

ОСНОВЫ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

ОСНОВЫ ИТ

Концевой Михаил Петрович

- 144 часа (74 аудиторных часа)
- 2 и 3 семестры
- 2 семестр: 10 ч. лекций + 16 ч. лаборатор-х,
- 3 семестр: 14 ч. лекций + 18 ч. лаборатор-х
- **Экзамен (январь 2024 г.).**

О допуске к экзамену!

3. ПОРЯДОК ДОПУСКА СТУДЕНТОВ К ТЕКУЩЕЙ АТТЕСТАЦИИ

24. Студенты допускаются к сдаче зачета (дифференцированного зачета), экзамена по учебной дисциплине при условии выполнения и защиты ими всех расчетных, расчетно-графических, лабораторных и иных работ, курсовых проектов (курсовых работ), предусмотренных учебно-программной документацией соответствующей учебной дисциплины в текущем семестре.

Недопуск студента к зачету (дифференцированному зачету) или экзамену по учебной дисциплине осуществляется решением соответствующей кафедры. Решение кафедры о недопуске студента (выписка из протокола заседания кафедры) должно быть предоставлено в деканат не позднее дня, предшествующего текущей аттестации. До выдачи экзаменатору зачетно-

Работаем с Пособием



пособие конечного

 Все

 Видео

 Новости

 Карты

 Карты

Результатов: примерно 50 500 (0,40 сек.)

lab314.brsu.by › kmp-lite › kmp2 ▾

Пособие Концевого М.П. для всех

Вектор развития информационных технологий - создание более умных вещей для все более глупых людей. kmp.

308 аудитория гл. корпуса

- Лимит - 500 Мб в неделю,
- Доступ в Интернет возможен с любой рабочей станции сети университета.
- Посещение сайтов, не имеющих отношения к учебной деятельности, запрещается.
- Все действия пользователей в сети регистрируются и регулярно просматривается администраторами.

Распоряжение №7 от 10.01.2018

О соблюдении порядка проведения текущей аттестации.

Обратить особое внимание на:

- недопустимость освобождения студентов от ответа на вопросы экзаменационного билета;
- предоставление обучающемуся не менее 20 минут для подготовки ответа на экзамене;

Распоряжение №7 от 10.01.2018

О соблюдении порядка проведения текущей аттестации.

Обратить особое внимание при выставлении отметки по учебной дисциплине на:

- успеваемости студентов в течение семестра.
- результаты промежуточного контроля.

МАШИННЫЙ
(КОМПЬЮТЕРНЫЙ)
ПЕРЕВОД

Disclaimer

Всё услышанное и увиденное
является исключительно
плодом Вашего собственного
воображения (перевода)

НАПИСАНО

КОМПЬЮТЕРНЫЙ ПЕРЕВОД

ПРОЧИТАНО

КОМПЬЮТЕРНЫЙ ПЕРЕВОД
ТЕКСТА

С ОДНОГО ЕСТЕСТВЕННОГО ЯЗЫКА
НА ДРУГОЙ ЕСТЕСТВЕННЫЙ ЯЗЫК
ДЛЯ СИМУЛЯЦИИ

ВЫПОЛНЕНИЯ УЧЕБНОГО ЗАДАНИЯ

Всё есть перевод

- слушание (может быть понято как перевод)
- говорение
- чтение
- письмо
- мышление
- редактура корректура....
- генерация по запросу...

Перевод - игра воображения

По результатам исследования одонго английского университета, не имеет значения, в каком порядке распознаются буквы в слове.

22.10
2023



Elon Musk ✓ X
@elonmusk



I will give them a billion dollars if they change their name to Dickikipedia

[Перевести пост](#)

Elon Musk ✓ X @elonmusk · 5 ч



Wikipedia is not for sale.

A personal appeal from Jimmy Wales

Please don't scroll past this 1-minute read.

I'm sorry to interrupt, but it's Sunday,

5:41 PM · 22 окт. 2023 г. · 6,8 млн просмотров

22.10
2023



Лучший
офлайн
переводчик
сделает **ж** нужную
работу за тебя!

**Бессрочный
переводчик
PROMT**
по цене годовой
лицензии:

~~5 900~~ **₽ 3 300** **₽**

Bottom left navigation icons: bookmark, camera, and menu.

Предложение действует с 12 по 22 октября 2023 года.

Переводчик PROMT Master NMT

Такой разный перевод

Авторизованный, Авторский, Адаптированный,
Адаптивное транскодирование, Адекватный,
Прагматический, Семантико-стилистический,
Аннотационный, Аспектный, Аутентичный,
Бинарный, Буквальный, Диахронический,
Вольный, Грамматическая трансформация,
Дезиративно- адекватный, Информативный,
Интерсемиотический, ...

Такой разный перевод

Калькирование, Компенсация,
Консультативный, Косвенный,

Лексико-грамматический,

Множественное, соответствие, Нулевой,
Окказиональное, соответствие, Официальный,
Переводческая, трансформация, Специальный,
Трансмутация, Транспозиция, Эквивалентный,
Трансформационный, Экспликация....

Компетентное мнение

Палажченко

Павел Руслановч,
переводчик МСГ (англ,
нем, фр, исп. ит.),
писатель, создатель
«Несистематического
словаря», фрилансер



Лексико-грамматический перевод

Палажченко Павел Русланович

- Процесс массового, школьного, изучения языка (не в двуязычной семейной среде) есть сочетание запоминания слов, освоения грамматики и учебного перевода.
- Это не эффективный, но неизбежный подход в условиях отсутствия иноязычной среды.
- Такие упражнения непродуктивны и создают неверное представление о сути перевода.

Лексико-грамматический перевод

Ридра Вонг (Вавилон-17, Сэмюэль Дилэни):

- Это не шифр. Это язык.
- Неизвестный язык может быть понят без перевода.
- Я запоминала языки как большинство людей запоминает мелодии популярных песен.
- Иногда, когда очень хотелось получить ответ, в голове что-то начинало складываться и внезапно все, что я знала, соединялась в одно целое и я могла не напрягаясь читать текст, говорить об этом...

Лексико-грамматический перевод

Платон:

- Человека ничему новому научить невозможно.
- Душа еще до своего вселения в телесную оболочку уже всё знала, да только потом всё забыла.
- Однако забыла не окончательно — если ей как-то об этом напомнить, она потихоньку будет забытое припоминать.

Каждый из нас имеет успешный опыт освоения языка без всякого межъязыкового перевода...

Лексико-грамматический перевод

основан на грамматических правилах
проверяем на соответствие правилам
может быть осуществлен по правилам

а значит:

формализуем и вычисляем

при условии логичных (разумных) правил

Языковая реформа

- упрощения и очищение языка (письменности) от непродуктивных пережитков.
 - русский язык (1708/10, 1918),
 - немецкий язык (1901/02, 1996/1998),
 - от *The new Euro language* до *Euro English*
 - испанский (1754, 1803, 2010)
 - португальский (1910, 1943, 1971)
 - французский язык (2017)

Правила неправильны

Плунгян Владимир Александрович

- Более того, и словари и грамматики теперь нужны не традиционные, а нового поколения.
- То есть не просто словари и грамматики, а словари такого-то корпуса и грамматики такого-то корпуса

Компетентное мнение

Палажченко Павел Русланович

- Такие упражнения непродуктивны и **создают неверное представление о сути перевода.**



Компьютерный перевод!

