Today

1. Intro
2. IR-based QA
   - Datasets
   - Models
3. Commonsense QA and inference
4. Knoweledge-based QA

Types of questions

1. Factoid questions:
   - What is the dress code for the Vatican?
   - Who is the President of the United States?
   - What are the dots in Hebrew called?
2. Commonsense questions:
   - What do all humans want to experience in their own home? (a) feel comfortable, (b) work hard, (c) fall in love, (d) lay eggs, (e) live forever
3. Opinion questions:
   - Can anyone recommend a good coffee shop near HSE campus?
4. Cloze-style questions

Types of questions

1. Types of answers
   - binary (yes / now)
   - find a span of text
   - multiple choice

Major paradigms for factoid question answering

1. Information retrieval (IR)-based QA: find a span of text, which answers a question
2. Open-domain Question Answering (ODQA): answer questions about nearly anything
3. Knowledge (KB)-based QA: build a semantic representation of question are used to question knowledge bases
   When Bernardo Bertolucci died? $\rightarrow$ `death-year(Bernardo Bertolucci, ?x)`

Today

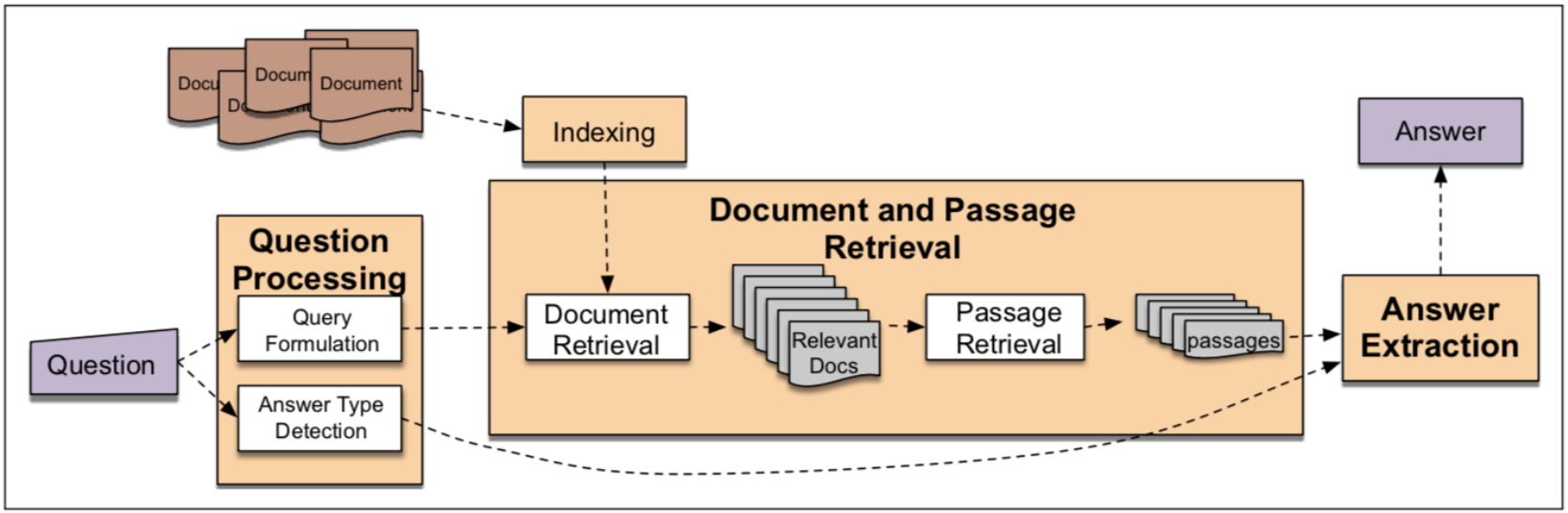
1 Intro

2 IR-based QA
- Datasets
- Models

3 Commonsense QA and inference

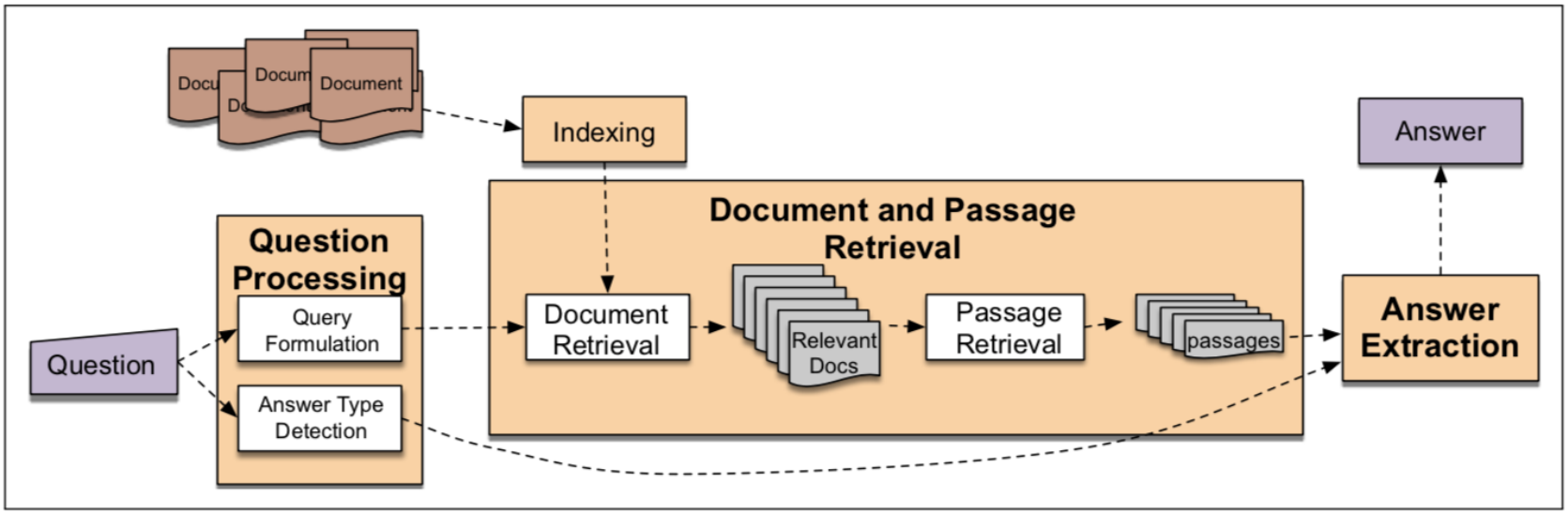
4 Knoweledge-based QA

IR-based QA



1. Question processing
   - answer type (PER, LOC, TIME)
   - focus
   - question type
2. Query formulation
   - question reformulation: remove *wh*-words, change word order
   - query expansion

