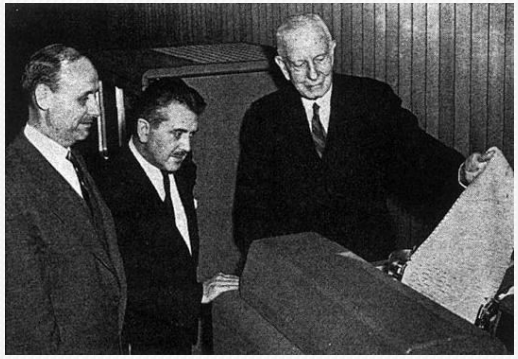
PROMT®

Как улучшить современный машинный перевод?

Понятие и история машинного перевода

Машинный перевод (МП) - процесс автоматического перевода текстов с одного ЕЯ на другой с помощью компьютерных алгоритмов.



Джорджтаунский эксперимент

- 7 января 1954 года, Нью-Йорк, штаб-квартира IBM
- 250 словарных записей,

6 грамматических правил перевода

- более 60 предложений

(с русского на английский)

Сегодня

- Множество компаний и систем МП
- Сотни тысяч словарных записей,

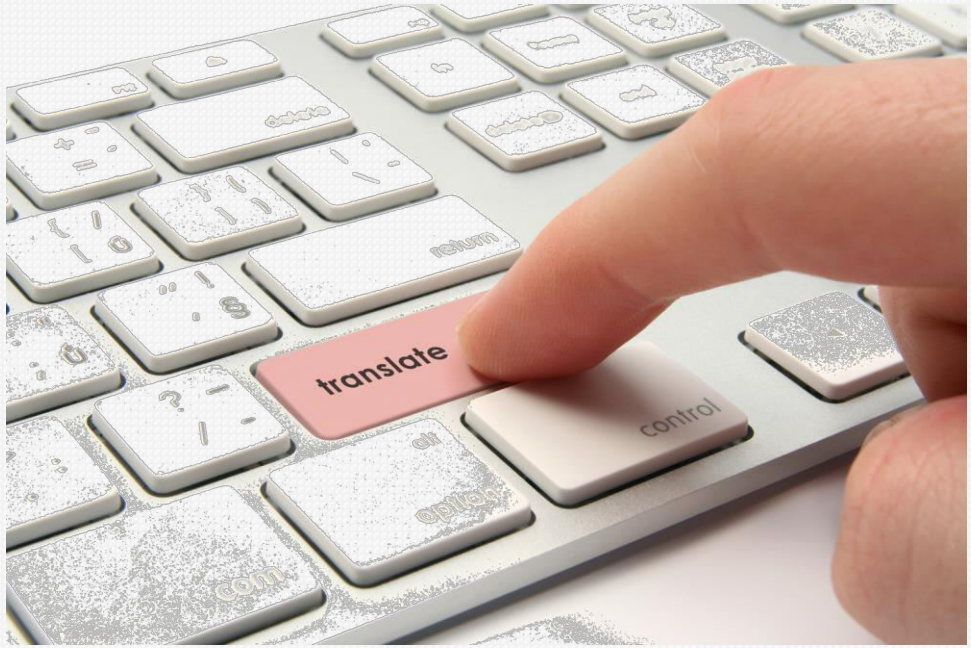
тысячи аналитических алгоритмов

- Различные подходы к МП
- Онлайн-сервисы

(Google, Microsoft Bing, Translate.ru)

- 100 млн. запросов на перевод на

Translate.ru в месяц



Для чего используется МП?

Пользователи

- Джистинг
- Личная переписка
- Новости
- Контент интернет-магазинов

Компании

- Локализация контента
- Оптимизация работы отделов переводов и переводческих бюро
- Деловая переписка

Основные подходы к МП

- **Аналитический / Основанный на правилах**

(Rule-Based Machine Translation, RBMT)

- **Статистический**

(Statistical Machine Translation, SMT)

- **Гибридный**

(Hybrid Machine Translation, HMT)

- **Нейронный**

(Neural Machine Translation, NMT)

Системы, основанные на правилах (RBMT)

Системы RBMT анализируют текст и осуществляют перевод на базе встроенных словарей и набора правил.

