АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ФАКТОВ ИЗ ТЕКСТОВ ДОСЬЕ: ОПЫТ УСТАНОВЛЕНИЯ АНАФОРИЧЕСКИХ СВЯЗЕЙ

AUTOMATICAL EXTRACTION OF FACTS FROM TEXTS OF PERSONAL FILES: EXPIRIENCE IN ANAPHORA RESOLUTION

Ермаков А.Е. *(ermakov@metric.ru), ООО "Эр Си О" (http://www.rco.ru)*

Доклад описывает опыт решения задачи автоматического извлечения фактов из текстовых документов особого стиля – досье. Описываются использованные для поиска фактов средства на основе синтаксического анализатора и синтактико-семантических шаблонов. Особое внимание уделяется закономерностям организации дискурса, использованным для установления анафорических связей.

Введение

Доклад посвящен задаче извлечения фактографической информации из текстовых документов особого стиля, к которому можно отнести биографии, протоколы, сводки и прочие документы, назначение которых состоит в лаконичной передаче совокупности фактов о некоторых объектах. В исследованном нами корпусе в фокусе внимания авторов всегда находилась персона или организация, вследствие чего этот класс документов наиболее точно характеризуется термином "досье". Автору доклада не известны описания практических разработок в области анализа текста досье, равно как и лингвистические исследования особенностей текстов такого стиля.

Ключевой особенностью текста досье является высокая плотность таких связей между словами, которые не выражаются грамматическими средствами – анафорических связей. Большинство предложений в подобных текстах либо бессубъектно (*Родился в 1958 году. Работает директором ООО "Ромашка"*), либо номинативно (*1958 года рождения. Директор ООО "Ромашка"*), либо разорвано в списках, каждый элемент которых в свою очередь представляет набор предложений – вложенное мини-досье (*Является совладельцем следующих* ***предприятий****: - **ООО "Одуванчик"**, ИНН 500103819710, зарегистрировано в 2001 году. Заявленный вид деятельности – собаководство…. – **ООО "Лютик"**, ИНН 500204519555, основано в 2005 году. С 2006 года занимается… - **ООО "Тюльпан"**, ИНН* ...). В остальном текст является совершенно нормальным и мало чем отличается, скажем, от текстов СМИ – досье часто содержит полные, достаточно сложные предложения, так что его машинный анализ представляет собой задачу, требующую привлечения полного арсенала средств компьютерной лингвистики.

Постановка задачи компьютерного анализа текстов досье требует распознавания и классификации описанных в них фактов, извлечения участников–фигурантов фактов, с последующим преобразованием информации в записи БД в соответствии с требуемой схемой. На рисунке 1 представлена логическая схема организации тех фактографических данных, которые мы извлекали в соответствии с требованиями заказчика.