1. Фундаментальные проблемы лингвистики, компьютерных наук, математики, психологии
2. Важнейшая часть современной глобальной информационно-коммуникационной инфраструктуры
3. Многомиллиардный бизнес...

Светлана Соколова

- председатель совета директоров ПРОМТ:
За 2022 год на Translate.ru переведено
69 000 000 000 слов
в день - 190 000 000 слов



~~Подсчитываем~~ Переводим

- $69\ 000\ 000\ 000 / 365 = 190\ 000\ 000$
- 190 000 000 слов в день
- 950 000 страниц в день
- 8 страниц – в день (норма переводчика)
- 31 717 дней (87 лет) работы одного переводчика
- **76 120** переводчиков (работающих ежедневно)

~~Подсчитываем~~ Переводим

Google Translate работает со 103 языками
и ежедневно переводит

140 000 000 000 слов

Эквивалентно ежедневной работе

56 000 000 переводчиков

Беларусь: 4 400 000 Россия: 71 000 000

Корректурa = Перевод

- пользователи ошибаются в **10–15%** запросов (анализ логов поисковых систем)
- **250 000 000** запросов в сутки проверяются поисковым спелл-чекером Яндекс
- **30 000 000** запросов **переводятся** спелл-чекером Яндекс ежедневно

250 000 стр. – 30 000 переводчиков!

Перевод голоса в диагноз

ORIGINAL ARTICLE | [VOLUME 1, ISSUE 4, P534-544, DECEMBER 2023](#)

Acoustic Analysis and Prediction of Type 2 Diabetes Mellitus Using Smartphone-Recorded Voice Segments

[Jaycee M. Kaufman, MSc](#)   • [Anirudh Thommandram, MASc](#) • [Yan Fossat, MSc](#)

[Open Access](#) • Published: October 17, 2023 •

DOI: <https://doi.org/10.1016/j.mcpdig.2023.08.005>

 PlumX Metrics

10 сек. вместо анализа крови

ОРИГИНАЛЬНАЯ СТАТЬЯ | [ТОМ 1, ВЫПУСК 4, P534-544, ДЕКАБРЬ 2023 Г.](#)

Акустический анализ и прогнозирование сахарного диабета 2 типа с использованием голосовых сегментов, записанных со смартфона

[Джейси М. Кауфман, магистр](#)   • [Анируд Томмандрам, MASc](#) • [Ян Фоссат, магистр](#)

Открытый Доступ • Опубликовано: 17 октября 2023 года •

DOI: <https://doi.org/10.1016/j.mcpdig.2023.08.005>

 Показатели PlumX

Глобальный мир перевода

Machine Translation Market Size to Touch USD 4,069.5 Million By 2030

Acumen Research and Consulting recently published report titled “Machine Translation Market and Region Forecast, 2022 - 2030”

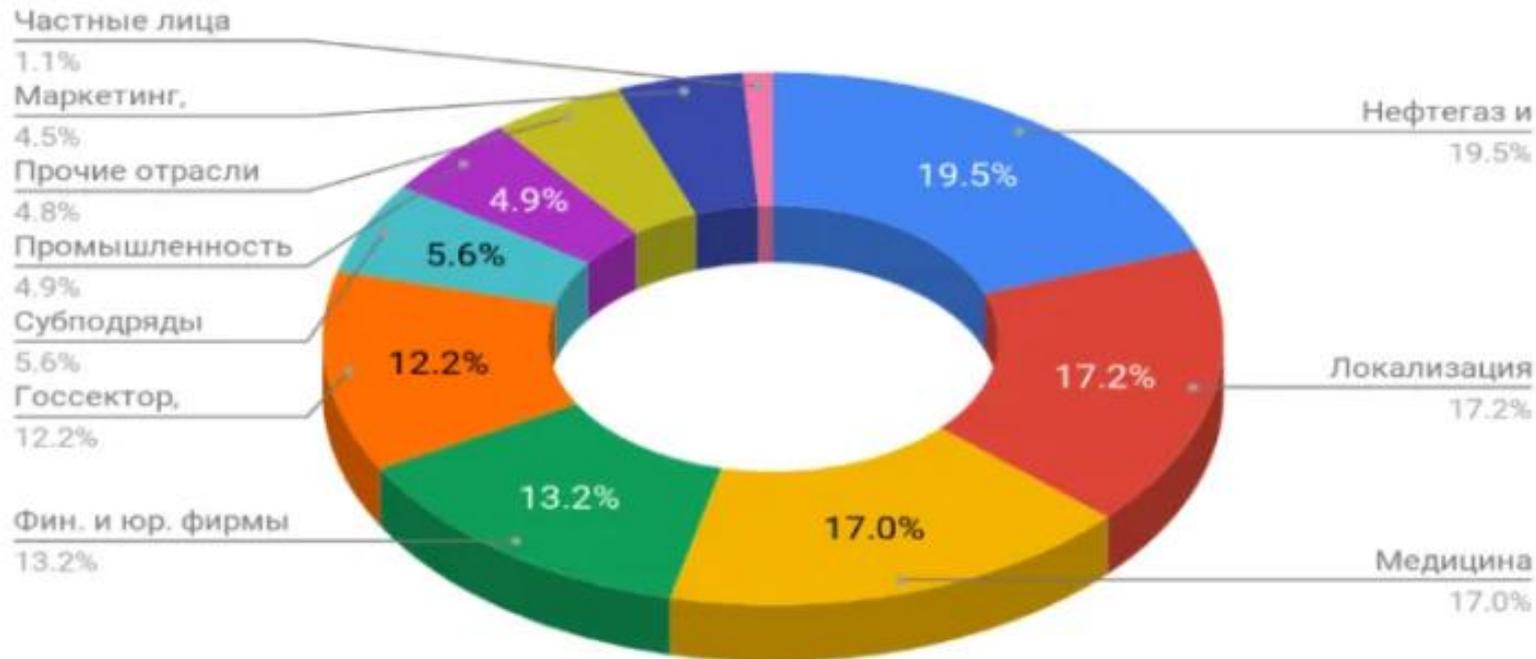
January 24, 2023 18:00 ET | Source: [Acumen Research and Consulting](#)

Глобальный мир перевода

SYSTRAN (2020):

