СИСТЕМА СЕМАНТИЧЕСКОЙ РАЗМЕТКИ КОРПУСА ТЕКСТОВ В ОГРАНИЧЕННОЙ ПРЕДМЕТНОЙ ОБЛАСТИ[1]

Загорулько М. Ю. (zagulko_maxim@yahoo.com),
Кононенко И. С. (irina_k@cn.ru),
Сидорова Е. А. (lena@iis.nsk.su)

Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск, Россия

Рассматривается технология объектно-ориентированной экспертной разметки корпуса текстов, предназначенная для извлечения знаний при построении информационных систем для конкретных предметных областей. Исследуются методы и программные средства объектно-ориентированного аннотирования корпусов текстов с целью выявления терминологии и способов представления универсальных ситуаций и отношений. Предложены общие принципы терминологической разметки и разметки универсальных ситуаций и отношений, которые легли в основу разметки коллекции текстов по катализу. Представлена разработанная система экспертной семантической разметки текстов, описаны её пользовательский интерфейс, функционал и архитектура. Описаны перспективные направления использования разметки корпуса: терминологическое наполнение предметных словарей на основе терминологически размеченных фрагментов текста, создание семантико-синтаксических моделей для извлечения фактов из текста.

Ключевые слова: разметка корпусов, семантическая разметка, объектно-ориентированное аннотирование, специализированный корпус текстов

[1] Работа выполняется при финансовой поддержке Президиума РАН (Интеграционный проект СО РАН № 15/10 «Математические и методологические аспекты интеллектуальных информационных систем»).

SYSTEM FOR SEMANTIC ANNOTATION OF DOMAIN-SPECIFIC TEXT CORPORA

Zagorulko M. Ju. (zagulko_maxim@yahoo.com),
Kononenko I. S. (irina_k@cn.ru),
Sidorova E. A. (lena@iis.nsk.su)

A. P. Ershov Institute of Informatics Systems, Novosibirsk, Russia

A system for universal annotation of text corpus by an expert is presented that contributes to extraction of domain knowledge within the framework of developing information systems in specific domains. The technique and software tools for annotation of text corpora allow expert to carry out two types of semantic annotation: 1) identify text fragments in which the domain concepts represented by special terms actually appear (term annotation) and 2) identify text fragments (often discontinuous) that correspond to domain relations or situations including their participant structure (event annotation). The general principles and schemes of term and event annotation have been formulated and tested for the domain of heterogeneous catalysis on base of the hierarchy of term classes chosen beforehand. The system, its functional, architecture, and user interface are described. Two main directions of usage of semantically annotated texts are discussed to be as follows: automatic construction of domain lexicons that associate terms with their linguistic and semantic properties; semi-automatic generation of semantic-syntactic patterns for event extraction.

Key words: text corpora annotation, semantic annotation, object-oriented annotation, domain-specific text corpora.

Введение

При создании лингвистических ресурсов для использования в информационных системах, ориентированных на конкретную область знаний, необходима инструментальная среда исследования корпуса текстов [3], применимая для работы экспертов. Такая среда позволила бы эксперту сопоставлять фрагменты текста заданным понятиям или категориям в соответствии с моделью предметной области (ПО). В отличие от лингвистической разметки корпуса (морфологической, синтаксической и т. п.), используемой многими исследователями [1], семантическая разметка специализированного корпуса ориентирована на конкретную предметную область, и должна производиться аннотаторами в соответствии с предварительно разработанными и согласованными с экспертами принципами разметки [8].

На сегодняшний день основным форматом представления семантической разметки является текст с тегами (xml, rdf, wiki и т. п.), помечающими начало и конец выделяемых фрагментов, и атрибутами, описывающими признаки

фрагмента [2]. Несмотря на несомненные достоинства, связанные с развитием стандартов разметки тэгами и наличием средств их визуализации и обработки, данный подход имеет ряд недостатков, таких как:

- сложность выделения разрывного фрагмента,
- сложность описания связей между фрагментами,
- невозможность выделить описание многоатрибутных объектов предметной области (проекцию объекта на текст),
- система признаков, используемая аннотатором, «размыта» по тэгам и атрибутам,
- неэффективность программной обработки по сравнению со специализированными форматами и т. д.

