THE UNIVERSITY OF ARIZONA

TOWARDS THE ADVANCEMENT OF OPEN-DOMAIN TEXTUAL QUESTION ANSWERING METHODS

Fan Luo

Nov 17, 2022

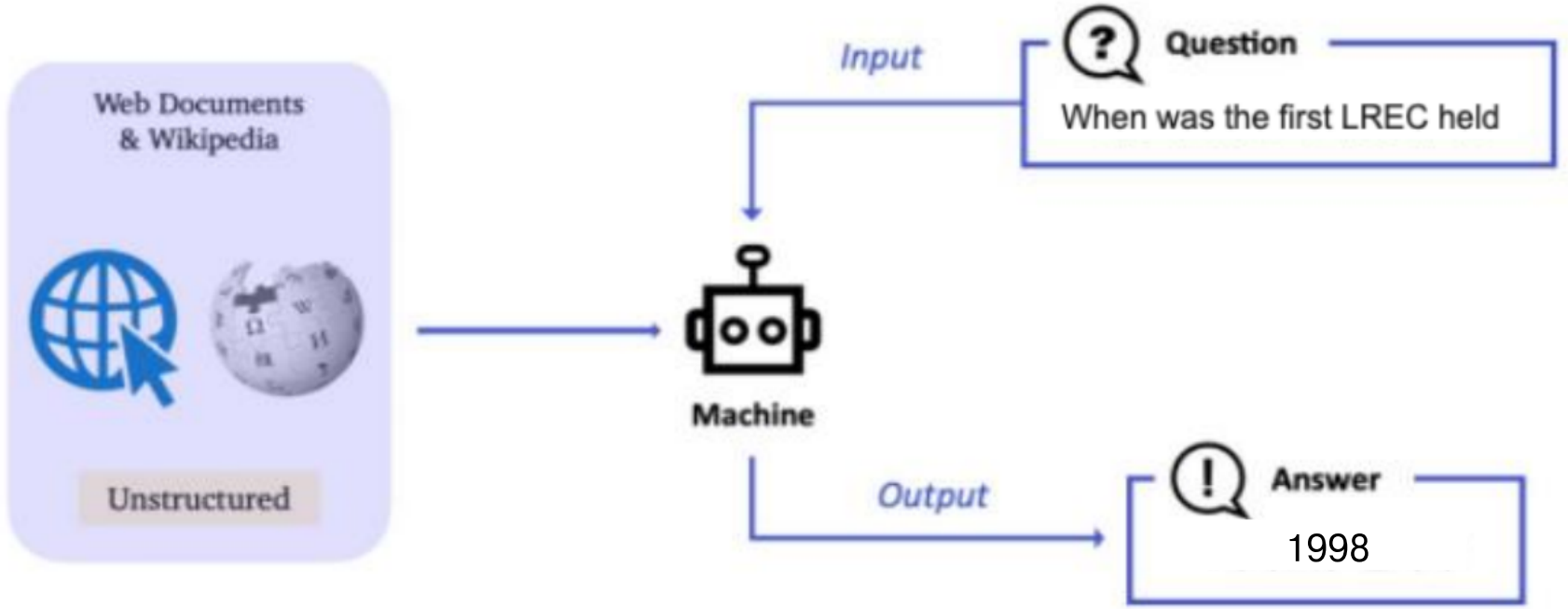
Dissertation Committee:

Mihai Surdeanu, Lila Bozgeyikli, Joshua A Levine, Chicheng Zhang

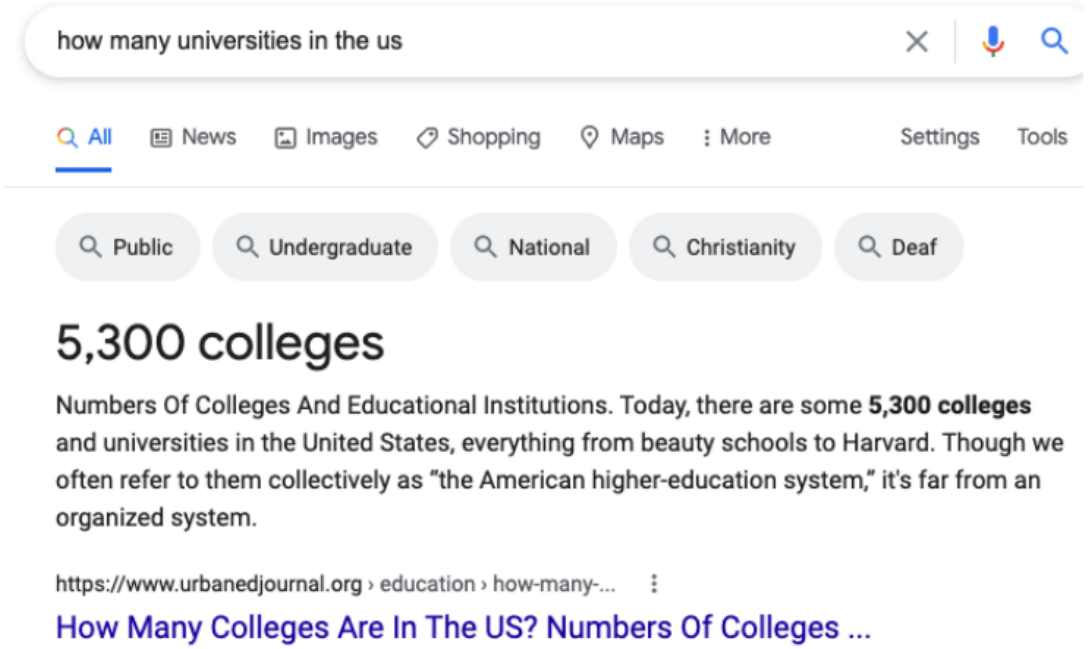
Computer Science

Overview

Open-Domain Question Answering (ODQA)



- Modern question answering applications
  - Search engines evolve to handle question queries
  - Digital assistants address multi-turn QA
  - Business analytics service adopt natural language QA interface



QA categories

Open-domain QA vs. Closed-domain QA

Textual QA vs. Knowledge-based QA vs. Table-based QA

A history of open-domain textual QA

- **Simmons et al. (1964)** was the first to explore answering questions based on matching dependency parses of a question and answer

- **Murax (Kupiec 1993)** aimed to answer questions over an online encyclopedia using IR and shallow linguistic processing

- **The NIST TREC QA track** begun in 1999 first rigorously investigated answering fact questions over a large collection of documents

- **IBM's Jeopardy! System (DeepQA, 2011)** used an ensemble of many methods

- Many neural approaches after 2015… (more later)

https://www.cs.princeton.edu/courses/archive/spring20/cos598C/lectures/lec10-open-qa.pdf

IBM's Watson and Jeopardy! Challenge

IBM Watson defeated two of Jeopardy's greatest champions in 2011

Sample questions:

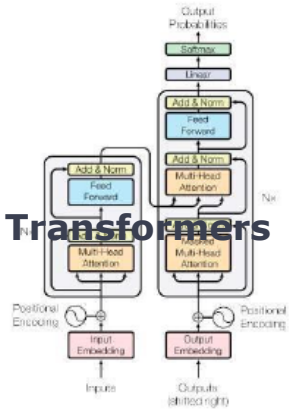
Q: Even a broken one of these on your wall is right twice a day

A: clock. Watson got it correctly.

