002-Azarova_Grebenkov_Lando:_Layout 1 12.05.2008 22:00 Page 11

ИСПОЛЬЗОВАНИЕ МАРКЕРОВ АКТАНТНЫХ ПОЗИЦИЙ ПРИ АНАЛИЗЕ ДЕЛОВЫХ ТЕКСТОВ ДЛЯ РАСШИРЕНИЯ ЛОГИЧЕСКОЙ СХЕМЫ ПРЕДМЕТНОЙ ОБЛАСТИ

THE CONTEXT SCHEMA OF PREDICATE ARGUMENTS FOR AUTOMATIC EXPANSION OF A DOMAIN ONTOLOGY

Азарова И.В. *(ivazarova@gmail.com),* ***Гребеньков А.С.*** *(shurix@grebenkov.ru),*
Ландо Т.М. *(tatiana.lando@gmail.com)*
Санкт-Петербургский государственный университет

В докладе представлено описание системы анализа деловых текстов Фактус для ограниченной предметной области с использованием синтактико-семантических шаблонов и маркеров актантных позиций. Обсуждается возможность расширения логической схемы заданной предметной области на основании анализа текста и создания таким образом «открытой понятийной системы».

1. Введение

На Кафедре математической лингвистики Санкт-Петербургского государственного университета был выполнен ряд исследований [1, 2] по автоматическому анализу текста с использованием имеющихся в распоряжении коллектива ресурсов – компьютерного тезауруса RussNet и формально-грамматического парсера Rus4IR. Была представлена пропозициональная модель семантического анализа текста [2], однако для решения практических задач необходимо прояснить ряд вопросов.

Одной из наиболее важных проблем является то, каким образом соотносится идеографическая структура тезауруса RussNet и логико-понятийная схема некоторой предметной области. RussNet в силу методики построения [3] фиксирует общую для различных функциональных стилей активную часть современного русского языка. В традиции построения wordnet-словарей существует представление о том, что терминологические «расширения» базовой структуры будут естественным образом присоединяться к имеющей основе, создавая более подробную понятийную классификацию [9], хотя и отмечалось, что семантическая интерпретация словосочетаний (особенно терминологических) требует особого подхода. Поскольку маловероятно исчисление всех комбинаций компонентов, образующих терминологические и нетерминологические словосочетания, необходимо понять, каким образом и в каком объеме следует предварительно фиксировать основные понятия предметной области, как их комбинировать и дополнять; каким образом будут соотносится понятия логической схемы предметной области и идентификаторы конкретных объектов (так называемые «именованные сущности»)?

Еще одним нерешенным вопросом является то, как представлять содержание текста: в пропозициональной структуре, предложенной ранее, или в какой-то иной форме [8, 5].

Для прояснения перечисленных вопросов была создана модельная система анализа деловых текстов Фактус, которая сама по себе имеет практическое значение, но при этом должна позволять естественно расширять как имеющуюся структуру понятий, так и вводить новые структуры. В качестве предметной области мы выбрали «назначения на должности управляющего персонала в компаниях», она довольно часто используется для иллюстрации возможности анализа текстов [4, 6] и имеет собственную информационную ценность.

2. Модель предметной области
2.1 Логическая схема предметной области

Для создания логико-понятийной схемы предметной области были исследованы тексты (25 текстов в среднем по 1 000 знаков) по данной тематике (CNEWS.ru и подобные). Поскольку тексты содержали разнообразные сведения, было отобрано «ядро», которое впоследствии может быть расширено, − это собственно данные о назначении руководителей высшего звена в компаниях различного типа. Другие сведения: о профильной продукции компаний, выпуске новых видов продукции, участии в конкурсах и различных проектах, а также биогра-