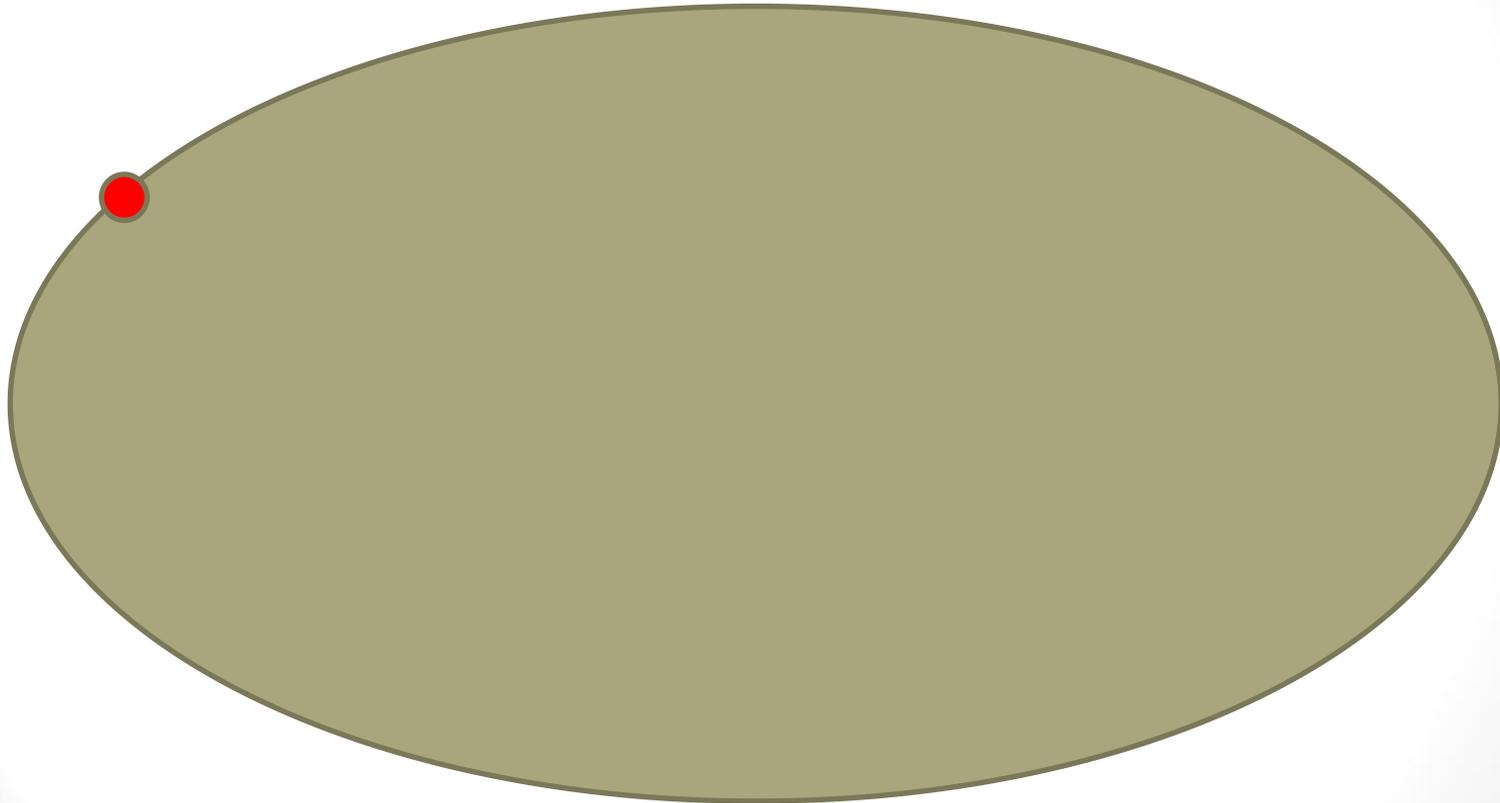
- 50% мирового рынка переводов
- **70% заказчиков** продуктов SYSTRAN - большие корпорации и государственные органы.
- 57,1% дохода от 10 основных клиентов.
- 28% - от интернет-компаний
- **2% - от обычных пользователей.**

Рынок машинного перевода



привет из 2020

Машинный перевод сегодня



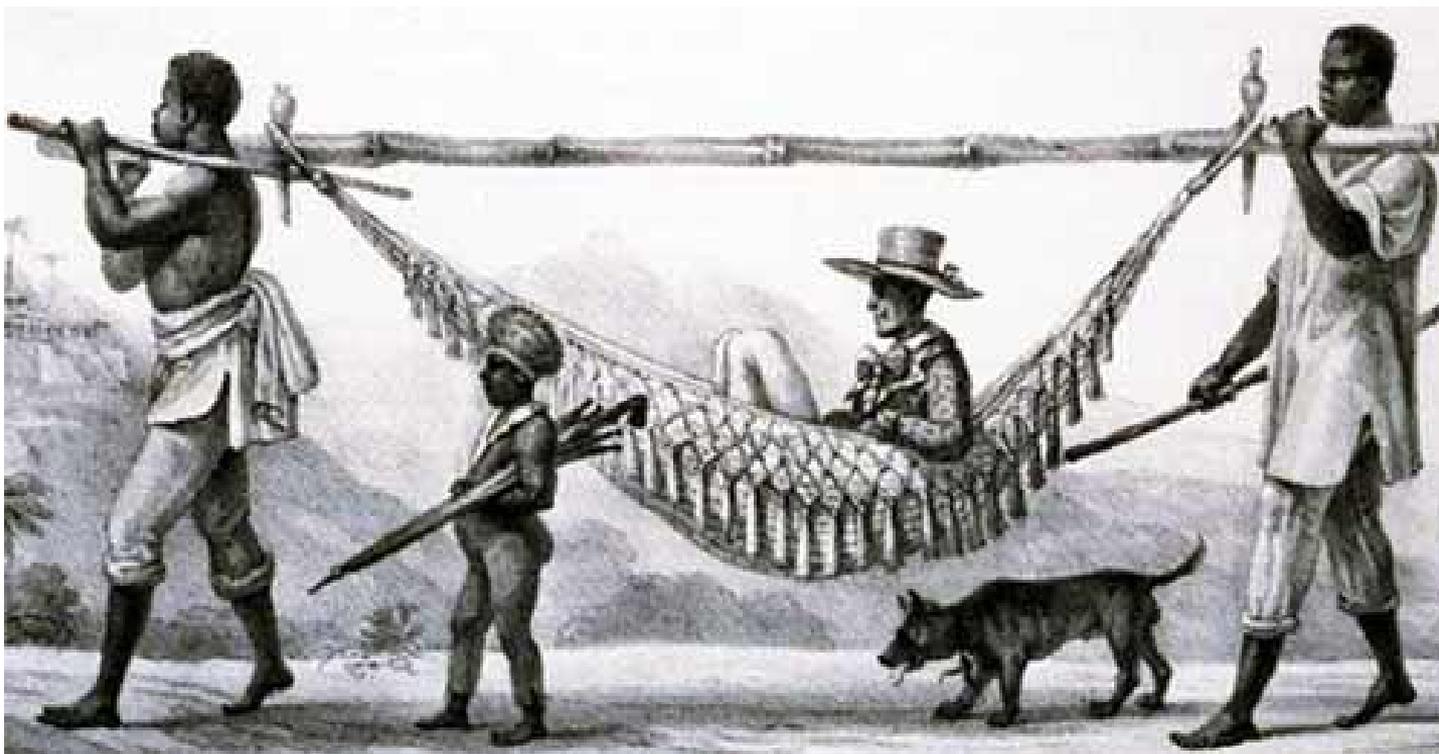
Страда во имя жизни



Решение проблем!



