Автоматическая обработка текстов

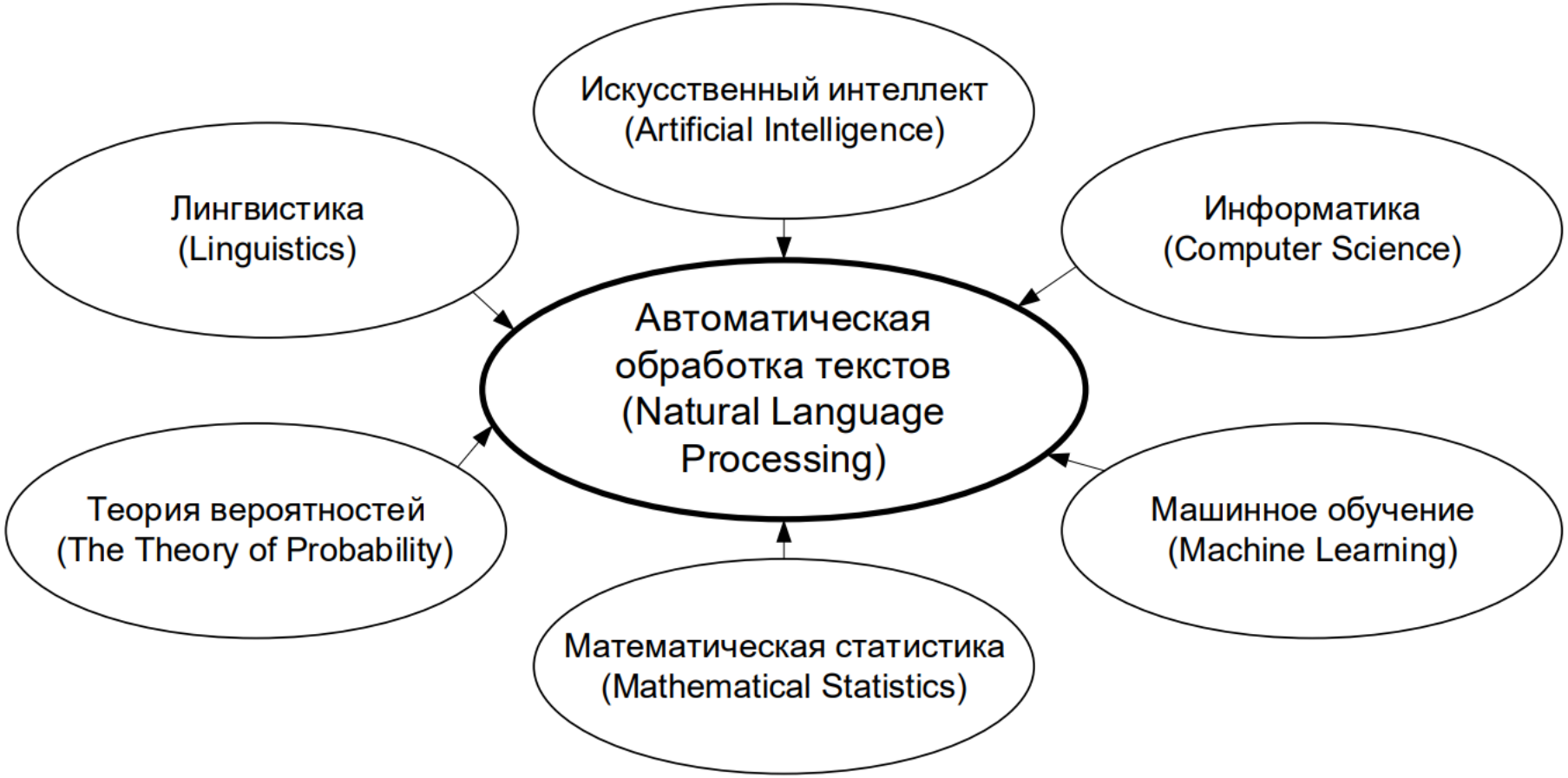
Введение

2025

Основные определения

- **Автоматическая обработка текстов** (АОТ) – междисциплинарную область, которая объединяет лингвистику, информатику и искусственный интеллект для решения задач анализа, понимания и генерации текстов на естественном языке
- **Natural Language Processing** (NLP) – общее направление искусственного интеллекта и математической лингвистики, которое изучает проблемы компьютерного анализа (понимание) и синтеза (генерация) текстов на ЕЯ
- **Компьютерная лингвистика** – научное направление в области математического и компьютерного моделирования интеллектуальных процессов, которое ставит своей целью использование математических моделей для описания естественных языков
  В отличие от NLP, которое фокусируется на практических приложениях, компьютерная лингвистика углубляется в основополагающие лингвистические теории и модели

