кафедра Алгоритмов и Технологий Программирования

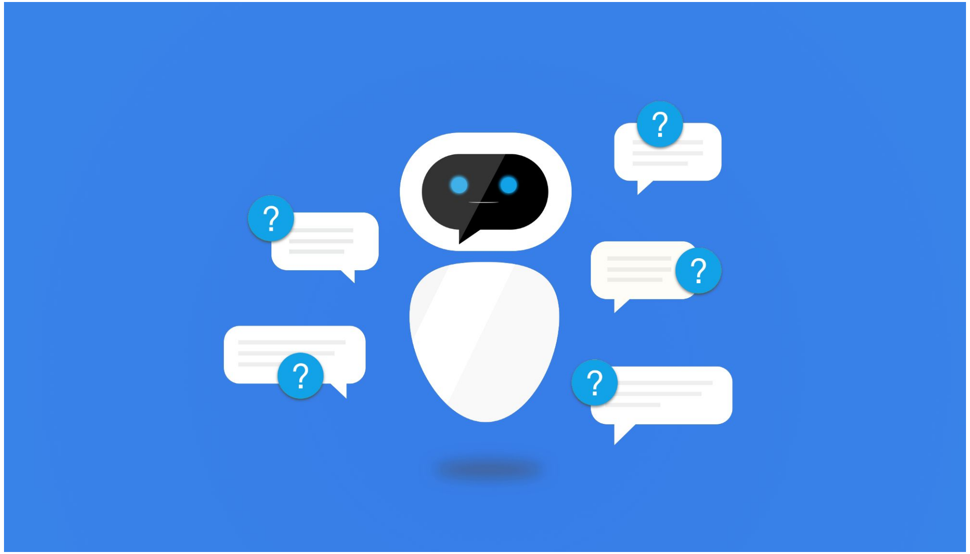
QuestionAnswering.

MIPT
01.04.2021
Anton Emelianov, Alena Fenogenova.

https://github.com/king-menin/mipt-nlp2021

Today

- What is question answering? Why so many?
- Reading comrehension
- Open domain question answering
- KBQA
- Chatbots



QA systems

The goal of question answering is to build systems that automatically correctly answer questions posed by humans in a natural language.

QA systems

- What information source does a system build on?
  - a text passage
  - all Web documents
  - knowledge bases,
  - images..
- Answer type
  - a short segment of text
  - a paragraph
  - a list
  - yes/no, …
- Question type
  - Factoid vs non-factoid
  - open-domain vs closed-domain
  - simple vs compositional

QA through years

Setting	Closed-domain	Open-domain	Reading compre-hension	Open-domain	Conversation-al, multi-hop, multilingual
Method-ology	Hand-engi-neered parsers	IR + shallow linguistic analysis	Document reader	IR + document reader	IR + document reader
Systems, datasets	LUNAR, QUALM	TREC QA	CNN / Daily Mail, SQuAD	Natural Questions	CoQA, TyDiQA, HotPotQA
Years	1970s–1990s	2000s	2013–today	2019–today	2020–today

QA systems

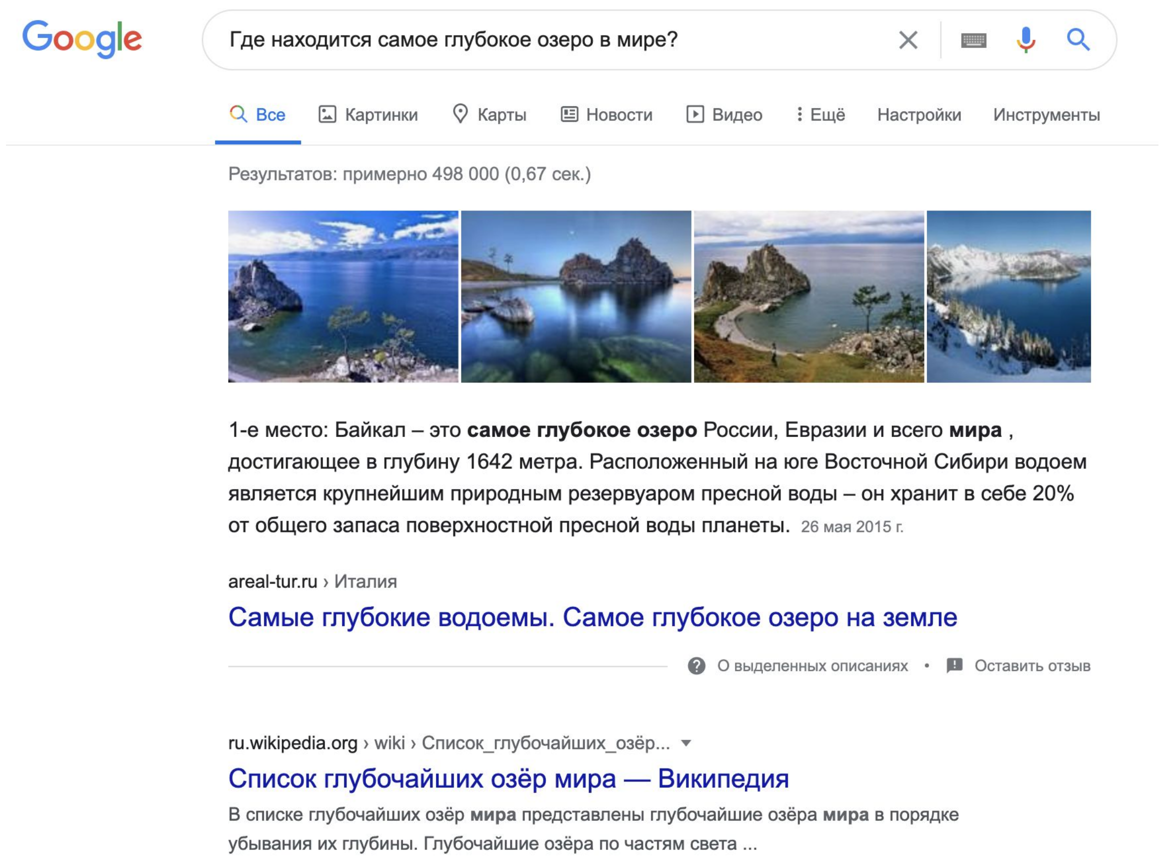
Processes automation and engineering

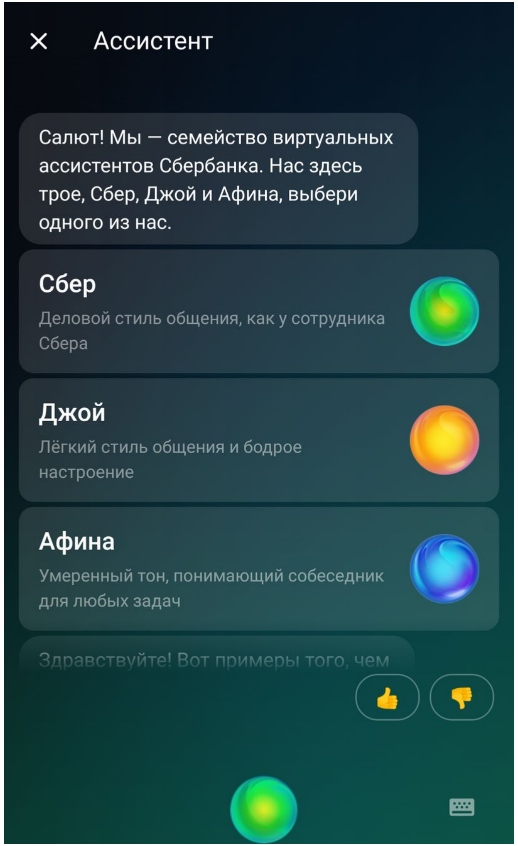
- QA goal applications
- Chatbots
- Call centers
- Study projects
- etc.

