

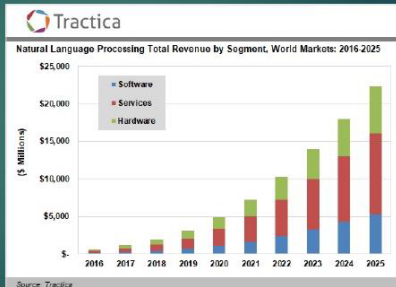
Введение

Алексей Сорокин

МГУ им. М. В. Ломоносова
весенний семестр 2022–2023 учебного года
Межфакультетский курс “Введение в компьютерную
лингвистику” 15 февраля, занятие 1

Рынок обработки естественного языка

Natural Language Processing Market to Reach \$22.3 Billion by 2025



Коммерческие приложения

- Пользовательские интерфейсы:
 - Голосовые помощники.
 - Чатботы.
 - Автоматическая техподдержка.
- Интерфейсы человек-компьютер:
 - Преобразование текст ↔ речь.
 - Исправление опечаток и грамматических ошибок.
 - Автоматический перевод.
- Извлечение информации из текста:
 - Автоматическая классификация (тематика, интенты).
 - Распознавание именованных сущностей.
 - Анализ тональности.
- Образовательные приложения.

Chatbot Landscape 2017 by KeyReply

The infographic is a circular chart divided into three main sections: **Marketing**, **Support**, and **Managed**. The central circle contains logos for popular messaging apps: WhatsApp, WeChat, Telegram, Kik, Line, and SMS. The outer ring lists various chatbot providers, including:

- Marketing:** Verbio, interactions, chatsuite, botmetrics, noHold, next IT, clustaar, wit.ai, api.ai, dashbot.io, botimize, Converse AI, MindiQ, motion ai, TALBOT, speech i, IBM Watson, Amazon Lex, Drift, FLOWXO, reply.ai, chatly PEOPLE, LUIS, and many others.
- Support:** Microsoft, CNTK, Chainer, Caffe, torch, dmlc, mxnet, theano, Apache SINGA, OPERABLE, ARTIFICIAL SOLUTIONS, COGNICOR, SERVICE FRIEND, DigitalGenius, inbenta, agentbot, KORE, [24]7, KASISTO, KeyReply, K, and others.
- Managed:** im, person, MINDFIELD, snaps, pypestream, TouchCommerce, creativevirtual, ChatGrid, PullString™, CONVERSABLE, Chatfuel, OCTANE AI, ManyChat, Surveybot, SEQUEL, gupshup, botfuel, Expedia, lyR, eBay, hiPUNK, KAYAK, skyscanner, L'ORÉAL, MARCÓN 5, Forbes, TechCrunch, TOMMY HILFGER, ESEPHORA NORDSTROM, BURBERRY, TED, CNN, WSJ, CapitalLand, Disney, esurance, Permod Record, WHOLE FOODS, gov.sg, PIZZA HUT, HSBC, OCBC Bank, TATA CAPITAL, MasterCard, RBS, KLM, DBS, HYATT, leumi card, EVERNOTE, BOSCH, republie, vodafone, ILLUMINATE, Deutsche Post, VW, hp, J.W. MARRIOTT, STAPLES, SONY, Alibaba.com, J.P.Morgan, Bank of America, Capital One, and others.

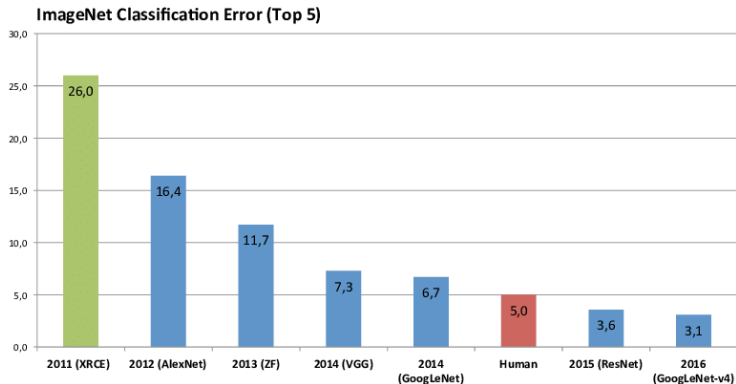
Эволюция алгоритмов

- До 1990-х: model-driven (модели, правила, словари).
- 1990-2005: статистические.