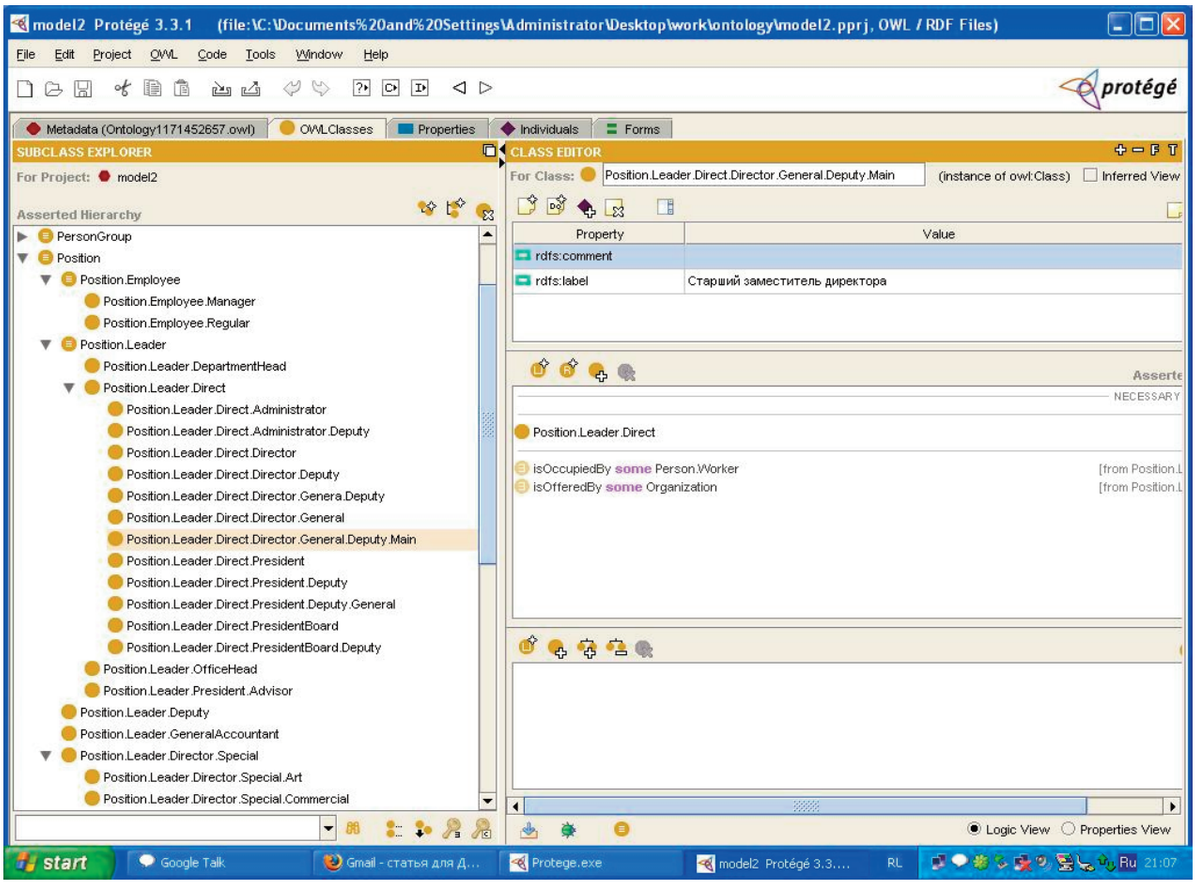
002-Azarova_Grebenkov_Lando:_Layout 1 12.05.2008 22:00 Page 12

фические сведения о руководителях – при необходимости будут использованы для расширения границ предметной области.

Мы учитывали, что возможны варианты онтологического представления предметной области, в которых могут быть акцентированы в зависимости от целей моделирования:

1. приближенность к объективной картине мира;
2. соответствие описания толковым словарям, энциклопедиям или другим ресурсам;
3. удобство обработки итоговой онтологии;
4. потенциальная возможность автоматически расширять онтологию.