Подтипы RBMT :

- интерлингвистические (исходный текст → языконезависимое представление → текст перевода)
- трансферные (исходный текст → языкозависимое представление → текст перевода)

Преимущества и недостатки RBMT

Преимущества

- Синтаксическая и морфологическая точность
- Стабильность и предсказуемость результата
- Возможность тонкой настройки на предметную область

Недостатки

- Трудоемкость и длительность разработки
- Необходимость поддерживать и актуализировать лингвистические базы данных
- «Машинный» акцент в тексте перевода

RBMT перевод

This is the first time I flew with Allegiant. I am a nervous flyer so am very aware of everything the plane had to offer.

↘

«Это первый раз, когда я был с Преданным. Я - возбужденный летчик, так что очень знаю обо всем, что самолет предложил.»

В душе я хиппи

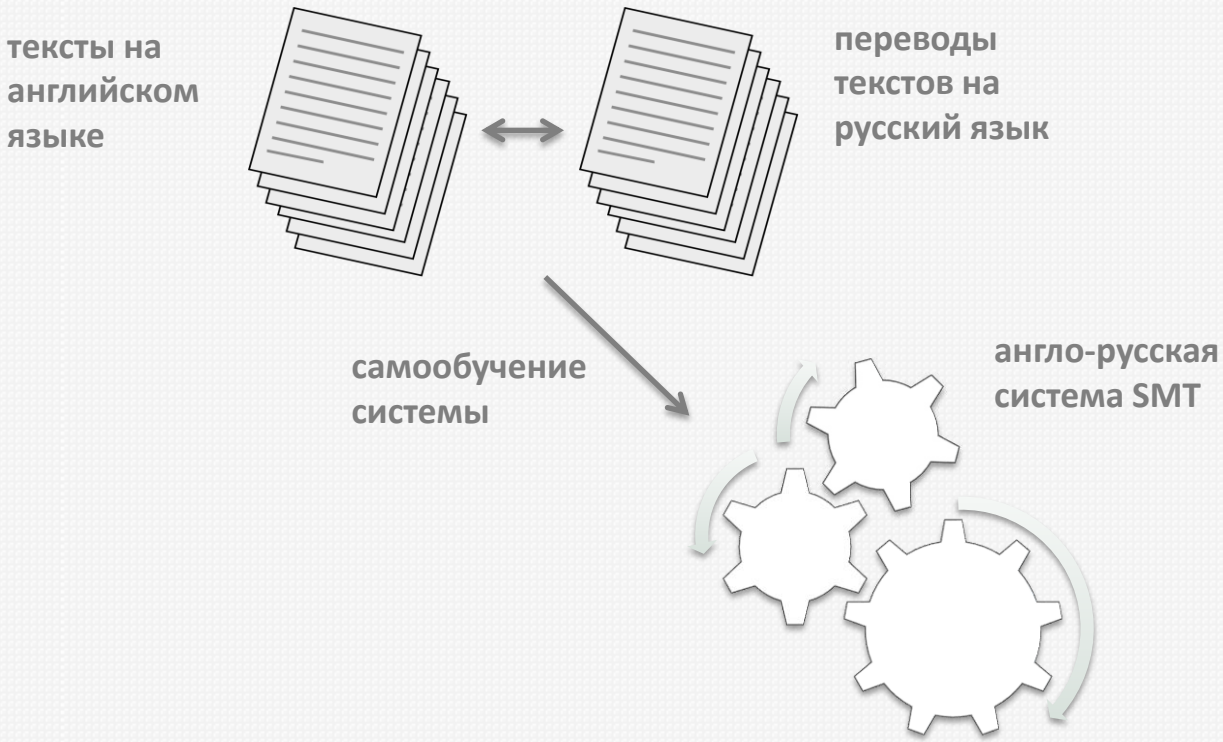
↘

“In a shower I hippie”

Системы статистического МП (SMT)

Принцип работы статистических систем МП – статистический анализ.

Система самостоятельно обучается на больших объемах параллельных текстов (на языке оригинала и языке перевода).



Преимущества и недостатки SMT

Преимущества

- Легко создается, при достаточном объеме параллельных текстов
- Универсальная технология (можно применить к любой языковой паре)
- «Гладкий» перевод

Недостатки

- Нужны большие объемы параллельных текстов
- Ориентирован на конкретную предметную область
- Несвязность перевода
- «Непредсказуемость» перевода

SMT перевод

Donald J. Trump @realDonaldTrump · 27 мин.