Альтернативой является использование «внешнего» аннотирования, синхронизированного с текстом [7]. В этом случае описание фрагмента создается отдельно от текста и связывается с текстом указанием позиций его начала и конца. Развивая данный подход, можно создавать описания сущностей (сколь угодно сложные) и связывать их с текстом, указывая позиции начала и конца фрагментов, сопоставляемых с той или иной частью структуры описываемого объекта. Такой подход позволяет осуществлять *объектно-ориентированную разметку* текста и в значительной степени расширяет возможности использования разметки корпуса для создания лингвистических ресурсов, ориентированных на анализ текстов предметной области (терминологических словарей, шаблонов и правил анализа).

В данной работе предложен подход к семантической разметке текста, позволяющий сопоставлять объектно-ориентированные представления сущностей фрагментам текста. Приводимые иллюстрации основаны на результатах проекта по созданию специализированного семантически размеченного корпуса текстов по катализу [4].

1. Принципы семантической разметки

При построении информационных систем неизбежно возникает задача автоматизации процесса извлечения экспертных знаний о предметной области и ее подъязыке — системе понятий и отношений между ними, способах представления сущностей и типовых ситуаций предметной области. Такая задача эффективно решается методами корпусной лингвистики, то есть путем создания и исследования специализированного корпуса текстов, представляющего собой достаточный объем снабженных экспертной интерпретацией лингвистических данных, который может служить основой формирования системы автоматического анализа текстов, т. е. выступать в роли обучающего корпуса. В состав корпуса текстов отбираются фрагменты из справочной и учебной литературы, научные статьи и рефераты, посвященные определенной тематике.

При создании специализированных корпусов текстов обычно производится лингвистическое аннотирование (морфологическое, синтаксическое),

не зависящее от ПО и осуществляемое автоматически и/или вручную. Семантическая разметка, напротив, предметно ориентированна, поскольку определяется онтологией ПО и производится экспертами [8]. Процессу семантической разметки специализированного корпуса текстов предшествует достаточно длительный (2–3 месяца) предварительный этап совместной работы экспертов, лингвистов и разработчиков системы, в рамках которого происходит обмен компетенциями, выработка и согласование признаков и принципов разметки. Результатом этого этапа является инструкция по семантическому аннотированию. Речь идет о двух видах семантического аннотирования:

- терминологическая разметка, которая в первую очередь предназначена для фиксации в тексте имен понятий ПО,
- разметка отношений (или ситуаций, представляющих собой многоместные отношения), в которых размеченные сущности выступают в определенных семантических ролях.

Ниже изложены основные принципы аннотирования корпуса текстов по катализу и результаты предварительных экспериментов по разметке.

Терминологическая разметка фиксирует не только присутствие в тексте наименований сущностей ПО, но и особенности использования общеупотребительной лексики в данном подъязыке. Предложены следующие принципы терминологической разметки:

- Определять признаки, соответствующие типам и подтипам сущностей в соответствии с иерархией признаков.
- Определять максимальный текстовый фрагмент, представляющий сущность. *(нитрил акриловой кислоты)*. При этом остается возможность указывать и вложенные фрагменты, представляющие сущности (*нитрил акриловой кислоты*).
- Сопоставлять аннотацию самой сущности, а не всей синтаксической группе, в которую она входит *(в реакцию сочетания вступают кислородсодержащие продукты CH3OH, CH2O)*.
- Размечать все ссылки на сущности, в том числе анафорические замены.
- Использовать при необходимости разрывные фрагменты (*после контакта CH4 с оксидами ряда металлов*), в том числе при перечислении с сочинительным сокращением: *конденсация гликолевого и глицеринового альдегидов*

Иерархия признаков для терминологической разметки текстов по катализу позволяет пометить фрагмент текста как Вещество или конкретный подкласс веществ (Элемент, Соединение и др.). В иерархию признаков для веществ внесены и лексические показатели Роли (ролевая лексика *катализатор, реагент, реакционная смесь, реактант, продукт* и т. п.).