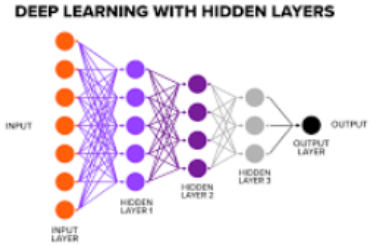
Q: Its largest airport is named for a World War II Hero; its second largest for a World War II Battle

A: Chicago. Watson didn't get it correctly.

Recent Advancements in NLP



Transformers



HUGGING FACE

Open Research Questions

01 | Complex questions

Retrieve and synthesize info from multi-resources

02 | Explainability

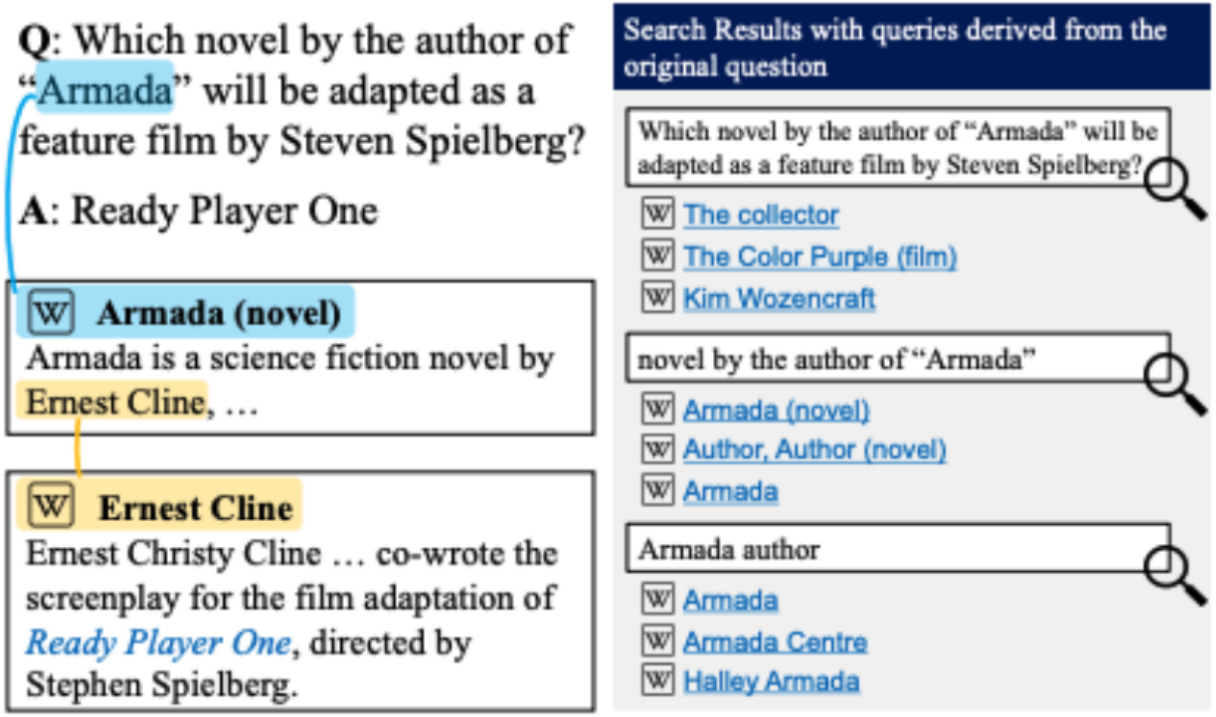
Explain with evidence

03 | Annotation cost

Less human labeling effort

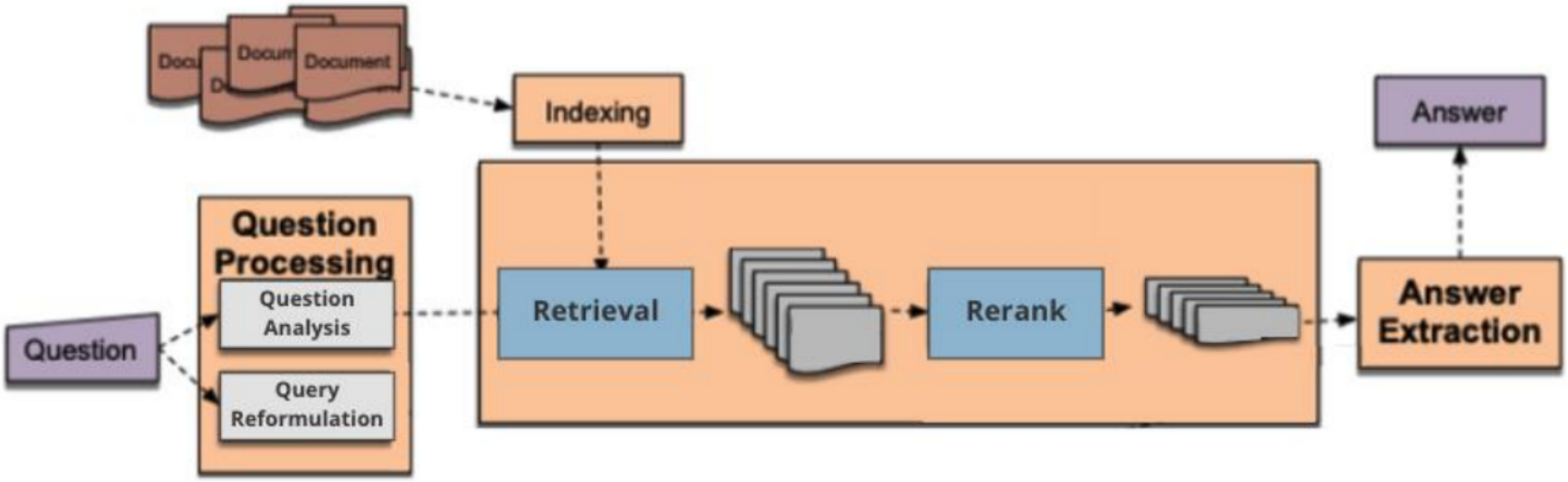
Challenges in Multi-Hop QA

- 'Multi-hops' refers to multiple pieces of information relevant to tackle the multi-hop question.
- Secondary hops are lexically or semantically distant to questions.
- More challenging to provide explainability for the answer prediction for multi-hop QA.