North Korea is looking for trouble. If China decides to help, that would be great. If not, we will solve the problem without them! U.S.A.

Язык твита: английский. Переведено с помощью bing Ошибка в переводе?

Северная Корея ищет неприятности. Если Китай решает помочь, что бы здорово. Если нет, то мы будем решать проблемы без них! РОССИЙСКАЯ ФЕДЕРАЦИЯ

4,2 тыс. 7,1 тыс. 16 тыс.

Гибридные системы МП (HMT)

Гибридные системы перевода совмещают в себе несколько технологий машинного перевода.

Наиболее популярные сегодня гибридные системы совмещают классический RBMT с современным SMT.

SMT компонент гибридной системы перевода автоматически обучается исправлять типичные ошибки и «нестройность» RBMT перевода.

Преимущества обоих подходов

RBMT

- Четкая синтаксическая структура
- Связность
- Стабильность

SMT

- Гладкость перевода
- Возможность быстрой настройки на предметную область

Гибридный перевод

Before proceeding further, every effort was made by senior staff to ensure that a friendly atmosphere prevailed.

RBMT перевод

Прежде, чем продолжиться далее, каждое усилие было приложено руководящим персоналом, чтобы гарантировать, что преобладала дружественная атмосфера.

Гибридный перевод

Прежде чем продолжить, руководящий персонал предпринял все усилия, чтобы преобладала дружественная атмосфера.

Гибридный перевод

(Иногда) недостатки обоих подходов

The legroom on the plane was tight, and there is no screen entertainment available.

RBMT перевод

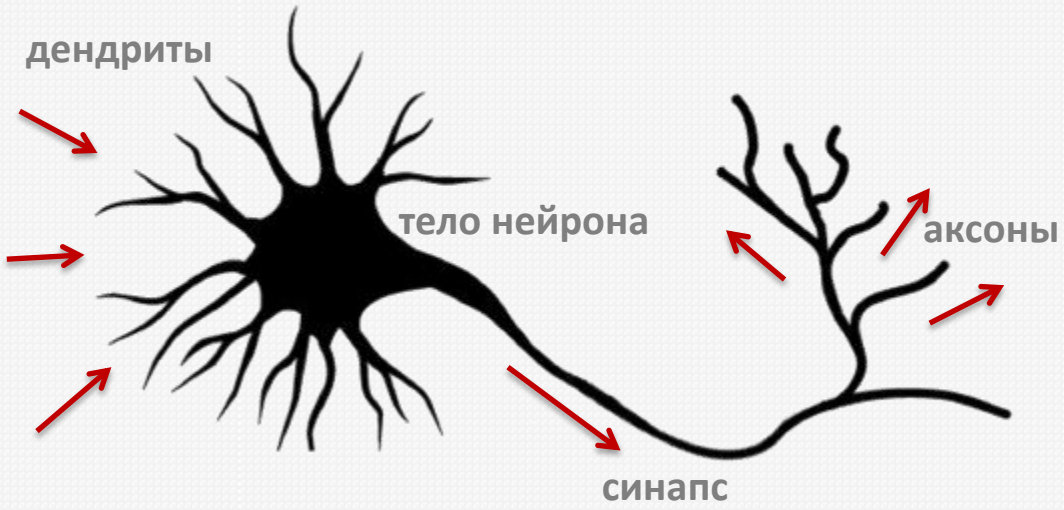
Место для ног в самолете было трудно, и нет никаких доступных развлечений экрана.

Гибридный перевод

Место для ног в самолете было мало, и нет никаких развлечений в туалете.

Нейронный МП (NMT)

Искусственные нейронные сети – попытка моделировать работу человеческого мозга.



