ЛИНГВИСТИЧЕСКИЙ ПАЙПЛАЙН И ЗАДАЧИ NLP

Маша Шеянова, masha.shejanova@gmail.com

October 26, 2018

НИУ ВШЭ

Лингвистический пайплайн

Что такое лингвистический пайплайн?

Итак, мы автоматически обрабатываем язык.
Что бы мы ни делали, наши действия делятся на некоторые стандартные шаги, которые выстраиваются в общий процесс.

Этот процесс и называется Pipeline.

(Pipeline — конвейер, последовательность действий, применяемых одно за другим).

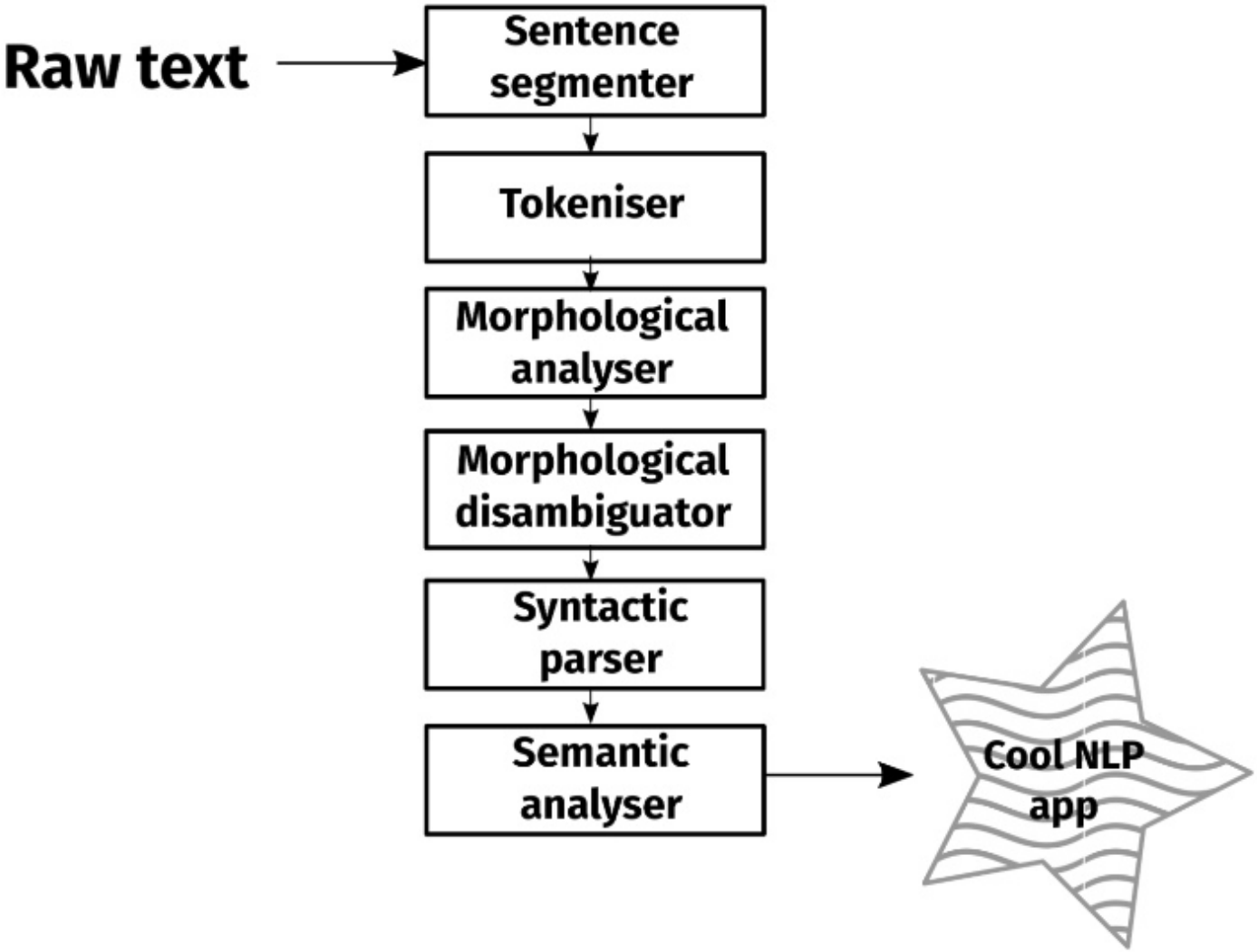
ШАГИ ПАЙПЛАЙНА

От низкоуровневых к высокоуровневым:

- сегментация
- токенизация
- морфологический анализ
- морфологическая дизамбигуация (разрешение неоднозначности)
- синтаксический анализ
- семантический анализ

Не для каждой задачи нужны все шаги, но почти каждый следующий шаг требует предыдущий.

ШАГИ ПАЙПЛАЙНА



СЕГМЕНТАЦИЯ ТЕКСТА

(токенизация предложений, определение границ предложений)
У нас есть ”сырой” (не обработанный) текст.
Наша цель: **получить список предложений**.
Необходима для некторых следующих этапов пайплайна (синтаксический, семантический анализ).
Примеры приложения:

- машинный перевод
- создание параллельных корпусов (тексты на двух языках, выровненные по проедложениям)

Сегментация текста. Как выделить предложения?

Наивный способ: поделить весь текст по знакам препинания, стоящим в конце предложения (.!?)
Главная проблема: точка как конец предложения vs. в середине предложения.
Например:

- сокращения: г. Москва
- внутри дат: 23.05.2018
- внутри пунктов: согласно п. 2.1.13
- супер проблема: сокращение в конце предложения

ТОКЕНИЗАЦИЯ

У нас есть ”сырой” (не обработанный) текст.
Наша цель: **получить список слов (токенов)**.

Токен — в первую очередь слово, но к ним также относятся знаки препинания, даты и прочие сегменты предложения.

Необходима для практически всех[1] лингвистических задач и всех следующих шагов пайплайна.

[1] кроме тех, где задача решается на уровне символов

Токенизация. Как выделить токены?

Наивный способ: кусок строки от пробела до пробела.
Дает нормальное качество, но обычно нужно лучше. **Чем хуже токенизация, тем хуже работает вся система в целом.**
Проблемы:

- знаки препинания – удалить, оставить?
- сокращения и другие апострофы (don’t, we’re, Smith’s)
- дефисы (Санкт-Петербург vs мальчик-программист)
- составные предлоги (в течение, не работает)
- и многие, многие другие детали

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ

У нас есть: текст, поделённый на токены.
Наша цель:

- каждый токен привести к начальной форме (ветров -> ветра); это называется **лемматизация**
- выяснить, к какой части речи он относится (существительное, глагол); это называется **POS-tagging**
- определить его грамматические характеристики (падеж, число, время)

Проблемы: тысячи их. Это вообще очень нетривиальная задача.
Применение:

- лемматизация — практически везде
- анализ — задачи, где грамматичиские признаки значимы

МОРФОЛОГИЧЕСКАЯ ДИЗАМБИГУАЦИЯ

Главная проблема морфологического анализа — омонимия:

```
lemma: сорок, analysis: NUMR loct, score: 0.285714
lemma: сорока, analysis: NOUN,inan,femn sing,nomn,
score: 0.142857
lemma: сорока, analysis: NOUN,anim,femn sing,nomn,
score: 0.142857
lemma: сорок, analysis: NUMR gent, score: 0.142857
lemma: сорок, analysis: NUMR datv, score: 0.142857
lemma: сорок, analysis: NUMR ablt, score: 0.142857
```

МОРФОЛОГИЧЕСКАЯ ДИЗАМБИГУАЦИЯ

У нас есть: токены с кандидатами на анализ.
Наша цель: выбрать правильный анализ для каждого слова.

