Recent Advances on Open Domain Question Answering

RSVP.ai

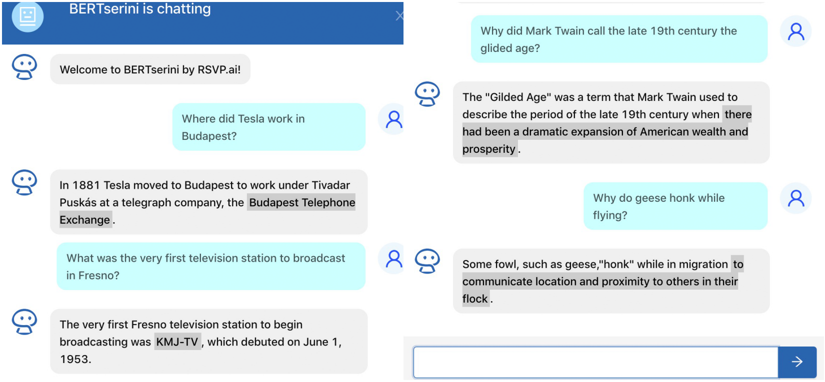
Victor Wei Yang

October 24, 2019

Table of Contents

Open Domain Question Answering

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019



Experimental Setup

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

- Problem Definition
  - **Machine Reading Comprehension (MRC)**: The question and the evidence document that might contain the answer to the question are given. The target is to find the answer from the document.

Experimental Setup

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

- Problem Definition
  - **Machine Reading Comprehension (MRC)**: The question and the evidence document that might contain the answer to the question are given. The target is to find the answer from the document.
  - **Open Domain Question Answering**: Only the question is given.

Experimental Setup

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

- Problem Definition
  - **Machine Reading Comprehension (MRC)**: The question and the evidence document that might contain the answer to the question are given. The target is to find the answer from the document.
  - **Open Domain Question Answering**: Only the question is given.
- Experiment Design
  - **Training**: train a MRC model on the human-annotated datasets (e.g. SQuAD and CMRC).

Experimental Setup

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

- Problem Definition
  - **Machine Reading Comprehension (MRC)**: The question and the evidence document that might contain the answer to the question are given. The target is to find the answer from the document.
  - **Open Domain Question Answering**: Only the question is given.
- Experiment Design
  - **Training**: train a MRC model on the human-annotated datasets (e.g. SQuAD and CMRC).
  - **Inference**: retrieve evidence documents using information retrieval toolkit. Then use the MRC model to read them!

Experimental Setup

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

- Problem Definition
  - **Machine Reading Comprehension (MRC)**: The question and the evidence document that might contain the answer to the question are given. The target is to find the answer from the document.
  - **Open Domain Question Answering**: Only the question is given.
- Experiment Design
  - **Training**: train a MRC model on the human-annotated datasets (e.g. SQuAD and CMRC).
  - **Inference**: retrieve evidence documents using information retrieval toolkit. Then use the MRC model to read them!
  - **Evaluation**: use exact match (EM) and partial match (F_1) scores between the prediction and ground truth answer as the evaluation metric.

BERTserini

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

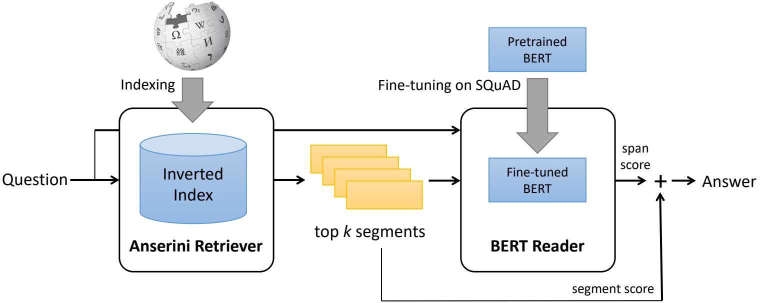


Figure: Architecture of BERTserini

BERTserini Pipeline

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

- Anserini-Retriever:
  - Filters out relevant documents
  - Gives paragraph scores.
  - BM25 Similarity:

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

BERTserini Pipeline

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

- Anserini-Retriever:
  - Filters out relevant documents
  - Gives paragraph scores.
  - BM25 Similarity:

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

- BERT-Reader:
  - Reads paragraphs by Anserini;
  - Predicts the answer spans;
  - Gives phrase scores.

