Очистка данных для тренировки моделей машинного перевода

Обработка параллельных корпусов

Задача

Параллельные корпуса

- Нет проблемы сбора данных
- Большие объемы (годы работы), 60-150k пар
- Накапливается мусор
- Невозможно чистить вручную

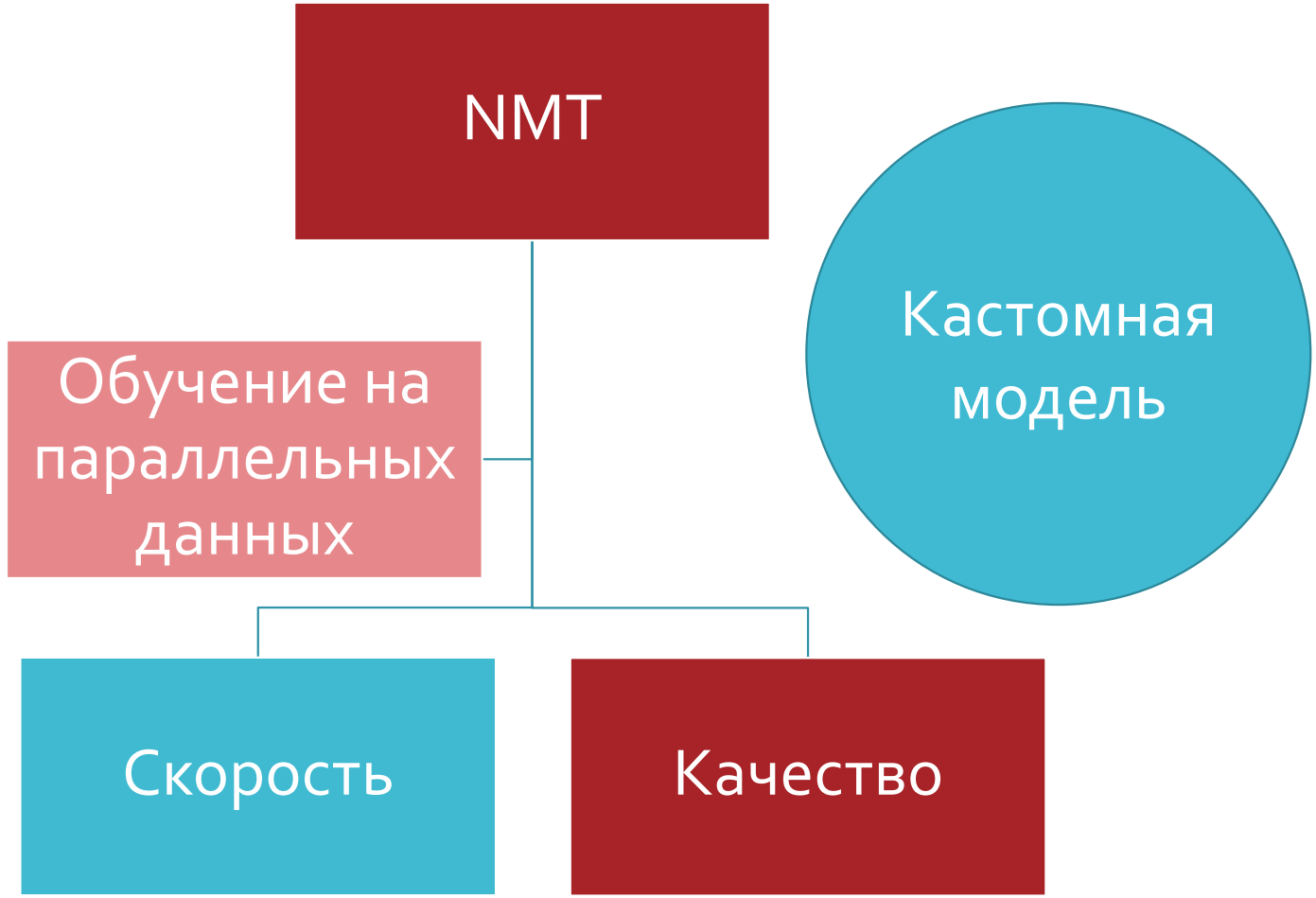
Задача

Идея

- Существующие инструменты не дают гибкости либо не рассчитаны на большие объемы данных
  - Только полные дубликаты
  - Недостаточно опций
- Возможность дальнейшего масштабирования для решения более сложных задач:
  - Более глубокая очистка данных для NMT
  - Извлечение терминологии
  - Составление двуязычных глоссариев

