COMP 3225

Natural Language Processing

Applications of NLP: Question Answering

Stuart E. Middleton

sem03@soton.ac.uk

University of Southampton

Overview

- Factoid Question Answering (QA)
- IR-based factoid QA
  - Machine Reading Comprehension (MRC)

- <break - discussion point>

- Knowledge-based QA
  - Graph-based QA
  - Neural Entity Linking
  - Neural Relation Detection and Linking
- Evaluation of Factoid QA

Factoid Question Answering (QA)

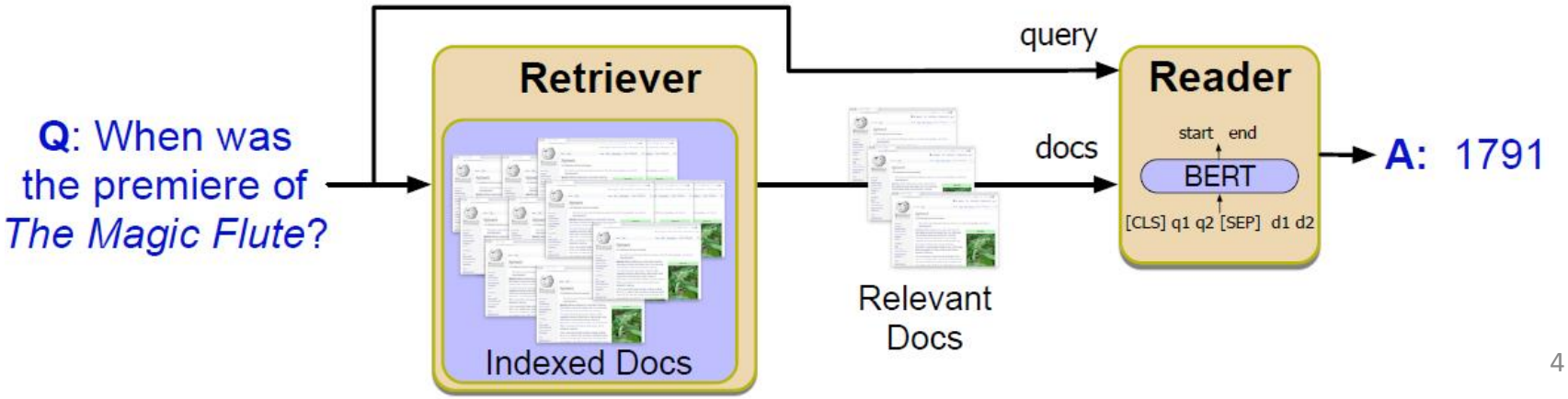
- Information-retrieval (IR) based QA (open domain QA)
  - Uses a large corpus found from web
  - Given a question Machine Reading Comprehension (MRC) extracts answers from spans of text in a corpus
- Knowledge-based QA maps question to a knowledge-base query to get answer
  - Uses a database of facts like DBpedia
- Other types of QA
  - Long-form QA uses with why type questions, long answers
  - Community QA uses QA pairs from sources like Quora or StackOverflow

IR-based factoid QA

- Given a question, return answer from text spans within a corpus of web documents

Question	Answer
Where is the Louvre Museum located?	in Paris, France
What are the names of Odin's ravens?	Huginn and Muninn
What kind of nuts are used in marzipan?	almonds
What instrument did Max Roach play?	drums
What's the official language of Algeria?	Arabic

- Retrieve and Read model
  - Retrieve relevant documents to query from an index
  - MRC to select best text span from relevant documents for answer



IR-based factoid QA

- MRC datasets
  - Tuples of (passage of text, question, answer)
  - SQuAD 1.1 dataset
    >> wikipedia docs + crowd-sourced QA pairs
    >> 100,000 QA pairs based on 500+ articles
  - SQuAD 2.0 dataset
    >> includes unanswerable questions
    >> 150,000 QA pairs (50,000 unanswerable)
  - HotspotQA dataset
    >> wikipedia docs + crowd-sourced QA pairs (multi-hop questions needing multi-docs to answer)
    >> 113,000 QA pairs
  - Natural Questions
    >> wikipedia docs + crowd-sourced QA pairs
    >> 300,000 QA pairs
  - TyDI QA
    >> multi-lingual wikipedia docs + crowd-sourced QA pairs
    >> 240,000 QA pairs