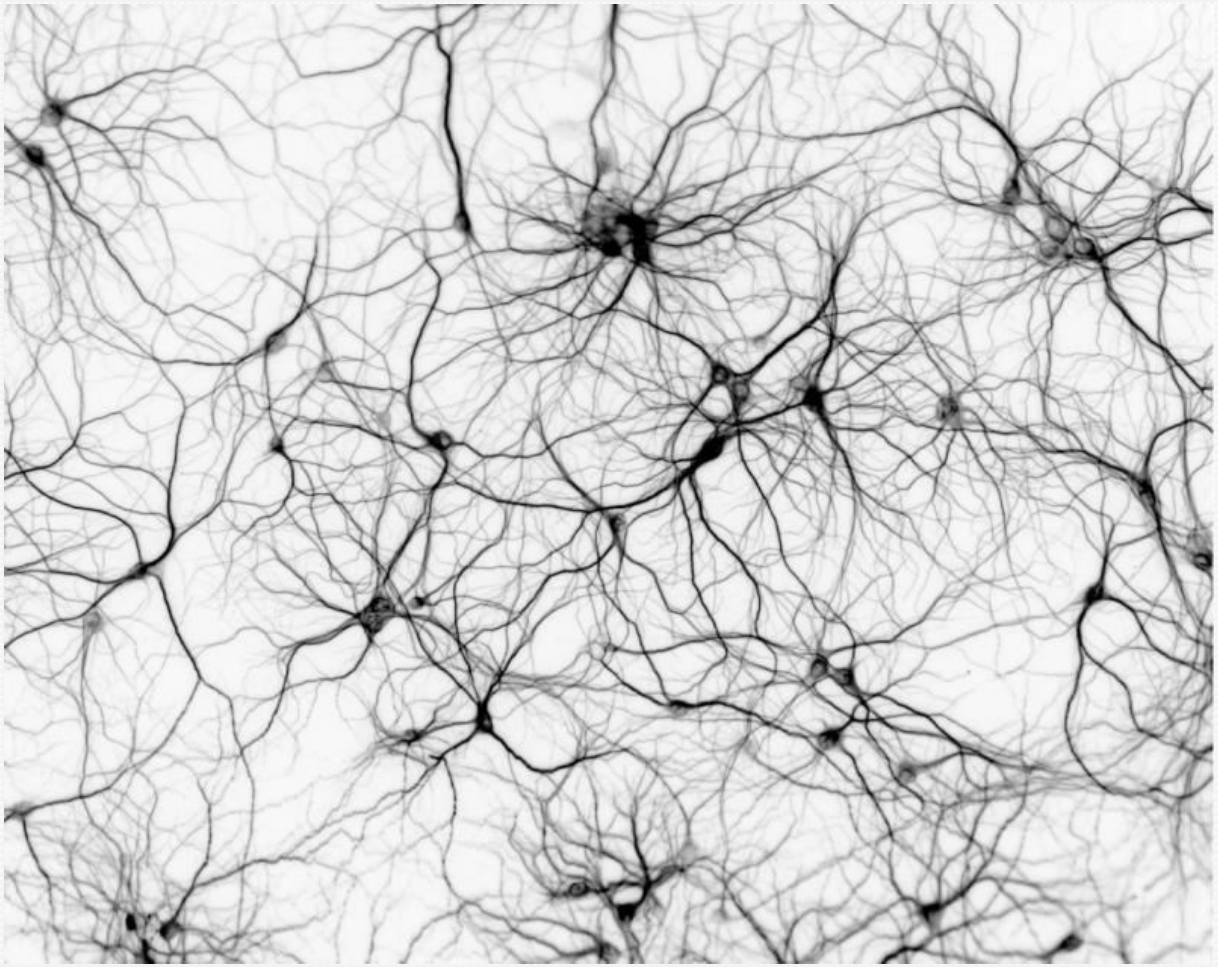
Нейрон принимает и передает электрический импульс (информацию)

Нейронный МП

Нейронные сети

50-500 миллиардов нейронов в человеческом мозге

До **10 тысяч** связей у одного нейрона с другими нейронами



Нейронный МП (NMT)