Место АОТ


Искусственный интеллект
(Artificial Intelligence)
Лингвистика
(Linguistics)
Информатика
(Computer Science)
Автоматическая обработка текстов
(Natural Language Processing)
Теория вероятностей
(The Theory of Probability)
Машинное обучение
(Machine Learning)
Математическая статистика
(Mathematical Statistics)

Ключевые конференции

- ACL – Association of Computational Linguistics
- NAACL – North American Association for Computational Linguistics
- EMNLP – Empirical Methods in Natural Language Processing
- COLING – International Conference on Computational Linguistics
- CoNLL – Conference on Computational Natural Language Learning
- EACL – European Association of Computational Linguistics

- Россия:
  - «Диалог»
  - AIST – Analysis of Images, Social Networks and Texts
  - AINL – Artificial Intelligence and Natural Language

Связь с прикладными областями

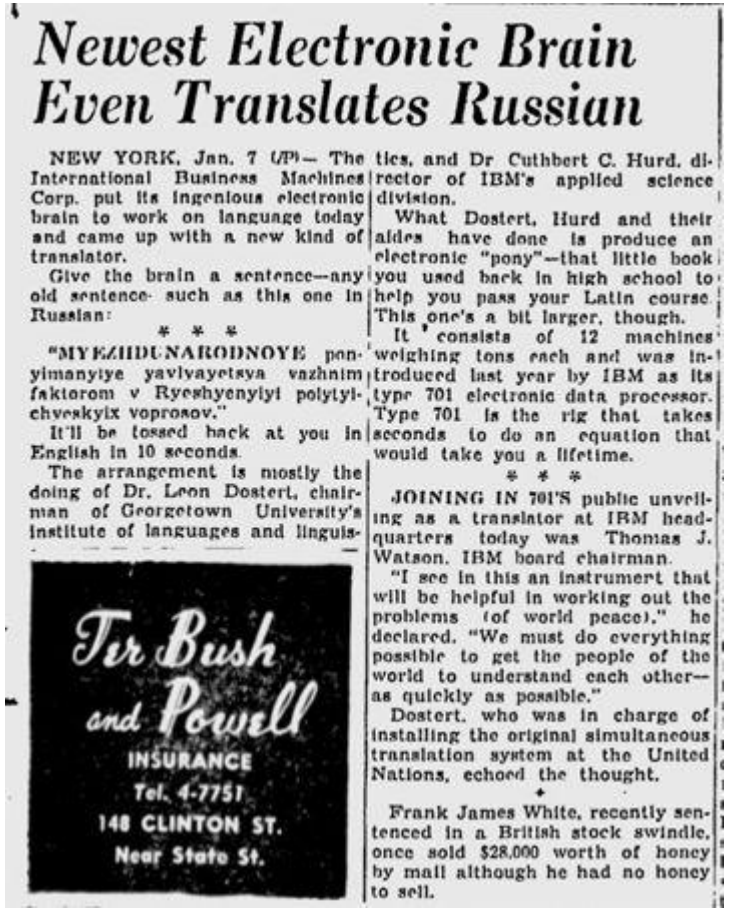
- Text mining
- Web mining
- Social media analysis
- Information retrieval (IR)
- Speech recognition
- Text-to-speech (TTS) и Speech-to-text (STT)
- Optical character recognition (OCR)
- Recommender systems

Исторический обзор

1950-е гг.

История автоматической обработки текстов началась в середине XX века

Джорджтаунский эксперимент 1954 года считается отправной точкой развития NLP. Компания IBM совместно с Джорджтаунским университетом публично продемонстрировала, как компьютер переводит более 60 предложений с русского языка на английский, используя ограниченный набор из 250 слов и 6 правил



Newest Electronic Brain Even Translates Russian

NEW YORK, Jan. 7 (AP)— The International Business Machines Corp. put its ingenious electronic brain to work on language today and came up with a new kind of translator.

Give the brain a sentence—any old sentence such as this one in Russian:

* * *

"MYEZHDUNARODNOYE ponyimanyiye yavlyayetsya vazhnim faktorom v Ryeshyenyiyi polityicheskyix voprosov."

It'll be tossed back at you in English in 10 seconds.