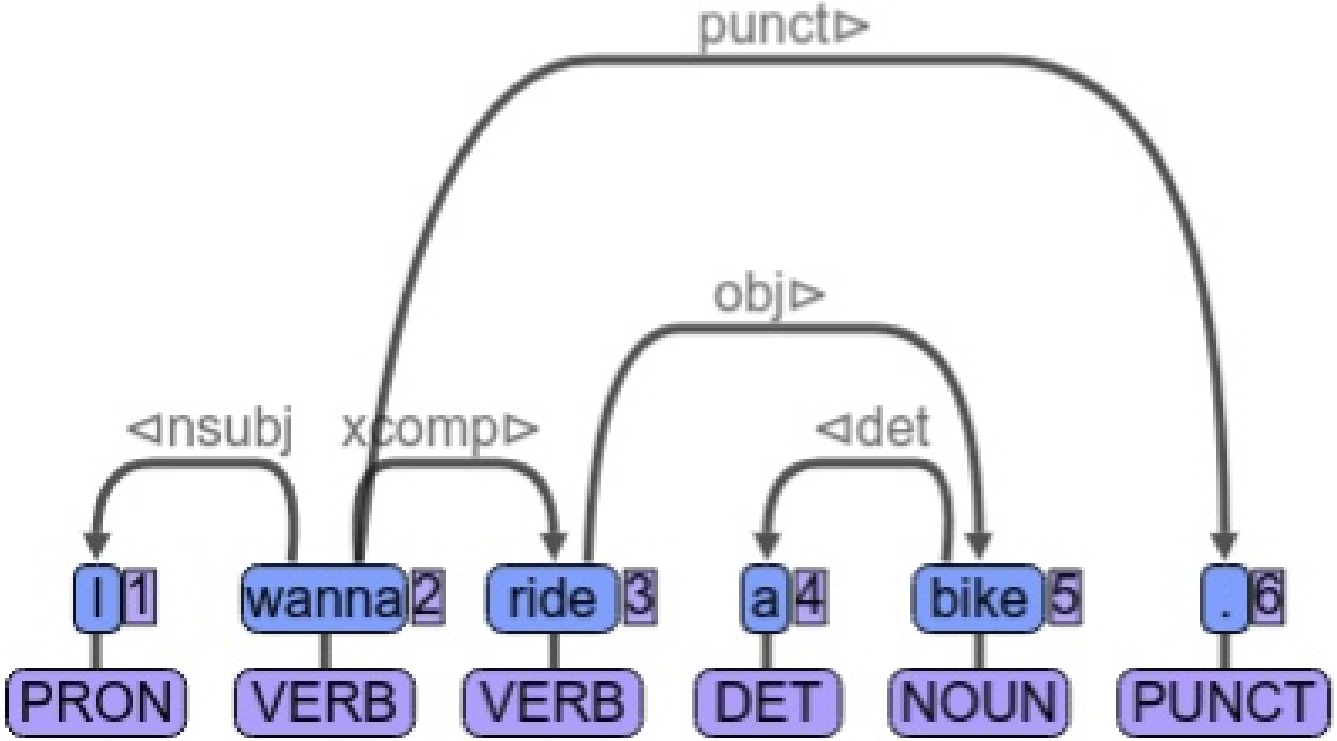
Подходы:

- основанные только на частотности разбора (e.g. pymorphy)
- основанные на контексте (e.g. mystem)

Морфологическая дизамбигуация — необходимый шаг после анализа, обычно морфологический анализ подразумевает и дизамбигуацию тоже.

СИНТАКСИЧЕСКИЙ ПАРСИНГ

У нас есть: результат работы предыдущих шагов пайплайна.
Наша цель (в теории зависимостей): для каждого слова найти слово, от которого оно зависит и какой тип связи.
То есть, построить **синтаксическое дерево**:



Семантический анализ. Word sense disambiguation

Как известно, неоднозначность бывает не только морфологической, но и семантической, ”по смыслу”.

Пример: велосипедный звонок vs. звонок с урока vs. Звонок (фильм, книга).

У нас есть: текст, побитый на токены, лемматизированный, возможно как-то ещё обработанный.

Наша цель: для каждого слова в тексте указать, в каком значении оно употреблено. Пример применения — поиск документов по запросу.

Семантический анализ. Semantic role labeling

У нас есть: результат работы предыдущих шагов пайплайна.
Наша цель: для каждого объекта в предложении указать, какую **семантическую роль** (например, действующее лицо, инструмент, цель и т.д.) он играет.