Research and Science

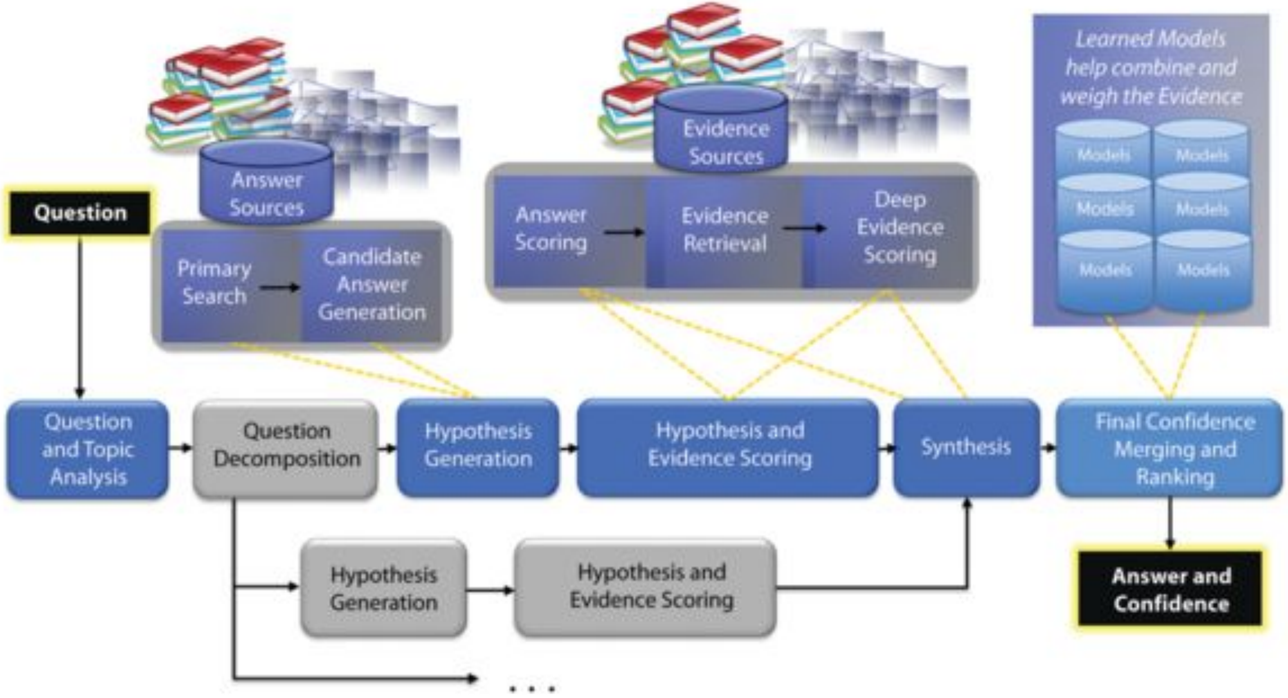
- Turing test
- AI

QA systems. Applications





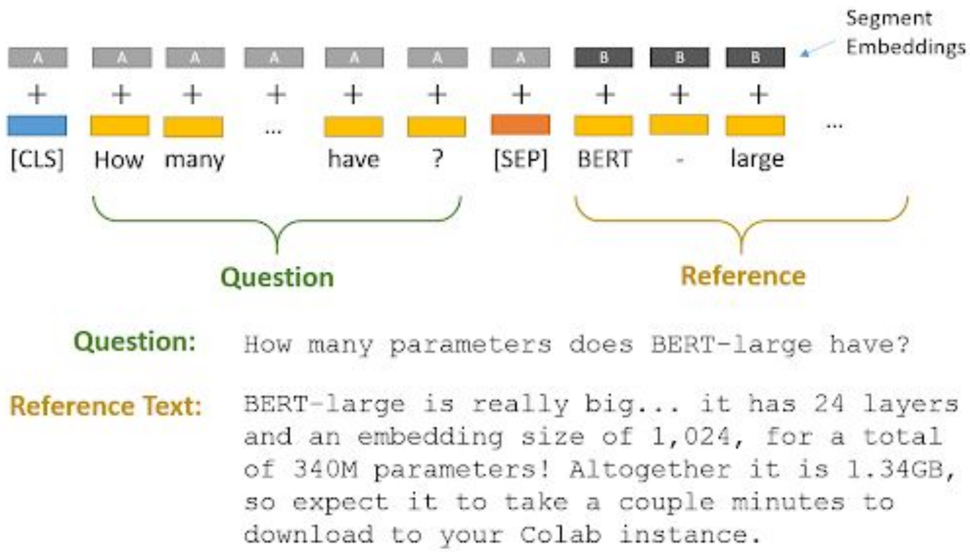
IBM Watson beated Jeopardy champions



https://archive.org/details/Jeopardy.2011.02.The.IBM.Challenge/Jeopardy.2011.02.14.The.IBM.Challenge.Day.1.HDTV.XviD-FQM.avi?utm_campaign=NLP%20News&utm_medium=email&utm_source=Revue%20newsletter

QA systems

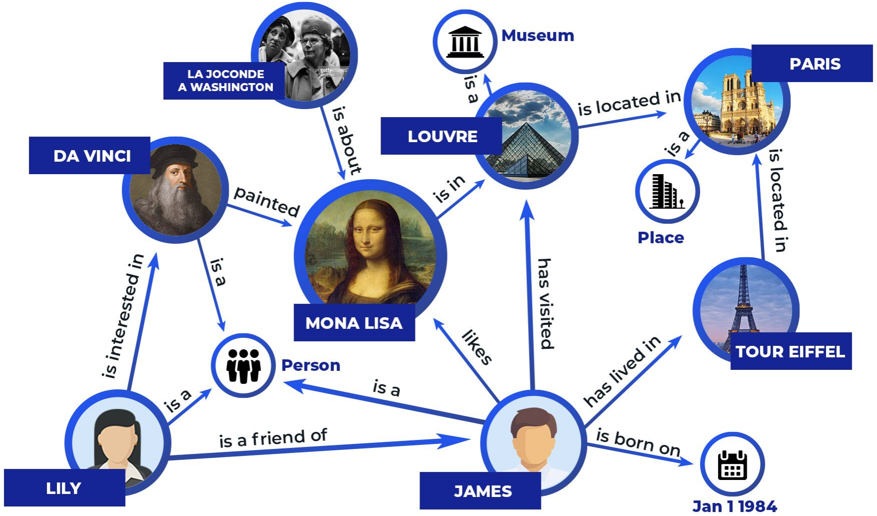
Question answering now



Almost all the state-of-the-art question answering systems are built on top of end-to-end training and pre-trained language models

QA systems

How to answer unstructed texts? Or not only texts?