РУЧНОЙ ПЕРЕВОД



МАШИННЫЙ ПЕРЕВОД

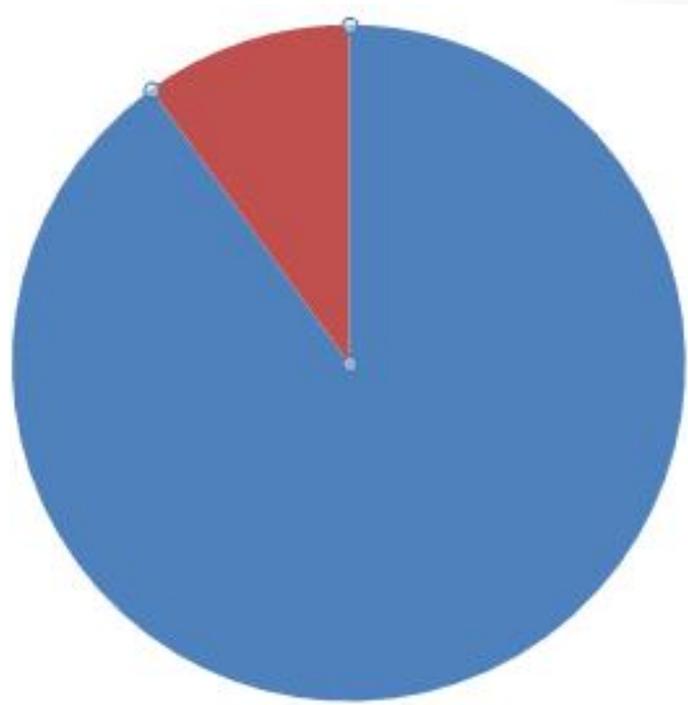


Два мира перевода

\$ 300-500

\$ 3000-5000

\$ 5000-10 000



Однако....

Исследование

"Банка данных заработных плат"

HeadHunter показало, что

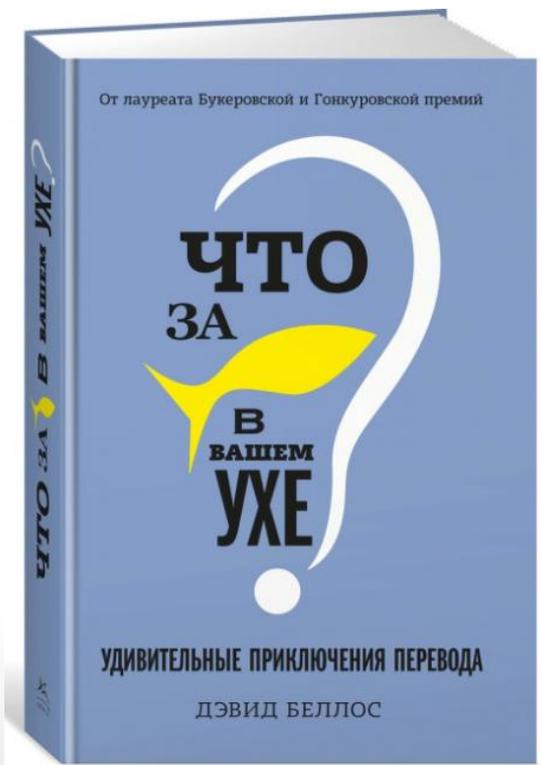
- **из-за автоматизации**
средняя заработная плата
учителей иностранных языков
сократилась на 12%

Автоматизируй это...

Переводчику платят не за перевод. А за отсутствие проблем с этим документом и этим переводом в будущем. **В. Афанасьев**

- От человека решение проблем,
- От машины требуется - решение задач (исполнение функций).

Babel fish



... рыба проскользнула в слуховые каналы.
Задыхаясь от ужаса, Артур выпучил глаза от удивления осознавая, что все еще слышит бульканье и завыванье, только теперь они каким-то образом преобразились в совершенно понятный английский язык.

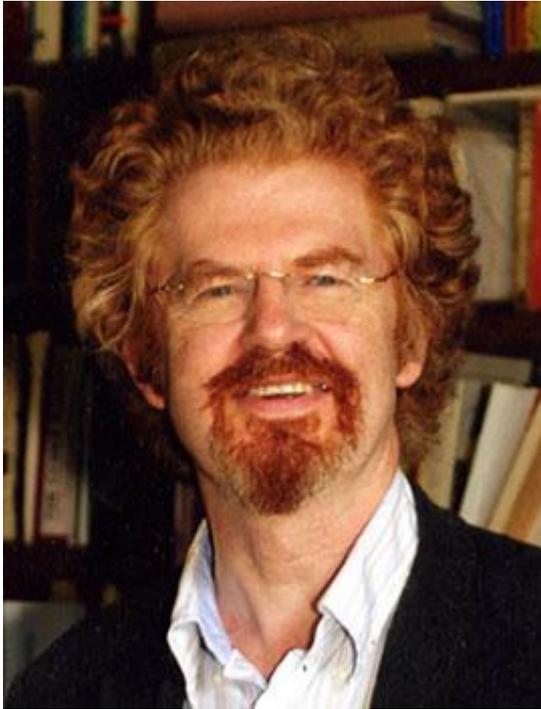
Babel fish 2020

- Google Pixel Buds
- Mymanu Klik
- WT2
- Bragi Dash Pro
- Pilot Waverly Labs



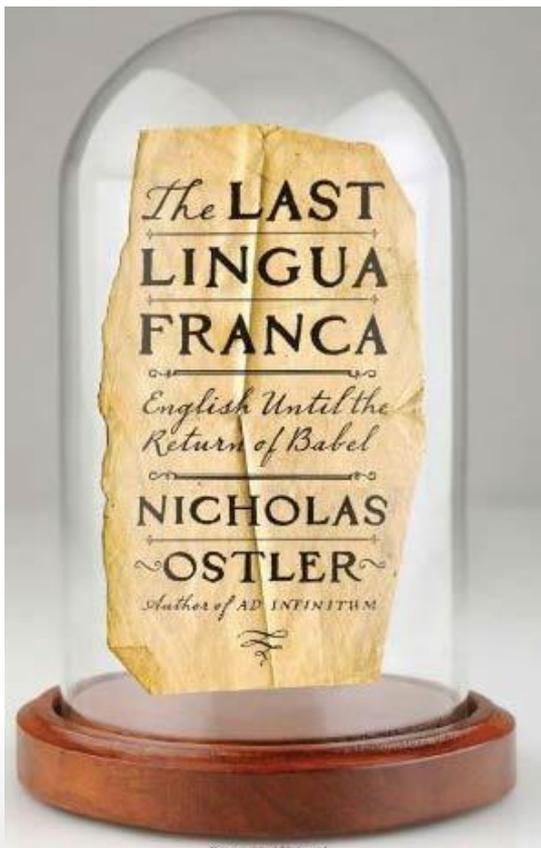
\$ 300

Nicolas Ostler



- Именно машинный перевод сформирует новый «всеобщий язык», lingua franca XXI века.

Nicolas Ostler



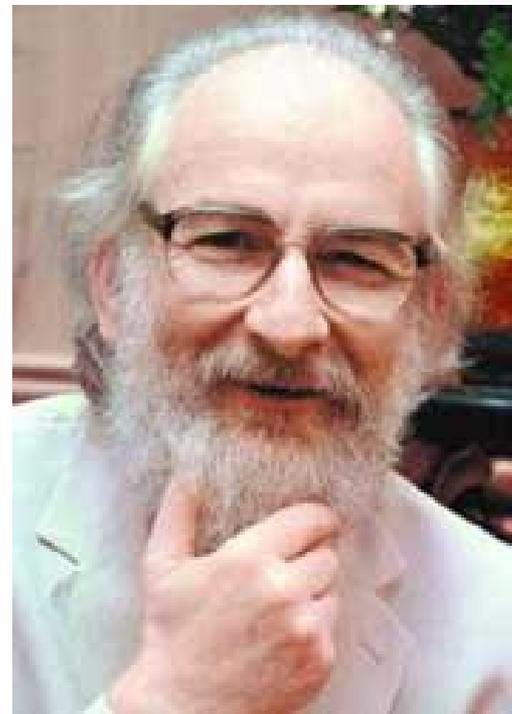
The Last Lingua Franca:
English until the Return to
Babel.