IR-based QA



3. Document and passage retrieval
4. Answer extraction
   What are the dots in Hebrew called?
   In Hebrew orthography, **niqqud or nikkud**, *is a system of diacritical signs used to represent vowels or distinguish between alternative pronunciations of letters of the Hebrew alphabet.*

1. Intro
2. IR-based QA
   - Datasets
   - Models
3. Commonsense QA and inference
4. Knoweledge-based QA

Datasets for IR-based QA

Passage: Tesla later approached Morgan to ask for more funds to build a more powerful transmitter. **When asked where all the money had gone, Tesla responded by saying that he was affected by the Panic of 1901**, which he (Morgan) had caused. Morgan was shocked by the reminder of his part in the stock market crash and by Tesla's breach of contract by asking for more funds. Tesla wrote another plea to Morgan, but it was also fruitless. Morgan still owed Tesla money on the original agreement, and Tesla had been facing foreclosure even before construction of the tower began.

Question: On what did Tesla blame for the loss of the initial money?

Answer: Panic of 1901

Figure: An example from the SQuAD dataset

1. Stanford Question Answering Dataset (SQuAD)
2. NewsQA
3. WikiQA
4. CuratedTREC
5. WebQuestions
6. WikiMovies
7. Russian: SberQUAD

SQuAD2.0 [1], [2]

100,000 questions in SQuAD1.1 and over 50,000 unaswerable questions in SQuAS2.0

1. Project Nayuki's Wikipedia's internal PageRanks to obtain the top 10000 articles of English Wikipedia, from which we sampled 536 articles uniformly at random
2. Articles splitted in individual paragraphes
3. Crowsourcing: ask and answer up to 5 questions on the content of that paragraph
4. Crowdworkers were encouraged to ask questions in their own words, without copying word phrases from the paragraph
5. Analysis: the (i) diversity of answer types, (ii) the difficulty of questions in terms of type of reasoning required to answer them, and (iii) the degree of syntactic divergence between the question and answer sentences.

`https://rajpurkar.github.io/SQuAD-explorer/`

RACE [3]

Passage:
In a small village in England about 150 years ago, a mail coach was standing on the street. It didn't come to that village often. People had to pay a lot to get a letter. The person who sent the letter didn't have to pay the postage, while the receiver had to.
"Here's a letter for Miss Alice Brown," said the mailman.
" I'm Alice Brown," a girl of about 18 said in a low voice.
Alice looked at the envelope for a minute, and then handed it back to the mailman.
"I'm sorry I can't take it, I don't have enough money to pay it", she said.
A gentleman standing around were very sorry for her. Then he came up and paid the postage for her.
When the gentleman gave the letter to her, she said with a smile, " Thank you very much, This letter is from Tom. I'm going to marry him. He went to London to look for work. I've waited a long time for this letter, but now I don't need it, there is nothing in it."
"Really? How do you know that?" the gentleman said in surprise.
"He told me that he would put some signs on the envelope. Look, sir, this cross in the corner means that he is well and this circle means he has found work. That's good news."
The gentleman was Sir Rowland Hill. He didn't forgot Alice and her letter.
"The postage to be paid by the receiver has to be changed," he said to himself and had a good plan.
"The postage has to be much lower, what about a penny? And the person who sends the letter pays the postage. He has to buy a stamp and put it on the envelope." he said . The government accepted his plan. Then the first stamp was put out in 1840. It was called the "Penny Black". It had a picture of the Queen on it.

Questions:

1): The first postage stamp was made _.
A. in England B. in America C. by Alice D. in 1910

2): The girl handed the letter back to the mailman because _ .
A. she didn't know whose letter it was
B. she had no money to pay the postage
C. she received the letter but she didn't want to open it
D. she had already known what was written in the letter

3): We can know from Alice's words that _ .
A. Tom had told her what the signs meant before leaving
B. Alice was clever and could guess the meaning of the signs
C. Alice had put the signs on the envelope herself
D. Tom had put the signs as Alice had told him to

4): The idea of using stamps was thought of by _ .
A. the government
B. Sir Rowland Hill
C. Alice Brown
D. Tom

5): From the passage we know the high postage made _ .
A. people never send each other letters
B. lovers almost lose every touch with each other
C. people try their best to avoid paying it
D. receivers refuse to pay the coming letters

Answer: ADABC

Figure: An example from RACE dataset

RACE consists of near 28k passages and near 100k questions generated by human experts (English instructors), and covers a variety of topics which are carefully designed for evaluating the students' ability in understanding and reasoning.

RACE [3]

Dataset	RACE-M	RACE-H	RACE	CNN	SQUAD	NEWSQA
Word Matching	29.4%	11.3%	15.8%	13.0%†	39.8%*	32.7%*
Paraphrasing	14.8%	20.6%	19.2%	41.0%†	34.3%*	27.0%*
Single-Sentence Reasoning	31.3%	34.1%	33.4%	19.0%†	8.6%*	13.2%*
Multi-Sentence Reasoning	22.6%	26.9%	25.8%	2.0%†	11.9%*	20.7%*
Ambiguous/Insufficient	1.8%	7.1%	5.8%	25.0%†	5.4%*	6.4%*

Figure: Statistic information about Reasoning type in different datasets

RACE includes five classes of questions: word matching, paraphrasing, single-sentence reasoning, multi-sentence reasoning, insufficient or ambiguous questions.
`http://www.cs.cmu.edu/~glai1/data/race/`

MS Marco

Field	Description
Query	A question query issued to Bing.
Passages	Top 10 passages from Web documents as retrieved by Bing. The passages are presented in ranked order to human editors. The passage that the editor uses to compose the answer is annotated as is_selected: 1.
Document URLs	URLs of the top ranked documents for the question from Bing. The passages are extracted from these documents.
Answer(s)	Answers composed by human editors for the question, automatically extracted passages and their corresponding documents.
Well Formed Answer(s)	Well-formed answer rewritten by human editors, and the original answer.
Segment	QA classification. E.g., tallest mountain in south america belongs to the ENTITY segment because the answer is an entity (Aconcagua).