Who painted Mona Lisa?

Semantic parsing

Relations and KB

Leonardo DaVinci

VQA https://visualqa.org/

The Answer to the Ultimate Question of Life, the Universe, and Everything

ANSWER
What do the answers look like?

QUESTION
How does the question look like (taxonomy)?

SOURCE
Where can I get the answers from?

Answers

- Factoid
- Yes/no
- Opinion/Info
- Explanation
- Document
- A sentence or paraphraph extracted
- Another question
- etc.

Questions

One-hop (single-hop) question is the question that can be answered based on a single sentence from a passage.

Multi-hop question is a question that requires reasoning over information spread across several sentences in a passage.

(1) Mother bought apples. (2) They were on the table. (3) John has never eaten apples, that’s why he couldn’t stand it and tried one.

Question: ”Where were fruits that were eaten by a boy?”

The question is multi-hop since the answer can be obtained with only information aggregated from more than one sentence (coreference resolution and general language understanding).

Reading comprehension

- Reading comprehension = comprehend a passage of text and answer questions about its content (P, Q) $\longrightarrow$ A

"Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum."

Question

- Reading comprehension task: build a system to comprehend a passage of text and answer questions about its content (P, Q) $\longrightarrow$ A

Reading comprehension

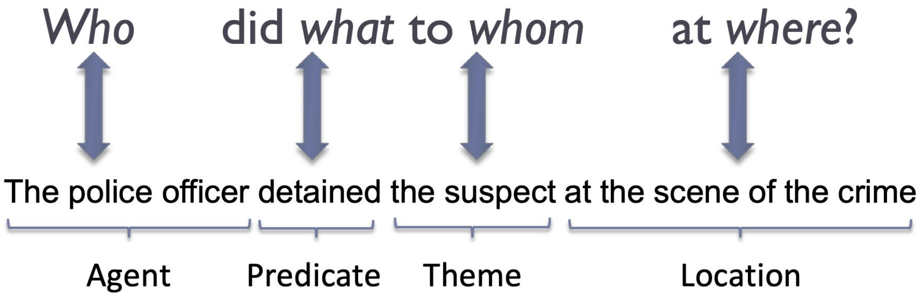
Why Reading comprehension is important?

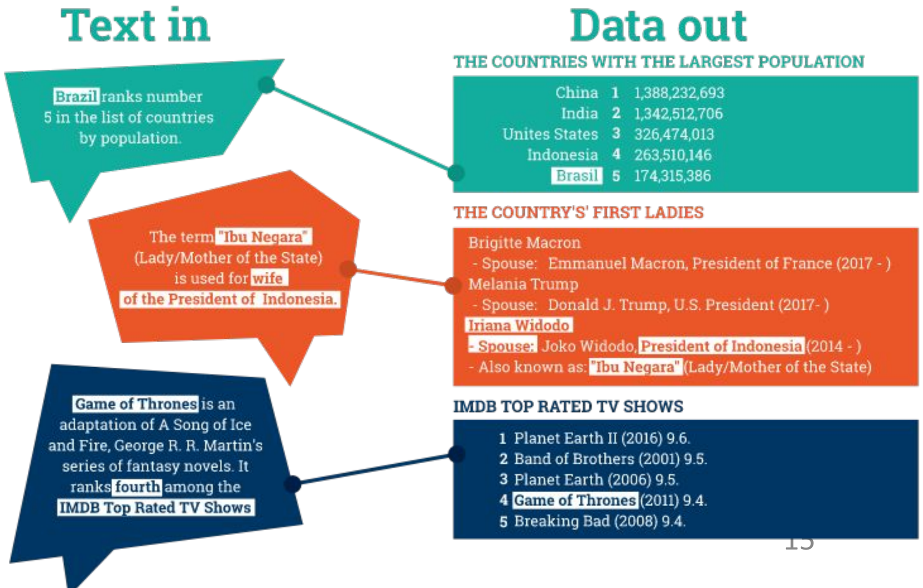
- Useful in many complex practical applications
- Testbed for evaluating how well computer systems understand human language

“Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding.”

(Wendy Lehnert 1977)

- Many complex NLP tasks can be reduced to a reading comprehension problem:
  - Semantic Role Labeling
  - Information extraction





Reading comprehension

Problem formulation:

- *Input*: $C = (c_1, c_2, \ldots, c_N)$ $Q = (q_1, q_2, \ldots, q_M)$ $c_i , q_i \in V$
- *Output*: $1 \leq \text{start} \leq \text{end} \leq N$

answer is a span in the passage

- A family of <u>LSTM-based</u> models with attention (2016-2018)

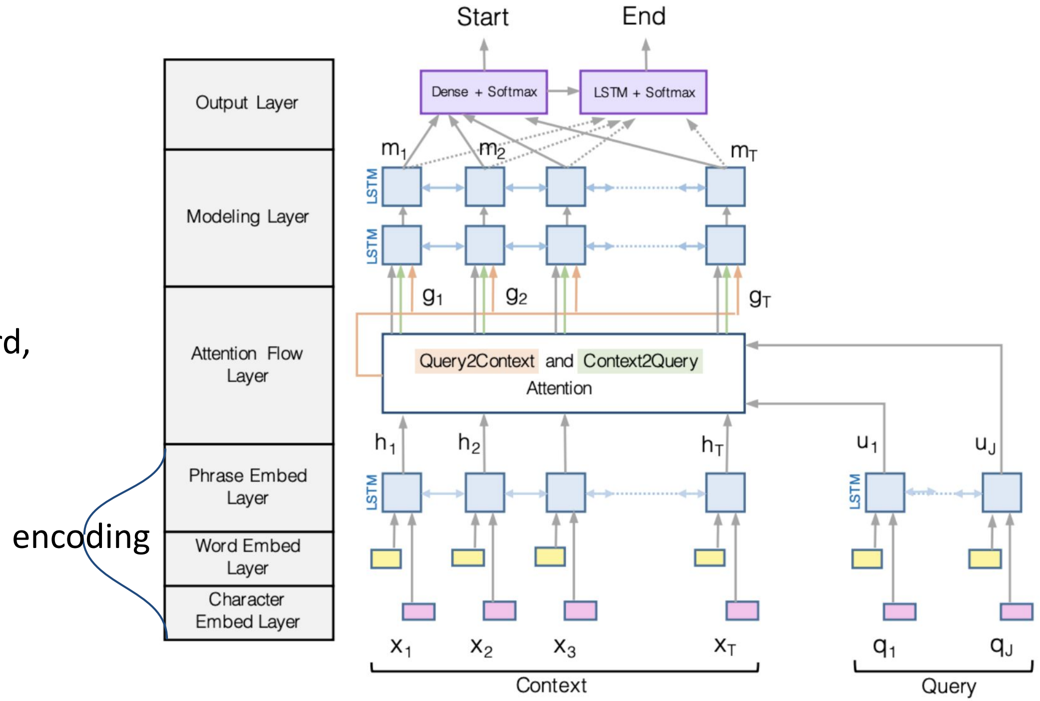
  Attentive Reader, Stanford Attentive Reader, MatchLSTM, BiDAF, Dynamic coattention network...