Задача

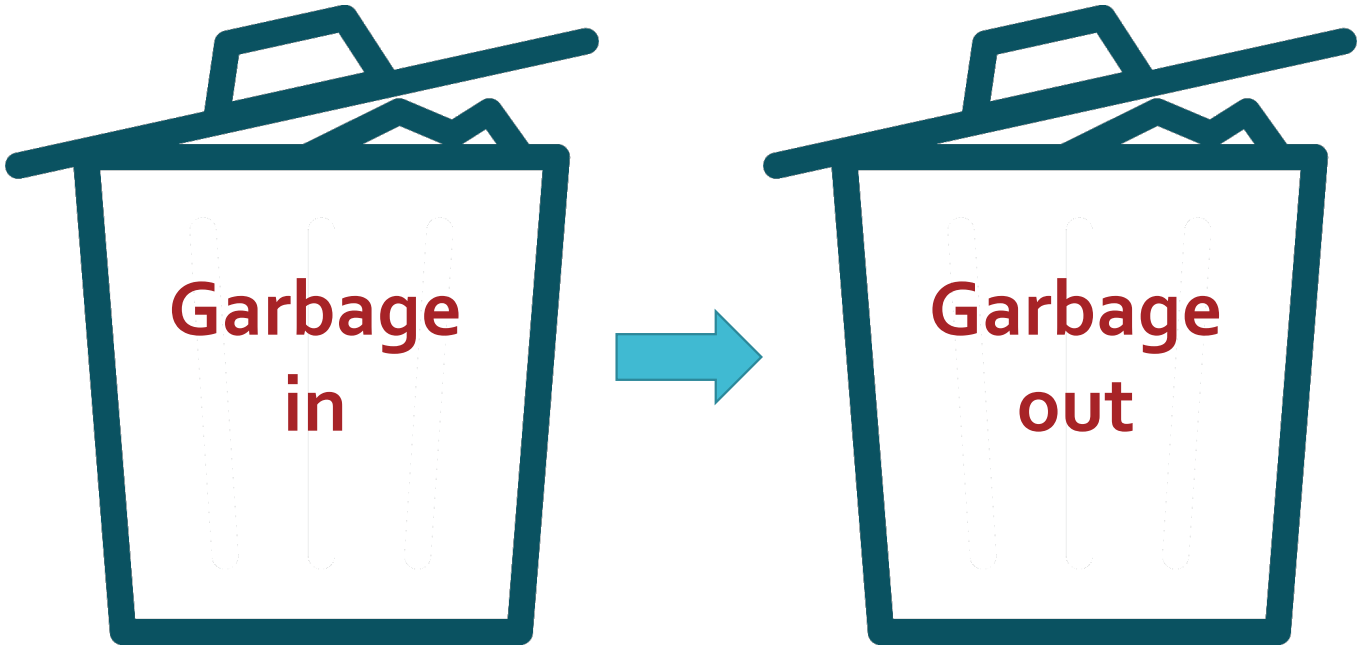
Нейронный машинный перевод



Задача

Обучение на параллельном корпусе

Quality over quantity



Задача

Очистка данных

1. **Удаление дубликатов**, в том числе неполных:
   - Различия в датах
   - Номера телефонов в разных форматах
   - Наличие разных ссылок
   - Разные значения полей
   - Лишние символы (пробелы, мягкие переносы)

Задача

Очистка данных

2. **Удаление мусора**:
   - Сегменты, состоящие из символов или цифр
   - Чрезмерно длинные предложения
   - Сегменты на языке, отличном от целевого

Задача

Очистка данных

3. **Устранение ошибок ввода**:
   - Опечатки в исходном тексте, которые приводят к появлению дубликата
   - Опечатки в переводе

Задача

Очистка данных

4. **Анонимизация**:
   - Имена (NER)
   - Номера телефонов
   - Адреса электронной почты
   - Другая идентифицирующая информация

Задача

Очистка данных

5. **Устранение несоответствий**:
   - Несоответствия в параллельных корпусах (поиск коллокаций и кандидатов на их перевод)

Подход

Этапы

1. **Анализ исходных параллельных корпусов**

Формат TMX — по сути XML.

```
<tu creationdate="20161221T125309Z" creationid="SONY-S\Svetlana"
changedate="20161221T130023Z" changeid="SONY-S\Svetlana" lastusagedate="
20161221T130023Z">
  <prop type="x-LastUsedBy">SONY-S\Svetlana</prop>
  <prop type="x-Context">8141944612297908607, 15989695196343197l6</prop>
  <prop type="x-ContextContent">FPGA-based, specifically: |  | На базе
  FPGA, а именно: | </prop>
  <prop type="x-Origin">TM</prop>
  <prop type="x-ConfirmationLevel">Translated</prop>
  <tuv xml:lang="en-US">
    <seg>x86/i64-based processor, specifically:</seg>
  </tuv>
  <tuv xml:lang="ru-RU">
    <seg>На базе процессора x86/i64, а именно:</seg>
  </tuv>
</tu>
```

Подход

Этапы

2. **Общий подход к решению задачи**

- Регулярные выражения для поиска данных
- Предварительная обработка:
  - ✓ Нижний регистр
  - ✓ Пунктуация (отдельная пунктуация в числовых данных)
  - ✓ Удаление незначащего текста в начале и конце строк
  - ✓ Замена незначащих для сравнения данных на символы
- Сравнение > построение индекса (словарь), множество

Подход

Этапы

3. **Взаимодействие с пользователем**

- Создание трех файлов:
  - Файл с отфильтрованной памятью
  - Файл с удаленными сегментами
  - Файл отчета
- Вывод прогресса для пользователя в консоль
- Запуск пользователем скрипта в терминале с указанием пути к файлу и параметров