Qi, Peng, et al. "Answering complex open-domain questions through iterative query generation." arXiv preprint arXiv:1910.07000 (2019).

Architecture of Modern Textual QA systems



Question Processor

(1) Analysis and reformulate questions

Retriever-Reranker

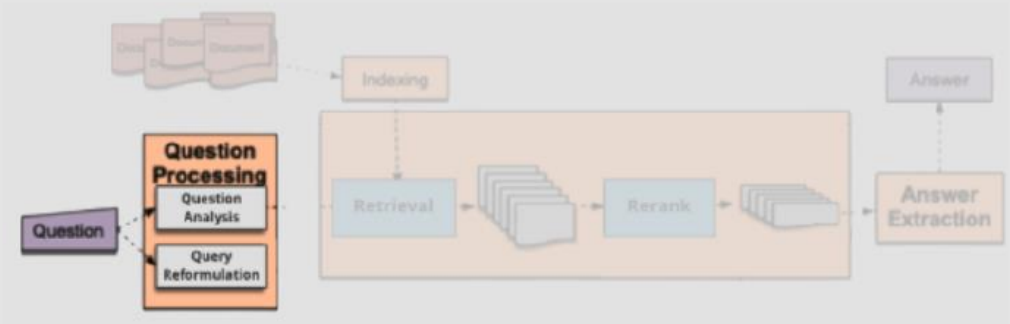
(2) Collect relevant context

Reader

(3) Extract the answer

My works

A STEP towards Interpretable Multi-Hop Reasoning: Bridge Phrase Identification and Query Expansion



- ✓Complex questions
- ✓Explainability
- ✓ Annotation

Bridge Phrase

- The film that has a score composed by Jerry Goldsmith: **Alien**
- Name of the executive producer of **Alien**:

  Ronald Shusett
- Not explicit lexical overlap between the answer sentence and the question
- Application: query expansion

Question: What is the name of the executive producer of the film that has a score composed by Jerry Goldsmith?

Supporting Document1: Alien (soundtrack)
The iconic, avant-garde score to the film "Alien" was composed by Jerry Goldsmith and is considered by some to be one of his best, most visceral scores. . .

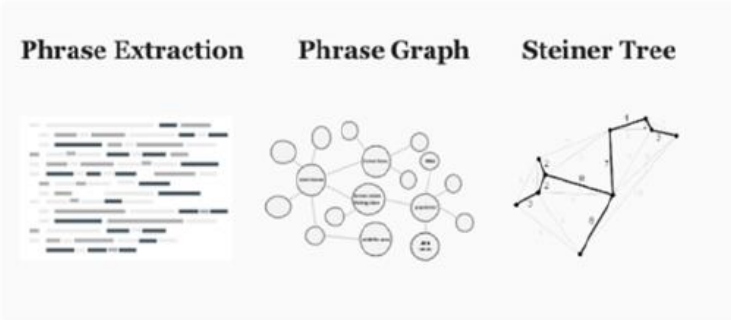
Supporting Document2: Alien (film)
Alien is a 1979 science-fiction horror film ... Dan O'Bannon, drawing upon previous works of science fiction and horror, wrote the screenplay from a story he co-authored with Ronald Shusett ... Shusett was executive producer.
. . .

Bridge Phrases: Alien

Answer: Ronald Shusett

Approach

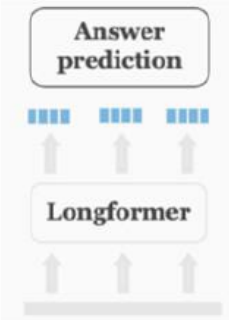
Input



Bridge Phrases Identification

Query Expansion

Evidence Retrieval



Reader

- Graph-based
- Unsupervised
- Modular

Noun Phrase Extraction

Question: [Three Men on a Horse]$_G$ is a [play]$_C$ by a [playwright]$_C$ born in which year?

Supporting Document 1: [***Three Men on a Horse***]$_T$
[Three Men on a Horse]$_G$ is a [play]$_C$ by [George Abbott]$_G$ and [John Cecil Holm]$_E$. . . .

Supporting Document 2: [***George Abbott***]$_T$
[George Francis Abbott]$_E$ ([June 25, 1887 – January 31, 1995]$_E$) was an [American theater producer]$_C$ and [director]$_C$, [playwright]$_C$, [screenwriter]$_C$, and [film director]$_C$ and [producer]$_C$ whose [career]$_C$ spanned [nine decades]$_E$.
. . .

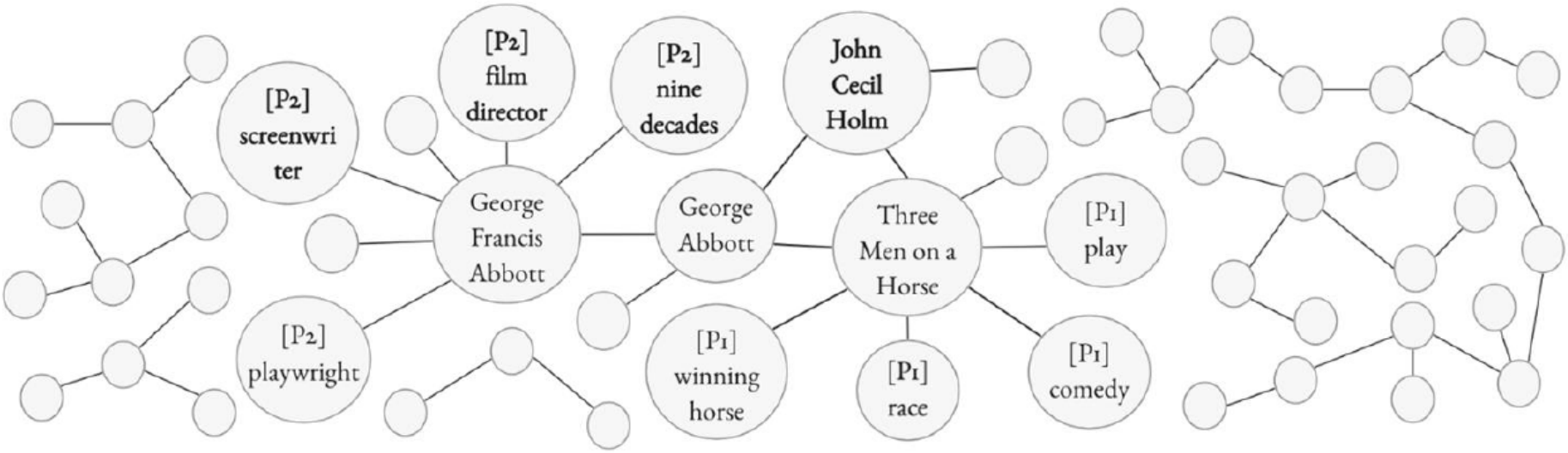
Quotation Extraction (Q)