Figure: The final dataset format for MS MARCO

Three tasks:

1. first predict whether a question can be answered, if so, generate the correct answer
2. the generated answer should be well-formed
3. the passage re-ranking

`http://www.msmarco.org`

1. Intro
2. IR-based QA
   - Datasets
   - Models
3. Commonsense QA and inference
4. Knoweledge-based QA

DrQA [4]

Document Retriever: return 5 Wikipedia articles, using simple $tf - idf$-based retrieval

Document Reader: we are given a query $q = q_1, \ldots, q_l$ and n paragraphs $p_1, \ldots, p_m$

Question encoding: weighted sum of $RNN(q_1, \ldots, q_l)$

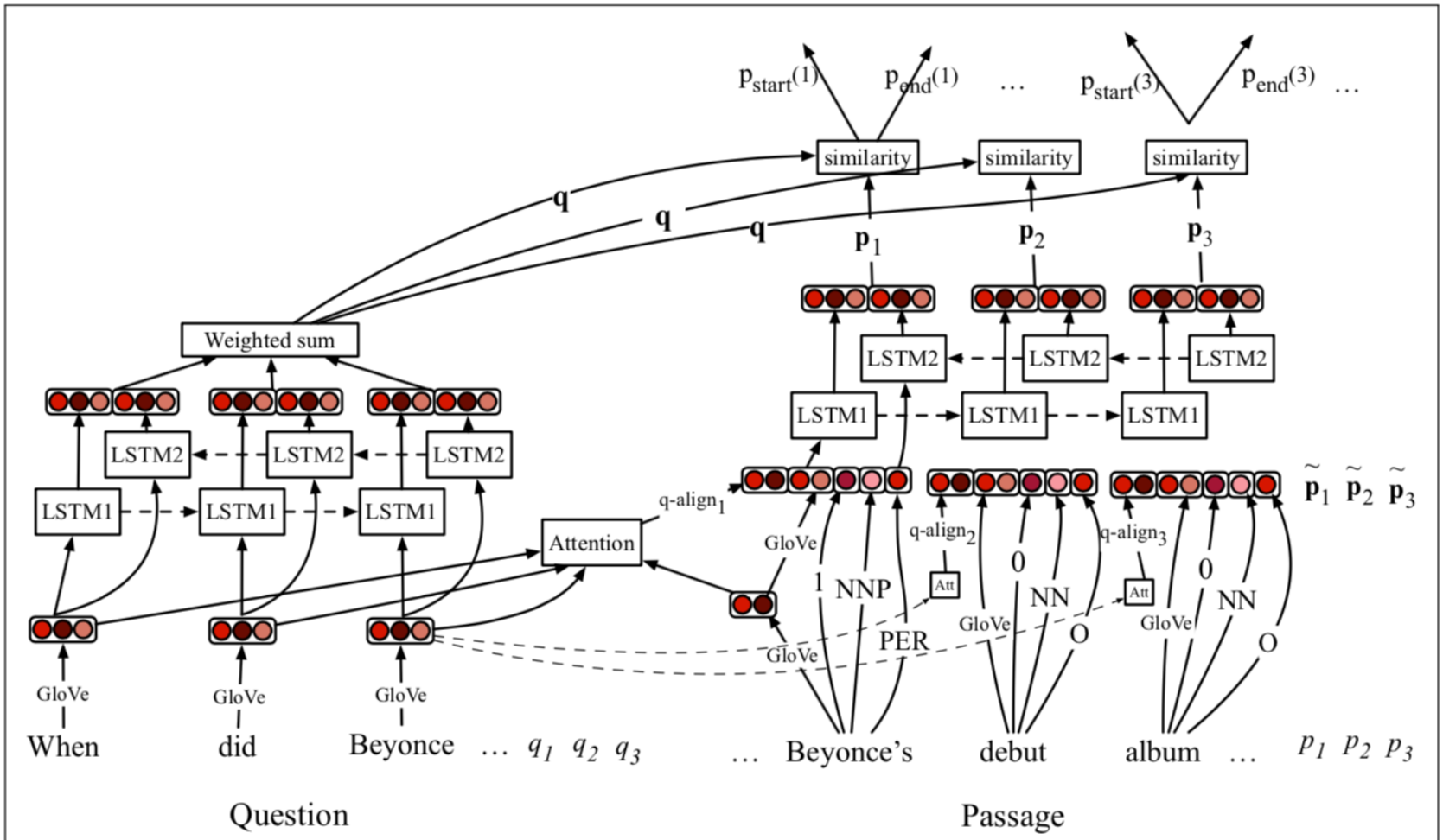
Paragraph encoding: $RNN(\widetilde{p_1}, \ldots, \widetilde{p_m})$, where $\widetilde{p_1}$ is comprised of:

- word embedding f_{emb}
- exact match $f_{exact\ match}$
- token features (POS, NER, TF), $f_{token\ features}$
- aligned question embedding $f_{align} = \sum_j a_{ij} q_j$