Кроме того, система признаков представляет классы предикатов:

- Реакции:
  - Химические реакции (*окисление, гидрировать, крекинг)*, подклассы которых учитывают валентностный потенциал: именные реакции

(*реакция Будуара*), реакции с инкорпорированным участником (*метанирование*) и др.;
  - Обобщенные реакции (Взаимодействие, Превращение, Получение);
- Лексические показатели ситуаций (в частности, представляющие взаимосвязи процессов и веществ в составе ситуации (такие как *катализировать*, *приводить к, использоваться в качестве/для*).

Разметка отношений (ситуаций) производится над терминологически размеченным текстом. Рассмотрим ситуации типа ПРОЦЕСС, описывающие процессы молекулярного взаимодействия в катализе:

Паровая конверсия метана в синтез-газ протекает на никелевом катализаторе.

Такая микроситуация представляется как многоместное отношение:

ПРОЦЕСС (Реакция, Реагент, Катализатор, Продукт)

где Реакция (*паровая конверсия*) — химическая реакция, характеризующаяся превращением одного или нескольких исходных веществ, выступающих в семантической роли Реагентов (*метан*), в отличающиеся от них по химическому составу или строению вещества, Продукты реакции (*синтез-газ*), с участием Катализатора (*никель*).

Ниже приводится пример разметки текста, в котором каждое вхождение термина и ситуации выделено идентификатором (Ti, Pj). Из примера видно, что *этилен* выступает в роли Продукта в ситуации П44 и в роли Реагента — в ситуации П46, таким образом, фрагменты, описывающие ситуации, в данном случае пересекаются:

Кроме того, этилен <T55>, *образующийся* <T56> *при окислительном превращении* <T57> *CH*4 <T58>, *можно почти полностью конвертировать* <T59> *в другие олефины* <T60> *и ароматические углеводороды* <T61> *на цеолитах* <T62>.

Процесс П44 (Реакция: T56 <Образование>, Продукт: T55)

… *этилен* <T55>, *образующийся* <T56> …

Процесс П45 (Реакция: T57 <Химическая реакция>, Реагент:T58)

…*этилен* <T55>, *образующийся* <T56> *при окислительном превращении*<T57> *CH*4 <T58>,…

Процесс П46 (Реакция: T59 <Превращение>, Реагент:T55, Продукт: T60, Продукт: T61, Катализатор: T62)

…*этилен* <T55>, *образующийся* <T56> *при окислительном превращении* <T57> *CH*4 <T58>, *можно почти полностью конвертировать* <T59> *в другие олефины* <T60> *и ароматические углеводороды* <T61> *на цеолитах* <T62>

Разработаны следующие принципы ситуационной разметки:

- Разделение терминологической и ситуационной разметок. Так, во фразе: *В данной работе мы исследовали влияние предварительного восстановления водородом платиновых и палладиевых катализаторов*
  терминологическая разметка:
  палладиевых <Элемент, Экземпляр> *катализаторов* <Роль>
  ситуационная разметка:
  палладиевых <Реагент, П24>
- Диагностирующим контекстом, позволяющим предположить наличие в тексте описания ситуации/отношения, является присутствие в нем соответствующего лексического предиката (для химических процессов — присутствие термина-обозначения конкретной или обобщенной реакции).
- Ситуация обычно выражена в рамках клаузы/предложения, выход за рамки клаузы/предложения возможен при анафорической замене, при этом в разметке указывается антецедент:
  первой стадией является синтез метанола, далее следуют его дегидратация
- При сочинении различаются множественные ситуации и ситуации с множественными участниками.
  Глубокое окисление метана на платиновых и палладиевых катализаторах, нанесенных на нитрид кремния (2 ситуации)
  Разложение сероводорода на элементную серу и водород (1 ситуация, 2 Продукта)
- Лексические показатели ситуаций не создают новых ситуаций, но фиксируют связи в рамках ситуации, позволяя определить потенциальных участников.
  Метанол подвергается превращениям, которые характерны для катализа высококремнистыми цеолитами.

2. Архитектура системы разметки текста

Внутренняя работа системы основывается на трех концептах: дерево признаков, разметка текста и список ситуаций (см. Рис.1).

