения; 2) инвентарь валентностей и семантические классы предикатов в патентах на способы частично повторяет эти характеристики патентов на устройство.

Дальнейшее рассмотрение предикатной лексики ФИС сосредоточено на предикатах, характеризующихся отличными от предикатов ФИУ свойствами. К ним относятся предикаты семантических классов «Изменение состояния», «Динамическое взаимодействие», «Получение нового объекта», «Выявление», «Воздействие», которые были тщательно исследованы.

На основании детального изучения функционирования этих предикатов отмечены следующие особенности, отличающие формулы изобретения патентов на способы от других объектов изобретения:

1) Значительно более частотными в ФИС (по сравнению с формами предикатов в ФИУ) являются формы герундия глагола и функционального существительного, посредством которых обозначаются цель и компоненты способа.

2) Цель и компоненты способа в ФИС выражаются с помощью предикатных конструкций (а не предметных существительных, как в ФИУ), принадлежащих, в большинстве своем, тем семантическим классам, которые не встречаются в ФИУ;

3) Лексическая семантика некоторых предикатных слов в ФИС включает обозначение действий (что предполагает наличие исполнителя действия, не выраженного в патентной формуле), в то время как семантика предикатов в ФИУ ограничивается обозначением отношений;

4) На синтаксическом уровне для предикатов в ФИС, обозначающих цель и компоненты способа, валентность Субъект (логически соответствующая исполнителю действия) не заполняется.

Знаменательная лексика аргументов предикатов представлена, в основном, частями речи, входящими в состав именных групп, являющихся преобладающим синтаксическим способом заполнения валентностей предикатов. Большую ее часть составляют существительные. В предметной области патентов на способ в фармакологии последние подразделены на следующие семантические классы:

---

[8] Шереметьева, С.О. Межуровневая организация текста патентной формулы США / С.О. Шереметьева, Е.А. Бородкина // Межуровневая организация текста в естественном языке: Межвузовский сборник научных трудов. – Челябинск: ЧГПИ, 1987. – С. 116-121; Sheremetyeva, S. A Flexible Approach to Multi-Lingual Knowledge Acquisition for NLG. In *Proceedings of the 7th European Workshop on Natural Language Generation* / P. St. Dizier (ed.). Toulouse, France. May 13-15, 1999. Pp. 106-115.

1) Вещество (*ligand, oxide*);
2) Единица измерения (*milligram, mole*);
3) Заболевание (*Alzheimer's disease, disorder*);
4) Клетка (*cell, mammalian cell*);
5) Орган (*bone, gland*);
6) Организм (*mammal, mouse*);
7) Параметр (*condition, pH*);
8) Ткань (*tissue, myocardium*);
9) Физический объект (*catheter, element*);
10) Формула (*alkyl, benz, indol*);
11) Другие (*amount, fingerprint*).

Семантически лексический состав аргументов предикатов в ФИС отличается от ФИУ и в значительной степени обусловлен предметной областью. Синтаксическая структура аргументов, представленных именными, предложными, наречными, герундиальными и инфинитивными группами, для формул изобретения на различные объекты изобретения практически идентична.

Служебные слова включают артикли, предлоги, союзы, относительные местоимения. Функции служебных слов в патентных формулах соответствуют их функциям в общеупотребительном языке.

Организация текста пункта формулы изобретения подобна для различных объектов изобретения. Текст пункта формулы представляет собой назывное предложение, состоящее из: 1) описания названия изобретения (цели способа); 2) описания составляющих компонентов изобретения (действий для достижения цели способа).

Довольно значительное сходство лексического состава и организации патентных формул на различные объекты изобретения дает основание переиспользовать модули автоматической обработки текста, настроенные на работу с ФИУ, для решения различных задач прикладной лингвистики на материале патентных формул на способ.

В **третьей главе** приводится описание разработанной автором модели извлечения информации из поискового массива ФИС по фармакологии, существенной составляющей которой является применение дополненной и настроенной на обработку ФИС процедуры лингвистического анализа текстов, использующейся первоначально для автоматической обработки текстов ФИУ.

