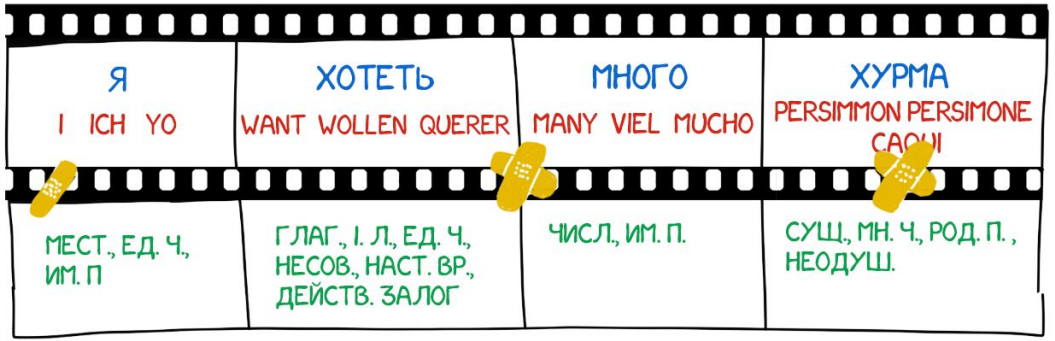
Развитие машинного перевода

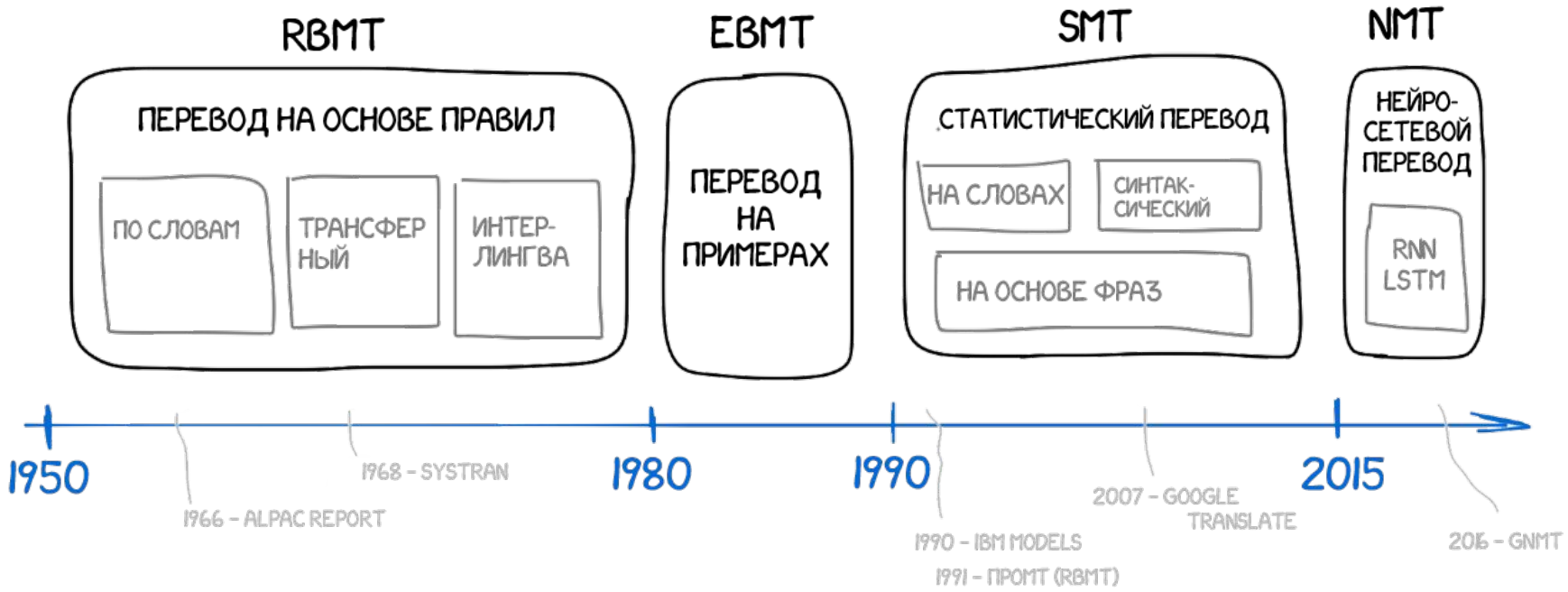
Балышев Артем

Машина Троянского

В 1933 году Советский ученый Пётр Троянский обращается в Академию Наук СССР с изобретённой им «машиной для подбора и печатания слов при переводе с одного языка на другой». Машина была крайне проста: большой стол, печатная машинка с лентой и плёночный фотоаппарат. На столе лежали карточки со словами и их переводами на четырёх языках. Машина Троянского впервые на практике реализовала тот самый «промежуточный язык» (interlingua).



КРАТКАЯ ИСТОРИЯ МАШИННОГО ПЕРЕВОДА