Phrase Grounding (G)

Named Entity Recognition (E)

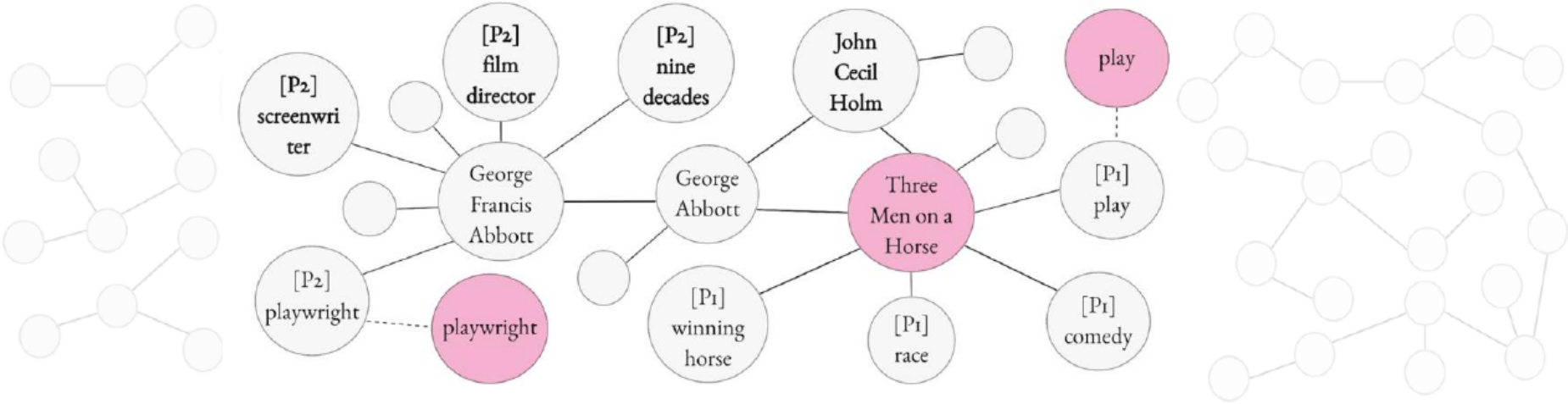
Noun Chunks Extraction (C)

Noun Phrase Graph Construction



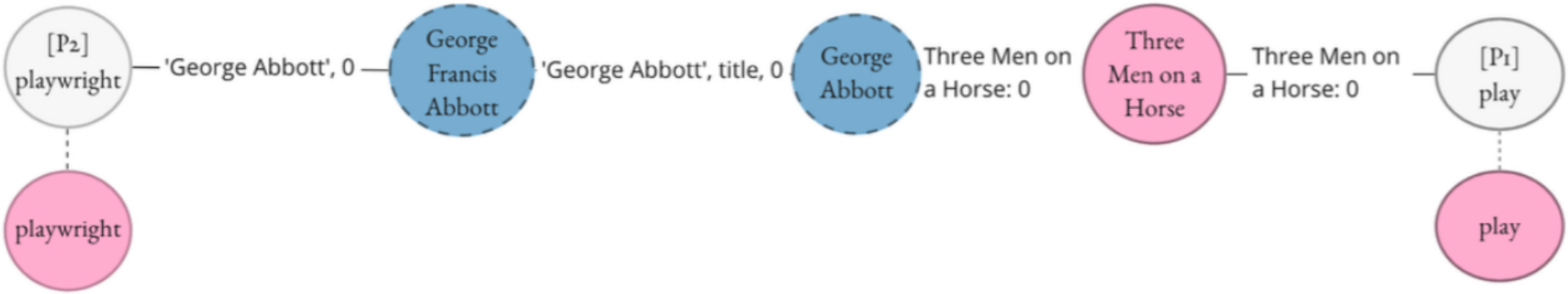
SENT-SENT	TITLE-SENT	TITLE-TITLE	Coreference
George Francis Abbott	Three Men on a Horse	Tomb Raider	Ronald Shusett
[P2] playwriter	[P1] play	2013 video game	Shusett

Question Phrase Identification & Graph Pruning



Question: [Three Men on a Horse]$_G$ is a [play]$_C$ by a [playwright]$_C$ born in which year?

STEP (Steiner Tree Phrases identification)



Algorithm: An approximate solution for the Steiner problem in graphs (Takahashi and others, 1980)

Steiner Tree: Minimum spanning tree of the sub-graph that contains all question phrases

Steiner Points: *'George Francis Abbott'* and *'George Abbott'*

Identified Bridge Phrases

Query Expansion and Retrieval

Question: Three Men on a Horse is a *play* by a *playwright* born in which year?

Supporting Document 1: Three Men on a Horse

Three Men on a Horse is a *play* by George Abbott and John Cecil Holm. . . .

Supporting Document 2: George Abbott

George Francis Abbott (June 25, 1887 – January 31, 1995) was an American theater producer and director, *playwright*, screenwriter, and film director and producer whose career spanned nine decades. **...**

Identified Bridge Phrases: *'George Francis Abbott' and 'George Abbott'*

Query Expansion: Three Men on a Horse is a play by a playwright born in which year, George Abbott, George Francis Abbott

Retriever: BM25 and MSMARCO cross-encoder

Experiments

Evidence Retrieval

Retrieval performance with/o our query expansion strategy

Answer Prediction

Accuracy of predicted answer using the retrieved context with/o our query expansion strategy

Bridge Phrases

Manual evaluation of the accuracy of identified bridge phrases

Post-hoc Explanation

Manual evaluation of the quality of the explanations

Dataset: HotpotQA (Yang et al., 2018) Development Set

- 5,918 bridge-type questions
  - Example: Alice David is the voice of Lara Croft in a video game developed by which company?
- 1,487 comparison-type questions
  - Example: Which American singer and songwriter has a mezzo-soprano vocal range, Tim Armstrong or Tori Amos?

Results: Evidence Retrieval

	Prec@2	Prec@3	Avg Prec	Recall@2	Recall@3	Recall@5	Recall@10	Recall@20
Cross Encoder	0.59	0.47	0.64	0.50	0.59	0.69	0.81	0.92
Cross Encoder w/ STEP	**0.64**	**0.51**	**0.69**	**0.54**	**0.65**	**0.75**	**0.85**	**0.94**
BM25	0.55	0.44	0.60	0.46	0.55	0.65	0.78	0.90
BM25 w/ STEP	0.60	0.49	0.66	0.51	0.62	0.72	0.84	0.93

When STEP is coupled with a retriever:

- BM25: traditional information retrieval model
- MSMARCO cross-encoder: a transformer-based neural dense retrieval model

evidence retrieval performance (evaluated against annotated supporting facts) increase

Results: Answer Prediction

		EM	F1
Baseline	Random 5	0.11	0.18
	Random 10	0.19	0.29
Ceiling	SF only	**0.56**	**0.77**
	Oracle Top 5	0.49	0.67
	Oracle Top 10	0.52	0.71

	EM	F1	
Question-only Top 5	0.33	0.48	Retrieved w/ Original question
Question-only Top 10	0.41	0.57	
STEP Top 5	0.4	0.55	Retrieved w/ Query expansion
STEP Top 10	0.45	0.62	

Reader: Longformer Fine-tuned with HotpotQA training data

Input: concatenating the question and context sentences

[CLS][Q] Question [/Q][SEP] [T] title_1 [/T] sent_{11} [/S] sent_{12} [/S] . . .[SEP] [T] title_2 [/T] sent_{21} [/S] sent_{12} [/S] . . .