BERTserini Pipeline

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

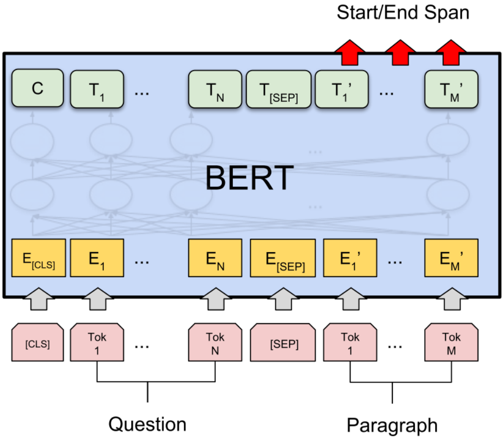


Figure: BERT for Question Answering

BERTserini Pipeline

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

- Anserini-Retriever:
  - Filters out relevant documents
  - Gives paragraph scores.
  - BM25 Similarity
- BERT-Reader:
  - Reads paragraphs by Anserini;
  - Predicts the answer spans;
  - Gives phrase scores.
- Aggregator:
  - Re-ranks the predictions according to weighted sum of scores.

$$S = (1 - \mu) \cdot S_{\mathrm{BM25}} + \mu \cdot S_{\mathrm{BERT}},$$

    where $\mu \in [0, 1]$ is a hyperparameter.

Text Segments

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

- **Article**: The 5.08M Wikipedia articles are directly indexed; that is, an article is the unit of retrieval.

Text Segments

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

- **Article**: The 5.08M Wikipedia articles are directly indexed; that is, an article is the unit of retrieval.
- **Paragraph**: The corpus is pre-segmented into 29.5M paragraphs and indexed, where each paragraph is treated as a "document" (i.e., the unit of retrieval).

Text Segments

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

- **Article**: The 5.08M Wikipedia articles are directly indexed; that is, an article is the unit of retrieval.
- **Paragraph**: The corpus is pre-segmented into 29.5M paragraphs and indexed, where each paragraph is treated as a "document" (i.e., the unit of retrieval).
- **Sentence**: The corpus is pre-segmented into 79.5M sentences and indexed, where each sentence is treated as a "document".

BERTserini

End-to-End Open-Domain Question Answering with BERTserini, NAACL 2019

Model	EM	R	F1
Dr.QA (Chen et al., 2017)	27.1	77.8	-
Dr.QA + Fine-tune	28.4	-	-
Dr.QA + Multitask	29.8	-	-
R^3 (Wang et al., 2017)	29.1	-	37.5
Kratzwald and Feuerriegel (2018)	29.8	-	-
Par. R. (Lee et al., 2018)	28.5	83.1	-
Par. R. + Answer Agg.	28.9	-	-
Par. R. + Full Agg.	30.2	-	-
BERTserini (Article, $k = 5$)	19.1	63.1	25.9
BERTserini (Paragraph, $k = 29$)	36.6	75.0	44.0
BERTserini (Sentence, $k = 78$)	34.0	67.5	41.0
BERTserini (Paragraph, $k = 100$)	**38.6**	**85.8**	**46.1**

Figure: Results on SQuAD development questions.

Table of Contents

Distant Supervision

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

- BERTserini only fine-tune BERT on the original SQuAD dataset, containing a total of **only 442 documents**.

Distant Supervision

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

- BERTserini only fine-tune BERT on the original SQuAD dataset, containing a total of **only 442 documents**.
- This contrasts with the diversity of paragraphs that the model will likely encounter at inference time in the retrieval-based setting.

Distant Supervision

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

- BERTserini only fine-tune BERT on the original SQuAD dataset, containing a total of **only 442 documents**.
- This contrasts with the diversity of paragraphs that the model will likely encounter at inference time in the retrieval-based setting.
- We create additional training examples by fetching paragraphs from the corpus using Anserini and give these paragraphs labels based on the ground truth answers provided.

Fine-tuning Order

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

- **SRC** + **DS**: Fine-tune BERT with all data, "lumped" together as a single, larger training set. In practice, this means that the source and augmented data are shuffled together.

- **SRC** + **DS**: Fine-tune BERT with all data, “lumped” together as a single, larger training set. In practice, this means that the source and augmented data are shuffled together.
- **DS** $\rightarrow$ **SRC**: Fine-tune the reader in stages, first on the augmented data and then the source dataset.