Машинный перевод на основе правил — Rule-based Machine Translation (RBMT)

Среди плюсов RBMT отмечают морфологическую точность (не путает слова), воспроизводимость (все переводчики получат одинаковый результат) и возможность обучить специальным терминам под предметную область.

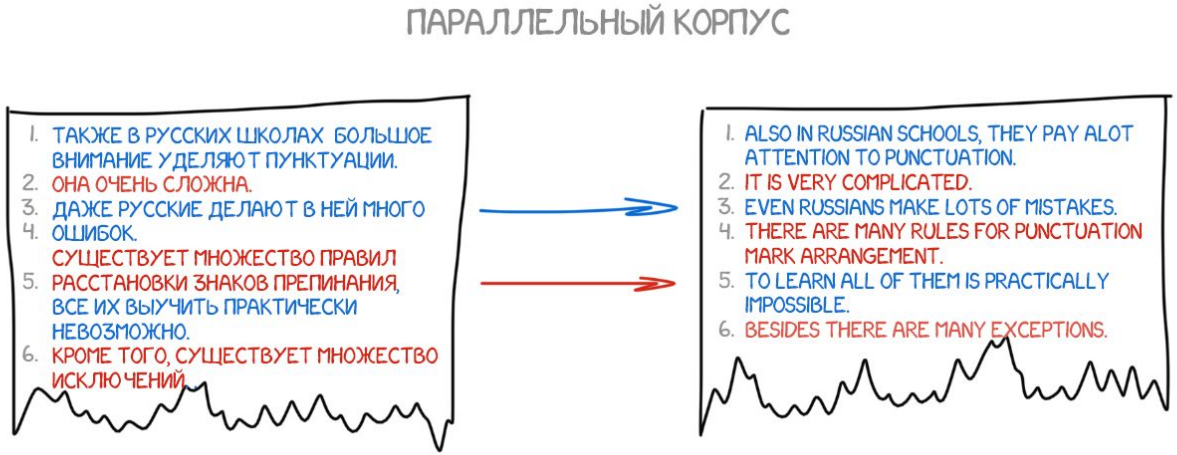
Минусы:

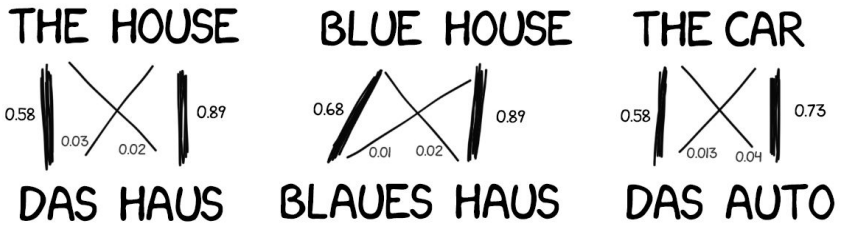
- исключения из правил языка - неправильные глаголы в английском, плавающие приставки в немецком, суффиксы в русском
- омонимия. Одно и то же слово может иметь разный смысл зависимости от контекста, а значит отличается и его перевод.