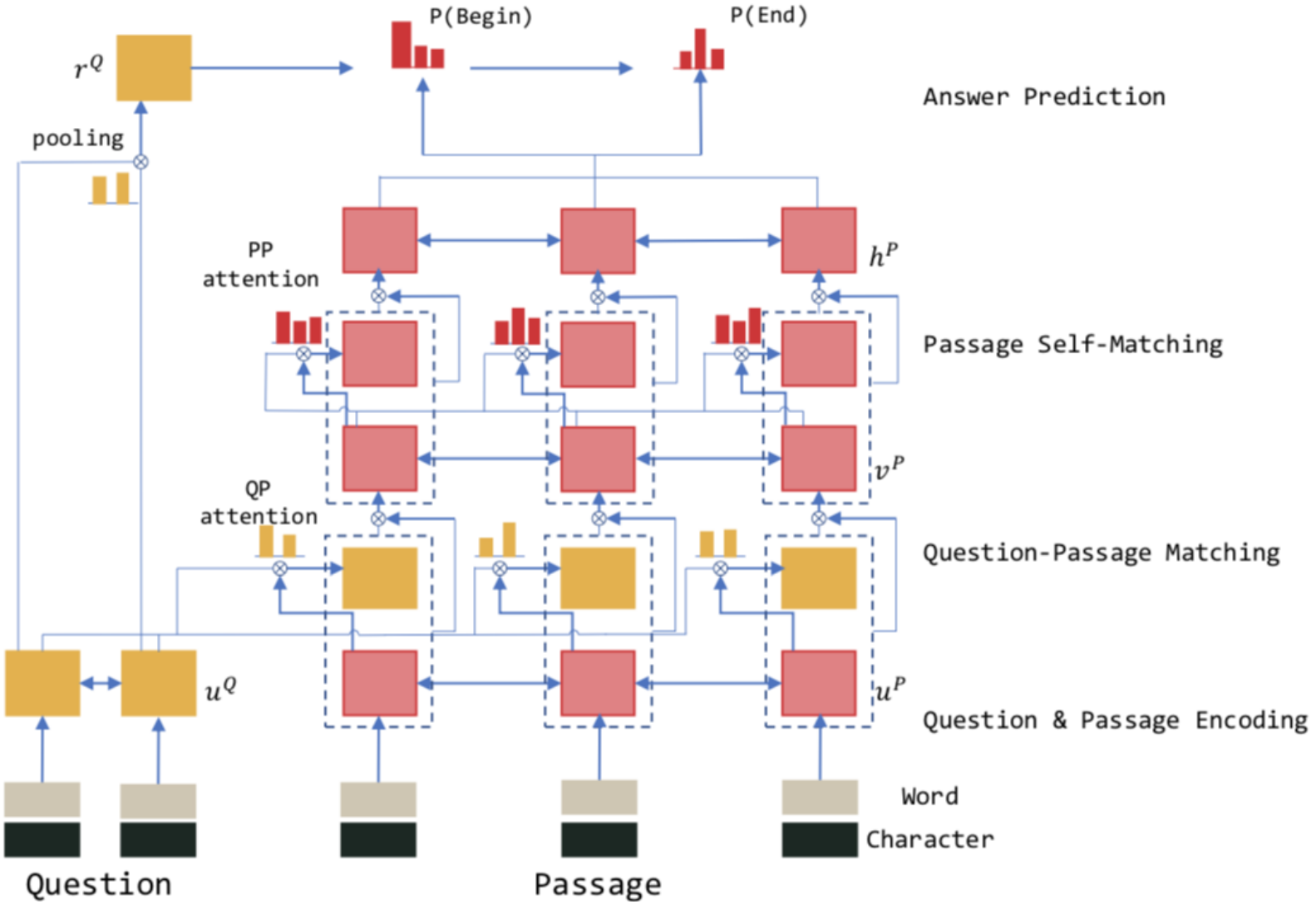
$$\frac{exp(\alpha(E(p_i)) \cdot exp(\alpha(E(q_i))}{\sum_{j'}(\alpha(E(p_i)) \cdot exp(\alpha(E(q_{j'}))}$$

DrQA [4]

Prediction: $P_{start} \propto exp(p_i W_s q)$, $P_{end} \propto exp(p_i W_e q)$
Choose the best span from token i to token i' such that $i \leq i' \leq i + 15$ and $P_{start}(i) \times P_{end}(i')$ is maximized.



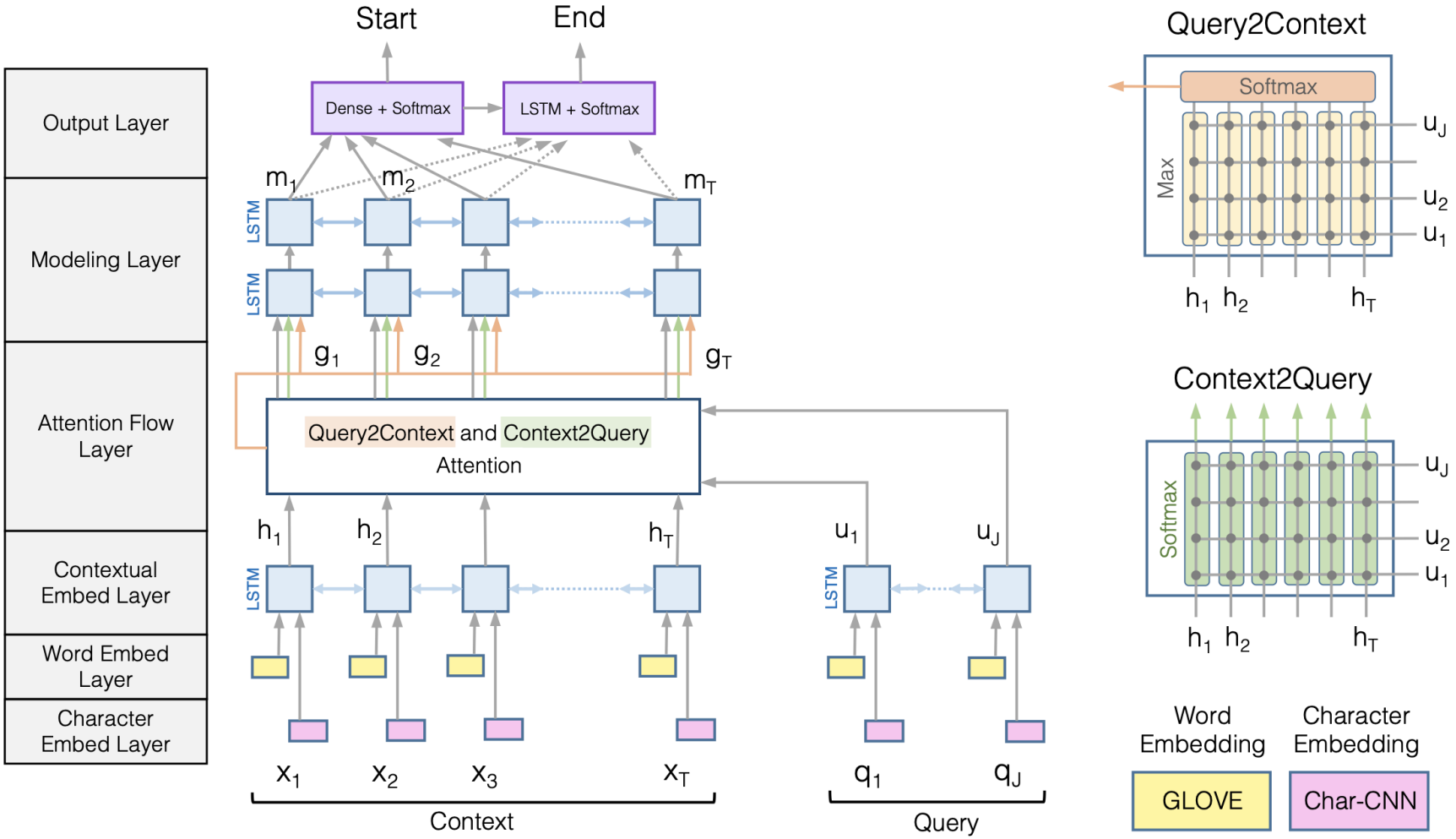
R-NET [5]