Модель извлечения информации состоит из следующих модулей:

I. Информационный (поисковый) массив документов: содержит патентные документы и их поисковые образы. Индексирование патентных документов осуществляется в результате переиспользования процедуры автоматического анализа текстов ФИУ для обработки текстов ФИС;

II. Модуль формирования поискового предписания: формирует поисковый образ запроса в ходе интерактивного опроса пользователя; в ходе опроса пользователь поэтапно задает параметры интересующего его описания изобретения;

III. Модуль выявления релевантных запросу документов: на основании предложенных метрик осуществляет сопоставление поискового предписания с

поисковыми образами документов информационного массива и определяет коэффициент их сходства;

IV.Модуль выдачи отобранной информации: на основании значений коэффициентов сходства ранжирует документы по уменьшению степени релевантности запросу и выдает список отсортированных таким образом патентных документов пользователю.

Лексикографический компонент модели извлечения информации включает: 1) лексикон, ориентированный на ограниченную предметную область; 2) тезаурус для сопоставления поисковых образов запроса (ПП) и документа (ПОД).

Лексикон, применяемый в нашей процедуре, аналогичен по структуре автоматическому словарю, используемому С.О. Шереметьевой для обработки текста ФИУ[9]. Одно вхождение соответствует лексеме. Информация о лексемах представляется в словаре на следующих уровнях:

1) лексико-семантический уровень:
   а) лемма (начальная форма);
   б) словоизменительная парадигма лексемы;
2) семантико-синтаксический уровень:
   а) семантический класс;
   б) модель управления предиката (только для предикатов);
   в) линейные формулы (только для предикатов).

В полях лексико-семантического уровня указывается морфологическая информация о лексеме. Словоформы лексемы задаются в пределах структуры парадигмы иконически. Каждой словоформе соответствует супертэг – метка, которая «сообщает о слове нечто большее, чем просто часть речи»[10]. Например, супертэг *~Pgcsg* обозначает, что словоформа является активным предикатом (Pg), принадлежит семантическому классу «Изменение состояния» (cs), имеет форму герундия (g).

На семантико-синтаксическом уровне значение поля «Семантический класс» соответствует одному из выделенных при анализе корпуса классов предикатов и знаменательных лексем. В поле «Модель управления» указывается набор валентностей из общего инвентарного списка, характерных для данного предиката, с синтаксическими способами их заполнения. В поле «Линейные формулы» показывается, в какой последовательности валентности предиката реализуются в тексте. Линейная формула представляет собой линейную цепочку обозначений валентностей предиката из его модели управления и символа X, соответствующего положению предикатного слова в цепочке.

Так, словарная статья для предиката *treating* в лексиконе имеет вид:

*Лемма*: treating

*Словоизменительная парадигма*:

Полная форма, причастие (~Pgcs): treating

---

[9] Sheremetyeva, S. Natural Language Analysis of Patent Claims. In *Proceedings of the Workshop on Patent Corpus Processing*. Sapporo, Japan. July 12, 2003. Pp. 66-73.

[10] Joshi, Aravind K., and B. Srinivas. Disambiguation of Super Parts of Speech (or SuperTags): Almost Parsing. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*. Vol. 1. Kyoto, Japan. August 1994. Pp. 154-160.

Полная форма, инфинитив (~Pgcsi): treat
Полная форма, герундий (~Pgcsg): treating
Краткая форма, ед. ч. (~Pgcss): treats
Краткая форма, мн. ч. (~Pgcsp): treat
Абсолютная форма (~Pgcsa): treating

*Семантический класс*: Изменение состояния

| *Модель управления*: | *Линейные формулы*: |
|---|---|
| O: (NP) | X-O |
| Ms: (("with" NP)) | X-O-Ms |
| Pr: (InfP) | X-O-Pr |
| T: ("over time") | X-O-T-Ms |

Тезаурус предметной области представляет собой лес иерархических деревьев, характеризующих исследуемую ПО Фармакология. В узлах деревьев располагаются понятия (концепты) ПО Фармакология. Они связаны между собой дугами, репрезентирующими меронимические отношения между ними. Концепт тезауруса объединяет термины предметной области, входящие в один класс условной эквивалентности. Каждому понятию тезауруса ставится в соответствие ноль или более терминов лексикона. Каждое вхождение лексикона соответствует одному понятию в тезаурусной иерархии.