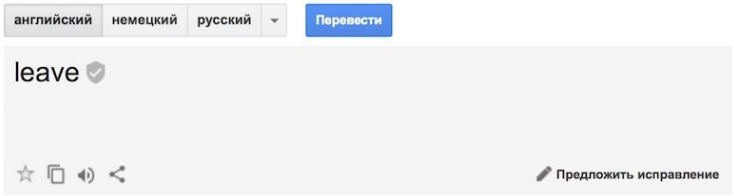
Машинный перевод на примерах — Example-based Machine Translation (EBMT)

Основной приницп: А что если не пытаться каждый раз переводить заново, а использовать уже готовые фразы?

Статистический машинный перевод — Statistical Machine Translation (SMT)







уходить: варианты перевода

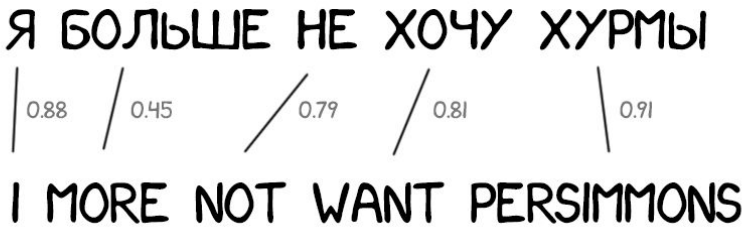
глагол

go	идти, ехать, ходить, ездить, уходить, проходить
go away	уходить, отойти, разойтись
depart	отступать, отклоняться, уходить, отправляться, уезжать, отбывать
leave	оставлять, покидать, уходить, уезжать, бросать, удаляться
exit	выходить, уходить, умереть
get out	выходить, выбираться, вылезать, убираться, уходить, выкарабкиваться
get away	избежать, уходить, ускользнуть, удирать, выбираться, отрываться
walk away	уходить, унести, обходить стороной, уводить, украсть
walk off	уходить, унести, одержать легкую победу, украсть, уводить
go off	уходить, сходить, выстрелить, уезжать, сбежать, проходить
withdraw	изымать, отзывать, уходить, забирать, увести, удаляться
take away	увезти, забрать, забирать, отнимать, унести, уходить
come away	уходить, отламываться, отходить
retire	уходить, удаляться, уходить в отставку, увольняться, отступать, уединяться
retreat	отступать, отходить, уходить, удаляться, отбросить назад
walk out	уходить, демонстративно покинуть, выходить
be off	уходить
buzz off	уходить, удаляться, смыться, улизнуть
scram	уходить, удирать
be away	отсутствовать, уходить, не быть дома
shove off	убираться, уходить, отталкиваться
toddle off	уходить
toddle	ковылять, учиться ходить, прогуливаться, бродить, уходить
draw away	отвлекать, удаляться, уводить, уходить, отрываться от противника
split	раскалывать, разбивать, расщепить, раскалываться, расщепляться, уходи

Статистический перевод по словам — Word-based SMT

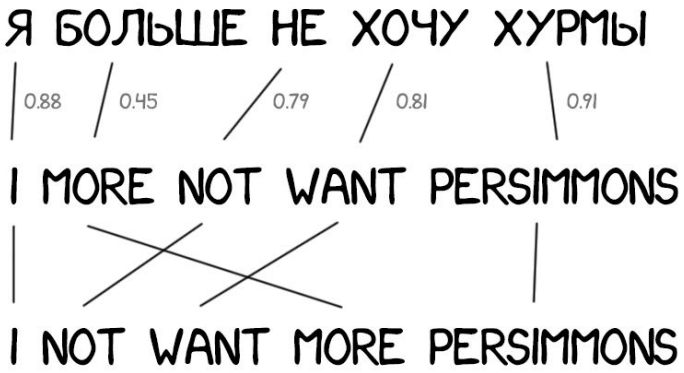
Model 1: мешок слов

Классический подход — делим всё на слова и считаем статистику. Никакого учёта порядка или перестановок. Из хитростей Model 1 умела разве что переводить одно слово в несколько. Der Staubsauger (пылесос) легко превращался в Vacuum Cleaner, но обратно уже как повезет.