Context sentences:

- **Random:** a set of k sentences randomly selected
- **Question-only:** top ranked sentences with/o query expansion (i.e., original question)
- **SF only:** the gold supporting sentences
- **Oracle:** top ranked sentences with query expansion, using oracle bridge phrases extracted directly from the ground-truth supporting sentences
- **STEP:** top ranked sentences with query expansion, using identified bridge phrases

Results: Bridge Phrases

	Question	Answer	Bridge Phrases (STEP)	Annotations
(1)	Ralph Hefferline was a psychology professor at a university that is located in what city?	New York City	Columbia University	(Correct, Correct)
(2)	According to the 2001 census, what was the population of the city in which Kirton End is located?	35,124	Boston	(Correct, Partial)
(3)	This Celtic ruler who was born in AD 43 ruled southeastern Britain prior to conquest by which empire?	Roman	Roman conquest of Britain	(Correct, Incorrect)

Manually evaluate the quality of identified bridge phrases

100 randomly selected questions

2 human annotators

3 annotations: correct, partial, incorrect

Average accuracy: 76.3%

Kappa agreement: 46%

Post-hoc Explanations

Question: In which city are the headquarters of the American research and scientific development company where Ravi Sethi worked as computer scientist located?

Answer: Murray Hill

Top ranked evidence candidates:

1. ***Bell Labs***: Its headquarters are located in Murray Hill, New Jersey, in addition to other laboratories around the rest of the United States and in other countries.

2. ***Ravi Sethi***: Ravi Sethi (born 1947) is an Indian computer scientist retired from Bell Labs and president of Avaya Labs Research.

3. ***Ravi Sethi***: He also serves as a member of the National Science Foundation's Computer and Information Science and Engineering (CISE) Advisory Committee.

4. ***Bell Labs***: Nokia Bell Labs (formerly named AT&T Bell Laboratories, Bell Telephone Laboratories and Bell Labs) is an American research and scientific development company, owned by Finnish company Nokia.

5. ***Ravi Sethi***: He is best known as one of three authors of the classic computer science textbook "", also known as the "Dragon Book".

1, 2, and 4 are the gold supporting facts

Nodes in the Steiner Tree:

Bell Labs, headquarters, Ravi Sethi, computer scientist, Avaya Labs Research, American research and scientific development company

Steiner Points: Bell Labs, Avaya Labs Research

Query expansion: In Murray Hill city are the headquarters of the American research and scientific development company where Ravi Sethi worked as computer scientist located, Bell Labs, Avaya Labs Research

Results: Post-hoc Explanations

Manually evaluate

44.5 (89%)

Quality of explanations

48 (96%)

Accuracy of post-hoc bridge phrases

50 random sampled questions

Top 10 candidate evidences

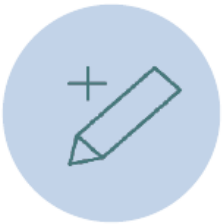
Takeaways

Bridge phrase identification

We introduce a graph-based strategy for the identification of bridge phrases for multi-hop QA;

Query expansion

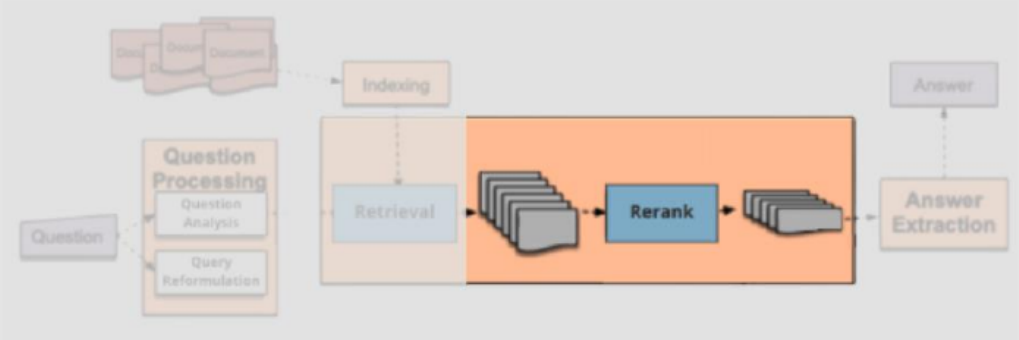
Identified bridge phrases can be used to expand the query used for improving evidence retrieval and answer extraction;



Post-hoc explanation

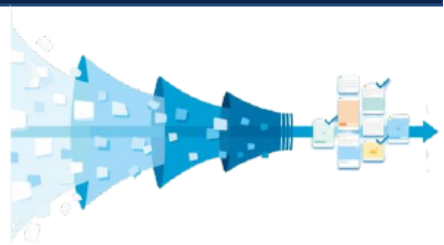
Post-hoc explanations can be made available to interpret answers provided.