Начальная классификация понятий предметной области была создана вручную, данные были представлены в формате OWL в редакторе Protégé 3.3.1. Обработка текстовых данных позволила зафиксировать так называемые «открытые» классы понятий с тем, чтобы далее использовать возможности автоматического дополнения



логической схемы, или онтологии, предметной области. Помимо прочих соображений, учитывалась возможность относительно простого соотнесения со структурой wordnet-словарей.

В качестве основных классов были выбраны те, которые имеют референциальную соотнесенность: ***Компания***, ***Человек***, ***Группа людей*** (коллективные органы управления, например, *совет директоров*). Еще одна группа понятий ***Должность*** – руководящие должности в компаниях – на первый взгляд может быть представлена в виде отношения между членами классов ***Компания*** и ***Человек***. В частности, названия руководящих позиций (*директор*, *заместитель* и проч.) представлены в Принстонском WordNet в структуре семантического дерева ***Person*** *(individual, someone, mortal, human, soul)*, а в SUMO [10] – класс ***Position*** входит в класс ***SocialRole*** и является классом атрибутов, указывающих область ответственности в некоторой организации. Связь элементов класса ***Должность*** задается предикатом «занимать позицию» (***occupiesPosition***), который имеет 3 аргумента, члены классов ***Человек***, ***Должность***, ***Организация***.

(instance occupiesPosition TernaryPredicate)
(domain occupiesPosition 1 Human)
(domain occupiesPosition 2 Position)
(domain occupiesPosition 3 Organization)

002-Azarova_Grebenkov_Lando:_Layout 1 12.05.2008 22:00 Page 13

(documentation occupiesPosition EnglishLanguage «(&%occupiesPosition ?PERSON ?POSITION ?ORG)

В нашем исследовании онтологическое представление используется в первую очередь для эффективной и специфической обработки данных, поэтому интерпретация понятий должностных позиций в духе SUMO, как класса, больше подходит для фиксации ядра понятийной схемы системы Фактус.

При поверхностном рассмотрении может показаться, что между классами ***Человек***, ***Организация*** и ***Должность*** есть довольно четкое различие: первые два имеют трудно перечислимое множество «экземпляров», которые соотносятся с реальными объектами мира, а должности – закрытый класс понятий. Однако, в реальных текстах ситуация не так проста. И хотя имеется ограниченное количество названий высших руководителей в компании (*генеральный директор, президент, управляющий*), для многих высших и средних руководящих должностей в компаниях имеются собственные структуры, которые довольно сложно поддаются «выравниванию», например: *директор по безопасности* vs *директор по безопасности и режиму*, *директор по информационной политике и связям с общественностью* vs *директор по информационным технологиям* и проч. Таким образом, текстовые реализации понятий для этих трех классов имеют сходные черты: они представляют собой «открытые» списки, среди которых есть небольшое количество частотных экземпляров и ряд не поддающихся исчерпывающему перечислению именованных сущностей.

В данной работе мы опирались на удобство обработки понятийной модели и ее потенциальную расширяемость, что в отдельных случаях «навязывало» нам некоторые решения. Например, в структуре класса ***Должность*** есть подструктура заместителей (*директора, директора по маркетингу* и проч.). В Принстонском WordNet различного рода заместители помещены в качестве гипонимов понятия «помощник» {assistant, helper, help, supporter}, то есть отдельно от поддерева с вершиной «руководитель» {leader}. В нашей модели мы посчитали неразумным создавать параллельную структуру заместителей для руководителей различного рода и сочли возможным при необходимости подчинять элементам класса ***Должность «***заместителей», которые вводятся в логическую структуру в виде стандартного расширения. Например, для понятия *заместитель креативного директора* будет автоматически создаваться узел онтологии Position.Leader.Director.Special.Art.Deputy на основе имеющегося понятия *креативный директор* Position.Leader.Director.Special.Art. Текстовые варианты названий должностей включаются в модель предметной области в редакторе Protégé 3.3.1 в качестве метки (label), поскольку для идентификации узлов онтологии используется латиница.