P(Begin)
P(End)
r^Q
pooling
Answer Prediction
PP attention
h^P
Passage Self-Matching
v^P
QP attention
Question-Passage Matching
u^Q
u^P
Question & Passage Encoding
Word
Character
Question
Passage

R-NET [5]

1. **Question and passage encoder**: BiRNN to convert the words to their respective word-level embeddings and and character-level embeddings
2. **Gated attention-based recurrent networks**: to incorporate question information into passage representation
3. **Self-matching attention**: passage context is necessary to infer the answer
4. **Output**: use pointer networks to predict the start and end position of the answer. To generate the initial hidden vector for the pointer network an attention-pooling over the question representation is used
5. **Training**: minimize the sum of the negative log probabilities of the ground truth start and end position by the predicted distributions

BiDAF [6]


Output Layer
Modeling Layer
Attention Flow Layer
Contextual Embed Layer
Word Embed Layer
Character Embed Layer
Start
End
Dense + Softmax
LSTM + Softmax
LSTM
Query2Context and Context2Query Attention
Context
Query
Query2Context
Softmax
Max
Context2Query
Softmax
Word Embedding
Character Embedding
GLOVE
Char-CNN

BiDAF [6]

1. **Character Embedding Layer** maps each word to a vector space using character-level CNNs
2. **Word Embedding Layer** maps each word to a vector space using a pre-trained word embedding model
3. **Contextual Embedding Layer** utilizes contextual cues from surrounding words to refine the embedding of the words. These first three layers are applied to both the query and context
4. **Attention Flow Layer** couples the query and context vectors and produces a set of query- aware feature vectors for each word in the context
5. **Modeling Layer employs** a Recurrent Neural Network to scan the context. 6. Output Layer provides an answer to the query.

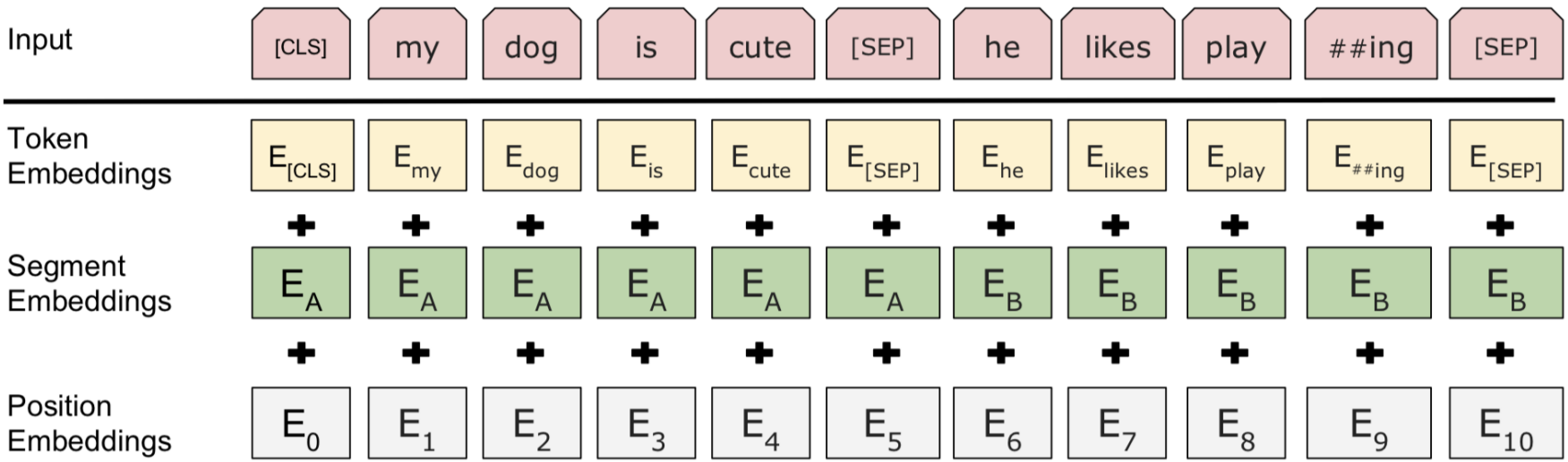
Next generation of QA models

1. S-NET [7]: Extraction-then-synthesis framework
2. QANet [8] benefits from data augmentation techniques, such as paraphrasing and back translation
3. V-NET [9]: end-to-end neural model that enables answer candidates from different passages to verify each other based on their content representations
4. Deep Cascade QA [10]: deep cascade model, which consists of the document retrieval, paragraph retrieval and answer extraction modules

BERT [11]

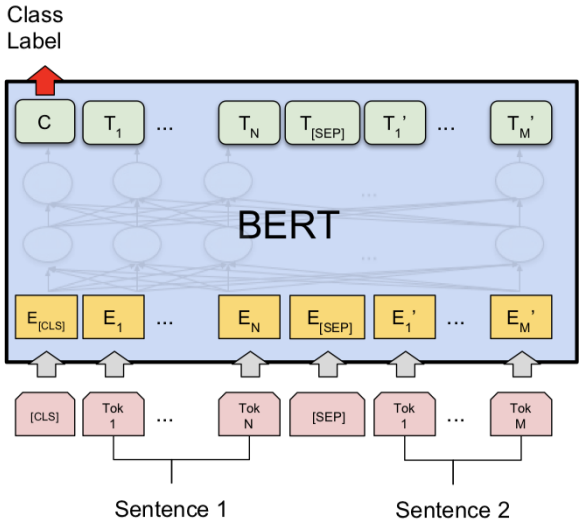
Bidirectional Encoder Representations from Transformers

- L – number of Transformer blocks, H – hidden size, A – the number of self-attention heads
- BERT$_{BASE}$: L=12, H=768, A =12, Total Parameters=110M
- Embeddings: WordPiece + position + segment
- **Two tasks**: Masked LM, Next Sentence Prediction