Fine-tuning Order

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

- **SRC** $+$ **DS**: Fine-tune BERT with all data, "lumped" together as a single, larger training set. In practice, this means that the source and augmented data are shuffled together.
- **DS** $\rightarrow$ **SRC**: Fine-tune the reader in stages, first on the augmented data and then the source dataset.
- **SRC** $\rightarrow$ **DS**: Fine-tune the reader in stages, on the source dataset and then the augmented data.

Fine-tuning Order

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

Model	EM	F_1	EM	F_1
	SQuAD		CMRC	
SRC	41.8	49.5	44.5	60.9
DS(+)	44.0	51.4	45.5	61.1
DS($\pm$)	48.7	56.5	48.3	63.9
SRC+DS($\pm$)	45.7	53.5	49.0	64.6
DS($\pm$) $\rightarrow$ SRC	47.4	55.0	45.6	61.9
SRC $\rightarrow$ DS($\pm$)	**50.2**	**58.2**	**49.2**	**65.4**

Table: Results exploring different approaches to combining source and augmented training data on the two datasets: SQuAD and CMRC.

Negative Sampling Strategy

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

- **Top-down**: We choose negative examples with the highest paragraph scores from the retrieved paragraphs.

Negative Sampling Strategy

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

- **Top-down**: We choose negative examples with the highest paragraph scores from the retrieved paragraphs.
- **Bottom-up**: We choose negative examples with the lowest paragraph scores from the retrieved paragraphs.

Negative Sampling Strategy

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

- **Top-down**: We choose negative examples with the highest paragraph scores from the retrieved paragraphs.
- **Bottom-up**: We choose negative examples with the lowest paragraph scores from the retrieved paragraphs.
- **Random**: We randomly sample negative examples from the retrieved paragraphs.

Negative Sampling Strategy

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

	SQuAD		CMRC	
	EM	F1	EM	F1
Top-down	49.2	57.2	**48.8**	64.5
Bottom-up	46.8	54.9	48.6	65.2
Random	**49.6**	**57.6**	48.6	**64.7**

Table: Effects of different negative sampling strategies on SQuAD and CMRC.

Sample Ratio

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

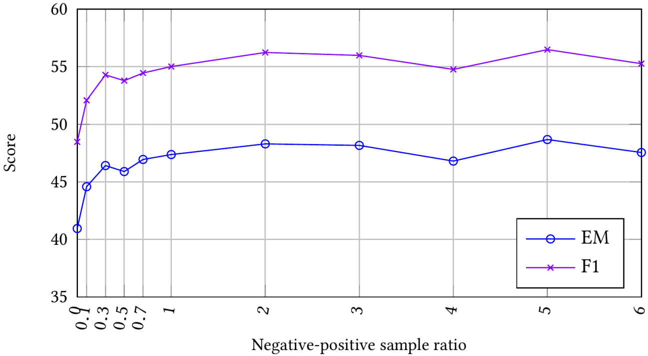


Figure: Effects of varying d, the positive–negative ratio of examples, on SQuAD.

Parameter Analysis

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

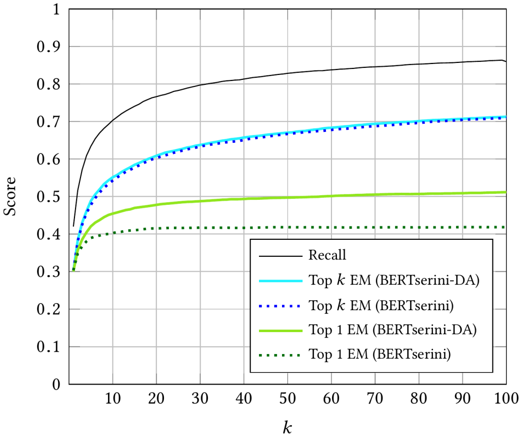


Figure: Effects of the number of retrieved paragraphs k on SQUAD

Sample Analysis

Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering

Question	**Answers from BERTserini***	**Answers from our augmented model**
Super Bowl 50 decided the NFL champion for what season?	Super Bowl XXXVII was an American football game between the American Football Conference (AFC) champion Oakland Raiders and the National Football Conference (NFC) champion Tampa Bay Buccaneers to decide the National Football League (NFL) champion for the **2002** season.	Super Bowl 50 decided the **2015** NFL Champion and was played at Levi's Stadium in Santa Clara, California on Sunday, February 7, 2016.

