Phrase-Indexed Question Answering: A New Challenge for Scalable Document Comprehension

Minjoon Seo, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, Hannaneh Hajishirzi
Question Answering?
Barack Obama (1961–present) was the 44th President of the United States. When was Obama born?
1961

"Barack Obama (1961-present) was the 44th President of the United States."

When was Obama born?
Extractive QA Datasets

• SQuAD (Rajpurkar et al., 2016)
• NewsQA (Trischler et al., 2016)
• TriviaQA (Joshi et al., 2017)
• QuAC (Choi et al., 2018)
• CoQA (Reddy & Chen & Manning, 2018)
• HotpotQA (Yang et al., 2018)
• And more...
Open-domain QA?
“Barack Obama (1961-present) was the 44th President of the United States.”

When was Obama born?
When was Obama born?

Model

1961

When was Obama born?

Question
4 Million documents
3 Billion tokens

0.1s / doc * 4M docs = 6 days!
When was Obama born? 1961

Choi et al., 2017; Chen et al., 2017; Clark & Gardner, 2017

TF-IDF, BM 25, LSA

Pipelined
When was Obama born?

1961

Choi et al., 2017; Chen et al., 2017; Clark & Gardner, 2017
When was Obama born?

1911

Choi et al., 2017; Chen et al., 2017; Clark & Gardner, 2017

Error propagation
Ideally...
When was Obama born? 1961

Information Retrieval

TF-IDF, BM 25, LSA

Model

When was Obama born? 1961
Model

When was Obama born?

? 

1961

End-to-end & elegant... But how?
Solution: Index phrases!
When was Obama born?

Nearest neighbor search

Document Indexing
- Locality Sensitive Hashing
- aLSH (Shrivastava & Li, 2014)
- ...
"Barack Obama (1961-present) was the 44th President of the United States."

Who is the 44th President of the U.S.?

When was Obama born?

Phrase encoding

Nearest neighbor search

Question encoding
\[ \hat{a} = \arg\max_a F_{\theta}(a, q, d) \]

\[ \hat{a} = \arg\max_a G_{\theta}(q) \cdot H_{\theta}(a, d) \]
Decomposability is a strong constraint
Phrase-Indexed QA (PIQA) Challenge

• Open-domain QA is **hard** to setup or evaluate
• Instead, **benchmark** on existing datasets (e.g. SQuAD)
• Create **two** models:
  • Phrase (document) encoder
  • Question encoder
• Phrase encoder **must** be **question-agnostic**, and vice versa
• Answer **must** be obtained via **nearest neighbor search** (NNS)
PI-SQuAD Evaluation
Is it too easy or too hard?
SQuAD v1.1

Red color is phrase-indexed.

- BERT (Devlin et al., 2018) - - - 92% F1
- SA+ELMo (Peters et al., 2018) - - - 86% F1
- SA+ELMo (Seo et al., 2018) - - - 64% F1
- Feature-based (Rajpurkar et al., 2018) - - - 50% F1

Decomposability gap
SQuAD v1.1

Red color is phrase-indexed.

- BERT (Devlin et al., 2018) 92% F1
- SA+ELMo (Peters et al., 2018) 86% F1
- Sparse+SA+ELMo 70% F1
- Match-LSTM (Wang & Jiang, 2017) 68% F1
- SA+ELMo (Seo et al., 2018) 64% F1
- Feature-based (Rajpurkar et al., 2018) 50% F1

*First neural model*
Phrase Representation Learning

• Not just about scalability, but also about comprehension
• **Standalone** representations of phrases (document)

PIQA can be viewed as:
• A **phrase embedding evaluation** method
  • Sentence embedding in SNLI (Bowman et al., 2015)
• Constructing a **memory of knowledge**
  • Memory Networks (Weston et al., 2014)
According to the American Library Association, this makes ...

... tasked with drafting a European Charter of Human Rights, ...
The LM engines were successfully test-fired and restarted, ...

Steam turbines were extensively applied ...
... primarily accomplished through the ductile stretching and thinning.

... directly derived from the homogeneity or symmetry of space ...
Demo on my Macbook

Corpus size: 300k Tokens (SQuAD dev set)
16 CPUs: 100s+
GPU: 10s+
A lot of things to do

• Closing the gap due to decomposability constraint
  • BERT (Devlin et al., 2018)?
• Reducing index storage (100TB+ for Wikipedia)
• Reducing phrase embedding dimension (1024)
• Extending to open-domain QA
• Analyzing phrase representations
• And more!
http://pi-qa.com

Thank you!