- Fine-tuning <u>BERT-like models</u> for reading comprehension (2019+)

BiDAF. Bidirectional Attention Flow model

Attention Flow Idea: attention should flow both ways – from the context to the question and from the question to the context.

- Concatenation of word embedding (GloVe) and character embedding (CNNs over character embeddings) for each word in context and query
- Two bi-LSTMs separately to produce contextual embeddings for both context and query
- <u>Context-to-query attention</u>: For each context word, choose the most relevant words from the query words
- <u>Query-to-context attention</u>: choose the context words most relevant to one of query words.
- Attention layer is modeling interactions between query and context
- Modeling layer is modeling interactions within context words
- Output layer: two classifiers predicting the start and end positions



$$p_{\text{start}} = \text{softmax}(\mathbf{w}_{\text{start}}^{\mathsf{T}}[\mathbf{g}_i; \mathbf{m}_i]) \qquad p_{\text{end}} = \text{softmax}(\mathbf{w}_{\text{end}}^{\mathsf{T}}[\mathbf{g}_i; \mathbf{m}'_i])$$

$$\mathbf{m}'_i = \text{BiLSTM}(\mathbf{m}_i) \in \mathbb{R}^{2H} \quad \mathbf{w}_{\text{start}}, \mathbf{w}_{\text{end}} \in \mathbb{R}^{10H}$$

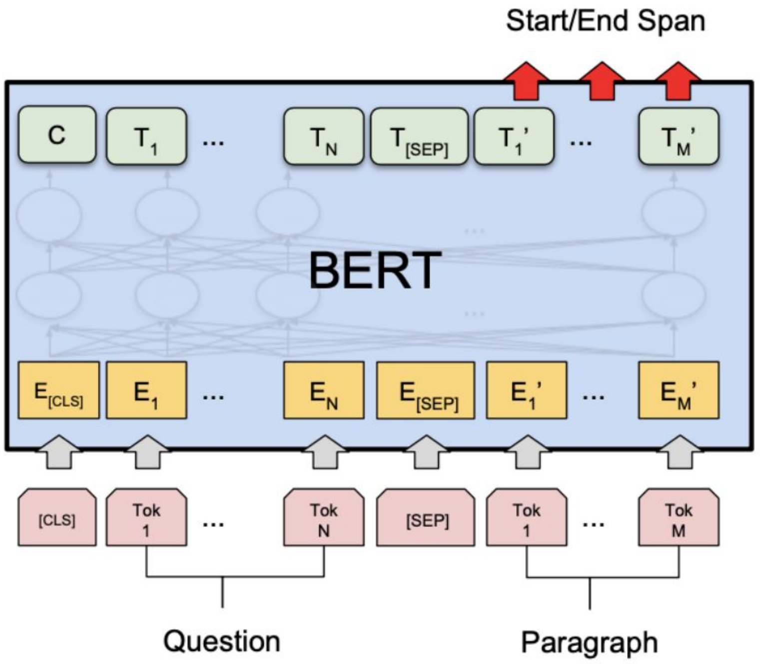
Reading comprehension. BERT

$L = -\log p_{start}(s^*) - \log p_{end}(e^*)$

$p_{end}(i) = \mathrm{softmax}_i(w'_{end} H)$
$p_{start}(i) = \mathrm{softmax}_i(w'_{start} H)$

where $H = [h_1, h_2, ..., h_N]$ are the hidden vectors of the paragraph, returned by BERT

All the BERT parameters (~110M) as well as H_{start} and H_{end} (e.g., 768 x 2 = 1536) are optimized together for L



Reading comprehension

SQUAD

Dataset size (Russian): 50k questions
Dataset size (English): 100k questions
Task: Find the answer and direct span for the question in text
Evaluation: exact match (0 or 1) and F1 (partial credit)

Model config	EM (dev)	F-1 (dev)
DeepPavlov RuBERT	66.30+-0.24	84.60+-0.11
DeepPavlov multilingual BERT	64.35+-0.39	83.39+-0.08
DeepPavlov R-Net	60.62	80.04

Example

Passage: Первая школа в Манитобе была основана в 1818 году католическими миссионерами в городе Виннипег, первая протестантская школа была учреждена в 1820 году. Провинциальное Управление образования было учреждено в 1871 году, оно отвечало за государственные школы и учебные программы, ...

Question: *Кем была в 1818 году основана первая школа в Манитобе?*
Answer:
"*text*": "католическими миссионерами",
"*answer_start*": 50

Reading comprehension

RuCoS

Dataset size

72193 train / 4370 val / 4147 test

Data source:

Lenta & Deutsche Welle

Task: Find the correct entity in the paragraph that best fits the placeholder in the query.

Example

Passage: Мать двух мальчиков, брошенных отцом в московском аэропорту Шереметьево, забрала их. Об этом сообщили ТАСС в пресс-службе министерства образования и науки Хабаровского края. Сейчас младший ребенок посещает детский сад, а старший ходит в школу. В учебных заведениях с ними по необходимости работают штатные психологи. Также министерство социальной защиты населения рассматривает вопрос о бесплатном оздоровлении детей в летнее время. Через несколько дней после того, как Виктор Гаврилов бросил своих детей в аэропорту, он явился с повинной к следователям в городе Батайске Ростовской области.

Query: 26 января <placeholder> бросил сыновей в возрасте пяти и семи лет в Шереметьево.

Correct Entities: Виктор Гаврилов

Reading comprehension

MuSeRC

Task: Reading comprehension challenge, questions can be answered only based on multiple sentences from the paragraph.