I. Информационный массив в разработанной модели извлечения информации состоит из двух частей: 1) текстовая часть: патентные документы; 2) «индексная» часть: наборы предикатно-аргументных структур, каждый из которых соответствует одному пункту патентной формулы документа (поисковые образы).

Формирование «индексной» части осуществляется в результате автоматического анализа текстовой части массива с помощью процедуры индексирования, основу которой составляет методика автоматического анализа, разработанная С.О. Шереметьевой[11]. Лингвистическая база знаний анализа включает: 1) лексикон, содержащий информацию о лексических единицах исследуемой ПО, и 2) грамматический компонент, включающий продукционные правила анализа текста ФИС. Процедура индексирования состоит из следующих этапов:

1) *Первичное разбиение текста*: в тексте пункта патентной формулы выделяются содержательные блоки его структуры. Разбиение осуществляется по формальным признакам на основании сопоставления со следующим шаблоном (в угловых скобках – содержательные блоки; в круглых скобках – необязательные элементы):

*A method for* <Цель способа> *(, said method) comprising (the step(s) (of))(:)*
<Компонент способа 1>*;*
<Компонент способа 2>*;*
…
<Компонент способа N-1>*; and*
<Компонент способа N> (,
*wherein* <Информация об участниках способа>).

[11] Sheremetyeva, S. Natural Language Analysis of Patent Claims. In *Proceedings of the Workshop on Patent Corpus Processing*. Sapporo, Japan. July 12, 2003. Pp. 66-73.

Кроме того, на основании знаков пунктуации и табуляции осуществляется разбивка на более дробные образования, используемые на дальнейших этапах для выделения текстовых границ аргументов и предикатных конструкций.

2) *Лексико-грамматический анализ текста*: включает следующие подэтапы:

а) *Приписывание каждой словоформе в тексте списка всех возможных супертэгов*: для каждой текстовой словоформы осуществляется поиск по лексикону совпадающих с ней иконически заданных форм слов. Супертэги, относящиеся к найденным словоформам, включаются в список потенциально соответствующих рассматриваемой текстовой словоформе меток.

б) *Выбор для каждой словоформы одного соответствующего ей супертэга из списка*: грамматический компонент этой стадии состоит из набора контекстуальных продукционных правил разрешения морфологической неоднозначности. На основании применения правил для каждой словоформы осуществляется выбор единственного супертэга.

3) *Семантико-синтаксический анализ текста*: включает следующие подэтапы:

а) *Восходящий анализ синтаксических конструкций*: грамматический компонент состоит из пяти наборов продукционных правил. Целью каждого из них является объединение текстовых элементов в блоки, соответствующие определенному типу синтаксических конструкций (именные, предложные, наречные группы, инфинитивные, герундиальные обороты). В каждом наборе правил описываются образцы – линейные последовательности супертэгов, репрезентирующие соответствующую синтаксическую конструкцию. Применение правил представляет собой распознавание по образцу: в случае, если в тексте встречается, например, описанная в одном из правил для именных групп последовательность супертэгов, то соответствующая часть текста объединяется в блок и помечается как именная группа (ИГ).

б) *Восстановление кореференции именных групп*: грамматический компонент включает два набора продукционных правил: 1. Выявление в ИГ головного существительного; 2. Определение антецедента для текущей ИГ. Первая группа правил применяется для всех именных групп с целью выделить в каждой главное смысловое слово. Вторая группа применяется для ИГ, потенциально имеющих антецедент в тексте. Формально это определяется по наличию в ИГ инициирующего определителя *the* или *said*.

в) *Поиск семантических зависимостей*: грамматический компонент этого этапа включает следующие наборы правил: 1. Выявление предикатов во входном тексте; 2. Определение для каждого предиката относящихся к нему синтаксических конструкций и соотнесение последних с соответствующим аргументом предиката. В результате применения правил все элементы текста организуются в фреймовые предикатно-аргументные конструкции.

4) *Преобразование аргументов предиката*: процедуры этого этапа осуществляют разложение сложных аргументов на более простые составляющие, исключение из состава аргументов служебных слов.