Эволюция алгоритмов

- До 1990-х: model-driven (модели, правила, словари).
- 1990-2005: статистические.
- С 2005: data-driven (машинное обучение, гибридные модели).
 - С 2015 – нейронные сети.
 - С 2019 – предобученные трансформерные нейронные сети

Качество алгоритмов



Качество модели

- Качество морфосинтаксического анализа (CoNLL 2018):

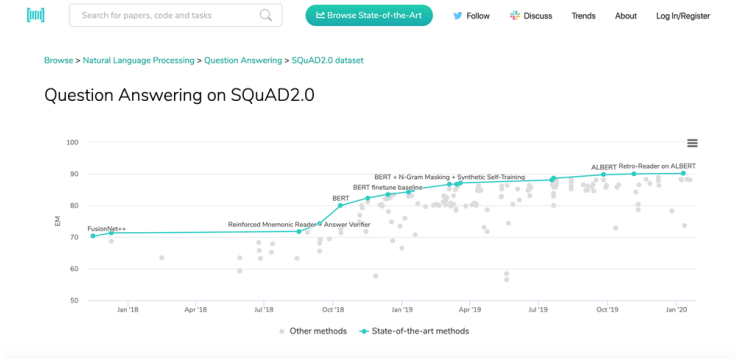
Корпус	Размер	Качество (тэги)	Качество (LAS)
English EWT	254K	93.94	84.57
Russian Syntagrus	1,103K	96.90	92.48
Spanish Ancora	549K	97.82	90.93
French Sequoia	70K	97.02	89.89

Качество модели

- Качество морфосинтаксического анализа (CoNLL 2018):

Корпус	Размер	Качество (тэги)	Качество (LAS)
English EWT	254K	93.94	84.57
English LinES	82K	93.69	81.97
Russian Syntagrus	1,103K	96.90	92.48
Russian Taiga	38K	88.00	74.24
Spanish Ancora	549K	97.82	90.93
French Sequoia	70K	97.02	89.89
French Spoken	34K	94.52	75.78

Прогресс качества ответов на вопросы на SQuAD 2.0



GPT-2

SYSTEM PROMPT
(HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.


MODEL
COMPLETION
(MACHINE-
WRITTEN, 10
TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.


Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.


ChatGPT




A car drives at 55 miles per hour for 7 hours. What is the total distance driven?




The total distance driven is 385 miles. (55 miles per hour * 7 hours = 385 miles)




How far away is it from where it started?




The distance from where the car started would be the same as the total distance driven, which is 385 miles.




But if it did not drive in a straight line?




If the car did not drive in a straight line, the total distance driven would still be the same (385 miles), but the distance from where the car started might be different, as it would depend on the specific route the car took. In order to determine the distance from the starting point one would need a map and the complete route the car took.



It drove north for 3 hours and then east for 4 hours.

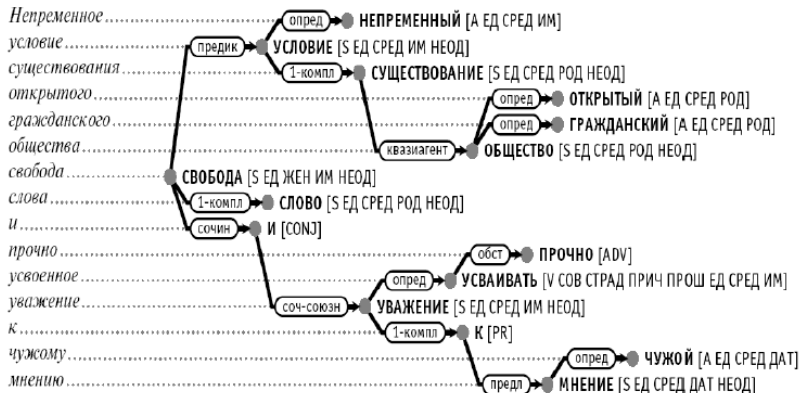


If the car drove north for 3 hours at 55 miles per hour, it would have traveled 165 miles north. And then, if it drove east for 4 hours at 55 miles per hour, it would have traveled 220 miles east. The distance from the starting point would be the hypotenuse of the right triangle formed by the north-south and east-west distances, which can be found by using the Pythagorean theorem, that is, the square root of $(165^2 + 220^2) = \text{sqrt}(72,725) \approx 269.8$ miles.