2.2. Фиксация «именованных сущностей»

Отдельной проблемой при моделировании предметной области является работа с именованными сущностями (named entities), то есть с идентификаторами отдельных индивидов, компаний и некоторых должностей. Формат OWL, в котором строится модель, поддерживает включение «экземпляров» или «индивидов» в онтологию, хотя возможно также создание внешних словарей-расширений для перечисления таких данных.

На первый взгляд, оба подхода вполне равноправны. При автоматическом анализе текста имена собственные желательно обрабатывать специальным образом, поскольку только частотные (и, как правило, исконные) могут быть перечислены заранее, поэтому мы считаем, что включение частичных перечней именованных сущностей в онтологию будет некорректным. Мы храним их в отдельных словарях-реестрах, а для связи конкретного имени собственного с нужным классом достаточно ввести дополнительное поле, в котором будет содержаться ссылка на понятийный узел онтологии (об этом подробно будет сказано ниже).

Особую проблему представляет отождествление именованных сущностей: текстовые варианты ссылок на лица и компании различаются. И если варианты упоминания лица достаточно предсказуемы (например: *г-н Иванов, Петр Иванов, Петр Сергеевич Иванов, П. Иванов, П.С. Иванов*), то варьирование названий компаний носит более сложный характер (например: *Siemens, Siemens CT, Siemens Corporate, Siemens Corporate Technology, Сименс* и т.д.). Разные ссылки на лицо в пределах указанной вариативности в одном тексте объединяются, при этом в качестве основного варианта выбирается наиболее развернутое именование (как правило, это первое упоминание в тексте), для остальных вариантов названий лиц в реестре даются ссылки на основной вариант. Для вариантов названия компаний вводятся обязательные и факультативные области (будут описаны в следующем разделе), отождествление имен производится при совпадении названия в основной части. Автоматическое расширение реестров именованных сущностей возможно при учете маркеров принадлежности к некоторому классу сущностей (маркеров актантных позиций), которые будут описаны в следующем разделе.

3. Маркеры актантных позиций

Центральное место при автоматическом анализе текстов в системе Фактус занимает выявление

002-Azarova_Grebenkov_Lando:_Layout 1 12.05.2008 22:00 Page 14

реализации предиката ***occupiesPosition***, который связывает между собой понятия 3 классов онтологической схемы, при этом текстовые реализации понятийных классов занимают актантные позиции. Каждый из рассматриваемых классов понятий (***Человек***, ***Должность***, ***Организация***) имеет определенные характеристики текстовых реализаций, на основании которых можно отнести то или иное слово или словосочетание к выражению понятия определенного класса. Совокупность таких признаков будем называть маркерами актантных классов, выделим среди них следующие типы (возможны также их комбинации).

1. Лексические маркеры

В качестве лексических маркеров выступают слова и выражения, которые явным образом задают «роль» актантной позиции, например, *господин, г-н, госпожа* и проч. для ***Человек***, *компания, в компании, концерн*, и проч. для ***Организация***, *на должность, на позицию, в должности, пост* и проч. для ***Должность***. В дальнейшем будем обозначать их как ролевые лексические маркеры. Кроме того, лексическими маркерами являются представленные в словаре компоненты текстовых реализаций для каждого понятийного класса, которые имеют собственную линейную структуру, включающую обязательные и факультативные компоненты. Позиция ***Человек*** задается указанием имени, отчества и фамилии, последний компонент является обязательным. ***Организация*** включает характеристику типа собственности компании и названия профиля организации, которые чаще выражены аббревиатурами (например, *ООО, ЗАО, КБ, ГПИ* и проч.), однако встречаются и в виде полных словосочетаний (*строительная компания, инвестиционный банк, коммерческий банк, акционерная аудиторская фирма* и т. п.), собственно названия – имени собственного в развернутом или сокращенном виде (*Ситроникс, Скай Линк* и т. п.) и указания территориальной принадлежности организации (*Уральский, Санкт-Петербургский, американский, в Дрездене* и проч.). Обязательным компонентом является имя собственное − название организации. Позиция ***Должность*** включает существительные, которые формируют ядро словосочетаний (*директор, управляющий, президент, заместитель* и т. п.) и зависимых слов – всевозможных атрибутов, представленных в виде определений и предложных конструкций (*старший вице-президент по продажам, директор по продаже профессиональных услуг* и проч.). Факультативные компоненты, оформляющие текстовое именование для класса ***Организация***, также используются как ролевые лексические маркеры: в первую очередь, это аббревиатуры типа *ООО, ЗАО* и т. д.