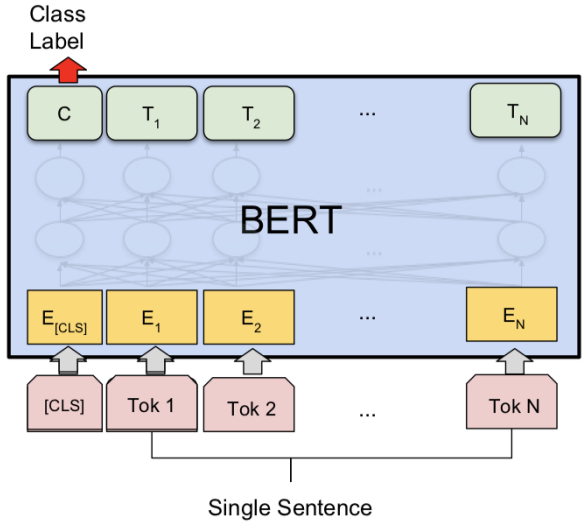


BERT [11]

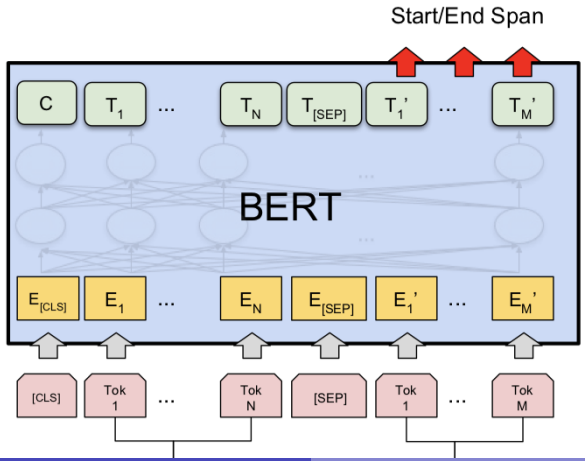
Bidirectional Encoder Representations from Transformers

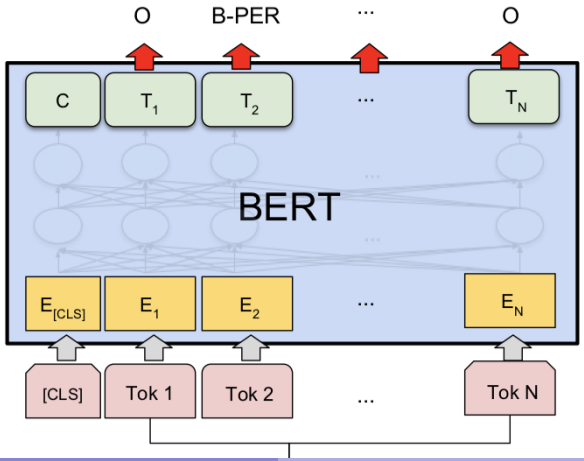


(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA





QA leaderboard

- SQuAD (SQuAD-explorer): ALBERT (ensemble)
- MS Marco (http://www.msmarco.org/leaders.aspxPassage Retrieval leaderboard: enriched BERT base + tips and tricks
- Race (Leaderboard): ALBERT (ensemble)

Today

1. Intro
2. IR-based QA
   - Datasets
   - Models
3. Commonsense QA and inference
4. Knoweledge-based QA

CommonsenseQA [12]

Where would I not want a fox?
👍 hen house, 👎 england, 👎 mountains,
👎 english hunt, 👎 california

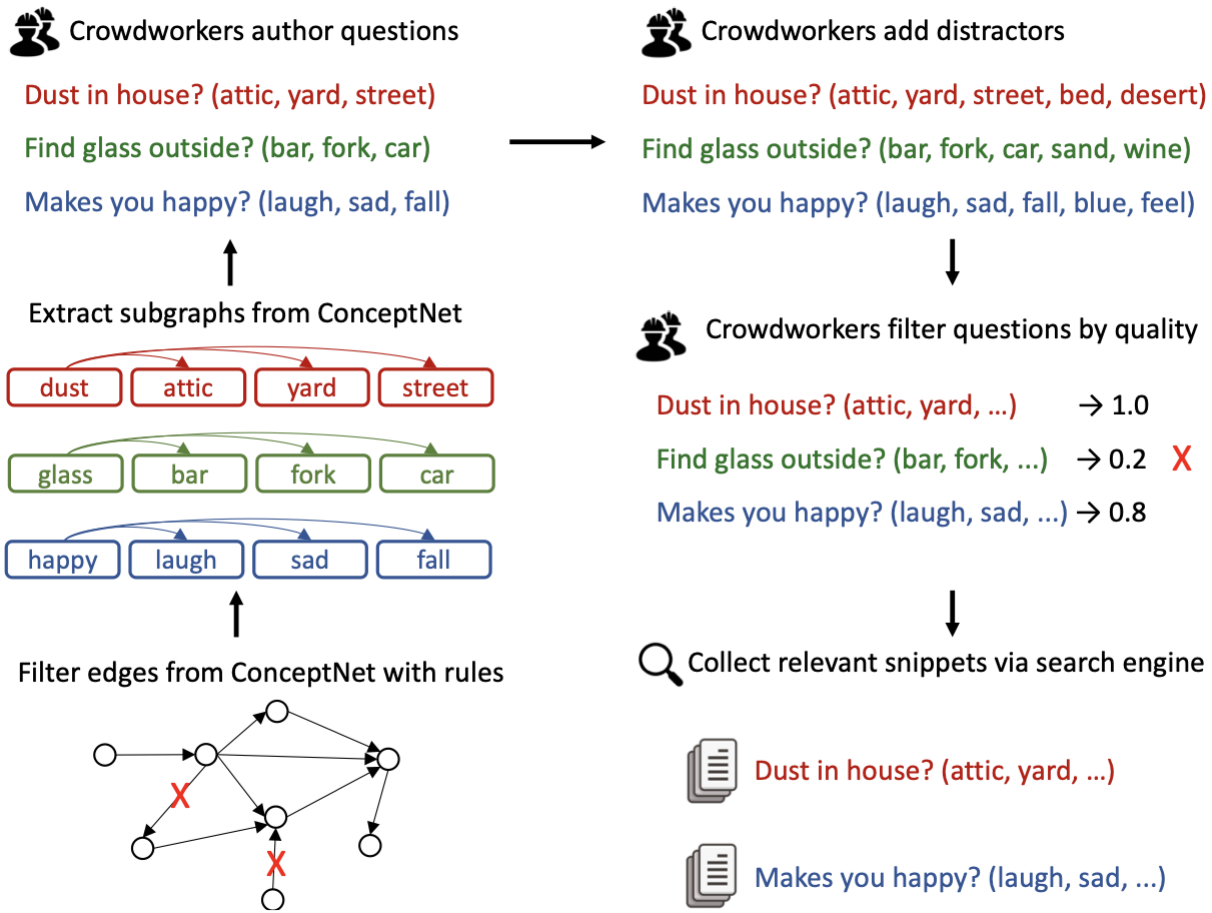
Why do people read gossip magazines?
👍 entertained, 👎 get information, 👎 learn,
👎 improve know how, 👎 lawyer told to