5) *Определение состава компонентных зон*: Под *компонентной зоной* мы понимаем набор предикатно-аргументных конструкций, включающий: а) один предикат, обозначающий *Цель* или *Компонент способа* (*формирующий* предикатом зоны); б) предикаты, обозначающие *Отношения* между участниками способа, причем в обозначаемых ими ситуациях принимают участие объекты, упомянутые в слотах структуры, которая соответствует формирующему предикату зоны. Грамматический компонент этапа включает правила, в результате применения которых на основании расположения предиката в тексте и информации о кореференции ИГ устанавливается принадлежность конструкции одной или более компонентным зонам.

Отмеченное лингвистическое сходство формул изобретения на различные объекты изобретения позволяет переиспользовать для анализа ФИС значительную часть грамматических правил, разработанных для автоматического анализа ФИУ, внеся в них некоторые изменения и дополнения.

В итоге выполнения процедуры индексирования текст патентной формулы представляется с помощью набора фреймовых предикатных конструкций, имеющих следующий вид:

(3 P4 Pgw *having* **P2**
1 “receptors$_1$” //<subj>
2 “different selectivity” //<dir-obj> [or]
“specificity for ligand$_2$” //<dir-obj>
),

где <u>в заголовке фрейма</u> по порядку: 3 – данный предикат обозначает *Отношения* между участниками способа; P4 – уникальный номер данного предиката в ФИС; Pgw – супертэг данного предиката (активный предикат, семантический класс «Меронимия», форма причастия); *having* – форма данного предиката в тексте; P2 – данный предикат принадлежит компонентной зоне с формирующим предикатом, имеющим уникальный номер P2;

<u>под строкой заголовка фрейма</u> показан список слотов данного предиката, заполненных текстовыми выражениями (в кавычках), которые реализуют его аргументы (в угловых скобках указана валентность, которую заполняет соответствующий аргумент; нижним индексом помечены существительные, кореферентные другим существительным в пределах ФИС).

II. Модуль формирования поискового предписания представляет пользователю интерфейс, структура которого заимствована из системы автоматического синтеза патентной формулы AutoPat[12]. Следуя предлагаемым этапам, пользователь имеет возможность описать свою информационную потребность, определяя последовательно цель, компоненты и отношения между участниками интересующего его способа. Описание каждого информационного элемента осуществляется в терминах предикатных конструкций: пользователь задает действие (предикат), репрезентирующее цель или компонент способа, и заполняет конкретными языковыми выражениями предлагаемые интерфейсом слоты (соответствующие ва-

[12] Sheremetyeva, S., S. Nirenburg, and I. Nirenburg. Generating patent claims from interactive input. In *Proceedings of the 8th International Workshop on Natural Language Generation (INLG'96)*. Herstmonceux, England. 1996. Pp. 61-70.

лентностям предиката). Для выбора конкретных выражений при заполнении слотов пользователь может обратиться к лексикону. На завершающем этапе формирования поискового предписания пользователь вручную восстанавливает кореферентные связи между обозначенными им участниками и определяет релевантность для поиска последовательности выполнения шагов-компонентов способа. Результатом определения информационной потребности является набор предикатно-аргументных структур, идентичный по своему строению конструкциям из «индексной» части информационного массива.

III. Модуль выявления релевантных запросу документов осуществляет сравнение составленного поискового предписания с «индексной» частью информационного массива. Лингвистическая база знаний данного модуля включает: 1) тезаурус предметной области; 2) набор коэффициентов, с помощью которых измеряется степень релевантности документов запросу.

Сравнение поисковых образов запроса и документа осуществляется на трех уровнях: 1) уровне запроса/документа в целом; 2) уровне компонентной зоны; 3) уровне предикатной конструкции.

Последовательно «спускаясь» по этим уровням (от более общего к более детальному представлению) осуществляется отбор из поисковых образов документов кандидатов для соответствующей уровневой единицы запроса:

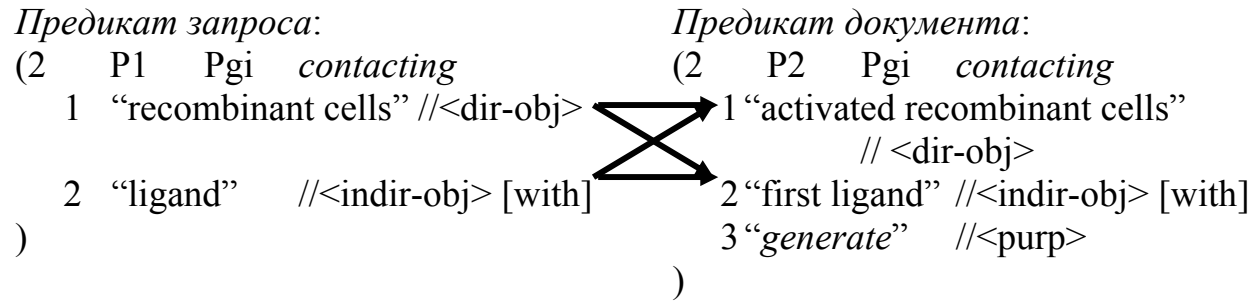
1) Для отбора кандидатов на уровне запроса/документа составляется два множества семантических классов формирующих предикатов компонентных зона: для запроса и для документа. Если пересечение этих множеств не пусто, документ признается потенциально релевантным запросу;

2) Для отбора кандидатов на уровне компонентной зоны в пределах документа-кандидата используются правила двух уровней: компонентная зона документа является кандидатом на соответствие компонентной зоне запроса, если формирующие предикаты двух зон принадлежат 1. одному классу условной эквивалентности; 2. одному семантическому классу. Переход к поиску кандидатов по правилу второго уровня происходит только в случае, если нет кандидатов, удовлетворяющих правилу первого уровня.

3) Для отбора кандидатов на уровне предикатно-аргументной конструкции в пределах компонентной зоны-кандидата используются следующие правила: а) отбор осуществляется только в пределах предикатов одной группы (формирующие – неформирующие); б) предикатная конструкция документа является кандидатом на соответствие предикатной конструкции запроса, если предикаты двух конструкций принадлежат: 1. одному классу условной эквивалентности; 2. одному семантическому классу. Степень сходства двух предикатных слов характеризуется коэффициентом Pred.

Определение релевантности документа запросу осуществляется посредством обратного «прохода» по уровням сравнения, в ходе которого для каждой единицы запроса на соответствующем уровне осуществляется сопоставление данной уровневой единицы с каждым из отобранных для нее кандидатов, и «наилучший» из кандидатов для каждой единицы ставится ей в соответствие.

Сопоставление предикатно-аргументных конструкций запроса и документа осуществляется по признаку лексического сходства предикатного слова и аргументов. Лексическое сходство предикатных слов определяется при отборе кандидатов. При определении лексического сходства аргументов рассматривается множество, представляющее собой всевозможные комбинации пар «аргумент предиката запроса–аргумент предиката документа», таких, что аргументы принадлежат одному и тому же типу. Среди аргументов можно выделить следующие типы: 1) адвербиальные группы; 2) предикатные конструкции; 3) именные группы. Схематично сопоставление аргументов двух предикатных конструкций можно представить как:




В примере первые и вторые аргументы конструкций запроса и документа принадлежат третьему типу аргументов и сопоставляются между собой. Третий аргумент документа ни с одним аргументом не сравнивается, так как принадлежит второму типу аргументных выражений, не представленному в запросе.

Сопоставление осуществляется в результате подсчета коэффициентов лексического сходства аргументов $Term_j$. Аргумент документа из списка тех, с которыми сопоставляется данный аргумент запроса и для которого получен максимальный коэффициент сходства с этим аргументом запроса, ставится в соответствие последнему. Коэффициент $Term_j$ определяется по следующим правилам:

1) для адвербиальных групп: $Term_j = 1$ в случае полного лексического совпадения; $Term_j = 0$ в противном случае;

2) для предикатных конструкций: $Term_j$ вычисляется рекурсивно по правилам определения сходства двух предикатных конструкций; $Term_j$ приравнивается значению коэффициента сходства конструкций, заполняющих валентности;

3) для именных групп: при расчете используются три коэффициента, определяющие сходство 1. головных существительных ИГ; 2. остального лексического состава ИГ; 3. заполняемых валентностей в предикатных конструкциях запроса и документа.