The arrangement is mostly the doing of Dr. Leon Dostert, chairman of Georgetown University's institute of languages and linguistics, and Dr Cuthbert C. Hurd, director of IBM's applied science division.

What Dostert, Hurd and their aides have done is produce an electronic "pony"—that little book you used back in high school to help you pass your Latin course. This one's a bit larger, though.

It consists of 12 machines weighing tons each and was introduced last year by IBM as its type 701 electronic data processor. Type 701 is the rig that takes seconds to do an equation that would take you a lifetime.

* * *

JOINING IN 701'S public unveiling as a translator at IBM headquarters today was Thomas J. Watson, IBM board chairman.

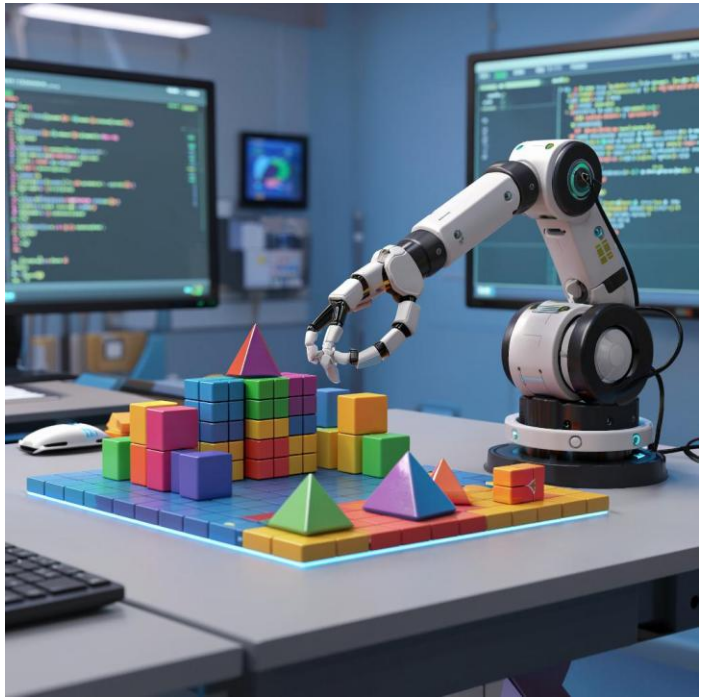
"I see in this an instrument that will be helpful in working out the problems (of world peace)," he declared. "We must do everything possible to get the people of the world to understand each other—as quickly as possible."

Dostert, who was in charge of installing the original simultaneous translation system at the United Nations, echoed the thought.

*

Frank James White, recently sentenced in a British stock swindle, once sold $28,000 worth of honey by mail although he had no honey to sell.

1960-1970-е гг.



В 1960-х годах появилась интерактивная система SHRDLU — парсер с небольшим словарем, который определял главные сущности в предложении

В 1970-х годах В. Вудс (William. Woods) с коллегами предложил расширенную систему переходов *Augmented transition network*— графовую структуру, использующую идею конечных автоматов для парсинга предложений

1980-1990-е гг.

После 1980-х годов для решения NLP-задач начали активно применяться алгоритмы машинного обучения

В 1990-х годах стали популярны n-граммы, а в 1997 году была предложена модель LSTM (Long-short memory)

2000-2010-е гг.

2007 году LSTM была реализована на практике

В 2011 году появился персональный помощник Siri от Apple, за которым последовали другие голосовые ассистенты

В том же году система Watson от IBM победила в игре «Jeopardy!».

2010-2020-е гг.

В 2013 году была представлена модель Word2vec, которая революционизировала представление слов в векторном пространстве

В 2017 году Google представил архитектуру Transformer в работе "Attention Is All You Need". Это привело к созданию моделей BERT (2018) и GPT (2018), которые установили новые стандарты в области NLP

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

2020-е гг.