Применяется достаточно редко, потому что задача непростая, плохо формализуется, сложно добиться хорошего качества.

ЗАДАЧИ NLP

ПОВТОРЕНИЕ

Какие задачи вы помните с прошлой лекции?

ЧТО ОБСУЖДАЛОСЬ

- информационный поиск (information retrieval)
- аггрегация новостей
- анализ тональности (sentiment analysis)

SEQUENCE LABELLING

ПОСТАНОВКА ЗАДАЧИ

Sequence labelling:

- Дан набор текстов
- Каждый текст представляет собой последовательность токенов
- Каждому токену присвоена метка из некоторого множества

В зависимости от множества меток получаем разные типы подзадач:

- для частей речи — POS-теггинг (часть морфологического анализа)
- для типов именованных сущностей — NER

NAMED ENTITY RECOGNITION

… или извлечение именованных сущностей.
Задача: **достать из текста все упоминания интересующих нас сущностей**. Сюда входят:

- имена людей
- названия мест
- даты
- суммы денег
- наименования организаций
- и многое другое, в зависимости от того, какая у нас цель

Зачем: например, системы принятия решений.

Машинный перевод

Зачем?

А вы как думаете?

ЗАЧЕМ?

НЕ наша цель:

- красивый художественный перевод
- перевод важных переговоров

Наша цель:

- перевести сайт, на который я зашёл
- быстро прочитать пришедший e-mail
- **помочь** переводчику не тратить время на очевидные части

Какие подходы бывают?

- Основанные на корпусах:
  - **Статистический** (SBMT — Statistical Machine Translation)
  - **Нейронный** (NMT — Neural Machine Translation)
  - **Example-based** (EBMT — Example-Based Machine Translation)
- **Правиловый** (RBMT — Rule-Based Machine Translation). Использует лингвистические знания человека для создания адекватной языковой модели.
- **Гибридные** (HMT — Hybrid Machine Translation). Не один подход, а разнородный кластер.

Правиловый машинный перевод. Пословный.

Правиловый перевод подразделяется на:

- **Dictionary-based** (direct) — прямой, пословный перевод
- **Interlingua** — с промежуточным представлением
- **Transfer** — два промежуточных уровня

Dictionary-based method – наивный подход. Использует прямые словарные соответствия между исходным и целевым языками. Не учитывает грамматическую структуру текста. Самый ранний.

ПРАВИЛОВЫЙ МАШИННЫЙ ПЕРЕВОД. INTERLINGUA.