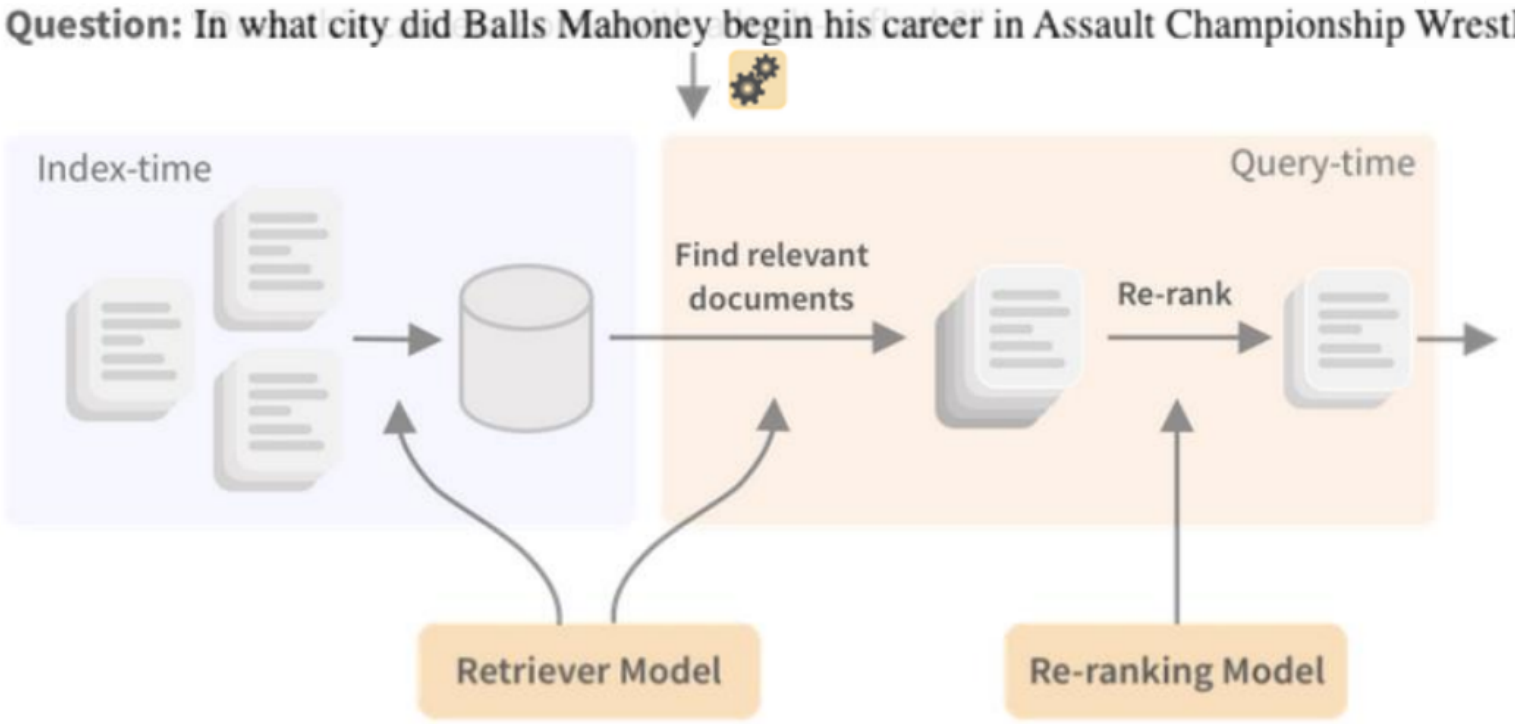
Divide & Conquer for Entailment-aware Multi-hop Evidence Retrieval



- ✓Complex questions
- ✓Explainability
- ✓ Annotation

Evidence **Retrieval and Rerank**





https://blog.griddynamics.com/question-answering-system-using-bert/

Evidence Ranking **Subtasks**

In this work, we propose to capture textual entailment in parallel with the semantic equivalence **with separate models**, which produce different and potentially conflicting rankings.

The goal is to combine them to figure out an **aggregated ranking** that promote gold evidence sentences to the top of the list.

Question: What nationality was James Henry Miller's wife?
Answer: American

Supporting Evidences

Ewan MacColl
(1) James Henry Miller (25 January 1915 – 22 October 1989), better known by James Henry Miller stage name Ewan MacColl, was an English folk singer, songwriter, communist, labour activist, actor, poet, playwright and record producer .

Peggy Seeger
(2) Margaret "Peggy" Seeger (born June 17 , 1935) is an American folksinger.
(3) She is also well known in Britain, where she has lived for more than 30 years, and was married to the singer and songwriter Ewan MacColl until his death in 1989.

Semantic Equality
James Henry Miller (Q) ≈ James Henry Miller (25 January 1915 – 22 October 1989), better known by James Henry Miller stage name Ewan MacColl (1)

Textual Entailment
American (2) ⊢ nationality (Q)
was married to (3) ⊢ wife (Q)

Base Models

Three off-the-shelf base models to capture the diverse relevance signals.

Sparse model

BM25

Statistical model relying on lexical overlap

Dense model

MSMARCO CE

transformers pre-trained for semantic search

Dense model

QNLI CE

transformers pre-trained for question-answer entailment

CE (Cross-Encoder): the standard BERT design that benefits from all-to-all attention across tokens in the input sequence.

MS MARCO: a large scale corpus consists of about 500k real search queries with 1000 most relevant passages (Bajaj et al., 2016)

QNLI: Question Natural Language Inference (QNLI) dataset introduced by GLUE Benchmark (Wang et al., 2018)

Base Models **Comparison**

14% questions with at least one evidence sentence is ranked within top-3 by BM25, but ranked beyond top-3 by MSMARCO CE and QNLI CE;

35% questions with at least one evidence sentence ranked within top-3 by QNLI CE, but ranked beyond top-3 by MSMARCO CE.

BM25	MSMARCO CE	QNLI CE	% Ques (k=3)	% Ques (k=5)
✓	✗	✗	14	10
✗	✗	✓	25	22
✗	✓	✗	20	16
✗	✓		33	30
✗		✓	38	35
	✓	✗	64	62
	✗	✓	35	33
✗	✗	✗	44	29

Percentage of questions with at least one evidence are ranked within top-k by a base model or not.

'✓' indicates that an evidence is ranked within top-k by the base model, while '✗' indicates that the evidence is ranked beyond top-k.

The three base models independently capture diverse relevance signals and are complementing each other

Ensemble Baselines

Techniques to combine the results from base models

Average ranking (AR)

Sents	R_{BM25}	$R_{MSMARCO}$	R_{QNLI}	Sum(R)	AR
S_1	1	1	5	7	1
S_2	4	3	2	9	3
S_3	3	4	6	13	5
S_4	5	6	3	14	6
S_5	6	5	1	12	4
S_6	2	2	4	8	2

AR simply sums all the ranks for each sentence, and re-rank all the sentences according to the **summation of ranks.**

Similarity Combination (SimCom)

$$QER_{q,s_j} = \begin{cases} \dfrac{\eta(\mathcal{BM}25_{q,s_j}) + \alpha \cdot \eta(STS_{q,s_j}) + \beta \cdot \eta(IS_{q,s_j})}{3} & \text{if } \mathcal{BM}25_{q,s_j} > 0 \\ \dfrac{\alpha \cdot \eta(STS_{q,s_j}) + \beta \cdot \eta(IS_{q,s_j})}{2} & \text{Otherwise} \end{cases}$$

Semantic Textual Similarity (STS) and Inference Similarity (IS) are scores from MSMARCO CE and QNLI CE.

SimCom calculates hybrid relevance scores though **a linear combination of scores** from base models

Entailment-Aware Ranking (EAR)

Goal

Combine **complementary relevance signals** captured by base models to retrieve candidate evidences for multi-hop questions.