Вопрос №1

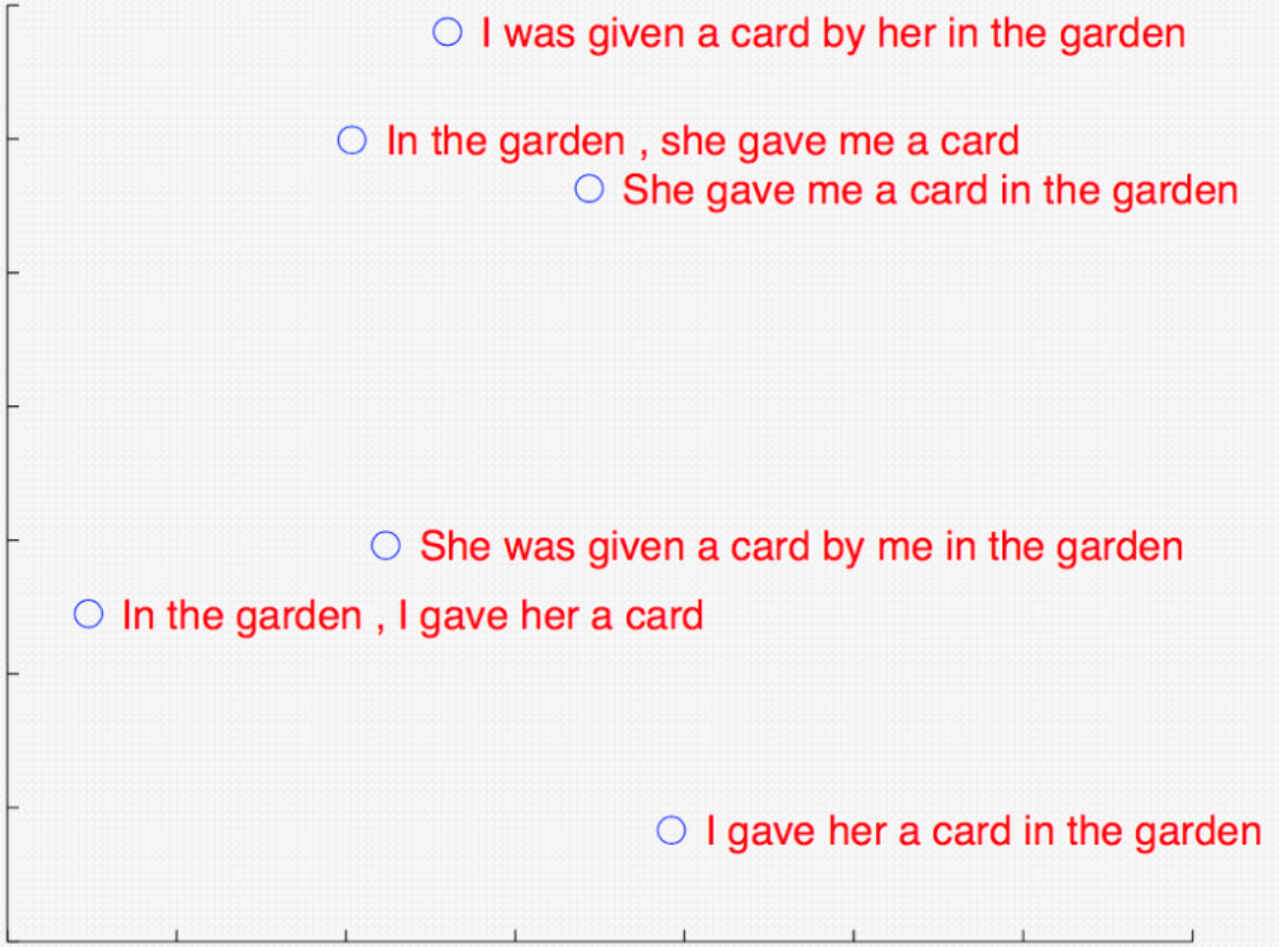
6289,64067 * (4698,76754 + 8960,0434 / 29,345) = ?