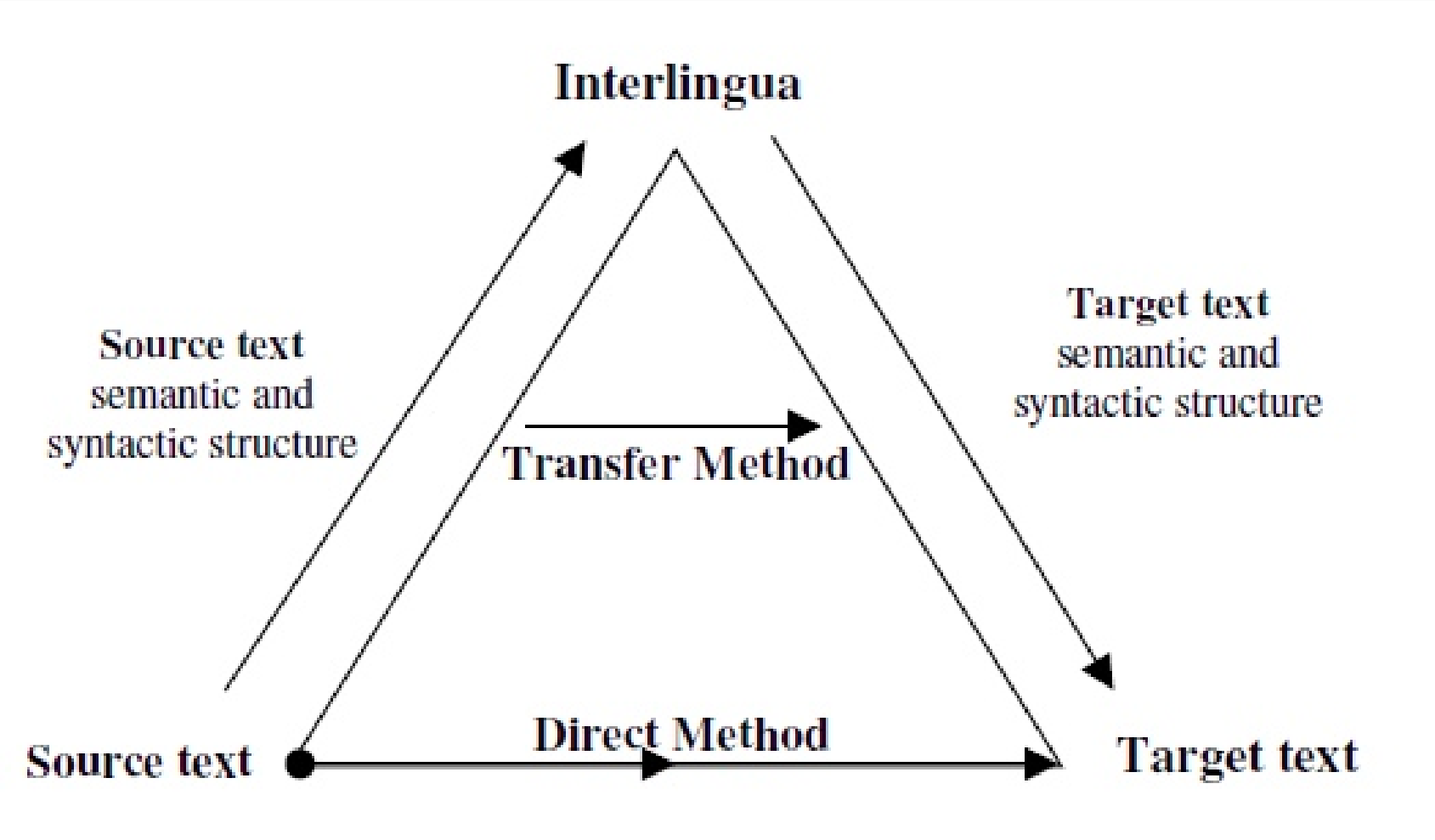
- использует **абстрактное глубинное представление** (интерлингву), не привязанное к конкретному языку
- основан на модели **Смысл** $\Leftrightarrow$ **Текст**, разработанной лингвистом Мельчуком
- хорош для **многоязыковых** (multilingual) систем

ПРАВИЛОВЫЙ МАШИННЫЙ ПЕРЕВОД. TRANSFER.

Transfer method: текст сначала преобразуется в проекцию, близкую к исходному языку, затем из неё – в проекцию, ориентированную на целевой язык. Бывает:

- **deep transfer**: каждое предложение имеет дерево разбора;
- **shallow transfer**: оперирует частями предложения (chunks).

ПРАВИЛОВЫЙ МАШИННЫЙ ПЕРЕВОД



CORPUS-BASED.

У нас есть параллельные корпуса:

Английский	Японский
How much is that red umbrella?	Ano akai kasa wa ikura desu ka.
How much is that small camera?	Ano chiisai kamera wa ikura desu ka.

С их помощью мы учим компьютер переводить предложения пользователя.

- **Статистический**
- **Нейронный**
- **Example-based**. Очень редкий. Не использует статистику. Переводы строятся на основе пропорциональных аналогий.

Статистический машинный перевод

Допустим, мы переводим строку A с исходного языка и хотим получить стрроку B — перевод. Максимизируем две вероятности:

1. что строка B является переводом строки A
2. что строка B появилась в целевом языке (языковая модель)

Для первого нам нужен **параллельный корпус**.
Для второго — корпус **целевого языка**.