Penguin in the UK, and
Bloomsbury/Walker Books in
the USA.

London and New York, 2010.

David Crystal

- Интернет может уже в ближайшем будущем уничтожить традиционный английский язык.
- Изменению подвергнется как произношение, так и письмо.
- Больше всего этому явлению подвергнуты дети.



Машинный перевод как ... перевод

Почему надо разбираться?

Традиционный (человеческий) хороший перевод возможен без знания того, что же такое перевод (знание важно и полезно, но не является необходимым)

Хороший компьютерный перевод, без знания того, что такое перевод (без построения модели перевода) - невозможен!

Почему надо разбираться?

Чтобы освоить язык и (ис-)пользоваться им, не нужно знать и понимать, ни что такое язык, ни как он устроен, ни как он работает... можно даже придумать сказку о языке и поверить в неё)

Чтобы создать АЛИСУ (разговорную машину, переводную машину, действующую модель языка) нужно понимать, модель чего создается!

Автономова Наталья Сергеевна

- Иногда думают, будто перевод - это школярская, второстепенная работа, например, одна из форм обучения иностранным языкам. Это свидетельство эпистемологической наивности.

«Познание и перевод», М., 2008



Автономова Наталья Сергеевна

- Перевод не только посредник в межкультурном и межъязыковом обмене, но и **условие возможности любого познания** в социальной и гуманитарной области...

«Познание и перевод», М., 2008



**ТЕХНОЛОГИИ
КОМПЬЮТЕРНОГО
АВТОМАТИЧЕСКОГО
ПЕРЕВОДА**

В ИСТОРИЧЕСКОМ КОНТЕКСТЕ

За пределами компьютеров

1947 год

Warren Weaver, директор Рокфеллеровского фонда:

- У меня перед глазами текст, написанный по-русски, но я собираюсь сделать вид, что на самом деле он написан по-английски и закодирован при помощи довольно странных знаков.
- Все что мне нужно — это взломать код чтобы извлечь информацию, заключенную в тексте.

1949 год **W. Weaver** заявил идею **Interlingua**.

Джорджтаунский эксперимент

7 января 1954 года

В Нью-Йорке, перевод с русского языка на английский с помощью IBM 701 (60 фраз, словарь из 250 слов по органической химии, 6 правил перевода).

Леон Достерт (переводчик президента Дуайта Эйзенхауэра, один из создателей модели синхронного перевода): В течение трёх-пяти лет проблема машинного перевода будет решена.

Анатолий Иванович Китов

1956 г **А.И. Китов**: Когда составлен машинный словарь и разработана система четких правил для работы машины, составление самой программы машинного перевода, несмотря на ее чрезвычайную громоздкость (тысячи команд), не представляет трудностей. Необходимо только избавиться от исключений и неоднозначностей.



Четыре технологических типа

- **Rule-based** - основаны на правилах описания языковых структур и их преобразования
- **Statistical-based** - основаны на статистических правилах
- **Neural Machine Translation (NMT)**
- **Hybrid-based**
- **LLM-Translation**

Rule-based

- фундаментально (неустранимо) ущербен для произвольных текстов
- позволяет получить в отдельных предметных областях (тематиках):
 - быстрый
 - эффективный
 - качественный
 - Предсказуемый (повторяемый)
 - черновой перевод

Statistical-based

В 2006 г. **Сергей Брин** попробовал перевести сообщение корейского пользователя через лицензированный переводческий сервис. Результат: «**Нарезанная сырая рыба обувь он хочет. Google зеленый лук штука!**».

Это стало началом разработки статистического переводчика **GoogleTranslate**

Statistical-based и Google Ocean

Книги еще один важный ресурс для

Google Translate

В 2002-2010 Google (за \$400 млн.) в рамках проекта **Google Ocean** оцифровала 129 864 880 книг.

На серверах Google более

60 петабайтов

оцифрованных книг на множестве языков.

Алексей Моисеенков



2016 представил мобильное приложение **Prisma**, которое позволяет преобразовывать изображения в стиле работ известных художников (с помощью самообучающейся нейронной сети на сервере).