Model 2: учёт порядка слов в предложении

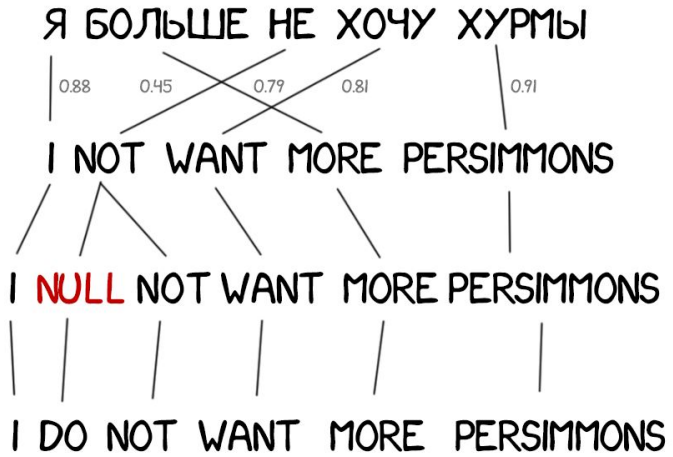
Отсутствие знаний о порядке слов в языках стало проблемой для Model 1. В некоторых он очень важен. Потому в Model 2 стали запоминать на каком месте появляется в переведённом предложении. Добавили промежуточный шаг — после перевода машина пыталась переставить слова местами так, как она думала будет звучать более естественно.



Model 3: добавление отсутствующих слов

Часто при переводе появляются новые слова, которых не было в оригинальном тексте. В немецком языке внезапно вылезают артикли, в английском вставляют глагол do где не попадя. «Я не хочу хурмы» → «I **do** not want persimmons. Чтобы решить эту проблему в Model 3 добавили два промежуточных шага:

1. Вставка маркеров (NULL-слов) на те места, где машина подозревает необходимость нового слова
2. Подбор нужного артикля, частицы или глагола под каждый маркер



Model 4: перестановки слов

Model 2 хоть учитывала порядок слов в предложении, но ничего не знала про перестановки слов между собой. Часто при переводе надо, например, поменять существительное и прилагательное местами. Тут сколько ни запоминай их порядок по всему предложению — лучше не станет. Потому в Model 4 стали учитывать еще и так называемый «относительный порядок». Если при переводе два слова постоянно менялись друг с другом — модель это запоминала.

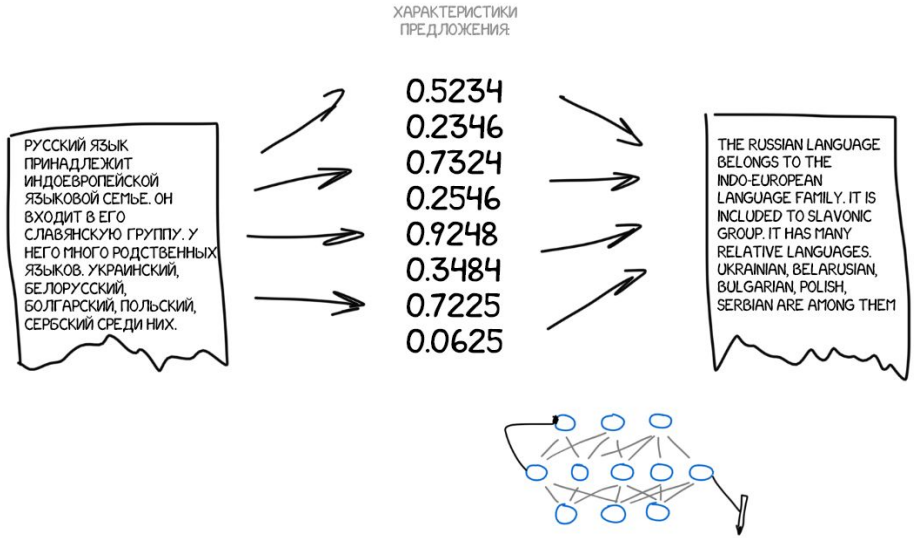
Статистический перевод по фразам — Phrase-based SMT