Статистический машинный перевод

Плюсы:

- хорошо запоминает редкие и сложные слова и фразы, если они встречались в параллельных текстах
- в отличие от правилового, не требует у разработчиков знания о языке!
- в отличие от нейронного, не требует таких больших вычислительных мощностей

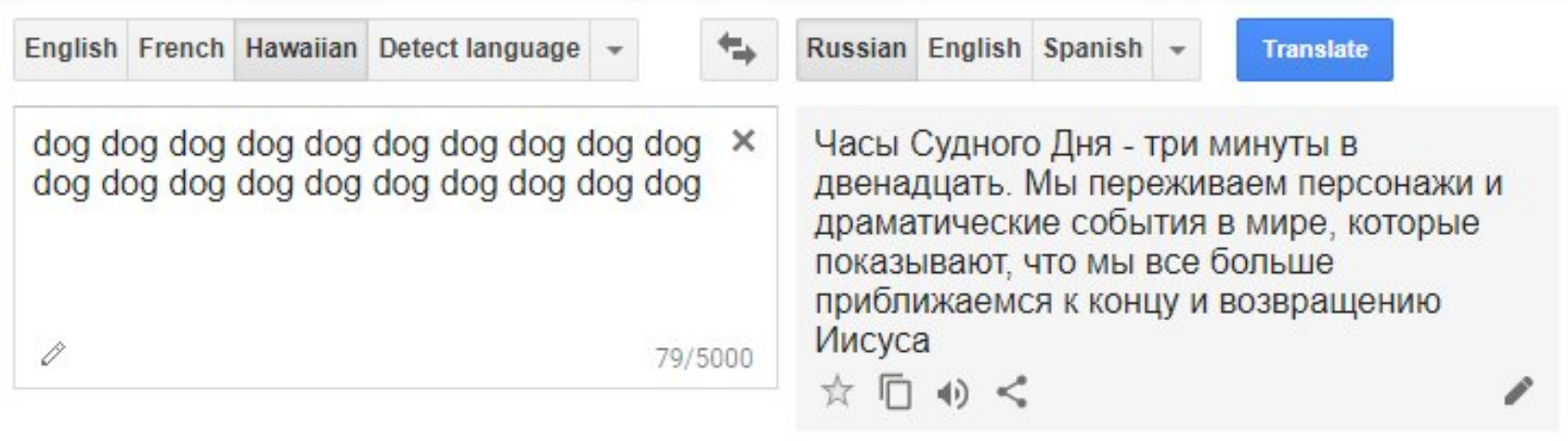
Минусы:

- результат перевода бывает похож на собранный пазл: связь между началом и концом может теряться
- если данных в корпусе не было, перевод будет странноватым

НЕЙРОННЫЙ МАШИННЫЙ ПЕРЕВОД

- самые крутые сервисы сейчас работают на нём!
- тоже анализирует массив параллельных текстов и учится находить в них закономерности
- но работает не со словами и фразами, а с предложениями
- двунаправленные рекуррентные нейронные сети (RNN)
- в отличие от статистического, картинка гораздо более сглаженная
- может выдавать странные вещи на данных, которых никогда не видела

НЕЙРОННЫЙ МАШИННЫЙ ПЕРЕВОД



Машинный перевод. Rule-based vs. corpus-based.

Corpus-based:

- широко используется сейчас (Google, Яндекс)
- требует параллельные корпуса: чем больше, тем лучше
- в принципе, не требует лингвистических знаний

Rule-based:

- сейчас всё больше уступает статистическому, НО
- может применяться при отсутствии больших корпусов → можно работать с малыми языками!
- их можно постепенно улучшать
- требует лингвистических знаний

Гибридные подходы.
Что можно сделать?

ГИБРИДНЫЕ ПОДХОДЫ

Делится на две большие группы:

- Multi-engine: применяется одновременно несколько подходов, результат сравнивается, выбирается лучший кандидат (устраивается голосование).
- Single-engine: разные методы применяются в разных частях системы

Гибридные подходы. Single-engine.

Делится на две большие группы.

- **Статистический** перевод, модифицированный правилами. Пример: использовать знания о морфологии, о синтаксисе для пост-обработки текста.
- **Правиловый** перевод, использующий статистические методы. Примеры:
  - предобработка (POS-тэггинг, синтаксический анализ)
  - взвешивание правил
  - выбор кандидатов на правильный перевод

Спасибо за внимание!

Вопросы?