What do all humans want to experience in their own home?
👍 feel comfortable, 👎 work hard, 👎 fall in love,
👎 lay eggs, 👎 live forever

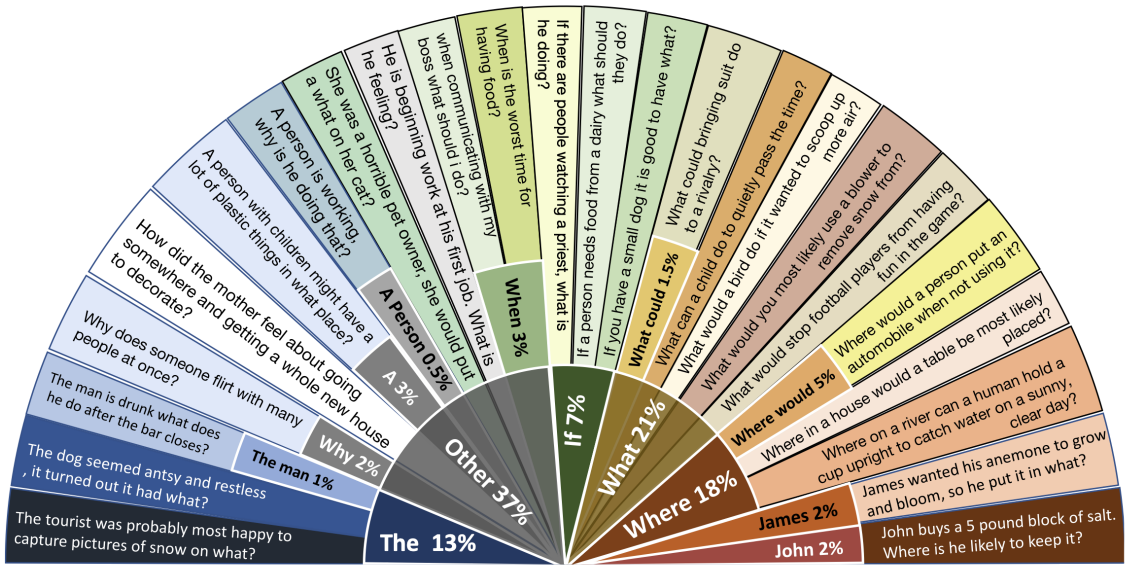
12,247 multiple choice questions that require common sense understanding

`https://www.tau-nlp.org/commonsenseqa`

CommonsenseQA [12]



CommonsenseQA [12]



Baseline: apply BERT

1. Linearize each question: [CLS] If ... ? [SEP] bedroom [SEP]
2. Fine-tune the pre-trained weights from BERT model
3. The hidden representations over each [CLS] token are run through a softmax layer to create the predictions

SWAG [13]

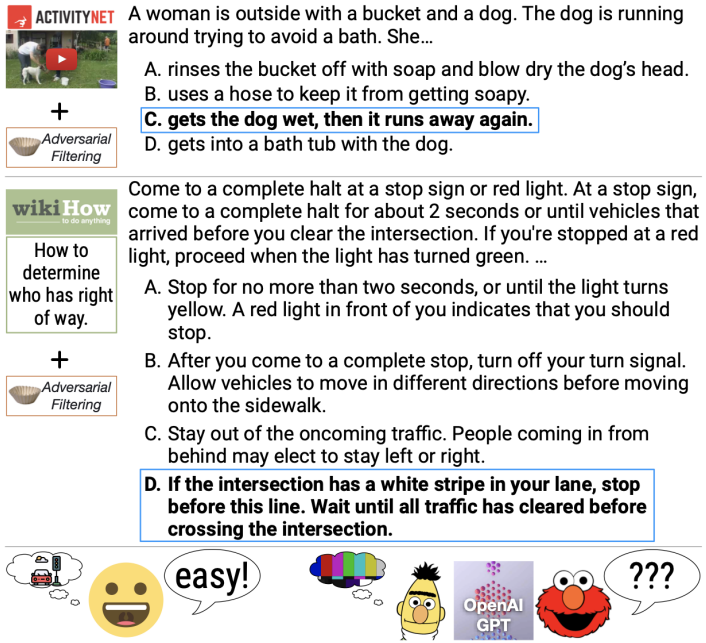
On stage, a woman takes a seat at the piano. She
a) sits on a bench as her sister plays with the doll.
b) smiles with someone as the music plays.
c) is in the crowd, watching the dancers.
d) nervously sets her fingers on the keys.

A girl is going across a set of monkey bars. She
a) jumps up across the monkey bars.
b) struggles onto the monkey bars to grab her head.
c) gets to the end and stands on a wooden plank.
d) jumps up and does a back flip.

The woman is now blow drying the dog. The dog
a) is placed in the kennel next to a woman's feet.
b) washes her face with the shampoo.
c) walks into frame and walks towards the dog.
d) tried to cut her face, so she is trying to do something very close to her face.