Подход

Этапы

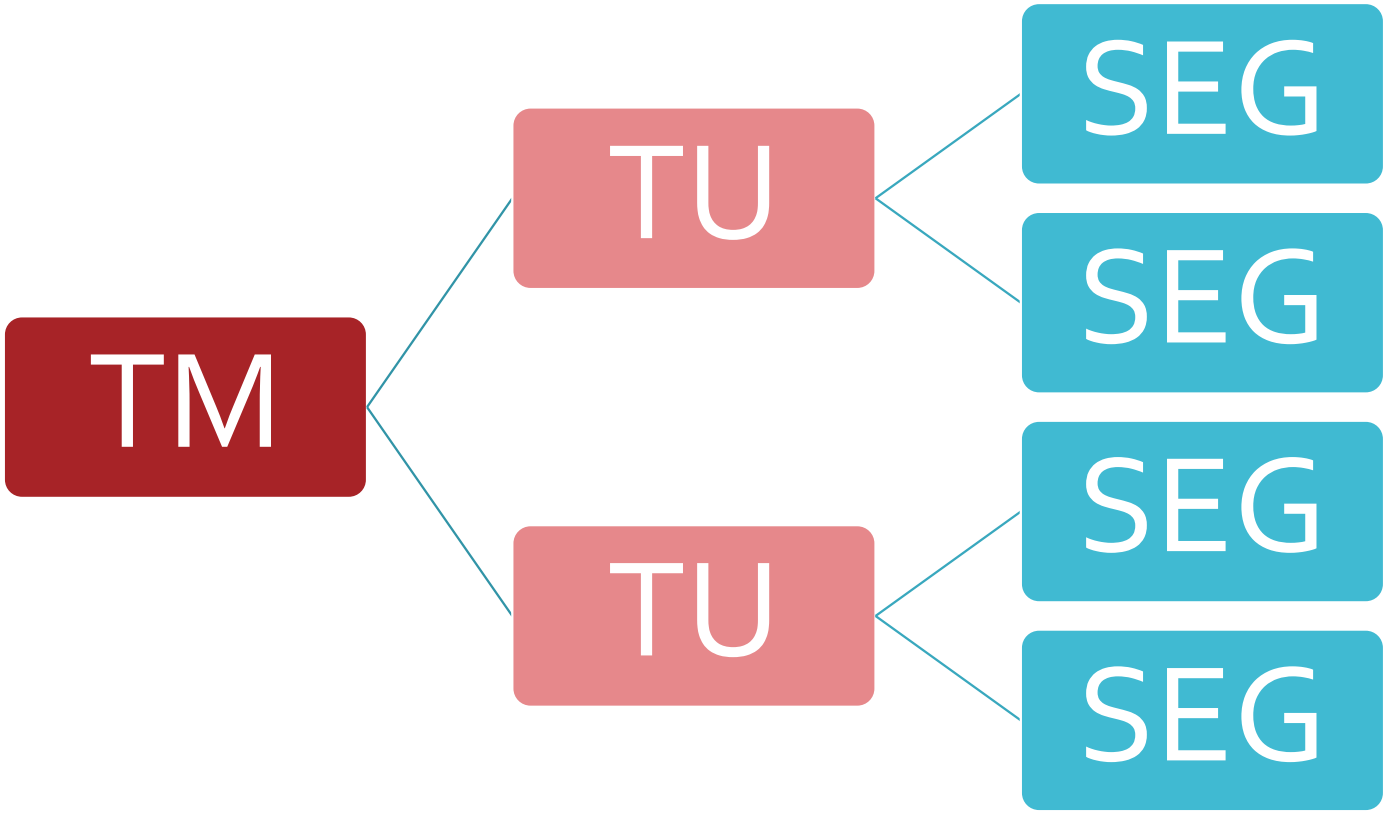
4. **Разработка**

- Функциональный метод
  - ✓ Регулярные выражения на максималках
  - ✓ Функция для предварительной обработки текста
  - ✓ Функция для замены незначащего текста

  Код можно посмотреть тут
- Объектно-ориентированный метод

Подход

4. **Разработка (с куратором)**



Подход

4. Разработка (с куратором)

Сложности

Основные сложности

1. Комплексная архитектура, много связей
2. Продумывание и обработка всевозможных исключений
3. Запись в XML
4. Запись в HTML
5. Отладка в ООП
   - Специальный параметр отладки
   - Запись особого тега в файл для проверки

Следующий этап

Что дальше?

1. Расширение интерфейса (опции критериев удаления)
2. Удаление дубликатов с опечатками (расстояние 1 символ)
3. Поиск опечаток в переводе
4. Извлечение терминологии
5. Составление двуязычных глоссариев…

Изученные материалы

Классный учебник по Python

- Справочник по языку Python3

Работа с XML-файлами в Python

- Создание и сборка XML-документов

Регулярные выражения

- Регулярные выражения - от простого к сложному

Интерфейс командной строки (аргументы для скрипта)

- Модуль argparse

Вывод информации в терминал

- Объекты stdin, stdout, stderr модуля sys в Python