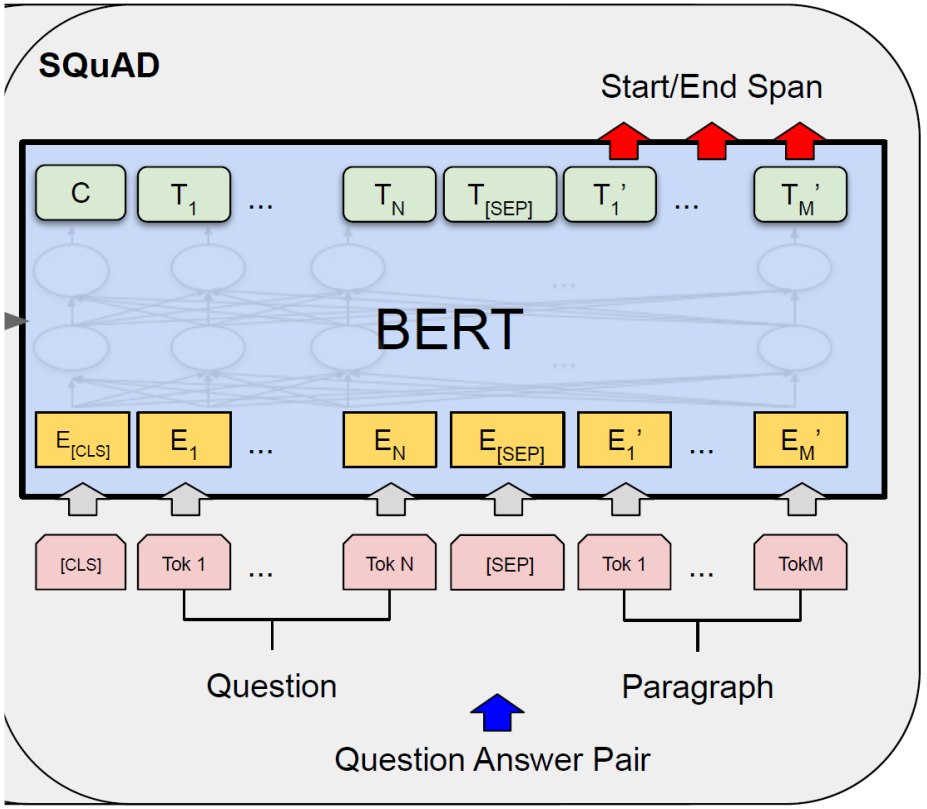
IR-based factoid QA

- MRC datasets
  - Tuples of (passage of text, question, answer)
  - SQuAD 1.1 dataset
    >> wikipedia docs + crowd-sourced QA pairs
    >> 100,000 QA pairs based on 500+ articles
  - SQuAD 2.0 dataset

Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".
Q: "In what city and state did Beyoncé grow up?" A: "Houston, Texas"
Q: "What areas did Beyoncé compete in when she was growing up?" A: "singing and dancing" Q: "When did Beyoncé release Dangerously in Love?" A: "2003"

IR-based factoid QA

- Machine Reading Comprehension (MRC)
  - Answer extraction is a span labelling task (extractive in this model)
  - X = question and passage = [CLS] question ... [SEP] paragraph ...
  - Y = answer start = is_start (0/1)
  - Y = answer end = is_end (0/1)

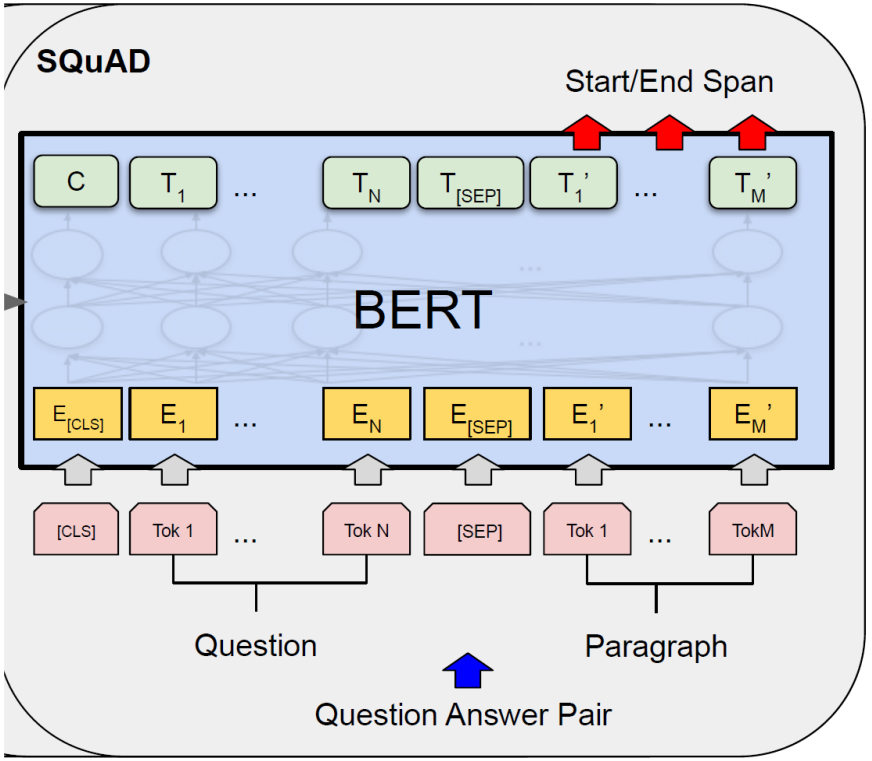


Cite: Devlin 2019, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/abs/1810.04805