Широкое распространение LLM

Появление мультимодальных LLM и LRM

Развитие автономных агентов

Язык

Понятие

- **Язык** — это структурированная система коммуникации, состоящая из грамматики и словаря, которая является основным средством передачи смысла людьми
- В лингвистике язык понимается как "естественный человеческий язык вообще и все языки мира как индивидуальные его представители". Лингвистика изучает не только существующие языки, но и человеческий язык как универсальное явление

- **Структурная** лингвистика рассматривает язык как систему взаимосвязанных элементов, где каждый уровень (фонетический, морфологический, синтаксический, дискурсивный) разбивается на минимальные единицы. Анализ включает создание инвентарей (фонемы, морфемы, лексические классы) для изучения их взаимосвязей в иерархии структур
- **Функциональная** лингвистика дополняет структурный анализ назначением семантических и других функциональных ролей каждой единице. Язык рассматривается как инструмент коммуникации, и лингвистические формы объясняются через их функциональную ценность
- **Когнитивная** лингвистика исследует взаимосвязь между языком и когнитивными процессами — восприятием, мышлением и памятью. Основная идея заключается в том, что язык управляется концептуальными метафорами, а концепты структурируются культурным познанием

Генеративный подход Хомского

- Язык как индивидуальное ментальное явление
  язык трактуется как внутренний, индивидуальный, ментальный объект, т.е. система знаний, реализованная в мозге конкретного носителя, а не как внешний социальный код или совокупность наблюдаемых корпусов
- Формальная математическая теория грамматики
- Врожденность языковой способности

Дискурсивный подход

- Изучние языка «в употреблении», где язык рассматривают как социально и когнитивно обусловленную практику
- методы включают анализ связности и макроструктур (*ван Дейк*), критический дискурс-анализ (*Фэркло*, *Водак*), конверсационный анализ (*Сакс*, *Шеглофф*, *Джефферсон*), этнографию коммуникации (*Хаймс*), британо-бирмингемскую школу (*Синклер–Коултхард*) и др.

Классификация

- Естественные языки — это языки, используемые для общения людей, которые характеризуются широкой сферой применения, гибкостью, открытостью и динамичностью. К ним относятся все национальные языки (русский, английский, китайский и др.).
- Формальные языки — это языки, в которых одинаковые сочетания знаков всегда имеют одинаковый смысл. К формальным языкам относятся:

Классификация

- Естественные языки — это языки, используемые для общения людей, которые характеризуются широкой сферой применения, гибкостью, открытостью и динамичностью. К ним относятся все национальные языки (русский, английский, китайский и др.).
- Формальные языки — это языки, в которых одинаковые сочетания знаков всегда имеют одинаковый смысл. К формальным языкам относятся:
  - Системы математических и химических символов
  - Нотная грамота
  - Азбука Морзе
  - Языки программирования
  - Десятичная система счисления

Классификации ЕЯ

Принято выделять три взаимодополняющих классификации естественных языков:

- генетическую (по родству),
- типологическую (по структурным признакам),
- ареальную (по географическим контактам и конвергенции)

Типологическая классификация ЕЯ

Объединение языков по сходным структурным параметрам независимо от родства;

Наиболее известна морфологическая типология:

- **Изолирующие** (аморфные): минимальная морфология, грамматические значения выражаются синтаксисом и служебными словами; классический пример — китайские языки
- **Агглютинативные**: последовательное присоединение аффиксов с относительно однозначными значениями и четкими границами морфем; примеры — тюркские, финский
- **Флективные** (фузионные): аффиксы часто многозначны, границы размыты; примеры — русский, латинский
- **Инкорпорирующие** (полисинтетические): включение аргументов в глагольный комплекс; примеры — эскимосско-алеутские, некоторые на-дене.

Уровни обработки текстов

Уровни обработки текстов

- **Морфологический**
  – I’m - I am
  – кошка-кошки, дно-?
- **Синтаксический**
  – Мне один черный кофе и один сладкий булка...
- **Семантический**
  лексическая и композиционная семантика
  – Сколько китайского шелка было экспортировано в Западную Европу в конце 18 века?
- **Прагматический (дискурс)**
  установление кореферентности (coreference resolution) - обнаружение всех выражений в тексте, которые ссылаются на одну сущность
  – Сколько тогда было штатов в США?

Задачи

Задачи NLP

- Определение языка (Language identification)
- Токенизация / сегментация на слова (Tokenization, Word segmentation)
- Обнаружение и исправление опечаток (Spell checking and correction)
- Морфологический анализ (Morphological parsing)
  - Лемматизация (Lemmatization)
  - Стемминг (Stemming)
  - Определение частей речи (Part-of-speech tagging)
- Разрешение лексической многозначности (Word sense disambiguation)
- Разрешение анафоры (Coreference resolution)

Задачи NLP

- Извлечение терминов/ключевых слов
  (Terminology extraction, Keywords extraction)
- Сегментация на предложения
  (Sentence segmentation, Sentence boundary disambiguation,
  Sentence breaking, Sentence boundary detection)
- Синтаксический анализ (Syntax parsing, Dependency parsing)
- Распознавание именованных сущностей
  (Named entity recognition, NER)
- Связывание именованных сущностей (Named entity linking)
- Извлечение отношений (Relationship extraction)