1. Коэффициент сходства головных существительных Head определяется:

а) Для названий химических соединений (помечаются при анализе супертэгом ~F): название разбивается на уровневые последовательности, каждая из которых состоит из элементарных корней (соответствующих названиям химических групп) и указателей на другие уровни (большие латинские буквы). Самый крайний справа из элементарных корней признается головным словом для уровневой последовательности. Например, название соединения

*2-[2-[4-(4-nitorbenzyloxy)phenyl]ethyl]isothioureamethanesulfonate*

включает 4 уровневые последовательности (головное слово подчеркнуто):

A = *2*-B-*isothioureamethanesulfonate*
B = *2*-C-*ethyl*
C = *4*-D-*phenyl*
D = *4-nitorbenzyloxy*

Каждая уровневая последовательность термина запроса сравнивается с каждой уровневой последовательностью термина документа. Далее по коэффициентам сходства определяется наилучшее соответствие уровневых последовательностей запроса и документа друг другу. Коэффициент сходства для уровневых последовательностей определяется как отношение числа элементарных корней, совпадающих в двух последовательностях, к общему числу корней в них.

б) Для остальных существительных: используется тезаурус. Сходство определяется как отношение номера уровня в иерархии для понятия, являющегося ближайшим общим предком для двух понятий, репрезентируемых сравниваемыми терминами запроса и документа, к большему из двух номеров уровней сравниваемых понятий. Например, для понятий, представляемых терминами *polypeptide* (уровень 5) и *complement* (уровень 4), первым общим предком в тезаурусной иерархии является понятие *combination* (уровень 2). Тогда коэффициент сходства двух терминов равен 2/5.

2. Коэффициент сходства остальной лексики Lex: в ИГ выделяется три группы характеристик, представленных отдельными зонами: а) количественные (выражены числительными или диапазоном значений); б) функциональные (выражены пассивным причастием); в) атрибутивные (остальные характеристики). Сравнение количественных характеристик включает определение, насколько диапазон значений в ИГ запроса включается в диапазон значений ИГ документа. Функциональные характеристики не оцениваются. Сходство атрибутивных характеристик определяется как доля общих элементов для ИГ запроса и документа. Коэффициент сходства Lex вычисляется как взвешенная сумма оценок сходства ИГ по указанным параметрам.

3. Коэффициент сходства валентностей аргументных выражений SemR: список валентностей разбит на две группы main (Субъект, Объект, Косвенный объект) и aux (остальные валентности). В зависимости от принадлежности аргументов запроса и документа к одной/разным группам коэффициент SemR принимает одно из списка произвольно заданных значений.

Коэффициент $Term_j$ для аргументов-ИГ вычисляется как взвешенная сумма трех описанных коэффициентов. Аргумент документа, получивший при сравнении с текущим аргументом запроса максимальное значение коэффициента $Term_j$, соотносится с последним, причем соответствующее значение $Term_j$ характеризует степень их сходства.

Коэффициент сходства аргументного состава Term двух предикатных конструкций в целом определяется как взвешенная сумма значений $Term_j$ для каждого аргумента рассматриваемой предикатной конструкции запроса.

Отбор одного из кандидатов для запроса на уровне предикатно-аргументной конструкции осуществляется на основании значения коэффициента $PredC_t$, кото-

рый вычисляется как $\max_k \left\{ \mathrm{Term}^k \times \mathrm{Pred}^k \right\}$, где $k$ – количество всех кандидатов (предикатно-аргументных конструкций документа) для данной предикатно-аргументной конструкции запроса. Если максимальное из произведений ниже заданного порогового значения, $\mathrm{PredC_t}$ принимает значение 0.

Сопоставление компонентных зон запроса и документа осуществляется на основании значения коэффициента $\mathrm{PredZone_j}$, который определяется как взвешенная сумма значений коэффициентов $\mathrm{PredC_t}$, найденных для всех предикатных конструкций рассматриваемой компонентной зоны запроса. Компонентная зона-кандидат документа с максимальным значением коэффициента $\mathrm{PredZone_j}$ ставится в соответствие рассматриваемой компонентной зоне запроса, причем соответствующее значение $\mathrm{PredZone_j}$ характеризует степень сходства двух зон.