With the same speed, driving in a straight line instead, how long would it take to get back to where it started?

Экспертная разметка



Готовые данные

- Лучше всего, когда можно взять готовые данные.
 - Машинный перевод: параллельные тексты (протоколы Европарламента/ООН, СМИ).

Готовые данные

- Лучше всего, когда можно взять готовые данные.
 - Машинный перевод: параллельные тексты (протоколы Европарламента/ООН, СМИ).
 - Диалоговые системы: диалоги техподдержки, диалоги в фильмах/книгах.
 - Оценка тональности: рейтинги фильмов/ресторанов/гостиниц.
 - Распознавание лиц: соцсети, СМИ.

Готовые данные

- Лучше всего, когда можно взять готовые данные.
 - Машинный перевод: параллельные тексты (протоколы Европарламента/ООН, СМИ).
 - Диалоговые системы: диалоги техподдержки, диалоги в фильмах/книгах.
 - Оценка тональности: рейтинги фильмов/ресторанов/гостиниц.
 - Распознавание лиц: соцсети, СМИ.
- Проблемы: поиск источников, наличие защиты, авторские права.

Готовые данные

- Лучше всего, когда можно взять готовые данные.
 - Машинный перевод: параллельные тексты (протоколы Европарламента/ООН, СМИ).
 - Диалоговые системы: диалоги техподдержки, диалоги в фильмах/книгах.
 - Оценка тональности: рейтинги фильмов/ресторанов/гостиниц.
 - Распознавание лиц: соцсети, СМИ.
- Проблемы: поиск источников, наличие защиты, авторские права.
- Данные можно извлечь из логов работы системы:
 - Информационный поиск: кликнул ли пользователь по ссылке.

Готовые данные

- Лучше всего, когда можно взять готовые данные.
 - Машинный перевод: параллельные тексты (протоколы Европарламента/ООН, СМИ).
 - Диалоговые системы: диалоги техподдержки, диалоги в фильмах/книгах.
 - Оценка тональности: рейтинги фильмов/ресторанов/гостиниц.
 - Распознавание лиц: соцсети, СМИ.
- Проблемы: поиск источников, наличие защиты, авторские права.
- Данные можно извлечь из логов работы системы:
 - Информационный поиск: кликнул ли пользователь по ссылке.
 - Исправление опечаток (в поиске): какую из ссылок выбрал пользователь.
 - Машинный перевод: было ли исправление предложенного перевода, скопировал ли его пользователь.

Неэкспертная разметка

- Сервисы для автоматической разметки.
 - Mechanical Turk.
 - Яндекс.Толока

Неэкспертная разметка

- Сервисы для автоматической разметки.
 - Mechanical Turk.
 - Яндекс.Толока
- Можно размечать низкоуровневые задачи:
 - Анализ тональности.
 - Оценка смысловой схожести слов.
 - Оценка качества текста (переводы, диалоги).

Неэкспертная разметка

- Сервисы для автоматической разметки.
 - Mechanical Turk.
 - Яндекс.Толока
- Можно размечать низкоуровневые задачи:
 - Анализ тональности.
 - Оценка смысловой схожести слов.
 - Оценка качества текста (переводы, диалоги).
- Сложности неэкспертной разметки:
 - Исполнители низкого качества.
 - Нужны детальные инструкции.

Неэкспертная разметка

- Возможные сложности при исправлении опечаток:
 - Неквалифицированный (неграмотный) разметчик.

Неэкспертная разметка

- Возможные сложности при исправлении опечаток:
 - Неквалифицированный (неграмотный) разметчик.
 - Вариативность написания (заимствования, имена собственные).
 - Нужно ли исправлять слэнг?
 - Нужно ли исправлять речевые ошибки?

Неэкспертная разметка

- Возможные сложности при исправлении опечаток:
 - Неквалифицированный (неграмотный) разметчик.
 - Вариативность написания (заимствования, имена собственные).
 - Нужно ли исправлять слэнг?
 - Нужно ли исправлять речевые ошибки?
- Меры улучшения качества разметки:
 - Тестовые задания.
 - Экспертная верификация.