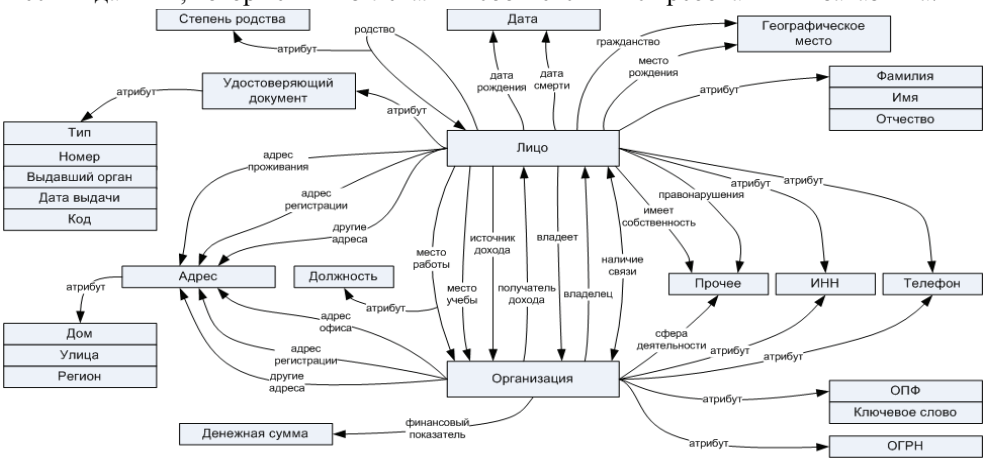

Рис. 1. Логическая схема организации извлекаемых фактографических данных

Абстрагируясь от конечной схемы хранения фактов в БД, которая определяется утилитарными соображениями, мы считаем, что результаты анализа текста должны быть описаны фреймовой моделью: каждый факт формирует запись в соответствующей таблице, имя которой определяет тип факта (например, “дата рождения”, “владеет предприятием”), имена столбцов – роли фигурантов факта (“атрибут”, “владелец”, “собственность”), а значения в полях таблицы – имена участников-фигурантов факта в соответствующих ролях (1*2 мая 1962, Александр Иванович Корейко, ООО “Василек”*).

Для решения поставленной задачи были использованы следующие средства:

1. Модуль распознавания особых текстовых конструкций (паспортных и регистрационных данных, адресов, телефонов, дат) на основании шаблонов, написанных на специальном формальном языке [1];
2. Синтаксический анализатор, определяющий лексико-грамматические характеристики элементов текста и преобразующий текст каждого предложения в семантическую сеть;
3. Алгоритм выделения фактов на основе распознавания требуемой конфигурации синтаксических связей между именами фигурантов факта – поиск фрагментов сети, удовлетворяющих заданным шаблонам [2];
4. Правила разрешения кореферентности имен собственных, в том числе анафорической, которые позволяют отождествить полные, краткие и местоименные обозначения персон и организаций [3];
5. Правила установления анафорической связи между свободной синтаксической или семантической валентностью, соответствующей опущенному в предложении фигуранту факта, и упоминанием одного из возможных референтов.

Средства анализа текста (1)-(3) уже были реализованы нами в линейке программных продуктов RCO (http://www.rco.ru), в то время как алгоритмы (4) явились предметом нового исследования и разработки, вследствие чего оказались в центре внимания настоящего доклада.

Поиск описаний фактов на основе синтаксических связей

Результатом синтаксического анализа каждого предложения текста является сеть синтактико-семантических отношений – семантическая сеть, представленная на рисунке 2.