IR-based factoid QA

- Machine Reading Comprehension (MRC)
- T = final hidden layer of BERT model for paragraph
- S = answer start embedding >> MLP = Dense layer (dim = 1)
- E = answer end embedding >> MLP = Dense layer (dim = 1)
- Model trained two outputs simultaneously (S and E)
- Loss = cross entropy (for both S and E)

- span score(i, j) = S . T_i + E . T_j
- best span = argmax(span_score(i, j))
  where j >= i



IR-based factoid QA

- Machine Reading Comprehension (MRC)
  - For no-answer questions passage will contain special [CLS] token only
  - BERT has 512 word seq limit
    >> use a sliding 512 word window over larger passage documents
    Start and end span tokens can then be from different 512 word windows
- Further reading
  - Original BERT paper
    Devlin 2019, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://arxiv.org/abs/1810.04805
  - Blog (fine tuning BERT for QA)
    https://medium.com/saarthi-ai/build-a-smart-question-answering-system-with-fine-tuned-bert-b586e4cfa5f5
  - Blog (TF code examples for BERT QA model)
    https://medium.com/swlh/fine-tuning-bert-for-text-classification-and-question-answering-using-tensorflow-framework-4d09daeb3330

Break

- Panopto Quiz - discussion point

- When would IR-based factual QA likely fail?

If retrieval requires access to pages in the deep web
For domains where answers involve specialist vocabulary
For questions where answers are contained across multiple documents

Break

- Panopto Quiz - discussion point

- When would IR-based factual QA likely fail?

If retrieval requires access to pages in the deep web
>> deep web is dynamically rendered and often hard to index and search, so retrieval phase may easily miss pages with answers

For domains where answers involve specialist vocabulary
>> as long as there are documents with answers it should be OK

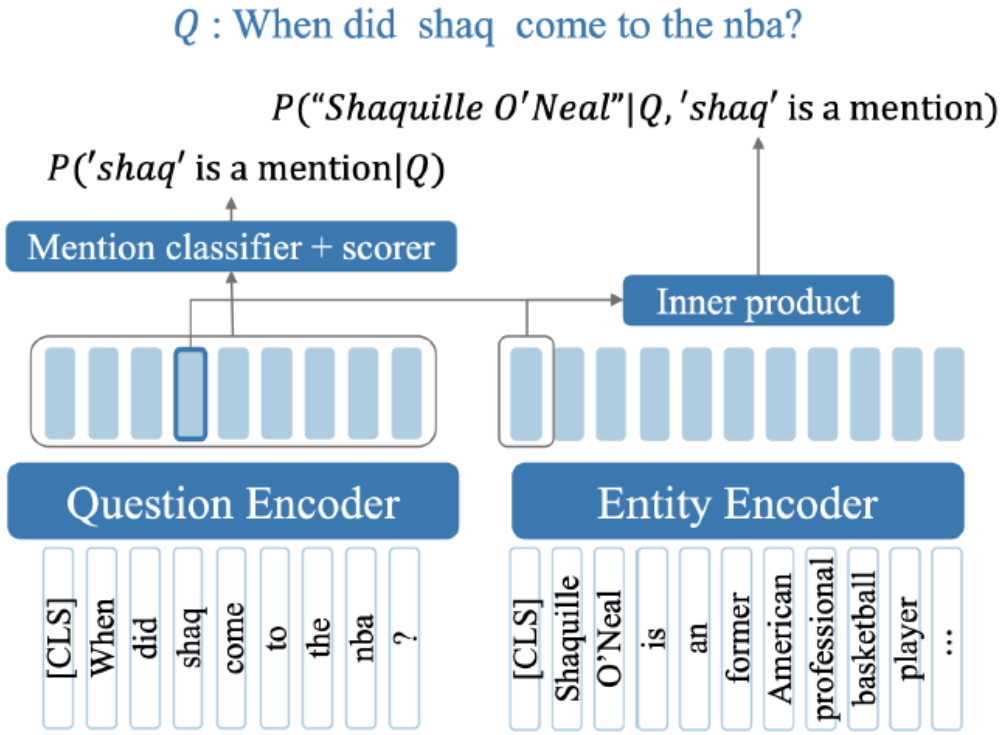
For questions where answers are contained across multiple documents
>> techniques like sliding input windows allow long documents, or several documents to be processed OK