BM25: $\{S_{a1}, S_{a2}, S_{a3}, S_{a4}, S_{a5}, S_{a6}, \ldots\}$

MSMARCO CE: $\{S_{a4}, S_{a3}, S_{a1}, S_{a6}, S_{a2}, \ldots\}$

QNLI CE: $\{S_{b1}, S_{b2}, S_{b3}, S_{b4}, S_{b5}, S_{b6}, \ldots\}$

Top ranked by semantic equivalence A = $\{S_{a1}, S_{a2}, S_{a3}, S_{a4}\}$

Top ranked by textual entailment B = $\{S_{b1}, S_{b2}, S_{b3}\}$

Pairs we consider are Cartesian product of two sets

$$\text{Pairs} = A \text{ Ч } B = \{(a, b) \mid a \in A \wedge b \in B\}$$

Score pairs against question with a ranker

$$(q,\ a \,||\, b)$$

Top scored sentence pair Sai, Sbj form a compositional relevant context covering both signals

Re-rank the rest against q || a || b

$$(q \,||\, a \,||\, b, S_i)$$

Idea

Jointly consider **pairs of top-ranked candidate evidence sentences** by base models with respect to semantic equivalence and textual entailment, respectively.

EARnest

Evidences for a multi-hop question should be intuitively related, and often logically connected via a shared named entity.

EAR

Entailment-Aware Retrieval

NEST

Named Entity Similarity Term

Question: Bordan Tkachuk was the *CEO* of a company that provides what sort of products?
Answer: IT products and services

Evidences:
1. ***Bordan Tkachuk***: Bordan Tkachuk is a British business executive, the former *CEO* of Viglen, also known from his appearances on the BBC-produced British version of "The Apprentice," interviewing for his boss Lord Sugar.
2. ***Viglen***: Viglen Ltd provides IT products and services, including storage systems, servers, workstations and data/voice communications equipment and services.

When scoring sentence pairs:

$$QER_{Earnest} = (1 + NEST) * Sim(q, s_i \| s_j)$$

Sim(): the scoring function of the reranker

NSET is a **binary switch**. That is, if the two sentences share one or more named entity, the promotion mechanism is activated, because they are more likely to be connected to form a coherent context.

Evidence Ranking **Results**

Models	P@3	P@5	MAP	R@3	R@5	R@10
Base models						
BM25	0.43	0.31	0.59	0.54	0.65	0.78
MSmarco	0.47	0.33	0.64	0.59	0.69	0.81
QNLI	0.33	0.25	0.46	0.43	0.52	0.65
Ensemble Baselines						
AR	0.43	0.31	0.61	0.55	0.66	0.83
SimCom	0.5	0.36	0.68	0.63	0.74	0.86
Our Approach						
EAR	0.53	0.36	0.71	0.66	0.76	0.86
EARnest	**0.55**	**0.38**	0.74	**0.7**	**0.78**	**0.87**

highest among the base models

ignoring other relevance signals

better than individual base models

ignoring interactions between the relevance signals

best model

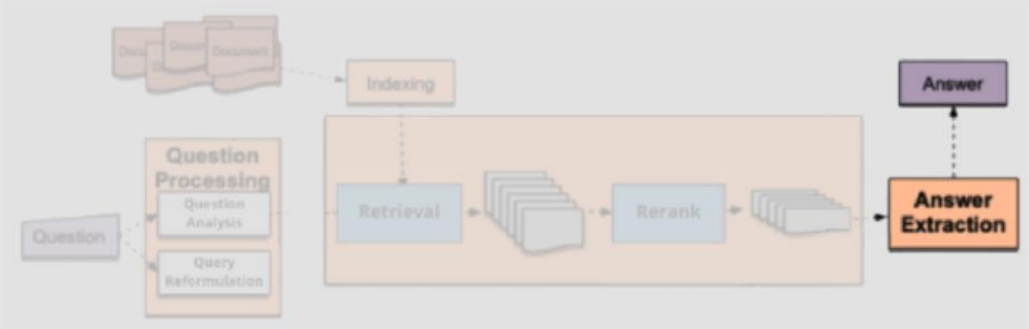
+ 10%

MAP

P@n and R@n fail to take into account the positions of the relevant sentences among the top n. **MAP**(Mean average precision) is a relative more importance metric to exam.

SimCom uses $\alpha = 3$ and $\beta = 1$, according to the grid search results on 10% of the full dataset.

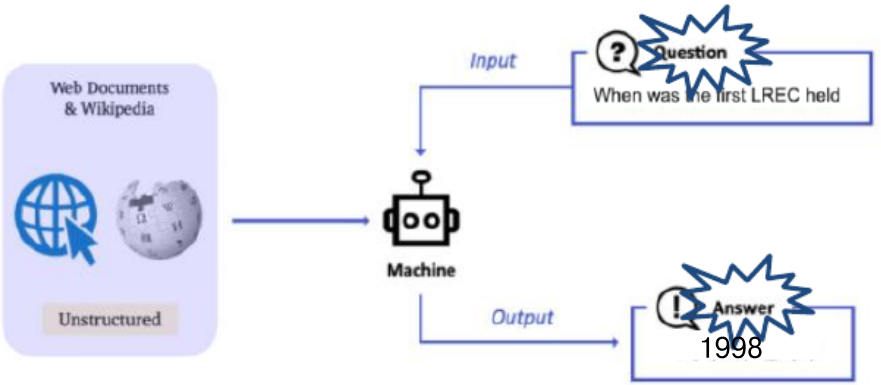
Learning Strategies for Question Answering with Fewer Annotations

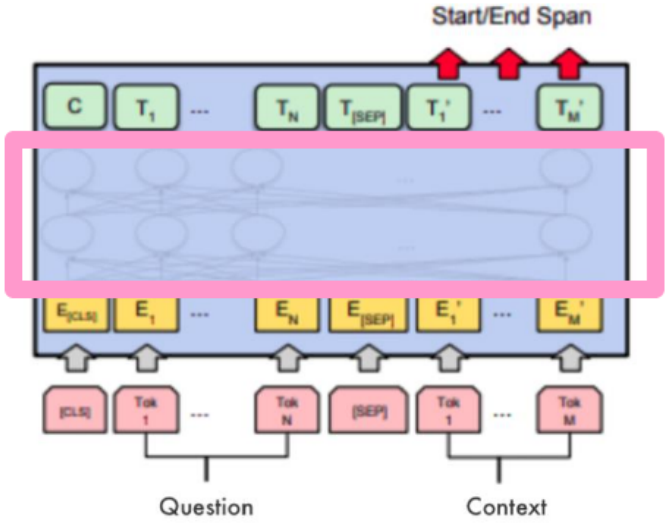


✓ Annotation

Answer extraction with a deep reader model

- A **deep neural network** to extract the answer to a question from the given context.
- **Challenges**: Suffer from "data hunger" and low robustness issues.