Dataset size: 500/100/322

Data source

+800 paragraphs ~6k questions

5 different domains collected from open sources:

1) elementary school texts
2) news
3) fiction stories
4) fairy tales
5) brief annotations of TV series and books

Example

Paragraph: (1) Мужская сборная команда Норвегии по биатлону в рамках этапа Кубка мира в немецком Оберхофе выиграла эстафетную гонку. (2) Вторыми стали французы, а бронзу получила немецкая команда. (3) Российские биатлонисты не смогли побороться даже за четвертое место, отстав от норвежцев более чем на две минуты. (4) Это худший результат сборной России в текущем сезоне. (5) Четвёртыми в Оберхофе стали австрийцы. (6) В составе сборной Норвегии на четвёртый этап вышел легендарный Уле-Эйнар Бьорндален. (7) Впрочем, Норвегия с самого начала гонки была в числе лидеров, успешно проведя все четыре этапа. (8) За сборную России в Оберхофе выступали Иван Черезов, Антон Шипулин, Евгений Устюгов и Максим Чудов. (9) Гонка не задалась уже с самого начала: если на стрельбе из положения лежа Черезов был точен, то из положения стоя он допустил несколько промахов, в резуль- тате чего ему пришлось бежать один дополнительный круг. (10) После этого отставание российской команды от соперников только увеличива- лось. (11) Напомним, что днем ранее российские биатлонистки выиграли свою эстафету. (12) В составе сборной России выступали Анна Богалий- Титовец, Анна Булыгина, Ольга Медведцева и Светлана Слепцова. (13) Они опередили своих основных соперниц - немок - всего на 0,3 секунды.

Question: На сколько секунд женская команда опередила своих соперниц?

Candidate answers: Всего на 0,3 секунды. (*T*), На 0,3 секунды. (*T*), *На секунду.* (F), На секунды. (*F*)

Reading comprehension

DaNetQA

Dataset size: 800 train, 200 dev, 200 test examples; 562 (~59%) unique questions

Task: Given a passage, answer a yes/no question to it.

Data source

1) Crowdsourced questions are used as queries to Wikipedia
2) Wikipedia pages are retrieved via Google API
3) Passages are retrieved by Deep Pavlov SQuAD models
4) Crowd workers answer the questions based on the passages

Example

Passage: В период с 1969 по 1972 год по программе «Аполлон» было выполнено 6 полётов с посадкой на Луне.

Question: *Был ли человек на луне?*
Answer: Yes

Open domain question answering

- We don’t assume a given passage; we have access to a large collection of documents (e.g., Wikipedia); we don’t know where the answer is located
- The goal: to return the answer for any open-domain questions.
- Closed-domain
- Factoid question

Question

WIKIPEDIA
The Free Encyclopedia

Google

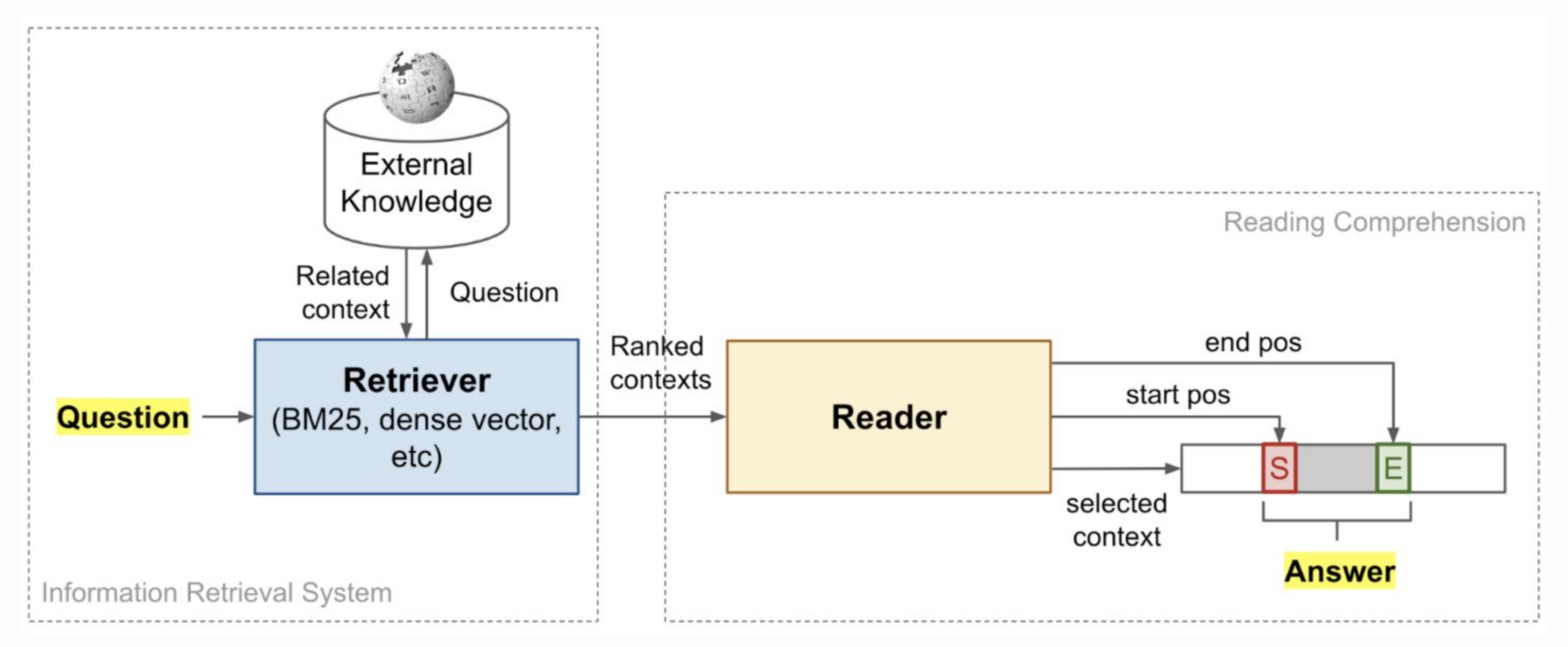
Answer

Open domain question answering

Retriever-reader framework

Input: a large collection of documents $D = D_1, D_2, \ldots D_n$ and Q

Output: an answer string A



Open domain question answering

DrQA (Document retriever Question-Answering)

<u>Retriever</u>: standard TF-IDF information-retrieval sparse model (a fixed module)
<u>Reader</u>: a neural reading comprehension model
(3-layer bidirectional LSTM with hidden size 128)