2. Пунктуационные маркеры

Четким пунктуационным маркером имени собственного компании являются кавычки, причем они выделяют как обязательный компонент названия, так и тип профиля/ владения организации, например, *ЗАО «Петер-Сервис», ОАО «Концерн «Ситроникс»*. Другим пунктуационным маркером являются скобки, которые используются для введения иноязычного названия организации или сокращенного обозначения, которое расценивается авторами текста как не вполне общепринятое, но в дальнейшем аббревиатура свободно используется в тексте: *ОАО «Концерн Научный Центр» (КНЦ), ОАО «Уралсвязьинформ» (URSI)*.

3. Графематические маркеры

Названия компаний и имена указываются в текстах в латинском написании чаще при пояснении в скобках и реже в основном тексте: *Алекс Адамопулос (Alex Adamopulos), компания Direct Tech Inc., Siemens AG*. Маркером имен собственных при обозначении лиц и организаций является написание с заглавной буквы, однако при этом необходимо отделять их от маркера начала предложения. Использование одних прописных в слове маркирует как аббревиатуры, так и собственные имена компаний: *ОАО МЕДУЗА*.

4. Маркер новизны

В том случае когда у предиката ***occupiesPosition*** есть незаполненные позиции, то графематически отмеченные слова и словосочетания (предположительно, это имена собственные), которых нет в словаре, получают дополнительный вес в качестве заполнителей актантных позиций. Для определения типа позиции большое значение имеют факультативные ролевые лексические маркеры (например, *ЗАО БМК-АВТО*). Такие текстовые фрагменты будут автоматически подключаться к словарю.

5. Синтаксические маркеры

В качестве частотной текстовой реализации предиката ***occupiesPosition*** выступают глаголы, которые предопределяют «конфигурации» актантных позиций (рамки валентностей) как в плане взаимного расположения актантов, так и способа морфо-семантического способа их оформления. Например, для пассивного залога глагола «назначить» (обычно в форме прошедшего времени *был назначен*) типичная конфигурация включает словосочетание-подлежащее, которое задает позицию ***Человек***, и дополнение в форме творительного падежа, обозначающее ***Должность***, при котором в форме предложной конструкции или генитива указывается ***Организация***: *Иван Петров был назначен директором ООО «Бригада»* или *Иван Петров был назначен директором в ООО «Бригада»*.

Словосочетания, заполняющие позицию ***Должность***, могут представлять ряды (*директор по информационной политике и связям с общественностью*), вариативность которых довольно велика, мы

002-Azarova_Grebenkov_Lando:_Layout 1 12.05.2008 22:00 Page 15

используем декомпозицию таких конструкций. Аналогичным образом декомпозиция применяется при анализе сложных описаний позиции ***Организация***, включающих факультативные компоненты в форме свободных словосочетаний (*ОАО Акционерная производственно-ювелирная компания «Золото Якутии»*).

4. Процедура анализа делового текста предметной области

При анализе текста в системе Фактус моделируется «заинтересованный взгляд», когда читатель пропускает неважное и выбирает то, что ему интересно. То есть это ситуация неполного семантического анализа текста.