Задачи NLP

- Анализ тональности и извлечение мнений (Sentiment analysis, Opinion mining)
- Распознавание эмоций (Emotion detection)
- Извлечение аргументации (Argumentation mining)
- Определение точки зрения автора текста (Stance detection)
- Распознавание связей между текстами (Recognizing textual entailment)
- Дискурсивный анализ (Discourse analysis)
- Определение стиля (Style detection)
- Определение авторства (Authorship attribution)

Задачи NLP

- Определение семантической близости (Semantic similarity)
- Тематическая классификация (Text categorization)
- Кластеризация текстов (Text clustering)
- Поиск заимствований (Plagiarism detection)
- Реферирование (Automatic summarization)
- Машинный перевод (Machine translation)
- Генерация текстов (Natural language generation)
- Вопросно-ответные системы (Question answering)
- Диалоговые системы (Dialog systems, chat bots)

Трудности

Трудности NLP

- Многозначность
  - Одно и то же выражение, форма, конструкция может означать разное
  - Разрешение неоднозначности может требовать знаний о мире, контекста и т.п.

Я траву **косил косой**,
Дождик вдруг пошел **косой**.
Бросил я тогда **косить**
И на Стешу стал **косить**.
Ну а Стеша, ох, краса,
Как огонь её **коса**!

Трудности NLP

Многозначность:

- морфологическая (часть речи)
  - *мой* (– нос, – руки)
  - *look* (look at me, have a look)
- синтаксическая
  - *мужу изменять нельзя*
  - *мать любит дочь*
- семантическая
  - омонимия (*ключ*) – семантически не связанные
  - полисемия (*платформа*) – семантически связанные
- прагматическая

Пример машинного перевода

- **Help для Windows 95**

  ... Мышь может неадекватно реагировать на щелчок по почкам. Но не спешите! Это могут быть физические проблемы, а не клоп Окон 95.

  Почистите вашу мышь.
  Отсоедините ее поводок от компьютера, вытащите гениталий и промойте его и ролики внутренностей спиртом.
  Снова зашейте мышь.
  Проверьте на переломы поводка.
  Подсоедините мышь к компьютеру.
  Приглядитесь к вашей прокладке (подушке) - она не должна быть источником мусора и пыли в гениталии и роликах.
  Поверхность прокладки не должна стеснять движения мыши.

* - https://tpc.ispras.ru/

Сложность языка

- Естественный язык:
  - –многозначен на всех уровнях
  - –сложное, едва уловимое использование контекста для передачи значения
  - –включает знания и рассуждения о мире
- Но обработка естественного языка может быть иногда очень простой
  - –использование грубых признаков часто позволяют достичь очень хороших результатов

* - https://tpc.ispras.ru/

Регулярные выражения

- Инструмент, который должен знать каждый IT-специалист
- Решает большинство встречающихся на практике задач
- Поддерживаются всеми современными редакторами текстов
- Примеры применения

- Обновить цену товара в прайс-листе:
  - для конкретного товара за 1000р. сделать 999.99р.
- Заменить все вхождения одного слова в тексте на другое
  - для части слова (Википедия -> Энциклопедия)
  - с учетом контекста
- Собрать базу e-mail для рассылки спама
- Найти нецензурные высказывания на форумах и сделать автогенератор ответов…

* - https://tpc.ispras.ru/

Литература

- Леонтьева Н.Н. Автоматическое понимание текстов (2006)
- Маннинг К., Рагхаван П., Шютце Х.
  Введение в информационный поиск (2011)
- Николаев И. и др. Прикладная и компьютерная лингвистика (2016)
- Большакова Е.И. и др. Автоматическая обработка текстов на естественном языке и анализ данных (2017)
- Макмахан Брайан, Рао Делип. Знакомство с PyTorch: глубокое обучение при обработке естественного языка (2020)
- Марков С. Охота на электро-овец. 2 тома (2024)

Литература

- Dan Jurafsky, James Martin – Speech and Language Processing (3rd ed., draft)
  - https://web.stanford.edu/~jurafsky/slp3/
- Christopher Manning, Hinrich Schutze – Foundations of Statistical Natural Language Processing (2000)
- Jacob Eisenstein – Natural Language Processing (2018)

- https://web.stanford.edu/class/cs224n/