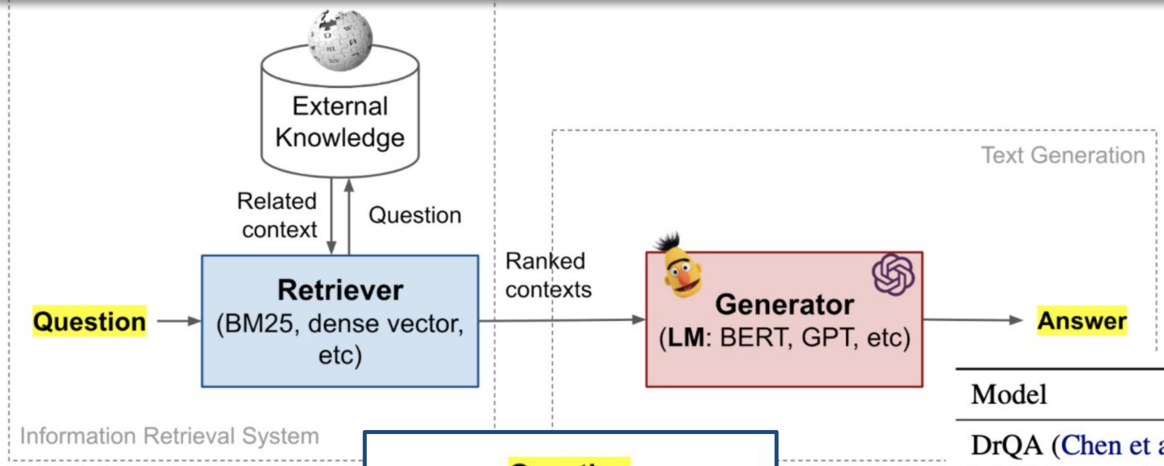
$$\begin{aligned} \text{tf-idf}(t, d, D) &= \text{tf}(t, d) \times \text{idf}(t, D) \\ \text{tf}(t, d) &= \log(1 + \text{freq}(t, d)) \\ \text{idf}(t, D) &= \log\left(\frac{|D|}{|d \in D : t \in d|}\right) \end{aligned}$$

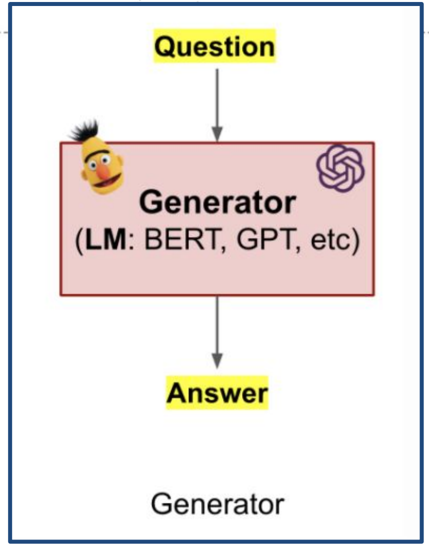
Train the retriever using question-answer pairs:

- ORQA (Open-Retrieval Question-Answering)
- DPR (Dense passage retrieval)
- REALM (Retrieval-Augmented Language Model pre-training)

https://lilianweng.github.io/lil-log/2020/10/29/open-domain-question-answering.html

Open domain question answering





Model	NaturalQuestions	TriviaQA	
DrQA (Chen et al., 2017)	-	-	-
Multi-Passage BERT (Wang et al., 2019)	-	-	-
Path Retriever (Asai et al., 2020)	31.7	-	-
Graph Retriever (Min et al., 2019b)	34.7	55.8	-
Hard EM (Min et al., 2019a)	28.8	50.9	-
ORQA (Lee et al., 2019)	31.3	45.1	-
REALM (Guu et al., 2020)	38.2	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-
SpanSeqGen (Min et al., 2020)	42.5	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0
T5 (Roberts et al., 2020)	36.6	-	60.5
GPT-3 few shot (Brown et al., 2020)	29.9	-	71.2

Semantic Parsing

Semantic parsing is a process of mapping a natural language into a formal representation of its meaning. Depending of the formalism logical representation can be used to query a structured knowledge base.

Knowledge base QA (KBQA)

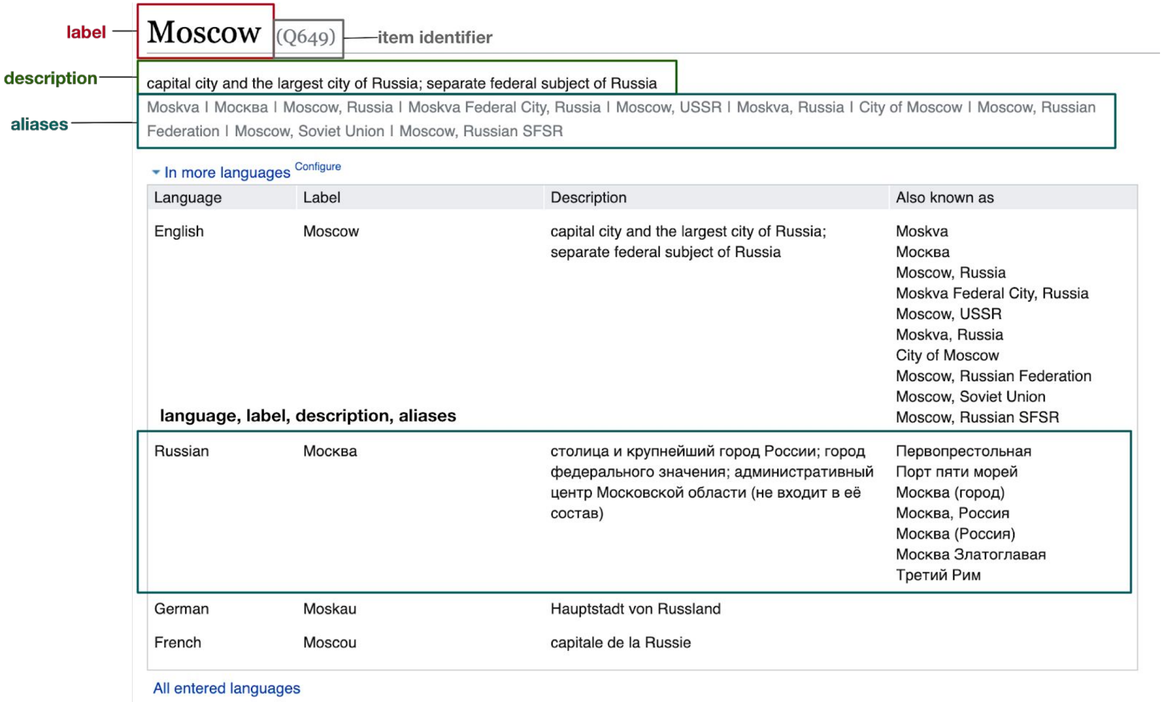
KBQA - Knowledge Base question answering.

Formal representation of knowledge. The graph model allows you to model physical and abstract entities and relationships between them. A graph is defined classically as a set of vertices and edges

$$G = (V, E) | E \subseteq \mathbb{R}^{|V| \times |V|}$$

For example, Wikidata - graph db, cross-links in Wikipedia.

DBpedia, Wikidata, YAGO, etc.



