Семантическое автодополнение

Сергей Василинец
Александр Удалов

Научный руководитель: Хитров Д.В.

Цель работы

Предыстория:

Набор текста на устройствах с тачскринами медлен и неудобен. Что делать?

Цели

- Исследование методов автодополнения
- Разработка системы автодополнения

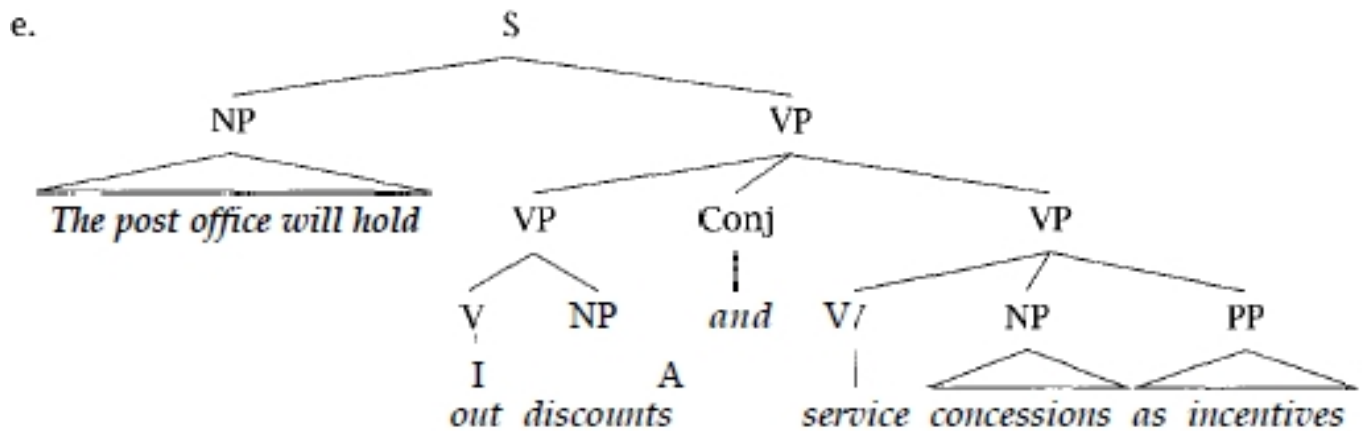
Методы

- Наивный
- Грамматический
- Statistical Inference: n-gram models over sparse data

Грамматический анализ

Этапы:

- Синтаксический анализ
  - Выделение подлежащего, сказуемого и т.п.
  - Возможные роли следующего слова



- Семантический анализ: категории слов

Statistical Inference

Подзадачи:

- Создание корпуса
  - Сбор статистики использования слов и словосочетаний
- Выбор оценочной функции
  - $p(w_1,....,w_n) = r$. Вероятность встретить подряд идущими слова $w_1\ w_2\ ...\ w_n$

Оценочные функции:

- Статистические
- Смешивающие

Наша система

- Statistical Inference
- Свободный корпус от Google Labs
- Пользовательская статистика
  - Отдельный корпус каждого пользователя
- Вероятные другие источники
  - Twitter

Пользовательская статистика

- Ежедневный словарный запас среднестатического человека невелик
- В его речи есть сленг и слова, характерные для узкого круга лиц

Статистика из Twitter

- Люди обсуждают мировые события
- Статистика, собранная из социальных сетей, должна выявить актуальные словосочетания

Ожидаемый результат: у автодополнений появляется временной контекст

Разочарование в Twitter Trends

Вывод: Большую часть времени люди пишут в Twitter бесполезные вещи.

Результаты тестов

- Литературный текст
- Угадывание следующего слова по 3 буквам:
  - первое предложенное: 55%
  - среди первых трёх: 73%
- По 1 букве:
  - первое: 38%
  - среди первых трёх: 56%
- 7-буквенное слово угадывается по 4 буквам в среднем
- 10-буквенное по 5.7 буквам в среднем

Пути улучшения и развития

- Источники для корпуса (форумы, чаты)
- Уточнение констант
- 4-граммы
- Создание андроид приложения на основе созданного веб-сервиса