Сопоставление на уровне запроса/документа в целом может осуществляться следующими способами: 1) степень сходства документа и запроса $\mathrm{Simil_{purp}}$ равна значению $\mathrm{PredZone_1}$, соответствующего цели способа; 2) степень сходства документа и запроса $\mathrm{Simil_{max}}$ равна максимальному из значений $\mathrm{PredZone_j}$, соответствующих компонентам способа; 3) степень сходства документа и запроса $\mathrm{Simil_{total}}$ определяется как обобщенный коэффициент, учитывающий: а) значения коэффициентов сходства всех компонентных зон, и б) соответствие указанной последовательности выполнений действий-компонентов способа в запросе и документе.

IV. Модуль выдачи информации представляет собой интерфейс, который предлагает пользователю ранжированный список ссылок на патентные документы, отсортированный в соответствии с рангом на основании значений одного из коэффициентов сходства на уровне запроса/документа ($\mathrm{Simil_{purp}}$, $\mathrm{Simil_{max}}$ или $\mathrm{Simil_{total}}$).

Предложенные правила сопоставления структурированных представлений запроса и документа дают возможность создать автоматическое приложение, осуществляющее извлечение текстов формул изобретений из патентных БД на основании глубокого лингвистического анализа и учета особенностей естественного языка. Такая система должна характеризоваться более тонким механизмом обработки реализации смыслов в языке и обладать, очевидно, большей семантической силой, чем любая система, использующая искусственный информационный язык.

Разработанная модель извлечения информации допускает дальнейшее развитие и может быть использована в направлении решения задач автоматизации патентных исследований, ключевым звеном которой является формальное выделение признаков изобретения. При доработке модели на основании процедуры сопоставления образов документа и запроса возможно автоматизировать анализ патентоспособности и патентной чистоты нового изобретения.

ОСНОВНЫЕ ПОЛОЖЕНИЯ ДИССЕРТАЦИОННОГО ИССЛЕДОВАНИЯ ОТРАЖЕНЫ В СЛЕДУЮЩИХ ПУБЛИКАЦИЯХ:

1. Бабина, О.И. Частотные характеристики семантических классов предикатов, встречающихся в формулах изобретения патентов на метод в фармакологии / О.И. Бабина // Международная научно-практическая конференция «Теория и методика преподавания языков в вузе»: Тезисы докладов / под ред. Е.Н. Ярославовой. (Челябинск, 15-17 декабря 2003 г). – Челябинск: Изд-во ЮУрГУ, 2003. – С. 141-142.

2. Бабина, О.И. Предикатная лексика формул изобретения патентов на метод / О.И. Бабина // Фундаментальные и прикладные исследования в системе образования: Материалы 2-й Международной научно-практической конференции / отв. ред. Н.Н. Болдырев. (Тамбов, 28 марта 2004 г). – Тамбов: Изд-во ТГУ им. Г.Р. Державина, 2004. – Ч. 4. – С. 62-65.

3. Бабина, О.И. Специфика процедуры автоматического анализа текстов патентов на метод / О.И. Бабина // Объединенный научный журнал. №33 (125). Декабрь 2004. – С. 62-66.

4. Бабина, О.И. Грамматические характеристики предикатов формулы изобретения патентов на метод / О.И. Бабина // Вестник ЮУрГУ. Сер. Лингвистика. – Челябинск: Изд-во ЮУрГУ, 2004. – №1. – С. 8-12.

5. Sheremetyeva, S. Meaning-Text theory for textual input analysis and proofing in a generation system / S. Sheremetyeva, O. Babina // Восток – Запад: Вторая международная конференция по модели «Смысл ⇔ Текст» / отв. ред. Ю.Д. Апресян, Л.Л. Иомдин. (Москва, 23-25 июня 2005 г). – М.: Языки славянской культуры, 2005. – С. 458-466.

6. Бабина, О.И. Семантическое сопоставление образов запроса и документа при автоматическом документальном поиске / О.И. Бабина // Наука и образование. IV международная научная конференция: Материалы конференции. (Белово, 2-3 марта 2006 г). – Кемерово: Изд-во КемГУ, 2006.

7. Бабина, О.И. Автоматический отбор релевантной информации из информационного массива патентных текстов / О.И. Бабина // Вестник ЮУрГУ. Сер. Лингвистика. – Челябинск: Изд-во ЮУрГУ, 2006. – №2. – С. 67-72.