Неэкспертная разметка

- Возможные сложности при исправлении опечаток:
 - Неквалифицированный (неграмотный) разметчик.
 - Вариативность написания (заимствования, имена собственные).
 - Нужно ли исправлять слэнг?
 - Нужно ли исправлять речевые ошибки?
- Меры улучшения качества разметки:
 - Тестовые задания.
 - Экспертная верификация.
 - Разметка с перекрытием.
 - Разные веса для разных пользователей.

Использование существующей разметки

- Удобно использовать существующие ресурсы:
 - Лингвистически аннотированные корпуса.
 - Онлайн словари-справочники.

Использование существующей разметки

- Удобно использовать существующие ресурсы:
 - Лингвистически аннотированные корпуса.
 - Онлайн словари-справочники.
- Пример – корпус TOEFL:
 - Автоматическое исправление грамматики / орфографии.
 - Оценка качества текста (essay scoring).
 - Определение родного языка.

Генерация данных

- Способы генерации данных:

- Генерация по шаблонам:

Фильм X мне показался очень Y -ым.

Генерация данных

- Способы генерации данных:
 - Генерация по шаблонам:

Фильм X мне показался очень Y-ым.
 - Автоматическое перефразирование, замена синонимами.
 - Автоматический перевод, двунаправленный перевод (back-translation).

Генерация данных

- Способы генерации данных:
 - Генерация по шаблонам:

Фильм X мне показался очень Y-ым.
 - Автоматическое перефразирование, замена синонимами.
 - Автоматический перевод, двунаправленный перевод (back-translation).
- Направленный поиск неразмеченных текстов (опечатки – поиск слов не из словаря).
- Если есть качественные языковые модели – дообучение на порождение данных с нужными свойствами.

Правиловые модели

- Правиловые модели – ответ в задаче определяется по правилам:
 - Анализ тональности: метка положительная, если есть положительно окрашенная характеристика (*вкусная еда*).

Правиловые модели

- Правилаовые модели – ответ в задаче определяется по правилам:
 - Анализ тональности: метка положительная, если есть положительно окрашенная характеристика (*вкусная еда*).
- Недостатки:
 - Правила усложняются: *еду в этом ресторане не назовёшь вкусной*.
 - Правила используют внешнюю информацию (морфологию, синтаксис, словари).

Статистические модели

- Статистические модели опираются на вероятностную модель.
- Пример – наивный байесовский классификатор:

$$p(y|x) \sim \prod_i p(x_i|y)p(y)$$

Статистические модели

- Статистические модели опираются на вероятностную модель.
- Пример – наивный байесовский классификатор:

$$p(y|x) \sim \prod_i p(x_i|y)p(y)$$

- Недостатки:
 - Опираются на жёсткие вероятностные предположения, не отражающие реальных данных.

Статистические модели

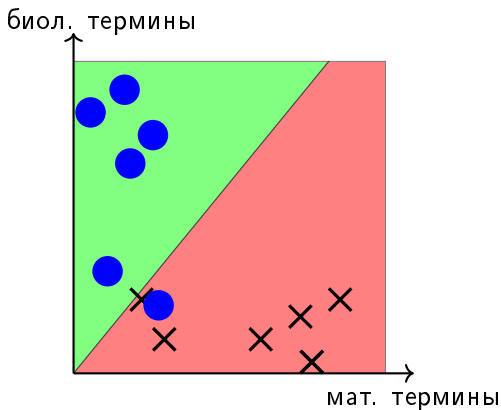
- Статистические модели опираются на вероятностную модель.
- Пример – наивный байесовский классификатор:

$$p(y|x) \sim \prod_i p(x_i|y)p(y)$$

- Недостатки:
 - Опираются на жёсткие вероятностные предположения, не отражающие реальных данных.
 - Признаки чувствительны к обучающему корпусу и алгоритмам извлечения.

Машинное обучение

Упрощённая тематическая классификация на 2 класса:



Машинное обучение

- Базовая модель – линейный классификатор:

$$\begin{aligned}f(x) &= \operatorname{sgn}(\langle w, x \rangle - w_0), \\f(x) &= \operatorname{argmax}_i \langle w_i, x \rangle - w_{i0}\end{aligned}$$

- Входным данным сопоставляется вектор признаков $[x_1, \dots, x_n]$.