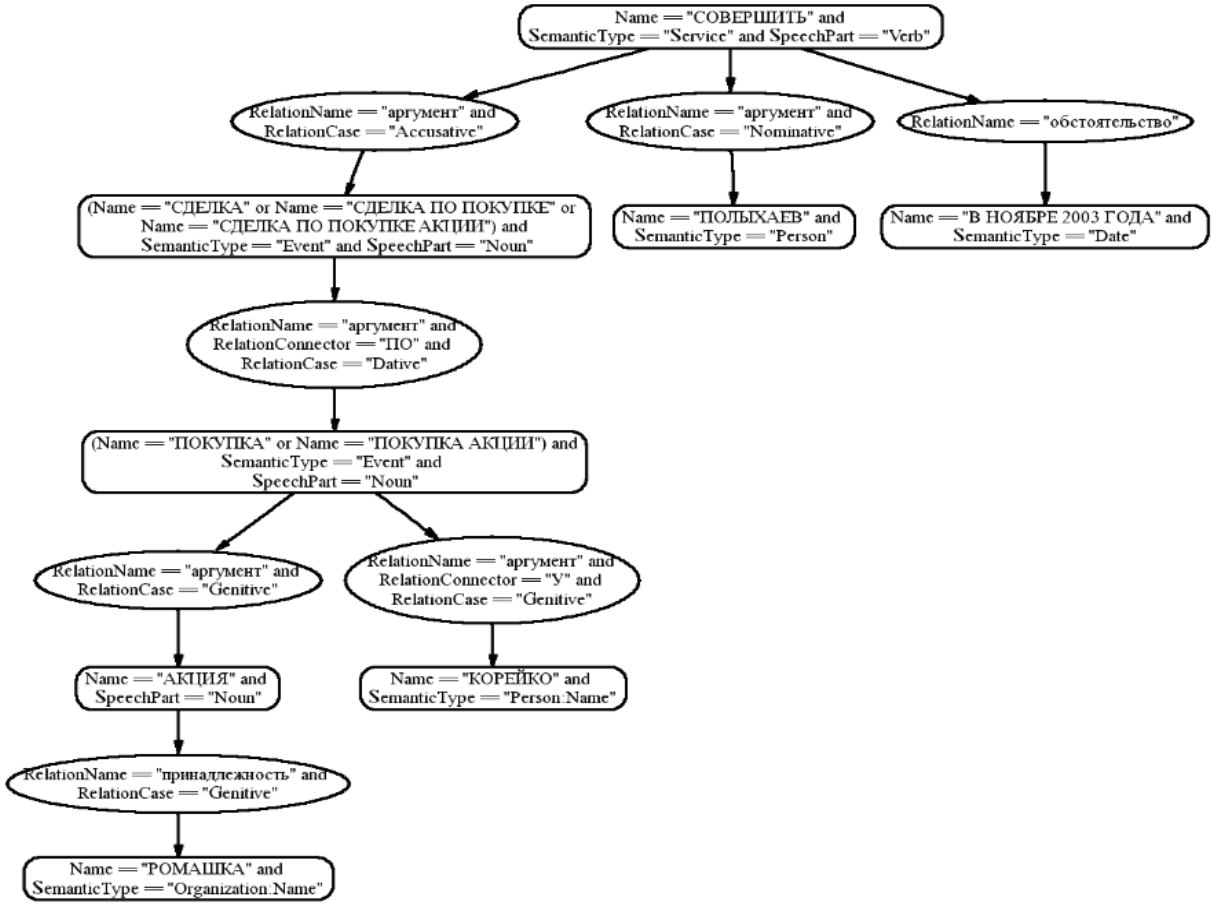

Рис. 2. Пример семантической сети, соответствующей предложению: В ноябре 2003 года Полыхаев совершил сделку по покупке акций ООО “Ромашка” у Корейко

Узлы и связи в сети имеют набор следующих основных атрибутов:

• SpeechPart – часть речи слова, соответствующего узлу.

• SemanticType – семантический разряд референта узла. Основные выделяемые разряды: именованная персона, организация, географическое место, артифакт, действие/состояние, предмет, одушевленный объект и пр.

• Name – строка текста, соответствующего узлу, в нормальной форме. Для именных групп может иметь несколько значений, которые представляют все цельные словосочетания, образованные от ключевого существительного в узле, например: *новый указ президента*, *указ президента*, *указ*. Для именованных объектов соответствует стандартизованному имени: *Корейко Александр Иванович*, *“Ромашка”*.

• RelationType – тип синтактико-семантической связи между узлами, например “*аргумент*”, “*атрибут*”, “*принадлежность*”, “*обстоятельство*”.

• RelationCase, RelationConnector – семантический падеж и коннектор (предлог, союз), при помощи которых устанавливается связь. Комбинация условий RelationCase + RelationConnector представляет принятый нами способ указания семантическим ролей.

Представление содержания текста в форме семантической сети позволяет абстрагироваться от многих особенностей его коммуникативной организации. Такая сеть инвариантна к синтаксической структуре предложения и порядку слов с точностью до структуры пропозиции, выбранной автором для описания ситуации. Например, конструкциям “*Корейко купил акции*” и “*акциях, купленных Корейко*” будут соответствовать одинаковые сети. В то же время пропозициям вида “*Корейко становится покупателем акций*” и “*покупка акций – дело рук Корейко*” будут соответствовать иные сети. Вследствие этого наша семантическая сеть является промежуточным уровнем представления между собственно семантической схемой ситуации и ее языковым описанием.

