МАШИННЫЙ ПЕРЕВОД

Маша Шеянова, masha.shejanova@gmail.com

August 6, 2018

НИУ ВШЭ

INTRO

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.

ПОДХОДЫ И ИСТОРИЯ

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.

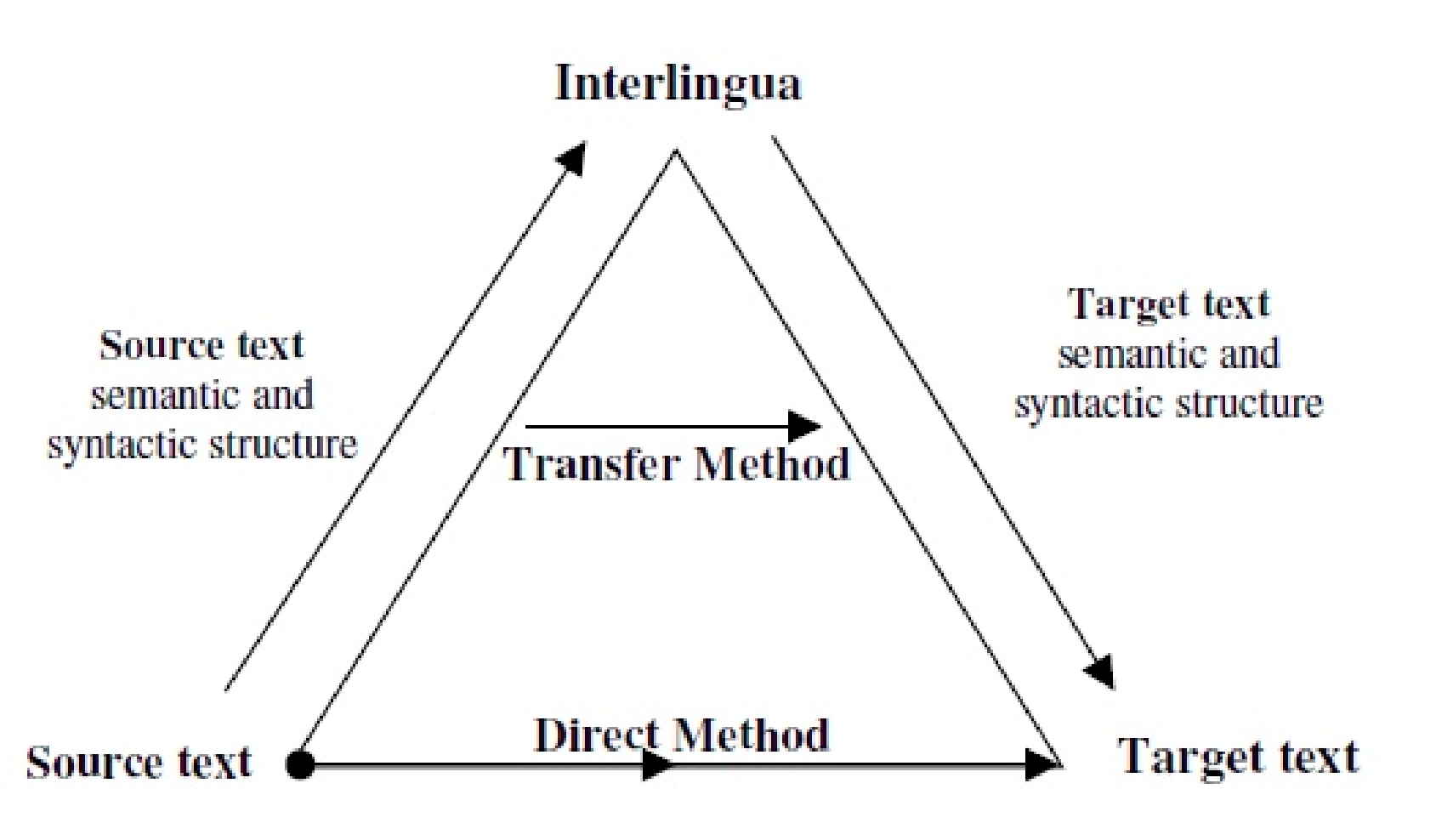
- использует **абстрактное глубинное представление** (интерлингву), не привязанное к конкретному языку
- основан на модели **Смысл** $\Leftrightarrow$ **Текст**, разработанной лингвистом Мельчуком
- хорош для **многоязыковых** (multilingual) систем

