Проекты

mathlingvo

NLP Seminar

- http://www.mathlingvo.ru/nlpseminar
- 1-2 раза в месяц
- Санкт-Петербург, 10-ая линия В.О.
- Видеотрансляция
  - Можно задавать вопросы докладчику
- Видеозапись

NLP Seminar - Весна 2011

- 16 апреля Методы теории графов для использование «слабого» знания. Александр Трусов, IBM
- 9 апреля Определение близости текстов с обучением на основе статистических данных.
- 26 марта Применение моделей глагольного управления и вероятностных правил при морфологической разметке русскоязычных текстов
- 19 марта Использование Гамма распределения при решении задачи классификации
- 5 марта Извлечение мнений из отзывов: простая стратегия, которая работает.

NLPSeminar - Контакты

- Татьяна Ландо

- Лидия Пивоварова

NLP Seminar

- Послушать доклад
  - Очно
  - По трансляции
  - В записи
- Сделать доклад
  - Напишите об этом организаторам

NLP Seminar - Контакты

- http://mathlingvo.ru/nlpseminar
- NLPseminar @ mathlingvo.ru
- http://twitter.com/NLPseminar
- http://groups.google.com/group/nlpseminar
- Tatiana Lando
  - tatiana.lando@gmail.com
- Lidia Pivovarova
  - lidia.pivovarova@gmail.com

OpenCorpora

- ПО для создания корпусов текстов с лингвистической разметкой

  code.google.com/p/opencorpora/

- Корпус текстов на русском языке

  OpenCorpora.org

OpenCorpora

- Уровни разметки:

- Метатекстовая
- Графематическая
- Морфологическая
  - Снятие омонимии (скоро)
- Синтаксическая, ... (когда-нибудь)
- «Типографская» *

OpenCorpora

- Морфологический словарь
  - Словарная база АОТ
  - Формализованная морф. модель + проверка качества
- 58 тыс. словоупотреблений *(на 10.07.11)*
  - Метатекстовая
  - Графематическая
  - Морфологическая (без снятия омонимии)
- Источники
  - «Частный корреспондент»
  - «Википедия»

OpenCorpora

- Тексты, разметку, код можно использовать
  - на условиях CC-BY-SA
  - скачать, бесплатно :)
- Можно присоединиться к работе
  - в том числе дистанционно
- Можно оценивать качество
  - в том числе независимо от авторов

OpenCorpora

- CC-BY-SA (Creative Commons)
  - Attribution
    - надо указывать авторство
  - Share Alike
    - надо публиковать производные произведения под такой же лицензией

OpenCorpora.org

OpenCorpora - задачи

- Добавление и разметка текстов
  - Разделение текстов на предложения и на слова
  - *Снятие морфологической омонимии**
- Разработка стандартов разметки
  - Морфология
  - Синтаксис (NE, локальный синтаксис, ...)
- Пополнение словаря
  - Новые слова
  - Связи между словами
  - Проверка словаря

OpenCorpora - задачи

- Разработка ПО (не лингвистика)
  - Сервер — PHP + MySql
  - Интерфейсы — PHP + JavaScript
    - Редактор словаря
    - Редактор синтаксической разметки
  - Статистика, экспорт — Perl, Python, C++, …
  - API — C++, Perl, Python, Java, …
    - API словаря
    - API токенизатора
    - Интеграция API в GATE

OpenCorpora - задачи

- Разработка ПО (лингвистика)
  - Поиск ошибок в разметке
  - Разделение на предложения
  - Установление связей в словаре
  - Снятие морфологической омонимии
  - Выделение именованных сущностей
  - Локальный синтаксис
  - «Большой синтаксис»
  - ...

OpenCorpora

- На сайте http://opencorpora.org
  - Статистика
  - Публикации
  - Корпус
- На сайте http://code.google.com/p/opencorpora
  - Исходный код
  - Обсуждение разработки
- На сайте http://mathlingvo.ru/NLPSeminar
  - Видеозапись лекции (май 2011)

OpenCorpora

- Василий Алексеев (интерфейсы)
- Светлана Бичинёва (лингвистика, стандарты, словарь)
  bichineva@opencorpora.org
- Виктор Бочаров (орг. руководство)
  bocharov@opencorpora.org
- Дмитрий Грановский (сервер, тех. руководство)
  granovsky@opencorpora.org
- Мария Николаева (интерфейсы)
- Наталья Остапук (лингвистика, словарь)
- Мария Степанова (лингвистика, словарь)