Table: Sample questions and answers

Unsolved Issues

Multiple Spans

Question: Which British general was killed at Khartoum in 1885?
Answer: Gordon
Context: In February 1885 Gordon returned to the Sudan to evacuate Egyptian forces. Khartoum came under siege the next month and rebels broke into the city, killing Gordon and the other defenders. The British public reacted to his death by acclaiming 'Gordon of Khartoum', a saint. However, historians have suggested that Gordon...

Figure: Noisy supervision can cause many spans of text that contain the answer, but are not situated in a context that relates to the question (red), to distract the model from learning from more relevant spans (green).

Unsolved Issues

Adversarial Examples

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Figure: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

Table of Contents

Paragraph Reranking

Ranking Paragraphs for Improving Answer Recall in Open-Domain Question Answering

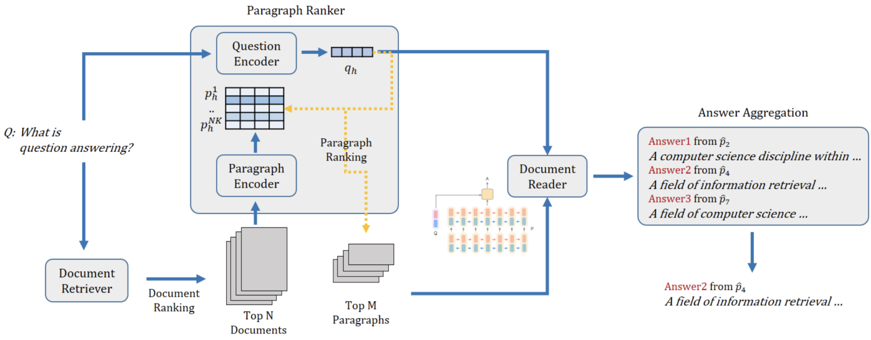


Figure: Open-domain QA pipeline with Paragraph Ranker

Answer Reranking

1. Evidence Aggregation for Answer Re-Ranking in Open-Domain Question Answering
2. Retrieve, Read, Rerank: Towards End-to-End Multi-Document Reading Comprehension

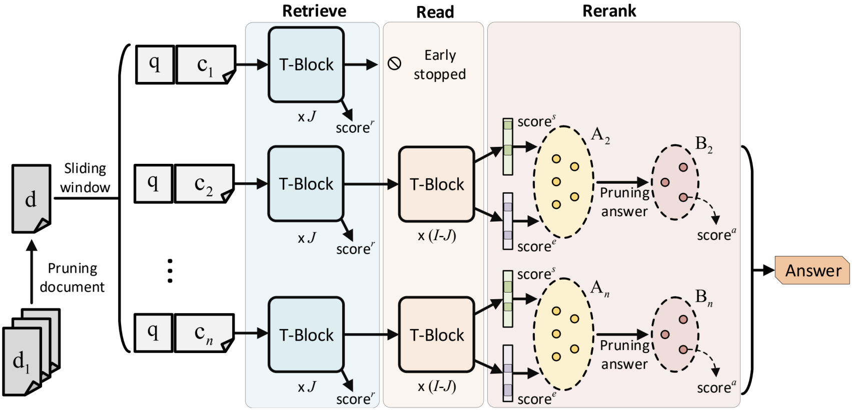


Figure: Retrieve-Read-Rerank QA Architecture

Weakly Supervised QA

A Discrete Hard EM Approach for Weakly Supervised Question Answering, ACL 2019

- Multi-mention reading comprehension
- Reading comprehension with discrete reasoning
- Semantic Parsing

Multi-mention reading comprehension (TriviaQA)
Question: Which composer did pianist Clara Wieck marry in 1840?
Answer: Robert Schumann
Document: Robert Schumann was a German composer and influential music critics. … Robert Schumann himself refers to it as “an affliction of the whole hand”. … Robert Schumann is mentioned in a 1991 episode of Seinfeld “The Jacket”. …. Clara Schumann was a German musician and composer. Her husband was the composer Robert Schumann. … Brahms met Joachim in Hanover, made a very favorable impression on him, and got from him a letter of introduction to Robert Schumann.