Для обучения он разбивал текст не только на слова, но и на целые фразы. Точнее N-граммы или фраземы — пересекающиеся наборы из N слов подряд. Машина училась переводить устойчивые сочетания слов, что заметно улучшило точность.

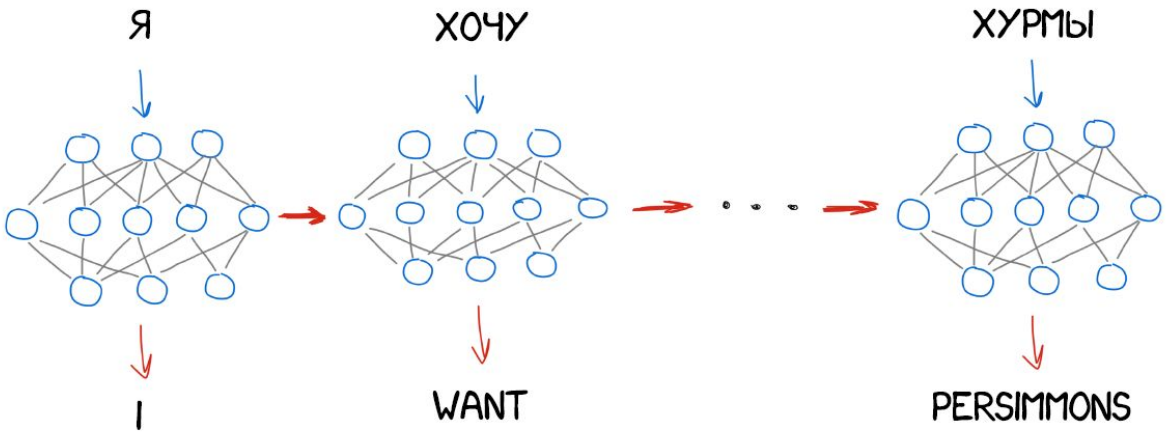
Нейронный машинный перевод — Neural Machine Translation (NMT)

Помните приложение Prisma, которое обрабатывало фото в стиле известного художника? Там не было особой магии — нейросеть обучили распознавать картины художника, а потом «оторвали» последние слои, где она принимает решение. Получившиеся наброски, по сути промежуточное представление сети, и было той самой стилизованной картинкой.

Первая нейросеть умеет только кодировать предложение в набор цифр-характеристик, а вторая только декодировать их обратно в текст. Обе понятия не имеют друг о друге, каждая знает только свой язык.