Целевая семантическая структура – извлечение тернарных предикатов ***occupiesPosition*** (назначение и увольнение руководящих сотрудников в компаниях) с заполненными позициями. Следовательно, основной принцип процедуры анализа состоит в том, чтобы вычленять и анализировать только релевантные отрывки текста, пропуская все, не относящееся к трехкомпонентной схеме. Необходимо смоделировать взгляд читателя, который выхватывает из текста определенные маркеры для понимания текста, учитывая то, что на вход анализатора подаются неподготовленные тексты (например, ленты новостей), которые могут не содержать интересующие нас факты, этот подход − пропускать все лишнее − является экономичным и оправданным, так как в тексте из нескольких абзацев может быть только одно предложение, содержащее интересующий нас факт. Таким образом, система работает как фильтр, реагирующий на определенные «ключи».

В соответствие с этой концепцией можно выделить несколько этапов анализа.

1. Предварительная обработка текста: фрагментация текста на предложения и словоформы, определение различных графематических маркеров.

2. Морфологический и словарный анализ, определение стандартных грамматических показателей; за счет контекстных связей частично разрешается грамматическая омонимия.

3. Синтактико-семантический анализ, результатом которого является выделение трехкомпонентных фактов.

Первые два этапа стандартны для решения любой лингвистической задачи, так что опишем подробнее этап синтактико-семантического анализа. В настоящее время единицей обработки является предложение, в дальнейшем планируется работа с абзацами и текстами. Основой этого этапа анализа является использование семантико-синтаксических шаблонов (аналогов рамок валентностей), в которых задаются схемы реализаций фактов. На сегодняшний день подготовлено порядка 100 подобных шаблонов. Структурно шаблоны состоят из следующих элементов:

1. Ключевые слова, или способы выражения предиката ***occupiesPosition***, которые выражены лексически: *назначить* (активная или квазипассивная конструкция), *уйти*, *перейти* и т. д.

2. Связанные с конкретной формой предиката конфигурации актантов, то есть синтаксических маркеров.

3. Шаблоны декомпозиции для комплексных описаний компонентов ***Человек***, ***Должность***, ***Организация***.

Предложение сравнивается с шаблонами, описывающими искомые факты, вначале проверяется наличие ключевых предикатов, затем − конфигураций актантов. Вычисляется мера соответствия предложения некоторому шаблону, поскольку оно может соответствовать нескольким шаблонам одновременно. В результате выбирается шаблон с наибольшей степенью соответствия, по которому в предложении вычленяются текстовые реализации компонентов ***Человек***, ***Должность***, ***Организация.***

Например, шаблон, описывающий предложение *Вася Пупкин был уволен с поста директора фирмы РиК* в нашей нотации будет выглядеть так

(1) person:nom VERB:уволить:pass PREP:c post:gen org:gen,

где VERB и PREP являются терминальными элементами,

person, post и org задают нетерминальные структуры реализаций актантных позиций ***Человек***, ***Должность*** и ***Организация***.

Шаблоны в отличие от контекстно-свободных грамматик не задают жесткого порядка слов, поэтому указанному шаблону будут соответствовать и другие предложения *С поста директора фирмы РиК был уволен Вася Пупкин* и *С поста директора фирмы РиК Вася Пупкин был уволен*.

Одной из проблем использования шаблонов является создание полных перечней возможных синтаксических структур. Решением является автоматизированное пополнение набора шаблонов. Если в тексте встречается предложение, частично реализующее некий шаблон, при этом лишь частично определены компоненты факта, то есть основание предполагать, что в этом предложении скрыт шаблон, не внесенный в список. В таком случае это предложение попадает на проверку лингвисту.

Текстовые фрагменты, не зафиксированные в онтологии и словарях-реестрах именованных сущностей, которые выделены с использованием актантных маркеров, то есть получившие «подтверждение» их принадлежности к классу, автоматически добавляются, расширяют онтологию (перечни экземпляров).