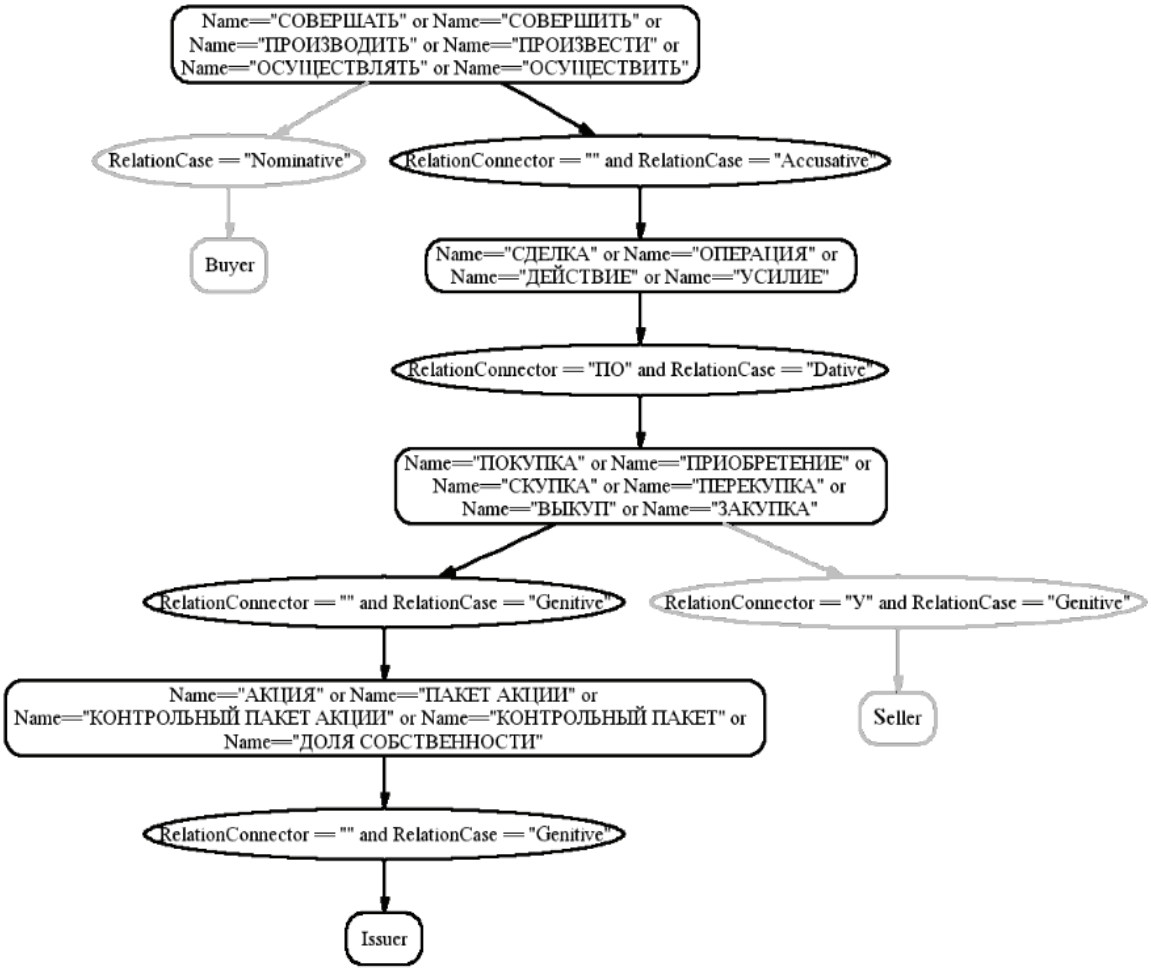

Рис. 3. Пример синтактико-семантического шаблона, распознающего факты, выраженные пропозицией вида “Покупатель совершает действие по приобретению у продавца акций предприятия”.

Для поиска конфигурации связанных слов, описывающей факт искомого типа, используется синтактико-семантический шаблон, который задается в виде сети, подобной искомой в тексте, но в ее узлах и связях при помощи логических выражений указываются условия, которым должны удовлетворять узлы и связи искомой сети (Рис. 3). Как правило, в некоторых узлах шаблона содержатся конкретные слова, которые должны присутствовать в описании искомого факта. Другие узлы, соответствующие искомым фигурантам, содержат метки, которые обозначают роли фигурантов. Такие узлы представляют валентности шаблона, подлежащие заполнению – соответствующие слова будут извлечены из текста при нахождении фрагмента сети, изоморфного шаблону.

Так, на рисунке 3 узлы, обозначенные метками Buyer, Issuer и Seller, представляют возможных фигурантов факта "покупка акций" в ролях "покупатель", "эмитент акций" и "продавец" соответственно. Светлые связи к фигурантам Buyer и Seller помечены как факультативные, так как соответствующие валентности могут оказаться свободными – референт Buyer'а может упоминаться ранее по тексту, а референт Seller'а может вообще не упоминаться в тексте.

В итоге, в результате анализа бессубъектной фразы "*В ноябре 2003 года совершил сделку по покупке акций ООО "Ромашка" у Корейко"* будет выделен факт типа "покупка акций" с фигурантами: Issuer = "*Ромашка*", Seller = "*Корейко*", Buyer = "?". Для установления имени фигуранта в роли Buyer будут использованы алгоритмы анафорического связывания, описанные далее.