Reading comprehension with discrete reasoning (DROP)
Question: How many yards longer was Rob Bironas' longest field goal compared to John Carney's only field goal?
Answer: 4
Document: … Titans responded with Kicker Rob Bironas managing to get a 37 yard field goal. … In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. The Chiefs answered with kicker John Carney getting a 36 yard field goal. Titans would retake the lead with Young and Williams hooking up with each other again on a 41 yard td pass. In the fourth quarter Tennessee clinched the victory with Bironas nailing a 40 yard and a 25 yard field goal.

41 - 37 ✗
41 - 37 ✗
40 - 36 ✓

Weakly Supervised QA

A Discrete Hard EM Approach for Weakly Supervised Question Answering, ACL 2019

In the weak supervision scenario, the model has access to x and $Z = \{z_1, z_2, \ldots, z_n\}$, and the selection of the best solution in Z can be modeled as a latent variable.

$$
\begin{aligned}
J_{\mathrm{MML}}(\theta|x, Z) &= -\log \quad \mathbb{P}(y|x;\theta) \\
&= -\log \sum_{z_i \in Z_{\mathrm{tot}}} \mathbb{P}(y|z_i)\mathbb{P}(z_i|x;\theta) \\
&= -\log \sum_{z_i \in Z} \mathbb{P}(z_i|x;\theta)
\end{aligned}
$$

3. SQL Query Generation (WikiSQL)

Question: What player played guard for Toronto in 1996-1997?
Table Header: `player, year, position, ...`
Answer (y): John Long
f: SQL executor
Z_{tot}: Non-nested SQL queries with up to 3 conditions
Z:
```
Select player where position=guard and year in toronto=1996-97
Select max(player) where position=guard and year in toronto=1996-97
Select min(player) where position=guard
Select min(player) where year in toronto=1996-97
Select min(player) where position=guard and year in toronto=1996-97
```

Weakly Supervised QA

A Discrete Hard EM Approach for Weakly Supervised Question Answering, ACL 2019

- The model computes the likelihood of each z_i given the input x with respect to θ, $\mathbb{P}(z_i|x;\theta)$, and picks one of Z with the largest likelihood:

$$\tilde{z} = \mathrm{argmax}_{z_i \in Z} \mathbb{P}(z_i|x;\theta)$$

Weakly Supervised QA

A Discrete Hard EM Approach for Weakly Supervised Question Answering, ACL 2019

- The model computes the likelihood of each z_i given the input x with respect to θ, $\mathbb{P}(z_i|x;\theta)$, and picks one of Z with the largest likelihood:

$$\tilde{z} = \mathrm{argmax}_{z_i \in Z} \mathbb{P}(z_i|x;\theta)$$

- Then, the model optimizes on a standard negative log likelihood objective, assuming $\tilde{z}$ is a true solution.

$$\begin{aligned} J_{\text{Hard}}(\theta|x, Z) &= -\log \mathbb{P}(\tilde{z}|x;\theta) \\ &= -\log \max_{z_i \in Z} \mathbb{P}(z_i|x;\theta) \\ &= -\max_{z_i \in Z} \log \mathbb{P}(z_i|x;\theta) \end{aligned}$$

Weakly Supervised QA

A Discrete Hard EM Approach for Weakly Supervised Question Answering, ACL 2019

Model	Accuracy Dev	Accuracy Test
Weakly-supervised setting		
REINFORCE (Williams, 1992)	< 10	
Iterative ML (Liang et al., 2017)	70.1	
Hard EM (Liang et al., 2018)	70.2	
Beam-based MML (Liang et al., 2018)	70.7	
MAPO (Liang et al., 2018)	71.8	72.4
MAPOX (Agarwal et al., 2019)	74.5	74.2
MAPOX+MeRL (Agarwal et al., 2019)	74.9	74.8
MML	70.6	70.5
Ours	**84.4**	**83.9**
Fully-supervised setting		
SQLNet (Xu et al., 2018)	69.8	68.0
TypeSQL (Yu et al., 2018b)	74.5	73.5
Coarse2Fine (Dong and Lapata, 2018)	79.0	78.5
SQLova (Hwang et al., 2019)	87.2	86.2
X-SQL (He et al., 2019)	**89.5**	**88.7**