002-Azarova_Grebenkov_Lando:_Layout 1 12.05.2008 22:00 Page 16

5. Выводы и перспективы исследования

В результате анализа модельной системы Фактус были получены некоторые ответы на поставленные в данной работе вопросы. Несмотря на то, что методика построения RussNet и онтологии предметной области Фактус опирается на анализ текстов, между структурами наблюдается довольно значительные отличия, которые, вероятнее всего, связаны с различием терминологического и нетерминологического употребления слов в деловом тексте и «усредненном», не имеющем четкой функциональной направленности тексте.

В отличие от чистых структур автоматического построения онтологий, в нашей системе реализован гибридный подход: первоначально ручная онтология расширяется автоматически, причем получаются плоские структуры. Также предполагается использовать синтаксические шаблоны совместно с порождающими грамматиками для описания структур компонентов. Мы надеемся, что этот подход позволит добиться значительного улучшения в поиске компонентов информации понятийной области.

Пропозициональная структура в качестве семантического представления имеет в большой степени теоретический характер, поскольку требует заполнения аргументных позиций, не всегда имеющих явного определения в тексте. Таким образом, она мало подходит для условий неполноты данных, особенно в ситуации неполной обработки текста. В такой ситуации более предпочтительной является опора на текстовое выражение информации – на то, что явно выражено в нем, и на формализованную систему фиксации информации, больше напоминая заполнение заранее заданных «слотов» − фрагментов информационных структур.

В настоящее время тестируется прототип системы, которая по завершению разработки будет включать в себя сразу несколько лингвистических механизмов анализа текста. Система проектируется таким образом, чтобы автономно обрабатывать большое количество реальных текстов деловой тематики.

В ближайшем будущем планируется соединить шаблонное представление структуры предложения или абзаца и порождающее описание локальных синтаксических групп (словосочетаний), для которых используется формализм порождающей грамматики AGFL [8]. Будет доработана схема объединения компонентов информации об одном факте или ряде фактов, которые представлены в двух или нескольких предложениях текста.

Список литературы

1. Азарова И.В., Иванов В. Л. Овчинникова Е. А. Использование схемы наследования рамок валентностей в тезаурусе RussNet для автоматического анализа текста. Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». М., 2006. С. 18−25.

2. Азарова И.В., Овчинникова Е.А. Семантическая интерпретация именных конструкций по корпусу русских текстов Труды Международной конференции «Корпусная лингвистика – 2006». СПб, 2006. С. 25−33.

3. Азарова И.В., Синопальникова А.А. Использование статистико-комбинаторных свойств корпуса современных текстов для формирования структуры компьютерного тезауруса RussNet // Труды международной конференции «Корпусная лингвистика 2004». 11−14 октября 2004 г. СПб., 2004. С. 5−15.

4. Ермаков А.Е. Поиск фактов в тексте // Мир ПК. 2005. N 2.

5. Железняков М.М., Невлева Т.Н., Новицкая И.М., Смирнова Л.Н., Цейтин Г.С. Опыт построения модели типа «текст -> действительность» с использованием ассоциативных сетей // Машинный фонд русского языка: предпроектные исследования, Институт русского языка АН СССР, М., 1988, 140−167.

6. Киселев С.Л., Ермаков А.Е., Плешко В.В. Поиск фактов в тексте естественного языка на основе сетевых описаний // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2004. М., 2004.

7. Azarova I.V., Ovchinnikova E. A., Ivanov V., Sinopalnikova A. A. RussNet as a Semantic Component of the Text Analyser for Russian. Proceedings of the Third International WordNet Conference. Brno, 2005. P. 19−28.

8. Koster C. H.A., Transducing Text to Multiword Units // Workshop on MultiWord Units MEMURA at the fourth International Conference on Language Resources and Evaluation, LREC-2004. Lisbon, Portugal, May 2004.

9. Harabagiu S.M., Moldovan D.I. Knowledge Processing on the Extended WordNet // WordNet: An Electronic Lexical Database / Ch. Fellbaum (ed.). MIT Press, 1998. P. 379–405.

10. Suggested Upper Merged Ontology (SUMO) // URL: http://www.ontologyportal.org/