Дерево признаков реализует представление иерархической системы признаков, где нулевая вершина является фиктивным признаком, не участвующим в разметке текста, а любой другой элемент дерева может одновременно выступать как в роли признака, так и в роли вершины поддерева признаков. Для хранения разметки текста используется хеш-таблица, в которой ключу соответствует признак, по которому размечен фрагмент текста, а значению соответствует начальная и конечная позиции в тексте и список экземпляров ситуаций. Список отношений (ситуаций) содержит их абстрактные описания, на основе которых создаются конкретные экземпляры.

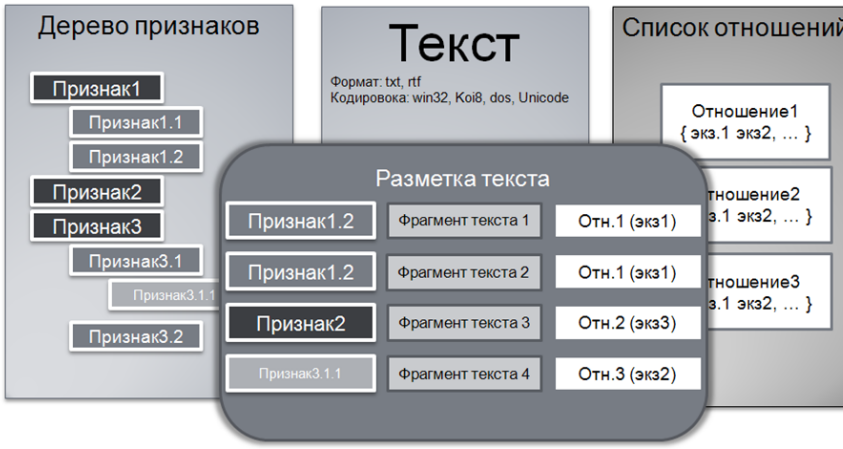

Рис. 1. Архитектура системы разметки текста

Разработанный инструмент позволяет пользователю формировать иерархию признаков, каждому признаку сопоставлять цветовую и стилевую схему разметки, которая используется при реализации функций визуализации разметки в тексте. Поддерживается раздельный просмотр разметки по признакам/ группам признаков.

Описание фрагментов вынесено в отдельную таблицу, в которой отражены позиции, текстовое представление, признаки и связи фрагментов. Таблица поддерживает навигацию в тексте, а также сортировку фрагментов по разным параметрам.

Для поддержки всех свойств разметки, а также для более эффективного дальнейшего использования для автоматизированного создания лингвистических ресурсов принято решение отказаться от стандартного формата хранения размеченного текста в виде текста с xml-тегами, помечающими начало и конец выделяемых фрагментов. Вместо этого создается аннотация, синхронизированная с исходным текстом (текст загружается из тестового или rtf-файла и в дальнейшем не меняется). Аннотация — это множество троек <признак, позиция, информация>, которые фиксируют, что определенная символьная последовательность в тексте (фрагмент) обладает определенными свойствами. В процессе работы аннотатора разметка динамически визуализируется.

Пользовательский интерфейс системы разметки текстовых корпусов, разрабатываемой для лингвистов и экспертов предметной области, должен быть легким (интуитивно понятным) как для опытного, так и для начинающего пользователя.

Система разметки предоставляет следующие возможности.

- Загрузка текста формата txt, rtf, поддержка кодировок win32, Koi8, dos, Unicode.
- Загрузка и сохранение размеченного текста в специальный формат mspr (Mars System Project).

- Просмотр и редактирование дерева иерархии признаков.
- Просмотр и редактирование фрагментов, приписанных определенному признаку.
- Просмотр всех (или части) размеченных фрагментов одновременно (в видимой части текста).
- Загрузка и сохранение размеченного текста; формат файла текстовый, например, xml-подобный.
- Сортировка списка размеченных фрагментов текста по позиции в тексте, по имени признака, по фрагменту текста.
- Обеспечение многослойной разметки.

На Рис.2 представлен пользовательский интерфейс системы разметки текста и продемонстрирована ситуационная разметка (терминологическая отфильтрована), при этом разметка ПРОЦЕССов частично перекрыта разметкой другими отношениями, например в первой фразе текста процессы *Окислительная конденсация метана* и *Синтез этана, этилена и других углеводородов* связаны отношением Реализации (отражающим способ осуществления химической реакции). Помимо ситуаций, в тексте отражены лингвистические отношения, например во фразе *…высокая стабильность CH4 затрудняет его переработку* светло-серым цветом показана анафорическая связь, где *CH4*- антецедент, а местоимение *его* — анафор.