Prisma
Алексей
Моисеенков



Monet → photo

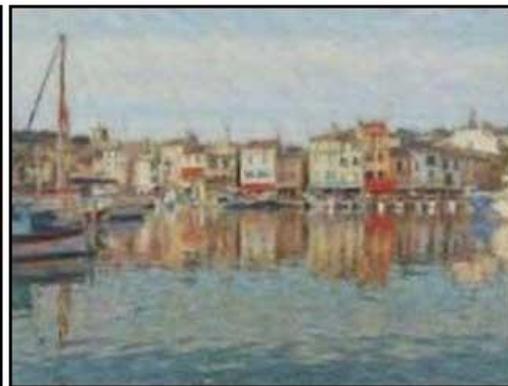


photo → Monet

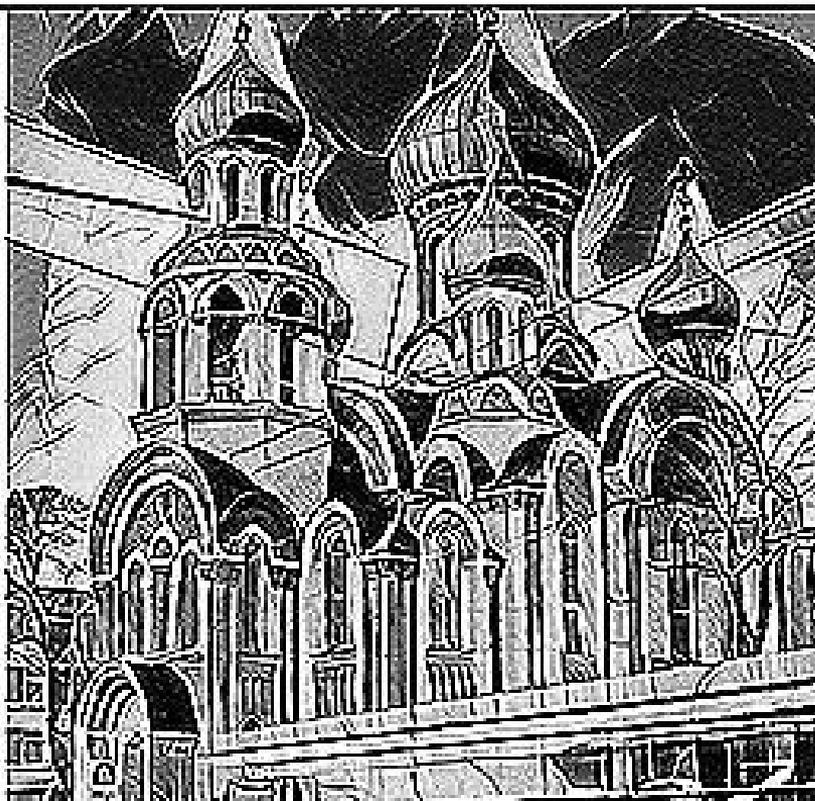
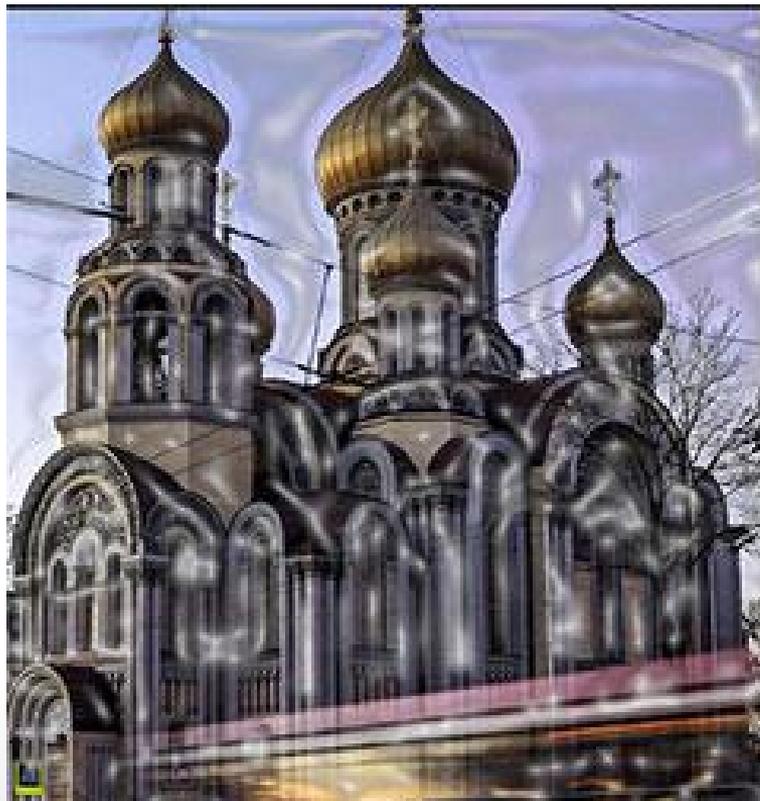
Алексей Моисеенков



Задача обучения нейросети — задача о подборе параметров. Важно количество фотографий (**текстов**) для обучения.

Важно посмотреть как влияет каждый пиксель (**слово**) на окончательный ответ и проследить корреляции между пикселями (**синтаксис**).

Язык – стиль описания реальности



Jeff Dean

11.03.2016 (на конференции Structure Data в Сан-Франциско):

- мой отдел совместно с командой переводов экспериментирует над созданием системы для перевода текстов с использованием обучаемых компьютерных алгоритмов и нейронных сетей.



Jeff Dean

11.03.2016 (на конференции Structure Data в Сан-Франциско):

- В будущем Google Translate будет производить намного более качественные переводы, по своей точности сравнимые с результатами труда профессиональных лингвистов.



Сказано - сделано

28.09.2016

Представлена **GMNT** (Google's Neural Machine Translation) – ... состоит из двух 8-слойных NN

Microsoft Bing (Neural-based machine translation)

Open Neural Machine Translation (SYSTRAN и др.)

NBMT в Baidu Translator

Время больших скоростей

Дмитрий Сошников (Microsoft):

- В ходе экспериментов процессор «прошили» на перевод с языка на язык, в результате роман «Война и мир» переводится с языка на язык за 2 секунды.



Нейросетевой переводчик Google

2013

- Статистический перевод offline в Google Translate

2018

- Нейросетевой перевод offline в Google Translate на бюджетных смартфонах
- Удвоение точности
- Нейросетевые языковые пакеты — по 35 Мб

Нейросетевой переводчик Яндекс

2018

- Нейросетевой перевод offline в Яндекс.переводчике с английского, немецкого, французского, турецкого, испанского, итальянского на русский и обратно.
- Нейросетевые языковые пакеты — по 43 Мб

Перевод с LLM

Computer Science > Computation and Language

arXiv:2304.03245 (cs)

[Submitted on 6 Apr 2023 (v1), last revised 22 May 2023 (this version, v3)]

Large language models effectively leverage document-level context for literary translation, but critical errors persist

Marzena Karpinska, Mohit Iyyer

Перевод с LLM

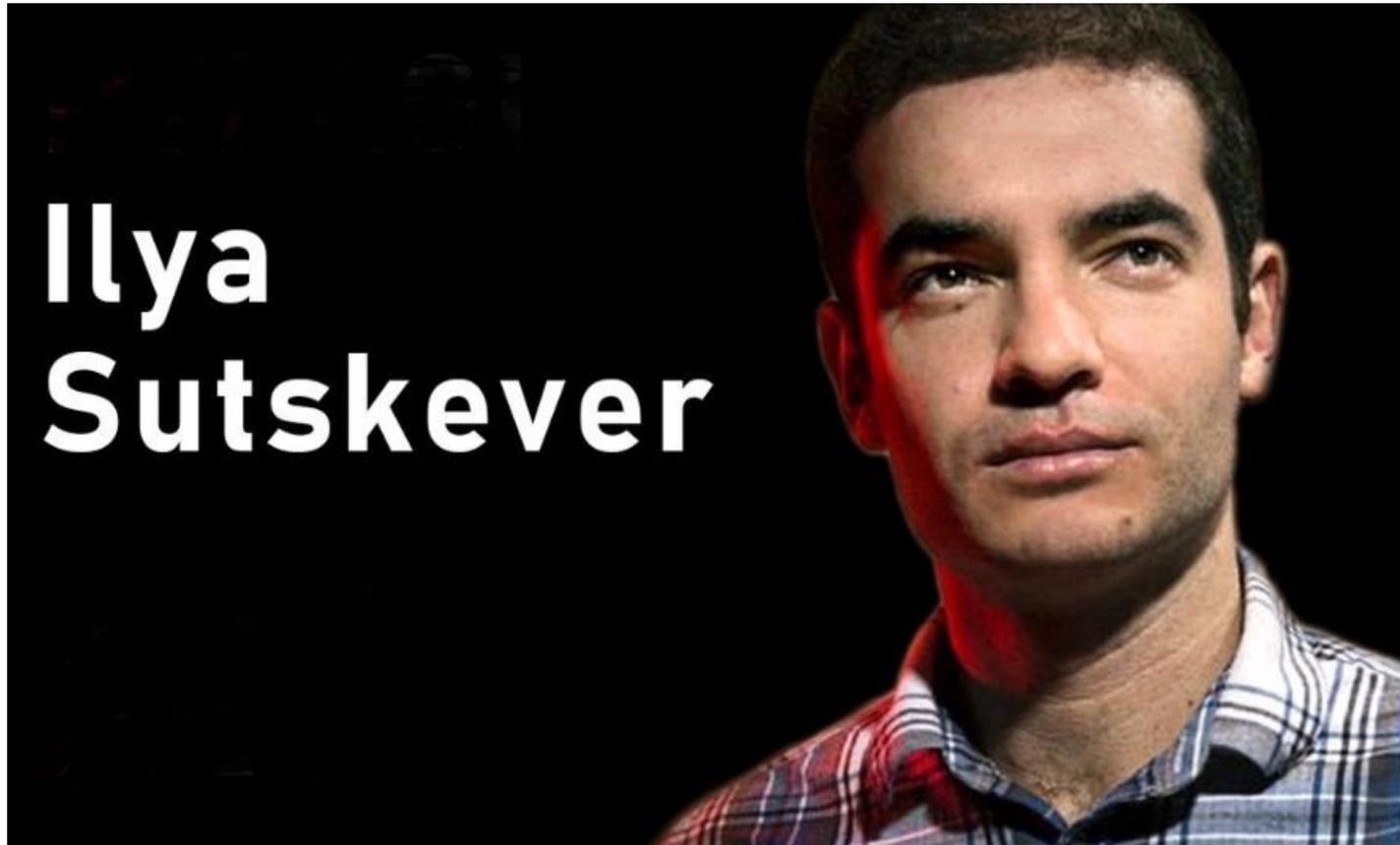
arXiv: 2304.03245 (cs)