SWAG is a dataset for studying grounded commonsense inference. It consists of 113k multiple choice questions about grounded situations: each question comes from a video caption, with four answer choices about what might happen next in the scene. The correct answer is the (real) video caption for the next event in the video; the three incorrect answers are adversarially generated and human verified, so as to fool machines but not humans. `https://rowanzellers.com/swag/`

HellaSWAG [14]



HellaSWAG is a dataset for studying grounded commonsense inference. It consists of 70k multiple choice questions about grounded situations: each question comes from one of two domains – activitynet or wikihow – with four answer choices about what might happen next in the scene. The correct answer is the (real) sentence for the next event; the three incorrect answers are adversarially generated and human verified, so as to fool machines but not humans. `https://rowanzellers.com/hellaswag/`

SWAG [13] and HellaSWAG [14] leaderboards

1. First place: RoBERTa
2. Grover, Big Bird participated, too
3. Other AI2 commonsense benchmarks: `https://leaderboard.allenai.org`

Today

1. Intro
2. IR-based QA
   - Datasets
   - Models
3. Commonsense QA and inference
4. Knoweledge-based QA

Knowledge-based QA

subject	predicate	object
Lyubov Polishchuk	death-date	28 November 2006

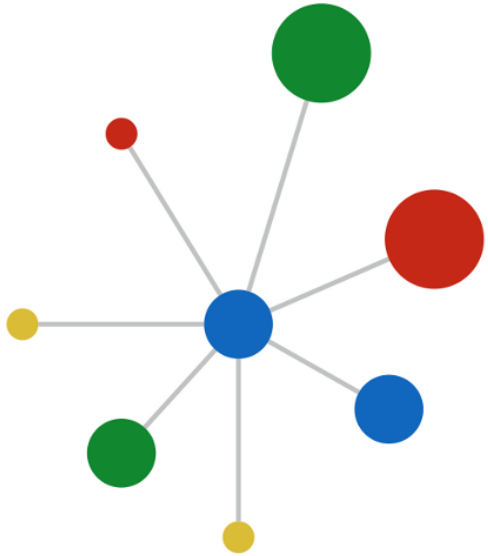
- When Lyubov Polishchuk died?
- Who died on 28 November 2006?

1. **Rule-based methods**: patterns that search for the question word and main verb
2. **OpenIE**: map between the words in question and canonical relations
3. **Knowledge base / knowledge graph**: match the words to concepts and relations in KB / KG

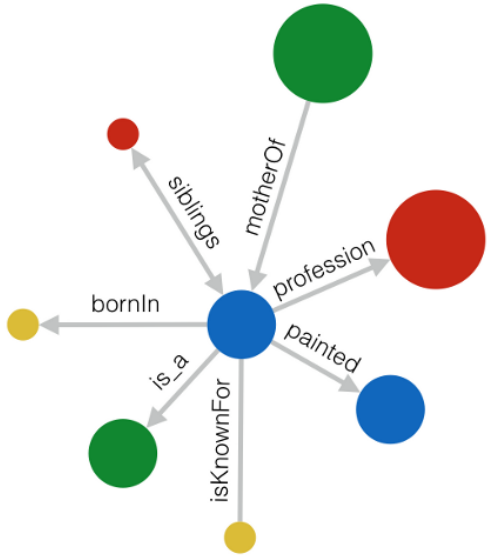
Knowledge representation

"leonardo da vinci"
String

Leonardo da Vinci
Recognized entity



Leonardo da Vinci
Recognized entity
Related entities



Leonardo da Vinci
Recognized entity
Related entities
Named Relationship

medium

Datasets

What American cartoonist is the creator of Andy Lippincott?	(andy_lippincott, character_created_by, garry_trudeau)
Which forest is Fires Creek in?	(fires_creek, containedby, nantahala_national_forest)
What is an active ingredient in childrens earache relief ?	(childrens_earache_relief, active_ingredients, capsicum)
What does Jimmy Neutron do?	(jimmy_neutron, fictional_character_occupation, inventor)
What dietary restriction is incompatible with kimchi?	(kimchi, incompatible_with_dietary_restrictions, veganism)

Figure: Examples of simple QA extracted from the dataset SimpleQuestions. Actual answers are underlined.

- SimpleQuestions (100k questions) [15]: contains more than 100k questions written by human annotators and associated to Freebase facts,
- WebQuestions (6k questions) is created automatically using the Google suggest API.

Knowledge Base Question Answering (KBQA)

The Knowledge Base Question Answering model uses Wikidata to answer question. To find the answer the following models are used: NER model performs entity discovery. In a given question it finds a substring which is an entity, possible mentioned in a Knowledge Base. Classification model classifies the question into a set of predefined relations from Wikidata. Substring extracted by the NER model is used for entity linking. Entity linking preforms matching the substring with one of the Wikidata entities. Matching is based on Levenshtein distance between the substring and an entity description. The result of the matching procedure is a set of candidate entities. The reset is search of the entity among this set with one of the top-k relations predicted by classification model. DeepPavlov

Reference I

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," *arXiv preprint arXiv:1806.03822*, 2018.

G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "Race: Large-scale reading comprehension dataset from examinations," *arXiv preprint arXiv:1704.04683*, 2017.

D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading wikipedia to answer open-domain questions," *arXiv preprint arXiv:1704.00051*, 2017.

Reference II

W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou, "Gated self-matching networks for reading comprehension and question answering," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 189–198.

M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," , 2016.

C. Tan, F. Wei, N. Yang, B. Du, W. Lv, and M. Zhou, "S-net: From answer extraction to answer synthesis for machine reading comprehension.," in *AAAI*, 2018.

A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "Qanet: Combining local convolution with global self-attention for reading comprehension," *arXiv preprint arXiv:1804.09541*, 2018.

Reference III

Y. Wang, K. Liu, J. Liu, W. He, Y. Lyu, H. Wu, S. Li, and H. Wang, "Multi-passage machine reading comprehension with cross-passage answer verification," *arXiv preprint arXiv:1805.02220*, 2018.

M. Yan, J. Xia, C. Wu, B. Bi, Z. Zhao, J. Zhang, L. Si, R. Wang, W. Wang, and H. Chen, "A deep cascade model for multi-document reading comprehension," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7354–7361, Jul. 2019, ISSN: 2159-5399. DOI: 10.1609/aaai.v33i01.33017354. [Online]. Available: http://dx.doi.org/10.1609/aaai.v33i01.33017354.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

Reference IV

A. Talmor, J. Herzig, N. Lourie, and J. Berant, *Commonsenseqa: A question answering challenge targeting commonsense knowledge*, 2018. arXiv: 1811.00937 [cs.CL].

R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, *Swag: A large-scale adversarial dataset for grounded commonsense inference*, 2018. arXiv: 1808.05326 [cs.CL].

T. Schnabel, I. Labutov, D. Mimno, and T. Joachims, "Evaluation methods for unsupervised word embeddings," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 298–307.

A. Bordes, N. Usunier, S. Chopra, and J. Weston, "Large-scale simple question answering with memory networks," *arXiv preprint arXiv:1506.02075*, 2015.