Машинное обучение

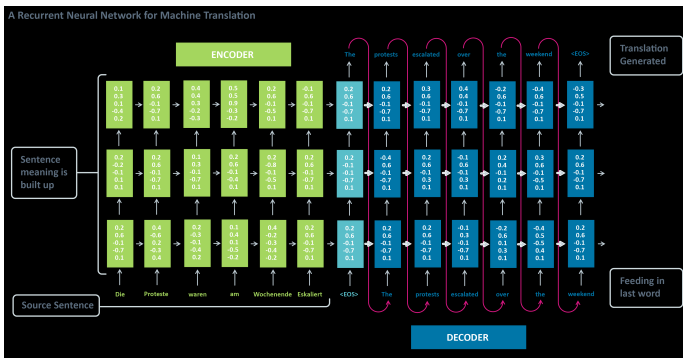
- Базовая модель – линейный классификатор:

$$\begin{aligned}f(x) &= \operatorname{sgn}(\langle w, x \rangle - w_0), \\f(x) &= \operatorname{argmax}_i \langle w_i, x \rangle - w_{i0}\end{aligned}$$

- Входным данным сопоставляется вектор признаков $[x_1, \dots, x_n]$.
- Недостатки:
 - Ручное конструирование признаков.
 - Сложно адаптировать к конкретной задаче.

Нейронные сети

- Нейронные сети: сложная архитектура, состоящая из базовых блоков:



Нейронные сети: компоненты

- Основные составляющие (до 2019 года):
 - Векторные представления слов (эмбединги).
 - Свёрточные сети.
 - Рекуррентные сети.
 - Энкодер-декодер компоненты.
 - Механизм внимания (attention).
 - Полносвязные сети (линейные классификаторы).
- Архитектура подбирается исходя из задачи, данных и вычислительных ресурсов.

Нейронные сети: предобучение на языковом моделировании

- Задача языкового моделирования – предсказание следующего слова в тексте (распределения вероятностей следующего слова).

Я вчера ел вкусную ?

Нейронные сети: предобучение на языковом моделировании

- Задача языкового моделирования – предсказание следующего слова в тексте (распределения вероятностей следующего слова).

Я вчера ел вкусную ?

- Она требует больших знаний о языке:
 - Морфология (если следующее слово существительное, то женского рода).
 - Графика (следующее слово с большой вероятностью кончается на -у).
 - Синтаксис.
 - Семантика (следующее слово “съедобное”).

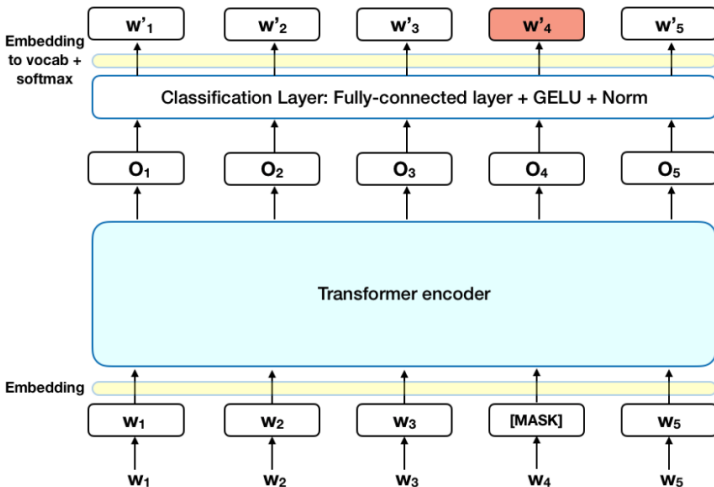
Нейронные сети: предобучение на языковом моделировании

- Задача языкового моделирования – предсказание следующего слова в тексте (распределения вероятностей следующего слова).

Я вчера ел вкусную ?

- Она требует больших знаний о языке:
 - Морфология (если следующее слово существительное, то женского рода).
 - Графика (следующее слово с большой вероятностью кончается на -у).
 - Синтаксис.
 - Семантика (следующее слово “съедобное”).
- Кроме того, для получения этой информации не нужны размеченные данные.