[Отправлено 6 апреля 2023 года (версия v1), последняя редакция 22 мая 2023 года (эта версия, версия v3)]

Большие языковые модели эффективно используют контекст на уровне документа для художественного перевода, но сохраняются критические ошибки

Marzena Karpinska, Мохит Айер

Создатель ChatGPT



Революция в КП 2014 года

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

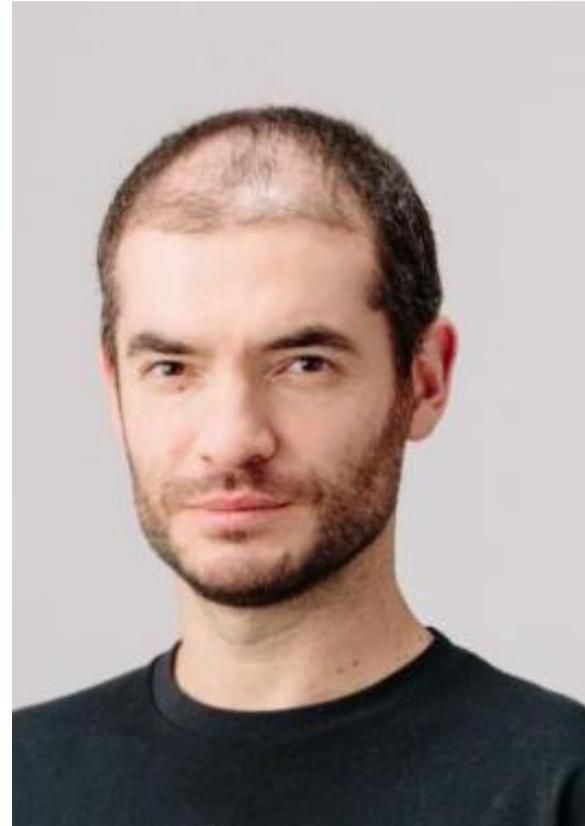
Quoc V. Le
Google
qvl@google.com

Abstract

Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever

Илья Суцкевер

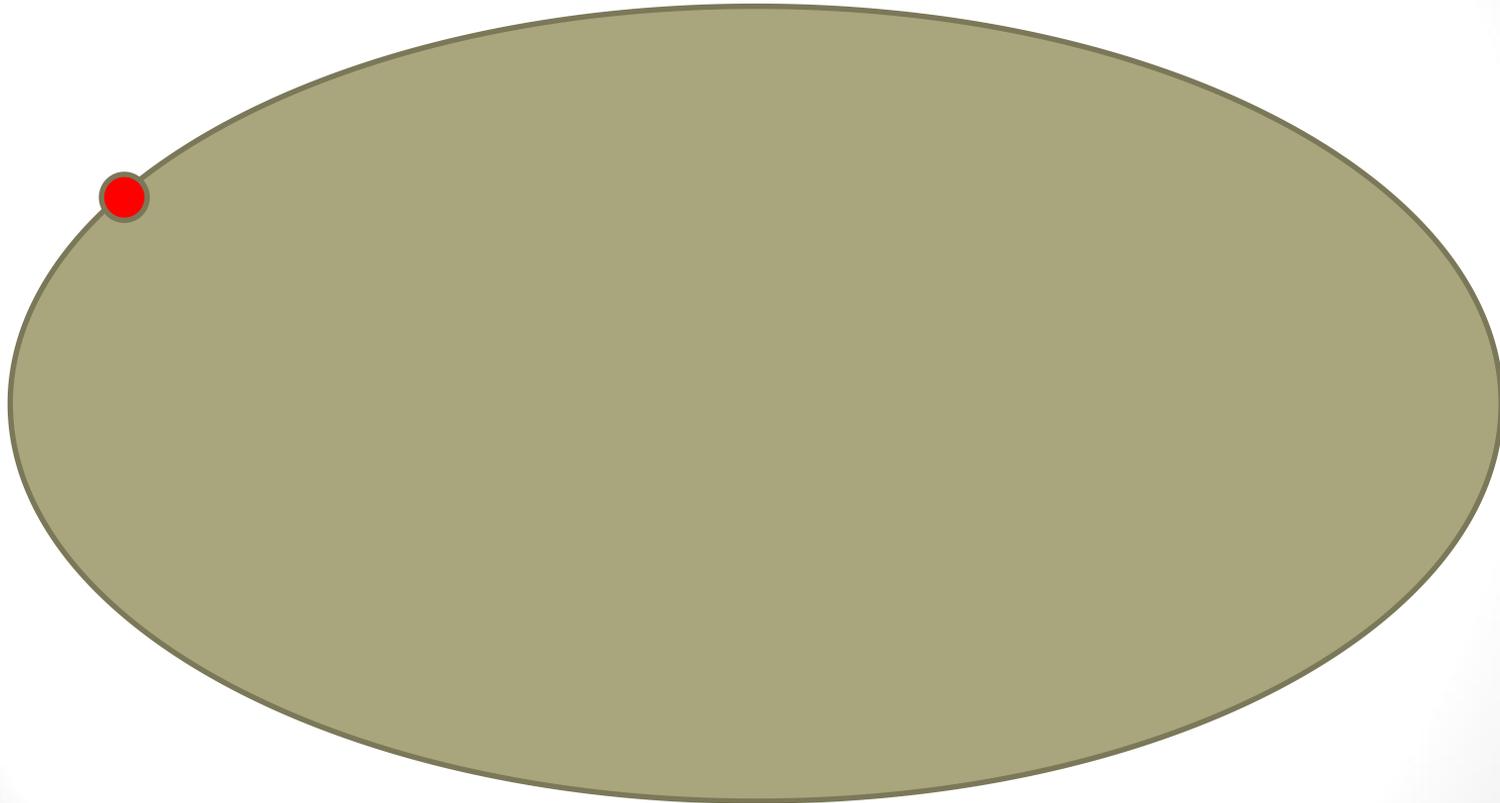
- Наша статья «Sequence to Sequence Learning with Neural Networks» (последовательное обучение нейронных сетей) вовсе не о переводе



Зачем это всё?

Расширение горизонта сознания,
как условие адекватной оценки ситуации
в преподавания языка в современной
(ПОСТ)информационной среде (развитие
способности задавать разумные вопросы
меняющемуся миру и себе самим...