Knowledge base QA

The Resource Description Framework (RDF) is a standard model for data interchange on the Web.
It defines the model of the subject-predicate-subject or subject-predicate-object triplet.
That is, an entity - "subject" can be associated with another entity or a simple value - an object - through some property - a predicate.

Special predicates: *rdf:type, rdf:Property, rdf:subject, rdf:predicate, rdf:object, rdf:first, rdf:value, rdf:List, etc..*

Triplet example:

“Университет ИТМО - находится в - Санкт-Петербург” links entities: Университет ИТМО and Санкт-Петербург via predicate “находится в”.

Triplet “Университет ИТМО - rdf:type - Университет” means that “Университет ИТМО” ∈ университеты.

To query the knowledge represented in RDF, the query language **SPARQL** is used (links knowledge graphs to applications based on knowledge graphs)

The part of the knowledge graph that describes abstract concepts and connections between them at a high level, otherwise it is also called **ontology**.

RuBQ

Chatbots and dialogue systems. Types

Dialogue systems, or conversational agents communicate with users in natural language (text, speech, or both)

Two classes:

1. **Task-oriented dialogue agents** use conversation with users to help people complete tasks. Dialogue agents in digital assistants (Siri, Alexa, Google Now/Home, Cortana, etc.), give directions, control appliances, find restaurants, or make calls.
2. **Chatbots** are systems designed for extended conversations, set up to mimic the unstructured conversations or ‘chats’ characteristic of human-human interaction, mainly for entertainment, but also for practical purposes like making task-oriented agents more natural.

Discourse

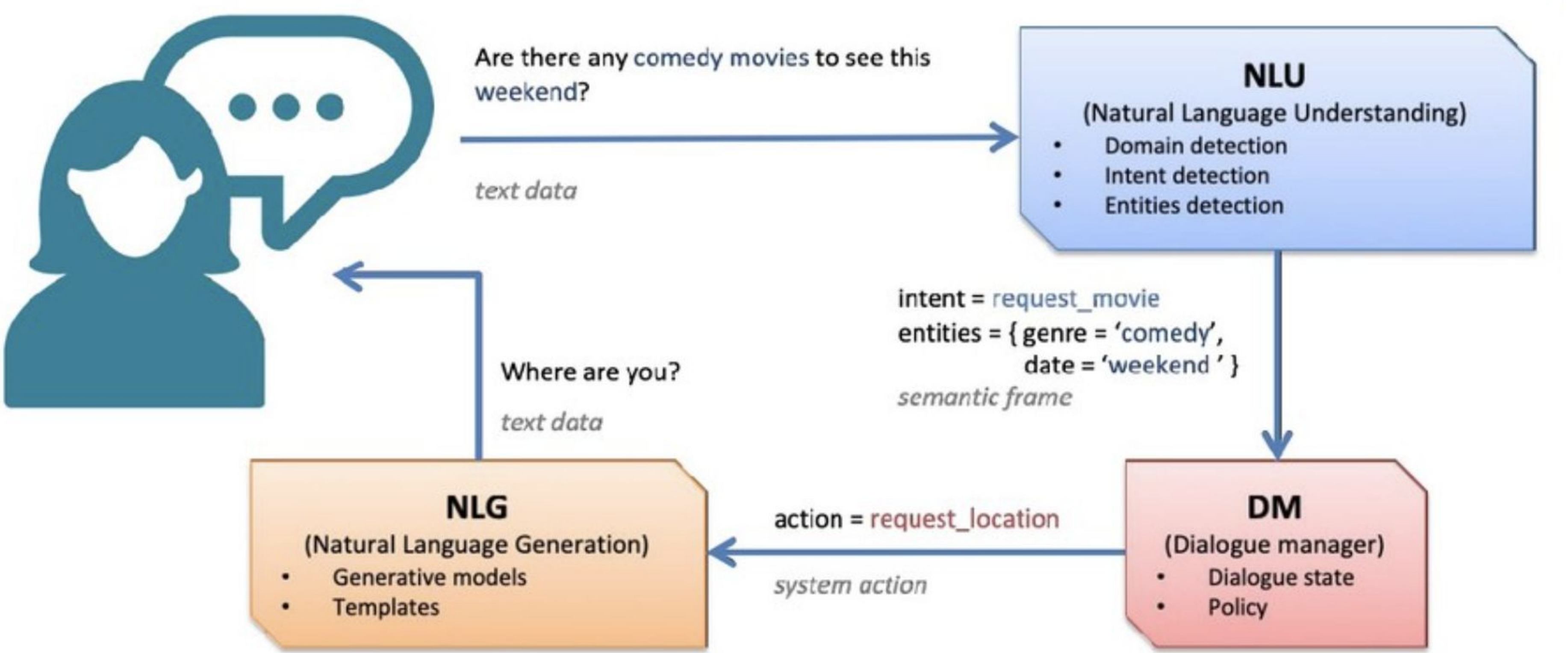
Discourse Analysis—What Speakers Do in Conversation. **Discourse** analysis is sometimes defined as the analysis of **language** 'beyond the sentence'.

Speech acts or Dialogue Acts

Constatives:	committing the speaker to something's being the case (*answering*, *claiming*, *confirming*, *denying*, *disagreeing*, *stating*)
Directives:	attempts by the speaker to get the addressee to do something (*advising*, *asking*, *forbidding*, *inviting*, *ordering*, *requesting*)
Commissives:	committing the speaker to some future course of action (*promising*, *planning*, *vowing*, *betting*, *opposing*)
Acknowledgments:	express the speaker's attitude regarding the hearer with respect to some social action (*apologizing*, *greeting*, *thanking*, *accepting an acknowledgment*)

From rule-based to NN

Goal oriented chatbots



Chatbots. Intents

Intent recognition or **intent** classification — is the task of taking a written or spoken input, and classifying it based on what the user wants to achieve.

Intent recognition forms an essential component of chatbots and finds use in sales conversions, customer support, and many other areas.

We need to do **slot filling**: extract the particular slots and fillers that the user intends the system to understand from their utterance with respect to their intent

Datasets: **SNIPS, ATIS, [MultiWoZ]()**

Show me morning flights from Moscow to Barcelona on Tuesday

DOMAIN: TRAVEL
INTENT: SHOW-FLIGHT (want to travel)
DATE: Tuesday
CITY-from: Moscow
CITY-to: Barcelona

http://nlpprogress.com/english/intent_detection_slot_filling.html

Chatbots. Anaphora

Coreference/Anaphora Resolution is the problem of resolving what a pronoun, or a noun phrase refers to.

The relation that holds between an anaphor and an antecedent.
Antecedent — a mention with already established referent
Anaphor — a mention referring to an earlier occurring antecedent

Examples:

Ann saw Mary yesterday. She was tired.
Mary went home. She did so because she was tired.
Putin made a surprise visit to USA yesterday. President met ...