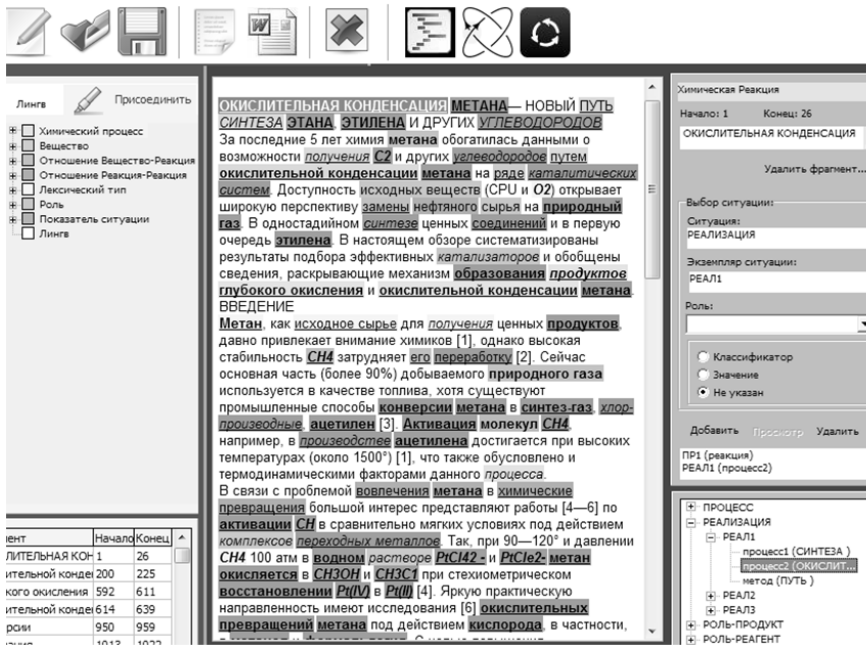

Рис. 2. Система разметки текста

3. Применение размеченного корпуса

Создание языковых ресурсов, ориентированных на автоматическую обработку текста, — довольно трудоемкий процесс, поэтому естественной является попытка автоматизировать их создание и начальное наполнение на основе аннотированного корпуса текстов (в первую очередь имеется в виду экспертная семантическая разметка).

Мы выделили следующие перспективные направления автоматизации процесса создания лингвистических ресурсов:

– Терминологическое наполнение предметных словарей;
– Создание семантико-синтаксических моделей для извлечения фактов из текста.

Первое направление достаточно изучено [5], однако с учетом наличия разрывных фрагментов и многословных терминов требует технологических пояснений. Для второго направления предложены основные идеи и принципы создания подобного ресурса.

3.1. Наполнение терминологических словарей

Наполнение предметного словаря на основе терминологически размеченных фрагментов текста осуществляется в несколько этапов (см. Рис.3).

- Перенос иерархии семантических признаков в словарь и согласование их с уже имеющимися в словаре признаками.
- Обработка текстовых фрагментов морфологическими и синтаксическими компонентами словарной технологии.
- Нормализация и формирование терминов (для многословных фрагментов фиксируется синтаксический шаблон или, если такой шаблон не найден, то фрагмент добавляется как несогласованный значимый словокомплекс).
- Снабжение терминов семантическими признаками в соответствии с размечаемыми признаками.

В ходе наполнения словаря возникают технические моменты, рассмотрение которых требует отдельного внимания:

(1) *Морфологическая и лексическая омонимия*. Поскольку эксперт не осуществляет морфологической и синтаксической разметки, то в словарь добавляются и снабжаются семантическими признаками все омонимы, соответствующие термину.
(2) *Универсальная лексика, входящая в состав словокомплексов* (многословных терминов). В этом случае словарь расширяется универсальной лексикой с пометкой о ее нетерминологичности.
(3) *Несловарная лексика, отсутствующая в универсальном словаре русского языка.* Неизвестные слова могут встречаться в качестве однословного

термина или входить в состав размеченного многословного фрагмента. В данном случае используется предсказание, которое строит гипотезу о принадлежности слова той или иной части речи, а также его морфологических и лексических признаках.

