ABBYY®

Автоматический поиск переводных словосочетаний

Новицкий Валерий,
компания ABBYY

Цели и задачи

- Задача:
  - Поиск переводных словосочетаний (словосочетаний и их переводов на другой язык) по корпусу выровненных параллельных текстов
- Цели:
  - Разработка алгоритма получения словосочетаний (с учётом ряда специфических требований)
  - Получение статистических данных для улучшения работы синтаксического анализатора
  - Расширение переводного словаря

Дополнительные требования и исходные данные

- Требования к получаемым словосочетаниям:
  - Синтаксическая связанность
  - Размер от 1 до 5 слов
  - Устойчивый перевод
  - Целостность (словосочетание не является частью другого, более полного словосочетания)
  - И т.д.

- Исходные данные и внешние механизмы
  - Корпус выровненных параллельных текстов
  - Синтаксический анализатор
  - Механизм пословного сопоставления синтаксических структур

Схема алгоритма

Фильтрация

- Задача: убрать случайные словосочетания
- Этапы фильтрации:
  - Предварительное удаление низкочастотных словосочетаний
  - Удаление вложенных/«внешних» словосочетаний
  - Разрешение неоднозначности перевода
  - Удаление известных (словарных) переводов
  - Финишная фильтрация по частоте
  - Сортировка результатов на новые словарные статьи и собственно переводные словосочетания

Результаты

- Корпус: ~4,2 млн. фрагментов
- На выходе: ~62 млн. уникальных словосочетаний
- После фильтрации: ~42 тыс. переводных словосочетаний

- Оценку полноты произвести затруднительно
- Оценка точности полученных результатов экспертом по выборке 100 случайных словосочетаний:
  - Хорошие – 67
  - Недостатки описания – 4
  - Недоработки алгоритма – 16
  - Другие – 12