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 A 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

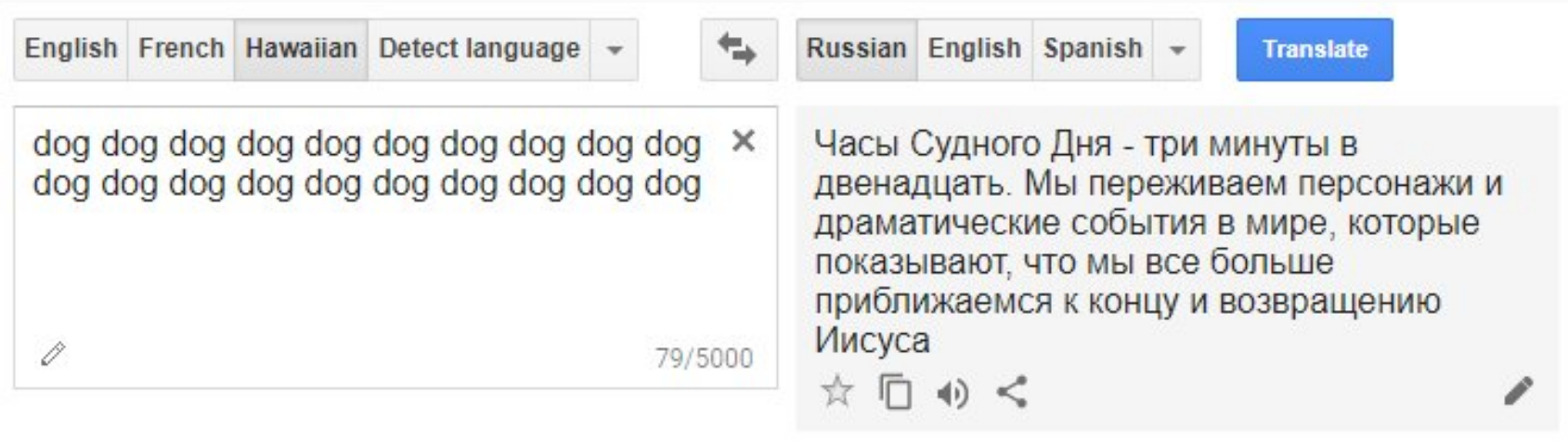
Минусы:

- 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

(очень) краткая история

- Первые идеи – 30-е годы XX века.
- Первый расцвет – середина XX века.
- 60-е годы XX века – разгромная критика ALPAC, разочарование и спад активности.
- 90-е годы: начало расцвета статистического перевода.
- с нулевых годов: разнообразие систем, общая доступность, преобладают статистический и смешанные подходы.
- настоящее время: разнообразие систем, общая доступность, преобладают статистический и смешанные подходы.

ОЦЕНКА

ОЦЕНКА

Ручная оценка, например, с помощью **round-trip translation**: переводим текст туда и обратно, смотрим что получилось.

Автоматическая оценка. Самые популярные:

- **BLEU**: чем ближе перевод к переводу профессионала, тем лучше
- **NIST**: модифицированная BLEU
- **Word error rate**: переводим текст, а потом правим руками до нормального перевода; основана на расстоянии Левенштейна

Примеры систем

GOOGLE TRANSLATE

- пожалуй, самый популярный сейчас
- разновидность: **нейронный**
- лучше всего работает перевод с английского и на английский
- есть API

ЯНДЕКС.ПЕРЕВОДЧИК

- тоже очень популярен, особенно для языков России
- Разновидность: **нейронный**
- предоставляет перевод на 95 языков
- есть API

APERTIUM

- малоизвестен, но для некоторых языковых пар превосходиит Google и Яндекс
- разновидность: **правиловый, shallow transfer**
- лучше всего работает для малых языков, особенно — для родственных
- open-source! =)

Спасибо за внимание!

Вопросы?