(4) *Разрывный фрагмент* рассматривается как синтаксическая группа, поэтому для него также формируется словокомплекс в соответствии с найденным синтаксическим шаблоном.

(5) *Буквенно-символьные конструкции* не являются элементами словаря, однако являются терминами в данной ПО и образуют синтаксическую и семантическую связь, входя в состав словокомплекса.

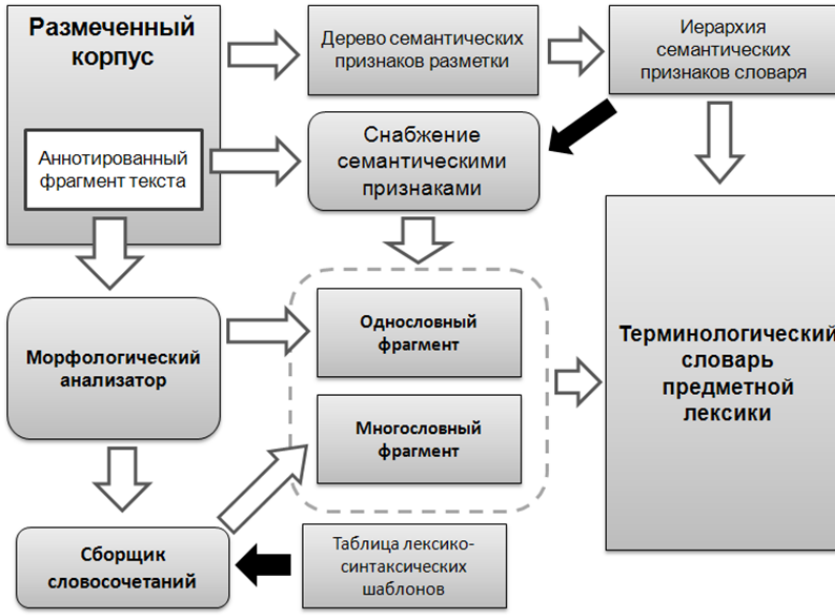

Рис. 3. Схема наполнения терминологических словарей

Наличие автоматически-пополняемого терминологического словаря, в случае, когда пополнение происходит параллельно с работой пользователя, позволяет интерактивно использовать новые термины для автоматизации дальнейшей работы эксперта. Т.е. словарный компонент может осуществлять автоматический поиск новых терминов при последующей разметке текстов. Ошибки, возникающие вследствие того, что либо фрагмент размечен неправильно, либо термин некорректно описывается в словаре, фиксируются и служат основанием для корректировки экспертом.

3.2. Семантико-синтаксические гипотезы

На основе разметки ситуаций или, в общем случае, разметки отношений можно фиксировать синтактико-семантические модели, или шаблоны (в частности, модели управления предикатных лексических единиц). Предполагается, что вначале создаются гипотезы в виде шаблонов, для которых накапливается статистика встречаемости и на основе накопленных данных впоследствии принимается решение о достоверности шаблона. Далее можно применять дополнительные методы обобщения шаблонов или непосредственно предоставлять результаты пользователю для обработки.

В качестве примера такого отношения и шаблонов его распознавания в тексте можно рассмотреть следующий:

Литература

1. *Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л.* и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003–2005. — М.:Индрик, 2005. — С.193–214.
2. *Бирюльцев Е. В., Елизаров А. М., Жильцов Н. Г., Иванов В. В., Невзорова О. А., Соловьев В. Д.* Модель семантического поиска в коллекциях математических документов на основе онтологий // Труды 12й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2010. –Казань, 2010. — С.296–300.
3. *Захаров В. П.* Корпусная лингвистика: Учебно-метод. пособие. — СПб, 2005. — 48с.
4. *Кононенко И. С., Сидорова Е. А.* Система семантической разметки корпуса текстов как инструмент извлечения экспертных знаний (на материале текстов по катализу) // Труды международной конференции «Корпусная лингвистика — 2011». — Санкт-Петербург, 2011. — С. 193–198.
5. *Ляшевская О. Н., Сичинава Д. В., Кобрицов Б. П.* Автоматизация построения словаря на материале массива несловарных словоформ // Браславский П. И. (отв. ред.), Интернет-математика — 2007: сборник работ участников конкурса научных проектов по информационному поиску. –Екатеринбург: Издательство Уральского университета, 2007. –С.118–125.
6. *Яковчук Е. И., Сидорова Е. А.* Обобщенные семантико-синтаксические модели в задачах обработки текста // Труды рабочего семинара «Наукоемкое программное обеспечение НПО-2011». Ершовская конференция по информатике. –Новосибирск: ИСИ СО РАН, 2011. –С.287–292.
7. *Blanco X.* Using NooJ for Multipurpose Analysis of Romance Languages Corpora // Труды международной конференции «Корпусная лингвистика–2008». – СПб., 2008. — С.40–44.
8. *Kim J. D., Ohta T., Tsujii J.* Corpus annotation for mining biomedical events from literature // BMC Bioinformatics. 2008. 9:10

References

1. *Apresjan Ju. D., Boguslavskij I. M., Iomdin B. L.* i dr. Syntactically and semantically annotated Russian corpus: current status and prospects [Sintaksicheski i semanticheski annotirovannyj korpus russkogo jazyka: sovremennoe sostojanie i perspektivy] Nacional'nyj korpus russkogo jazyka: 2003–2005 [Russian National Corpus: 2003–2005]. pp.193–214.
2. *Birjal'cev E. V., Elizarov A. M., Zhil'cov N. G., Ivanov V. V., Nevzorova O. A., Solov'ev V. D.* The model of semantic search in the collections of mathematical documents on the basis of ontologies [Model' semanticheskogo poiska v kollekcijakh matematicheskikh dokumentov na osnove ontologij] Trudy 12j Vserossijskoj nauchnoj konferencii «Èlektronnye biblioteki: perspektivnye metody i tekhnologii, èlektronnye kollekcii» — RCDL'2010 [Proceedings of the 12th Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections"]. Kazan', 2010, pp.296–300.
3. *Zakharov* (2005). Korpusnaia lingvistika [Corpus Linguistics]. Saint Petersburg, 2005, p. 48
4. *Kononenko I. S., Sidorova E. A.* Semantic annotation of text corpus as a means of expert knowledge acquisition (on the material of texts on catalysis) [Sistema semanticheskoj razmetki korpusa tekstov kak instrument izvlechenija èkspertnykh znanij (na materiale tekstov po katalizu)]. Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika — 2011» [Proceedings of the International Conference "Corpus Linguistics — 2011"]. Sankt-Peterburg, 2011, pp. 193–198.
5. *Ljashevskaja O. N., Sichinava D. V., Kobricov B. P.* Automating the construction of the dictionary on the material of an array of word forms not in the dictionary [Avtomatizacija postroenija slovarja na materiale massiva neslovarnykh slovoform] InBraslavskij P. I. (otv. red.), Internet-matematika — 2007: sbornik rabot uchastnikov konkursa nauchnykh proektov po informacionnomu poisku [Internet Math — 2007: a collection of works of participants of research projects competition in Information Retrieval]. Ekaterinburg: Izdatel'stvo Ural'skogo universiteta [Yekaterinburg: Ural University Publishing], 2007, pp.118–125.
6. *Jakovchuk E. I., Sidorova E. A.* Generalized semantic-syntactic models in text processing [Obobshchennye semantiko-sintaksicheskie modeli v zadachakh obrabotki teksta]. Trudy rabochego seminara «Naukoemkoe programmnoe obespechenie NPO-2011» Proceedings of the workshop "Science intensive applied software NPO-2011". Ershovskaja konferencija po informatike [Andrei Ershov Memorial Conference on Informatics]. Novosibirsk: ISI SO RAN, 2011, pp. 287–292.
7. *Blanco X.* Using NooJ for Multipurpose Analysis of Romance Languages Corpora. Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika–2008» [Proceedings of the International Conference "Corpus Linguistics-2008"]. Sankt-Peterburg, 2008, pp.40–44.
8. *Kim J. D., Ohta T., Tsujii J.* Corpus annotation for mining biomedical events from literature. BMC Bioinformatics. 2008, pp. 9–10.