Нейронные сети: BERT



Нейронные сети: BERT

- Глубокая (12 слоёв) архитектура на основе трансформера (механизма внимания).
- Преобразует последовательность индексов токенов в последовательность векторов.
- Дополнительный вектор для токена начала предложения (CLS-токен).

Нейронные сети: BERT

- Глубокая (12 слоёв) архитектура на основе трансформера (механизма внимания).
- Преобразует последовательность индексов токенов в последовательность векторов.
- Дополнительный вектор для токена начала предложения (CLS-токен).
- Предобучается на двух задачах:
 - Восстановлении пропущенного токена.
 - Проверке, идут ли два предложения друг за другом.

Нейронные сети: BERT

- Глубокая (12 слоёв) архитектура на основе трансформера (механизма внимания).
- Преобразует последовательность индексов токенов в последовательность векторов.
- Дополнительный вектор для токена начала предложения (CLS-токен).
- Предобучается на двух задачах:
 - Восстановлении пропущенного токена.
 - Проверке, идут ли два предложения друг за другом.
- Восстановление пропущенных слов – информация на уровне слов.
- Проверка следования предложений – задачи на уровне предложений / пар предложений.

GPT-2

SYSTEM PROMPT
(HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL
COMPLETION
(MACHINE-
WRITTEN, 10
TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

GPT-2

- GPT-2 – однонаправленная языковая модель на основе Трансформера.
- Она обучалась только на задачу предсказания следующего слова:
 - 1,5 миллиарда параметров.
 - Обучение на 8 миллионах разнообразных интернет-страниц (порядка 40Гб текста).

GPT-2

- GPT-2 – однонаправленная языковая модель на основе Трансформера.
- Она обучалась только на задачу предсказания следующего слова:
 - 1,5 миллиарда параметров.
 - Обучение на 8 миллионах разнообразных интернет-страниц (порядка 40Гб текста).
- Основное применение – few-shot learning: обучение на новую задачу по очень небольшому (10-100) количеству примеров.
- При этом задача должна быть сформулирована как языковое моделирование.

GPT-2: постановки задач

- Анализ тональности:

В этом ресторане отличная еда. sentiment= ?

- Переформулировка на языке GPT:

В этом ресторане отличная еда. Это X . – контекст

$p(X = \text{хорошо}) \geq p(X = \text{плохо}) \rightarrow \text{sentiment} = \text{positive}$

$p(X = \text{плохо}) \geq p(X = \text{хорошо}) \rightarrow \text{sentiment} = \text{negative}$

- Так можно переформулировать и другие задачи (ответ на вопросы, автоматическое реферирование, ...)

Нейронные сети: тенденции

- Предобучение на общем корпусе большого размера:
 - Распознавание изображений: ImageNet.
 - Обработка текста: языковое моделирование (GPT, BERT).

Нейронные сети: тенденции

- Предобучение на общем корпусе большого размера:
 - Распознавание изображений: ImageNet.
 - Обработка текста: языковое моделирование (GPT, BERT).
- Многоязычное обучение:
 - Google – общая архитектура перевода для всех языков, язык как дополнительные данные.
 - BERT – общая модель и представление данных для 104 языков.

Нейронные сети: тенденции

- Предобучение на общем корпусе большого размера:
 - Распознавание изображений: ImageNet.
 - Обработка текста: языковое моделирование (GPT, BERT).
- Многоязычное обучение:
 - Google – общая архитектура перевода для всех языков, язык как дополнительные данные.
 - BERT – общая модель и представление данных для 104 языков.
- Максимально простое представление входов: последовательность символов или BPE (Byte-Pair Encoding)

Нейронные сети: тенденции

- Предобучение на общем корпусе большого размера:
 - Распознавание изображений: ImageNet.
 - Обработка текста: языковое моделирование (GPT, BERT).
- Многоязычное обучение:
 - Google – общая архитектура перевода для всех языков, язык как дополнительные данные.
 - BERT – общая модель и представление данных для 104 языков.
- Максимально простое представление входов: последовательность символов или BPE (Byte-Pair Encoding)
- Дополнительное (динамическое) обучение с подкреплением на оценках человека.