На завершающем этапе машинного анализа факт "покупка акций" может быть проинтерпретирован в соответствии с заданной схемой (рисунок 1), в результате чего в БД будут помещены два факта типа "владеет", в которых владельцем предприятия-фигуранта Issuer будет выступать как Buyer, так и Seller, поскольку и покупатель, и продавец владели предприятием в определенные периоды времени.

Отметим, что в ряде случаев синтаксический анализатор не может установить связь между фигурантами факта, опираясь на заложенные в него общие правила русской грамматики (*Соучредитель ООО "Ромашка" (20%) – ЗАО "Сирень"*). Чтобы решить такие проблемы, в семантическую сеть добавляются связи особого типа (RelationType = next), которые просто связывают в цепочку идущие друг за другом в предложении слова и знаки препинания, причем "перепрыгивая" через синтаксически подчиненные слова в именных группах, что позволяет писать шаблоны, инвариантные к длине словосочетаний. В итоге совокупность узлов сети всегда представляет собой полносвязный граф, что позволяет написать шаблон, который извлекает из приведенного примера как соучредителя, так и его долю.

Для настройки шаблонов используется модуль с графическим интерфейсом, который позволяет строить семантическую сеть на основе эталонных фраз, т.е. обучать программу на примерах. После построения программой структурной основы шаблона — узлов и связей — лингвисту остается проставить ограничения и метки в элементах сети.

Поиск анафорических связей

Рассмотрим механизмы поиска тех факультативных фигурантов факта, которые не были найдены в предложении в результате применения синтактико-семантического шаблона.

Не всякая факультативная валентность факта допускает заполнение на основании анафорической связи – в шаблоне, приведенном на рисунке 3, анафорическая связь для валентности Seller не допустима. Для указания на возможность анафорической связи и на ее тип шаблон снабжается информацией - определенным узлам могут быть присвоены дополнительные метки:

- Object - маркирует факультативного фигуранта факта, который может упоминаться **далее** по тексту и с которым может быть установлена анафорическая связь. Так может быть маркирован фигурант Issuer в модификации рассматриваемого шаблона для пропозиции вида "*приобретение акций следующих предприятий: элемент списка 1, ... элемент списка k*", где референты для Issuer упоминаются далее в позициях тем в элементах списка, что и определяет сравнительно простой способ их поиска на практике.
- Subject – маркирует факультативного фигуранта факта, который может упоминаться **ранее** по тексту и с которым может быть установлена анафорическая связь. Такую метку в рассматриваемом шаблоне получает фигурант в роли Buyer, поиск упоминания которого может оказаться не тривиальным и описывается далее.

Первым этапом установления анафорической связи между пустой валентностью Subject и упоминанием референтной персоны/организации является проверка того, выражено ли описание факта бессубъектной или номинативной синтаксической конструкцией. Для этого используется дополнительная пара меток в шаблоне:

• Predicate – маркирует ключевой глагол в синтаксической конструкции, которая может быть бессубъектной (*В 2003 совершал операции по покупке…*). В шаблоне на рисунке 3 такая метка маркирует узел, описывающий синонимический ряд глаголов со значением “совершать”. Если соответствующий глагол найден и удовлетворяет требованиям бессубъектности (не имеет подлежащего и стоит в одной из допустимых грамматических форм), то упоминание референта для Subject следует искать в анафорической связи ранее по тексу.

• KeyNoun - маркирует ключевое существительное, идентифицирующее факт в конструкции, которая может быть номинативной (*2003 год – операции по покупке …*). В шаблоне на рисунке 3 такая метка маркирует узел, описывающий синонимический ряд существительных со значением “операция”. Если соответствующее слово найдено и удовлетворяет требованиям номинативности (не подчинено другому слову и стоит в одной из допустимых грамматических форм), то упоминание референта для Subject следует искать в анафорической связи ранее по тексту.

