Об одном статистическом методе пополнения морфологического словаря.

Черненьков Д. М.

Проблемы и задачи

- **Цель:** предсказание полных парадигм несловарной лексики, встречающейся в неразмеченном корпусе.
- **Исходные данные:** малый, неразмеченный корпус
- **Проблемы**
  - Оценка качества и анализ результатов
  - Большое число гипотез
  - Большое число признаков

Исходные данные

- Корпус: новости РБК 2003-2008, объем более 17,3 млн. словоупотреблений; неразмечен.
- Словарь системы “Crosslator 2.0”, 160 тыс. слов

Метод

- Разделение словаря на анализирующий и проверяющий
- Фильтрация гипотез по фиксированным порогам
- Кластеризация гипотез по леммам
- Машинно-обучаемый отбор гипотез

Сокращение числа гипотез

- Слияние парадигм
- Отсечение гипотез до кластеризации
- Отсечение гипотез после кластеризации

Выбор гипотез из кластеров

- Списковое ранжирование
- Оценочная функция:
  - Max log likelihood
  - Линейная комбинация признаков
  - Градиентный спуск
  - Пошаговый отбор признаков

Признаки

- Частотность словоформ гипотезы
- Число встреченных словоформ гипотезы
- Число словоформ в словаре с совпадающим постфиксом длины n и теми же грамматическими параметрами
- Частотность грамматических униграмм
- Частотность грамматических биграмм

Результаты

Число отобранных гипотез	Процент полностью угаданных парадигм
1	37%
2	49%
3	55%
4	59%
5	60%
6	60%
7	61%
8	62%
9	62%
10	62%

Наиболее значимые признаки

- Частоты биграммов (левый и правый контекст)
- Число найденных различных словоформ