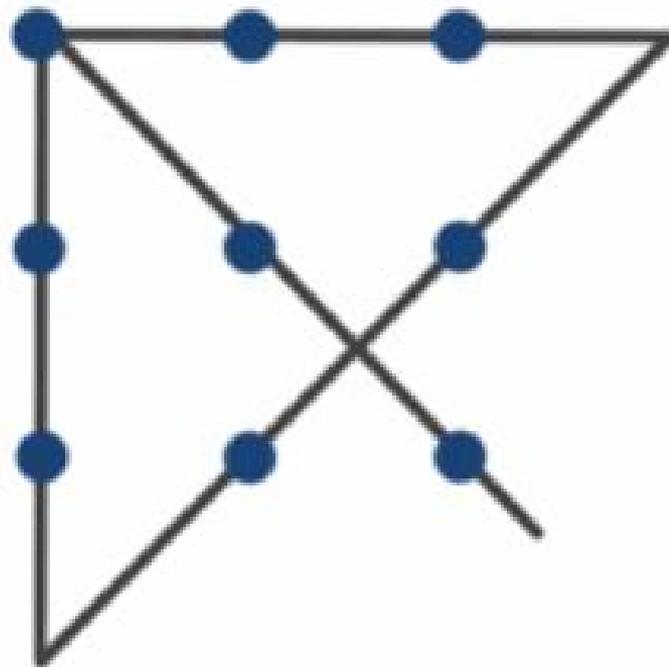
Вытеснит или нет?



Выходя за пределы



Выходя за пределы



МОРНИКИ

При обучении нейросети особым образом разбивают текст и делят даже слова на части.

NMT — не могут собрать текст, когда слова нет в их словарном запасе.

В этом случае GMNT пытается собрать его и склеить как раз по этим частям, если не выходит

КОНСТРУИРУЮТ НОВЫЕ СЛОВА

МОРНИКИ

Морники (от англ. good morning) Доброе утро
Как можно применять «Всем морники!»

Обонять (от англ. to sniff) Нюхать
Как можно применять «Пообоняй, пожалуйста,
это еще не прокисло?»

Сахариться (от англ. sugarcoat) Приукрашать
Как можно применять «Хватит сахариться, мы же
не укладываемся в дедлайн!»

МОРНИКИ

Сторонировать (глагол, от англ. reversal) Развернуть
Как можно применять «Сторонируйся — едем
обратно»

Тайма (сущ., от англ. time) Время
Как можно применять «Давай быстрее, у нас уже
совсем нет таймы»

Ушибистость (сущ., от англ. brutishness) Грубость,
Как можно применять «Этот парень какой-то
слишком ушибистый»

Грани «компьютерного перевода»

НОВЫЕ

ЛЮДИ

которые,
как всегда,
всё портят

- Процесс
- Деятельность
- Инструментарий
- Технологии
- Научная проблема
- Услуга (сервис)
- Бизнес
- Возможность
- Вызов (угроза)....

Целевая аудитория МП

- Подавляющее количество переводов осуществляется **не для людей**, а **для машин**
- Переводы для машин ~~лучше~~ эффективнее делают машины
- Переводы для людей работающих с машинами лучше делают люди работающие с машинами....

Перевод и компьютер

- Программирование - языковая деятельность, основанная на использовании различных языков и межязыковом переводе.
- Трансляция (translation) — преобразование программы, представленной на одном из компьютерных языков программирования, в функционально эквивалентную программу на другом языке.

Трансляции

- Понятие трансляции применимо и к другим компьютерным языкам (разметки, форматов данных, протоколов, описания аппаратуры, поисковых запросов).

Large Language Models (LLM)

Большая языковая модель

- LM Больших данных (7V)
- LM на основе Больших данных (7V)
- LM в мире Больших данных (7V)
- LM для обработки Больших данных (7V)

Big Data

Большие данные это данные, которые:

- огромны
- многообразны
- взаимосвязаны
- стремительно растут
- быстро меняются

7V

- Volume
- Velocity
- Variety
- Veracity
- Variability
- Visualization
- Value

- Объем
- Скорость
- Разнообразие
- Достоверность
- Изменчивость
- Визуализация
- Ценность

Эффективность технологии

Каждая технология на основе своих уникальных особенностей имеет области несомненного преимущества перед другими и, тем самым, все технологии нуждаются в развитии и совершенствовании.

Эффективность технологии

Оценка технологии должна учитывать:

- развитость технологии
- область ее функционального приложения
(адаптивный перевод, адекватный перевод, аутентичный перевод, буквальный перевод, художественный перевод и т.д.)
- число и характер языковых пар

Эффективность технологии

Оценка технологии должна учитывать:

- **трудоемкость использования** статистические технологии перевода предельно просты для пользователя. Технология Transfer также может быть применена без учета настроек перевода, но, в таком случае, качество результата резко снижается.

Эффективность технологии

Оценка технологии должна учитывать:

- **эвристичность**

Компьютерный перевод не только предоставляет инструментарий решения прикладных задач, но и открывает новые горизонты лингвистических исследований, формирует новое проблемное поле и меняет наше понимание языка.

Эффективность технологии

Каждая технология на основе своих уникальных особенностей имеет области несомненного преимущества перед другими и, тем самым, все технологии нуждаются в развитии и совершенствовании.

Комплементарные технологии

- информационного поиска
- систем управления базами данных
- статистического анализа
- фоносемантического анализа
- машинной памяти
- распознавания образов (OCR)
- data mining

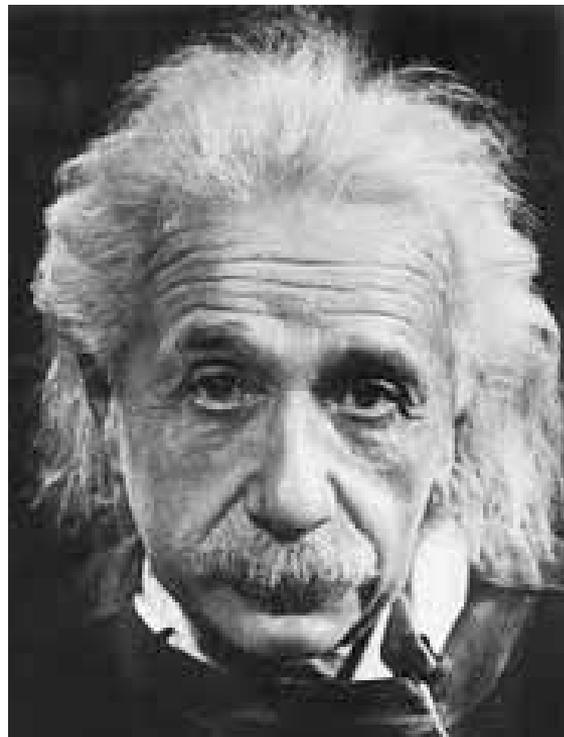
Комплементарные технологии

- синтеза речи по тексту,
- распознавания речи,
- автогенерации текста
- программирования
- метаданных (семантической разметки)
- интеграции данных
- Semantic Web technologies

Образование

Альберт Эйнштейн:

Образование, это то,
что остается,
когда забываешь все,
ему тебя научили.



Информационные технологии

- **технологии** создания, поиска, хранения, обработки, защиты, передачи ...информации.
- **технологии решения задач**.... на основе... информационной составляющей... с помощью ... информационных средств...
- **технологии выживания человека** в информационном обществе потребления...