Вторым этапом установления анафорической связи для валентности Subject является собственно поиск ближайшего по тексту упоминания (антецедента) подходящего референта из числа персон/организаций, удовлетворяющего тем законам построения дискурса, которые эмпирически были отобраны нами для текстов стиля “досье”:

1. Позиция в предложении. Никакой из антецедентов в позиции однородных членов предложения не может быть анафорически связан с фактами. Антецедент, являющийся второстепенным членом предложения, не может иметь анафорической связи с фактами за пределами своей синтаксической клаузы.
2. Тема предложения. За пределами своего предложения факт может быть связан только с антецедентом, входящим в тему предложения. В тему предложения включаются такие антецеденты, которые либо стоят в позиции подлежащего, если таковое найдено, либо не стоят после глагола и не стоят в скобках.
3. Тема параграфа. Факт можно связать с антецедентом, являющимся темой ближайшего предыдущего параграфа, пропустив те параграфы, в которых не удалось обнаружить тему. Темой параграфа считается упоминание персоны/организации в теме его первого предложения.
4. Скобки. Факты, излагающиеся внутри скобок, могут относиться только к антецеденту, стоящему в тех же скобках или непосредственно перед ними. К антецеденту, стоящему внутри скобок, могут относиться только факты, стоящие внутри этих же самых скобок.
5. Списки. Факт не может быть анафорически связан с антецедентом из другого параграфа в случае, когда оба параграфа являются элементами списка.

Одна важная проблема связана со сложностью распознавания списков специфического вида, элементы которых начинаются с нормальных слов, например: *Мать – Корейко Анна Захаровна…*, и следующий элемент списка *Отец – Корейко Иван Абрамович...* Проблема усугубляется тем, что каждый элемент списка зачастую состоит из нескольких параграфов, то есть досье содержит вложенные мини-досье. Невозможность проверить нарушение закона (6) может привести к тому, что *Корейко Иван Абрамович* станет отцом *Корейко Анны Захаровны*.

Дополнительно в ходе извлечения фактов проверяются тривиальные прагматические правила, которые запрещают соотносить с одним референтом более одного факта определенного типа (дата рождения, ИНН, паспортные данные и т.п.).

Заключение

Закономерности построения дискурса в текстах стиля “досье”, описанные в докладе и прошедшие экспериментальную проверку, опираются исключительно на коммуникативные особенности построения текста автором, связанные с формированием или удержанием фокуса внимания читателя. Как оказалось, эти вполне прозрачные законы достаточно строго соблюдаются авторами досье, в то время как в исследованных нами ранее текстах СМИ [3] коммуникативные законы построения дискурса зачастую нарушались из-за того, что авторы использовали в качестве предполагаемой опоры для разрешения анафоры семантические и прагматические компоненты дискурса. Вследствие этого установление анафорической связи в текстах СМИ требовало от нас особо жесткой стратегии принятия решений – кореферентным мог быть признан только ближайший по тексту антецедент подходящего семантического разряда, для которого к тому же не нарушались законы. Напротив, в текстах досье нам удается достоверно относить факты достаточно далеко – например, к такой персоне, от последнего упоминания которой факт отделяет упоминание десятка других персон.

Описанная в докладе схема выделения фактов из текстов досье воплощена в программном комплексе, который в настоящий момент проходит опытные испытания у заказчика. Мы ожидаем, что результаты будут признаны удовлетворительными и дальнейшая работа по улучшению и настройке алгоритмов будет продолжена. Основное направление работы – повышение полноты за счет написания новых шаблонов для фактов новых типов, а также для сравнительно редких способов выражения уже включенных фактов. Так, для извлечения фактографических данных в соответствии со схемой на рисунке 1 уже разработано более 100 шаблонов.

Список литературы

1. Ермаков А.Е., Плешко В.В., Митюнин В.А. RCO Pattern Extractor: компонент выделения особых объектов в тексте. // Информатизация и информационная безопасность правоохранительных органов: XII Международная научная конференция. Сборник трудов - Москва, 2003. - С. 312-317. (http://www.rco.ru/article.asp?ob_no=237)
2. 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. – Москва, Наука, 2004. – С. 282-285. (http://www.rco.ru/article.asp?ob_no=629)
3. Ермаков А.Е. Референция обозначений персон и организаций в русскоязычных текстах СМИ: эмпирические закономерности для компьютерного анализа // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2005. – Москва, Наука, 2005. - С.131-135 (http://www.rco.ru/article.asp?ob_no=2339)