Figure: Results on WIKISQL

Other Works

- **Transfer Learning**: *MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension*
- **Integration with Knowledge Bases**: *Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text*
- **Integration with Syntax Information**: *SG-Net: Syntax-Guided Machine Reading Comprehension*
- **Questions with Reasoning and Explanations**: *Dynamically Fused Graph Network for Multi-hop Reasoning*

Summary

- **BERTserini**
  - Anserini-retriever
  - BERT reader
  - Aggregator

Summary

- **BERTserini**
  - Anserini-retriever
  - BERT reader
  - Aggregator
- **Data Augmentation**
  - Fine-tuning Order
  - Negative Sampling Strategy
  - Sample Ratio

Summary

- **BERTserini**
  - Anserini-retriever
  - BERT reader
  - Aggregator
- **Data Augmentation**
  - Fine-tuning Order
  - Negative Sampling Strategy
  - Sample Ratio
- **Model Improvement**
  - Paragraph Reranking
  - Answer Reranking
  - A Discrete Hard EM Approach
  - Syntax-Guided Attentive Network

Other Works in the Past 2 Years

- **Information Retrieval and Semantic Matching**:
  - Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval (EMNLP)

Other Works in the Past 2 Years

- **Information Retrieval and Semantic Matching**:
  - Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval (EMNLP)
  - Document Expansion by Query Prediction (submitted to WWW)

Other Works in the Past 2 Years

- **Information Retrieval and Semantic Matching**:
  - Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval (EMNLP)
  - Document Expansion by Query Prediction (submitted to WWW)
  - Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models (SIGIR best short paper honorable mention)

Other Works in the Past 2 Years

- **Information Retrieval and Semantic Matching**:
  - Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval (EMNLP)
  - Document Expansion by Query Prediction (submitted to WWW)
  - Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models (SIGIR best short paper honorable mention)
  - Incorporating Compositional and Syntactic Structures Improves Semantic Similarity Modeling (EMNLP)

Other Works in the Past 2 Years

- **Information Retrieval and Semantic Matching**:
  - Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval (EMNLP)
  - Document Expansion by Query Prediction (submitted to WWW)
  - Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models (SIGIR best short paper honorable mention)
  - Incorporating Compositional and Syntactic Structures Improves Semantic Similarity Modeling (EMNLP)
  - Bridging the Gap of Relevance Matching and Semantic Matching with Hierarchical CoAttention Network (EMNLP)

Other Works in the Past 2 Years

- **Information Retrieval and Semantic Matching**:
  - Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval (EMNLP)
  - Document Expansion by Query Prediction (submitted to WWW)
  - Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models (SIGIR best short paper honorable mention)
  - Incorporating Compositional and Syntactic Structures Improves Semantic Similarity Modeling (EMNLP)
  - Bridging the Gap of Relevance Matching and Semantic Matching with Hierarchical CoAttention Network (EMNLP)
  - MultiPerspective Relevance Matching with Hierarchical ConvNets for Social Media Search (AAAI)

Other Works in the Past 2 Years

- **Information Retrieval and Semantic Matching**:
  - Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval (EMNLP)
  - Document Expansion by Query Prediction (submitted to WWW)
  - Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models (SIGIR best short paper honorable mention)
  - Incorporating Compositional and Syntactic Structures Improves Semantic Similarity Modeling (EMNLP)
  - Bridging the Gap of Relevance Matching and Semantic Matching with Hierarchical CoAttention Network (EMNLP)
  - MultiPerspective Relevance Matching with Hierarchical ConvNets for Social Media Search (AAAI)
- **Dialogue Analaysis**
  - End-to-End Neural Context Reconstruction in Chinese Dialogue (ACL workshop)

Other Works in the Past 2 Years

- **Information Retrieval and Semantic Matching**:
  - Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval (EMNLP)
  - Document Expansion by Query Prediction (submitted to WWW)
  - Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models (SIGIR best short paper honorable mention)
  - Incorporating Compositional and Syntactic Structures Improves Semantic Similarity Modeling (EMNLP)
  - Bridging the Gap of Relevance Matching and Semantic Matching with Hierarchical CoAttention Network (EMNLP)
  - MultiPerspective Relevance Matching with Hierarchical ConvNets for Social Media Search (AAAI)
- **Dialogue Analaysis**
  - End-to-End Neural Context Reconstruction in Chinese Dialogue (ACL workshop)
  - Detecting Customer Complaint Escalation with Recurrent Neural Networks and Manually-Engineered Features (NAACL)

Q & A

Thanks!