Вопрос №2

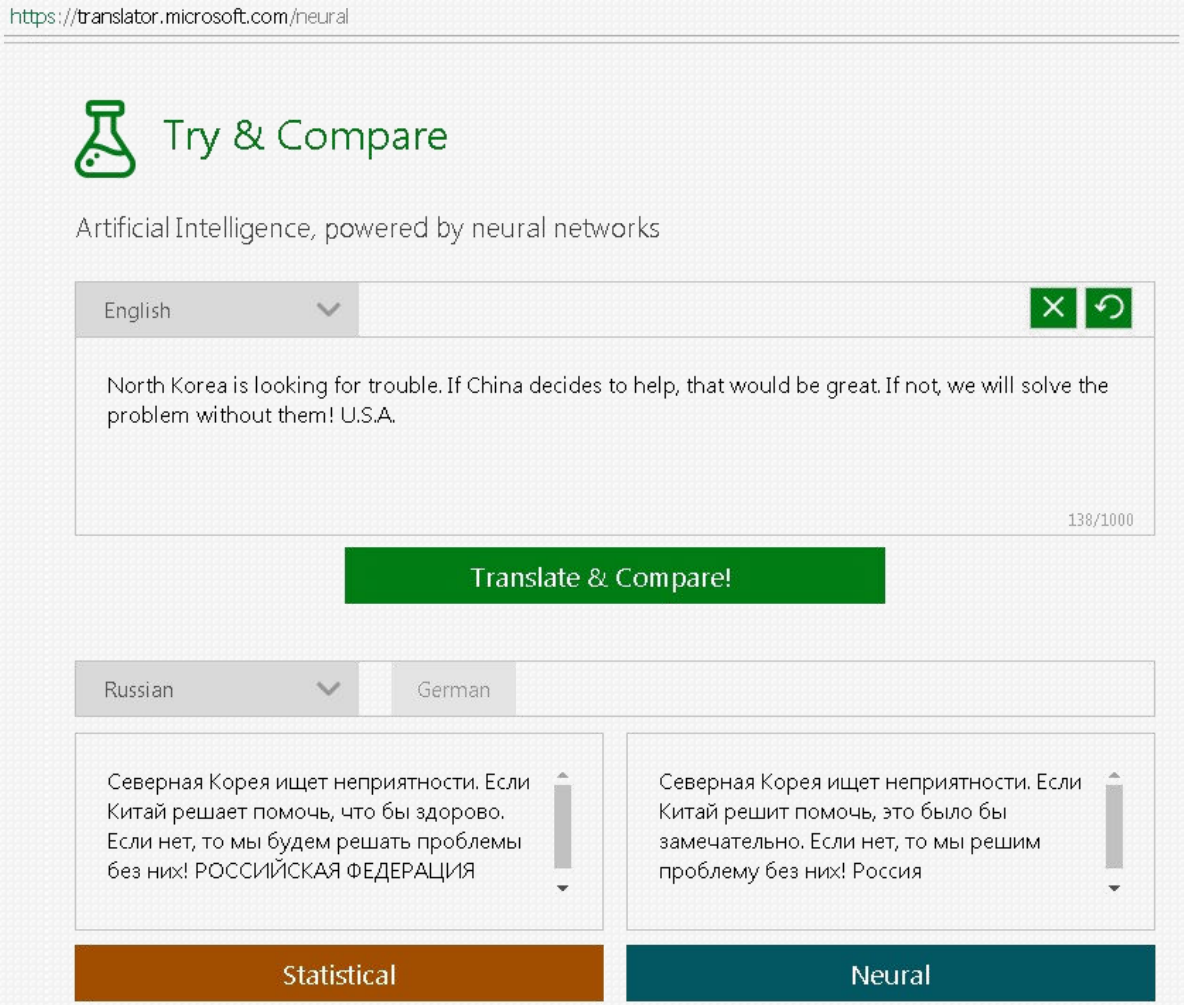
Где Барак Обама?

Как работает нейронный МП?

- Искусственный нейрон – это некоторая единица, которая принимает на вход число, преобразует его и выдает некоторое другое число.
- Предложение кодируется сетью нейронов в векторное представление – набор чисел.
- Вектор декодируется на язык перевода.



Нейронный МП не идеален



Перевод в PROMT

- Разработка с 1991 года
- 15 языков, 60 языковых пар
- RBMT перевод для английского, русского, других европейских языков
- Статистический перевод для казахского, финского, арабского, китайского, корейского, иврита, …
- Гибридный перевод для крупных клиентов, работающих с локализацией контента.

- Нейронный перевод – на стадии исследований / разработки.

Перевод в PROMT

RBMT

- Автоматическая обработка новостных источников и баз данных
- Пополнение словарных баз
- Улучшение алгоритмов

SMT

- Автоматический поиск и обработка источников параллельных текстов в Интернет
- Разработка и использование лингвистических правил

Перспективы МП

Philipp Koehn, 2016:

“[Ideally] we need a rule-based neural network-trained statistical machine translation system. How we are going to do that I have no idea right now.”

Спасибо за внимание!

Александр Молчанов

Руководитель отдела
разработки статистического и
гибридного
машинного перевода PROMT

Alexander.Molchanov@promt.ru