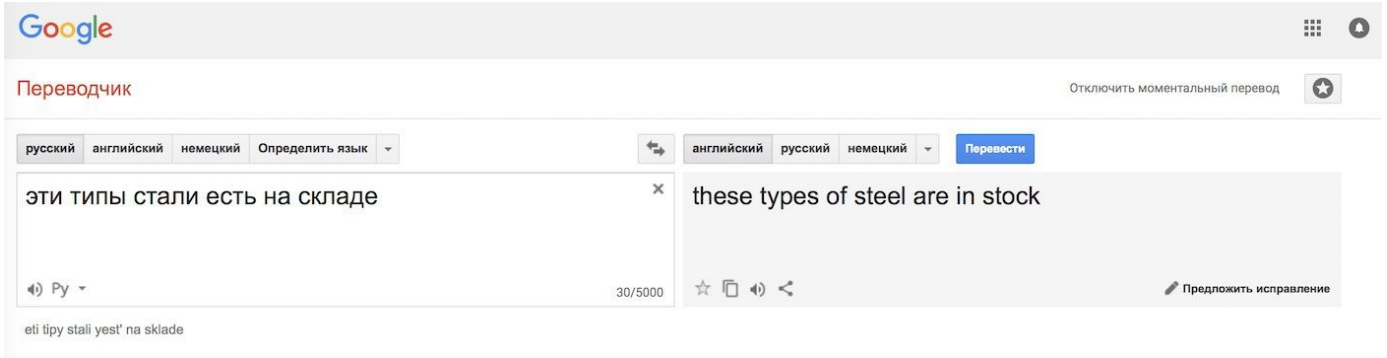


RNN сейчас применяют в: распознавание речи в Siri , подсказки слов на клавиатуре , генерация музыки и даже чатботы.



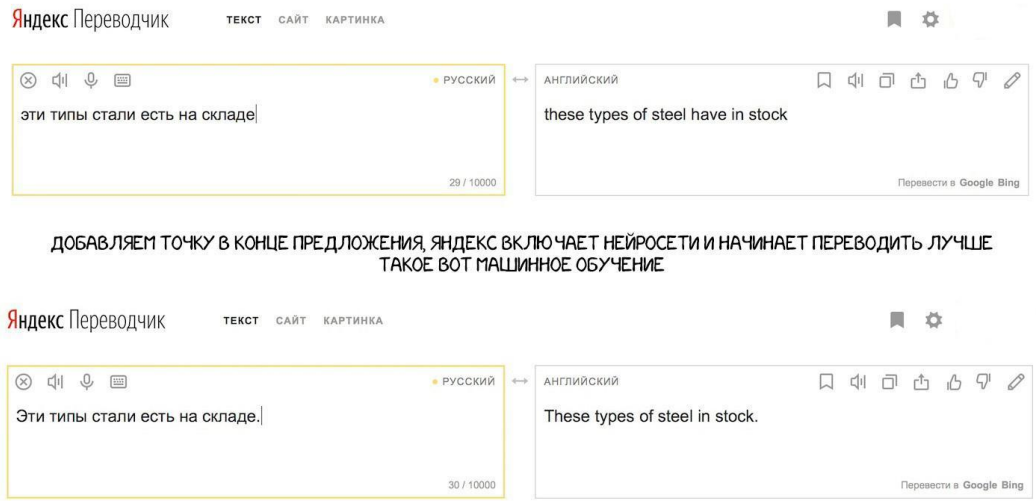
Google Translate (2016)

В 2016 году Google включил нейронный перевод девяти языков между собой, в 2017 был добавлен и русский. Google разработал собственную систему под нехитрым названием Google Neural Machine Translation (GNMT), состоявшую аж из 8-слойного RNN на входе и такого же на выходе и системы согласования контекста под названием Attention Model.



Яндекс Переводчик (2017)

Яндекс запустил свой нейросетевой перевод в 2017 году. Главным отличием они заявили гибридность. Переводчик Яндекса переводит предложение сразу двумя методами — статистическим и нейросетевым, а потом с помощью их любимого алгоритма CatBoost находит наиболее подходящий.



Заключение и будущее

Всех по прежнему будоражит идея «Вавилонской Рыбки» — синхронного перевода речи на лету. Google делала шаг в этом направлении, когда анонсировала Pixel Buds, но на поверку всё оказалось плохо. Синхронный перевод на лету отличается от обычного, ведь нужно знать места, когда начать переводить, а когда сидеть и слушать. Подходов к решению этой задачи найти не удалось.