Knowledge-based QA

- Given a question, return a database query to get the answer
- Graph-based QA
  - Database is a set of RDF triples (subject, predicate, object) such as DBpedia
  - Entity linking → Relation detection/linking → Database query for answer
- Entity linking is task of associating a mention in text with an ontology/database entry
  - X = question text
  - Y = [entity text span + entity URI] x N
- Non-neural entity linking
  - TAGME https://tagme.d4science.org/tagme/
  - Anchor dictionaries (wikipedia concept URI + text spans linking to this URI) + entity disambiguation
  - P(page|anchor entity) = co-occurrence in corpus (from counts) + relatedness
  - relatedness = entities who's pages share links in common

Knowledge-based QA

- Neural Entity Linking
  EQL model

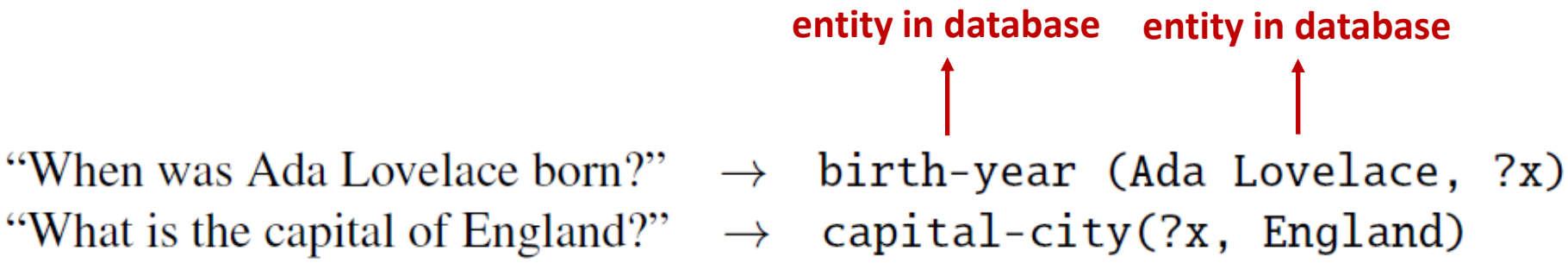


- X = (question, entity candidate desc)
  question encoder = BERT([CLS] question ... [SEP])
  entity encoder = BERT([CLS] entity title ... [ENT] entity desc ... [SEP])
- Y = entity mention classifier = is_start(0/1); is_end(0/1); is_entity(0/1)
  Y = entity linker = KB_entity (for entity mention in question)
- Jointly train entity mention classifier and entity linker
- The entity linker compares (a) entity mention embedding and (b) entity candidate embedding to provide a disambiguated KB_entity for each question entity text span

Cite: Li, B. Z. et. al. (2020). Efficient one-pass end-to-end entity linking for questions. EMNLP https://www.aclweb.org/anthology/2020.emnlp-main.522/

Knowledge-based QA

- Neural Relation Detection and Linking
  - For single relations can be done in same way as neural entity linking
  - X = question text
  - Y = [relation text span + relation URI] x N
- Relation detection
  - question text → relation mention
- Relation linking
  - relation mention → relation URI
- Compute answer by running query on database (e.g. DBpedia)



Evaluation of Factoid QA

- Mean Reciprocal Rank (MRR)
  - Assumes a ranked list of answers in order of confidence
  - Rank = highest-ranked correct answer

$$\text{MRR} = \frac{1}{N} \sum_{i=1 \text{ s.t. } rank_i \neq 0}^{N} \frac{1}{rank_i}$$

- Exact match
  - Percentage of predicted answers that exactly match gold answers
- F1
  - Bag of token overlap between predicted answer and gold answer
  - num_same = overlap
  - pred_toks = tokens of predicted answer
  - gold_toks = tokens of predicted answer

  - precision = 1.0 * num_same / len(pred_toks)
  - recall = 1.0 * num_same / len(gold_toks)
  - f1 = (2 * precision * recall) / (precision + recall)

Required Reading

- Information Extraction
  - Jurafsky and Martin, Speech and Language Processing, 3rd edition (online) >> chapter 23

Questions

- Panopto Quiz - 1 minute brainstorm for interactive questions

  Please write down in Panopto quiz in **1 minute** two or three questions that you would like to have answered at the next interactive session.

  Do it **right now** while its fresh.

  Take a screen shot of your questions and **bring them with you** at the interactive session so you have something to ask.