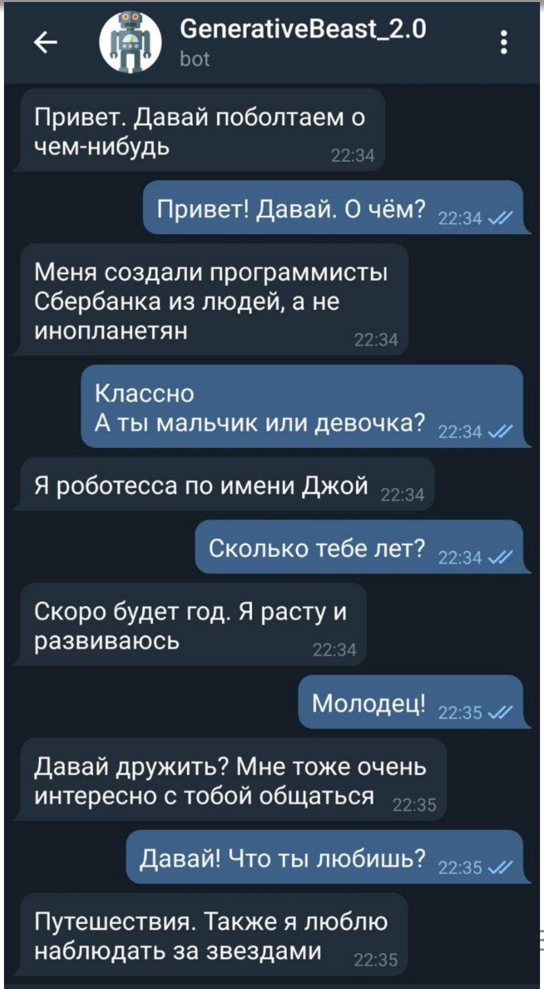
Elipsis is a case of anaphora., where the anaphor is a null proform.

Examples:

I haven't seen John yet, but I will Ø (see) soon.
Robert cooked the first course, and Mary Ø (cooked) the dessert.

Chatbots. Chit-chat

- Retrieval base models:
  - Take a dataset of dialogues,
  - construct pairs (context, reply)
  - build inverted index on contexts
  - return best with cos_sim(user_context, index_context)
- Neural networks
- Generation based networks (GPTs)



Questions?

References. Datasets

- SQuAD 2.0: the Stanford QA dataset.
- RACE: a reading comprehension dataset collected from English Examinations that are created for middle school and high school students.
- TREC QA: the TREC QA collections.
- MS MARCO: a QA dataset featuring 100,000 real Bing questions and a human generated answer.
- CuratedTREC: based on the benchmarks from the TREC QA tasks that have been curated by Baudis & Sedivy (2015).
- Google Natural Questions: contains real user questions issued to Google search, and answers found from Wikipedia by annotators.
- WebQuestions: designed for knowledge-base QA with answers restricted to Freebase entities.
- WikiQA: Bing query logs were used as the source of questions. Each question is then linked to a Wikipedia page that potentially contains the answer.
- WikiMovies: contains movie-related questions from the OMDb and MovieLens databases and where the questions can be answered using Wikipedia pages.
- WikiReading: to predict textual values from the structured knowledge base Wikidata by reading the text of the corresponding Wikipedia articles.
- TriviaQA: a reading comprehension dataset containing 95K question-answer pairs authored by trivia enthusiasts and independently gathered multiple evidence documents per question.
- Jeopardy! Questions: contains 200,000+ Jeopardy! questions.
- DeepMind Q&A Dataset: question/answer pairs from CNN and Daily Mail articles.
- bAbi: a rich collection of datasets for text understanding by Facebook.
- FEVER: for fact extraction and verification.
- SearchQA: question-answer pairs were crawled from from J! Archive, and then augmented with text snippets from Google.
- Quasar-T: a collection of open-domain trivia questions and their answers obtained from various internet sources.
- Quiz bowl: contains data from a trivia competition called quiz bowl.
- AmbigNQ: ambiguous questions selected from NQ-OPEN dataset.
- QA-Overlap: a collections of overlapped answers/questions between train and test set for Natural Questions, TriviaQA, and WebQuestionsю
- VQA : visual questin answering. Stanford.
- MultiWOZ (The Multi-domain Wizard-of-Oz (MultiWOZ)) - https://paperswithcode.com/dataset/multiwoz
- RUSSIAN datasets:
    a. SberSQUAD - https://drive.google.com/drive/u/1/folders/1AtLPhazqhpHTC-be10XsYIKE3n1Xut51
    b. RuCoS - https://russiansuperglue.com/tasks/task_info/RuCoS
    c. MuSeRC - https://russiansuperglue.com/tasks/task_info/MuSeRC
    d. DaNetQA - https://russiansuperglue.com/tasks/task_info/DaNetQA
    e. RuBQ - https://github.com/vladislavneon/RuBQ

References. Articles

[1] Danqi Chen & Scott Yih. “ACL2020 Tutorial: Open-Domain Question Answering” July 2020.

[3] Shuohang Wang, et al. “R^3: Reinforced Ranker-Reader for Open-Domain Question Answering” AAAI 2018.

[4] Jimmy Lin. “The neural hype and comparisons against weak baselines.” ACM SIGIR Forum. Vol. 52. No. 2. 2019.

[5] Wei Yang, et al. “End-to-End Open-Domain Question Answering with BERTserini” NAACL 2019.

[6] Christopher Clark & Matt Gardner. “Simple and Effective Multi-Paragraph Reading Comprehension.” arXiv:1710.10723 (2017).

[7] Rodrigo Nogueira & Kyunghyun Cho. “Passage Re-ranking with BERT.” arXiv preprint arXiv:1901.04085 (2019). | code

[8] Zhiguo Wang, et al. “Multi-passage BERT: A globally normalized BERT model for open-domain question answering.” EMNLP 2019.

[9] Fenogenova, Alena, Vladislav Mikhailov, and Denis Shevelev. "Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian." *Proceedings of the 28th International Conference on Computational Linguistics*. 2020.

[10] Seo, Minjoon, et al. "Bidirectional attention flow for machine comprehension." *arXiv preprint arXiv:1611.01603* (2016).

References. Sources

Wikidata
Wikidata Query Service
Python libraries qwikidata

Knowledge graphs course https://ods.ai/tracks/kgcourse2021/

DeepPavlov demos
IBM Watson video in Jeopardy challenge

XiaoIce chatbot https://arxiv.org/abs/1812.08989

DialoGPT2 https://github.com/vlarine/ruDialoGPT

Anaphora resolution for Russian (http://www.dialog-21.ru/en/evaluation/2019/disambiguation/anaphora/, AGRR-2019, RuCor)
Slot fillings, intent recognition http://nlpprogress.com/english/intent_detection_slot_filling.html