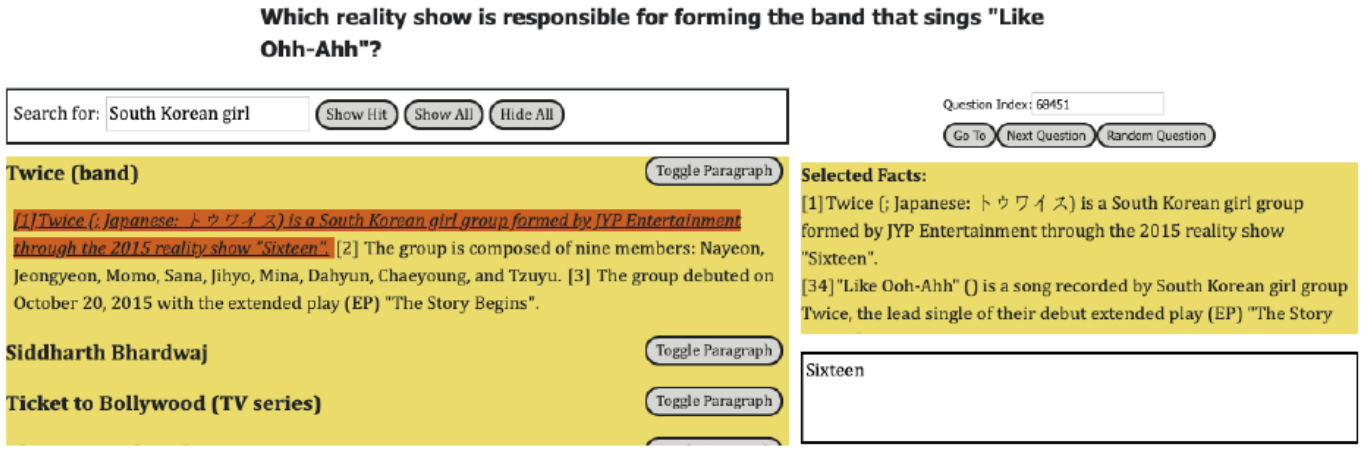




QA Dataset Annotation

- Costly
- Noisy

- Intensive manual labor
- Tedious and time consuming
- Low Agreement

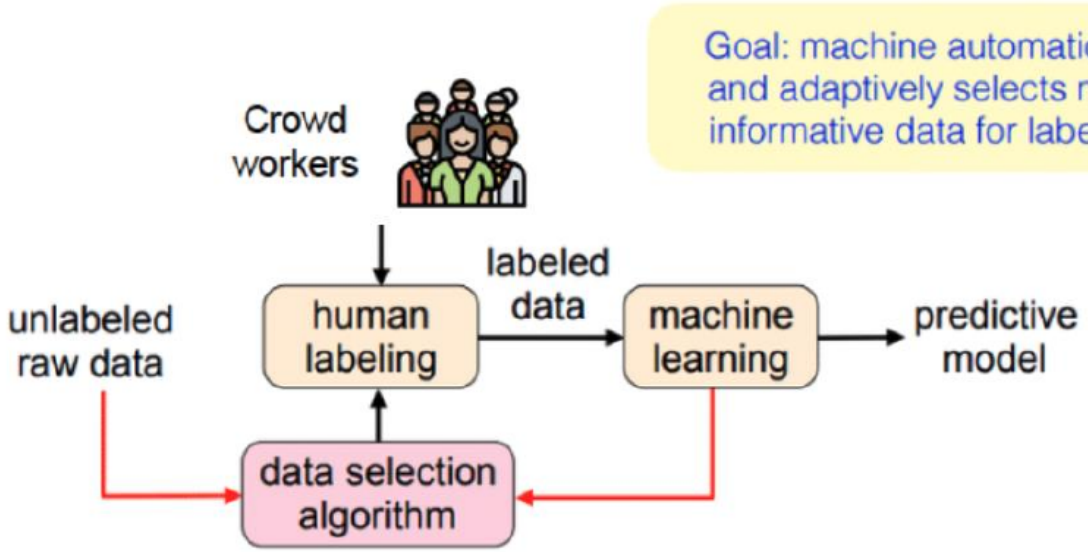


Objectives

Less Annotation

More Robust

Active Learning

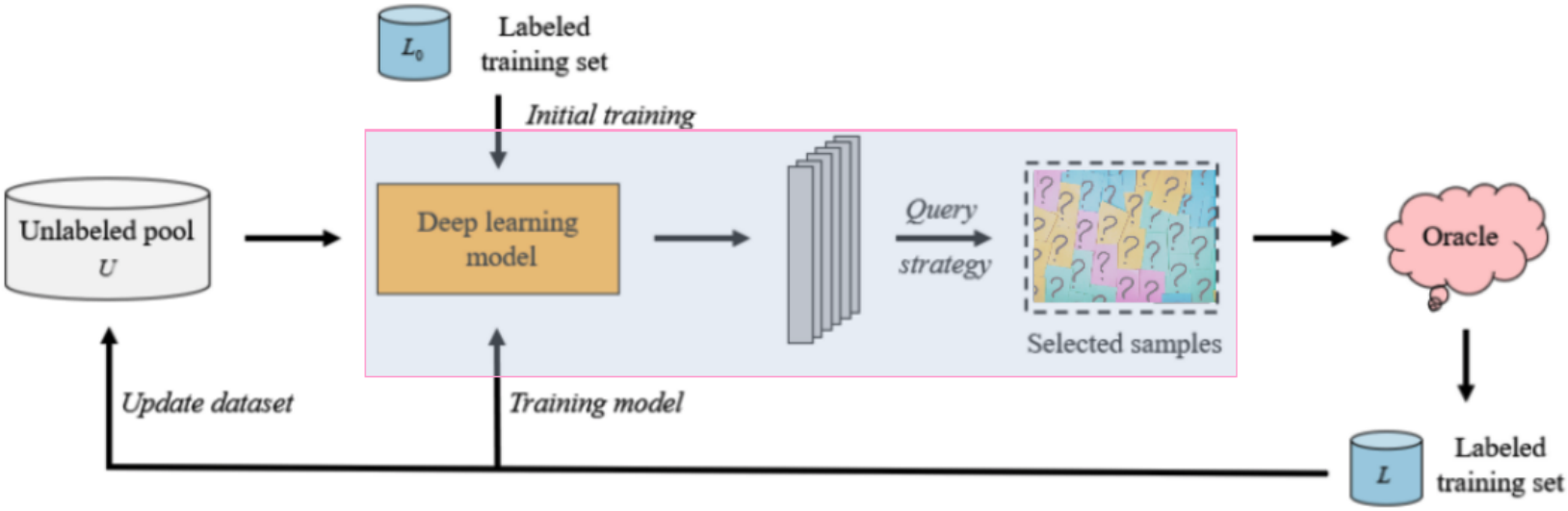


Query the most informative instances

Make much less annotation request

Output a relative good Model

DeepAL for QA task

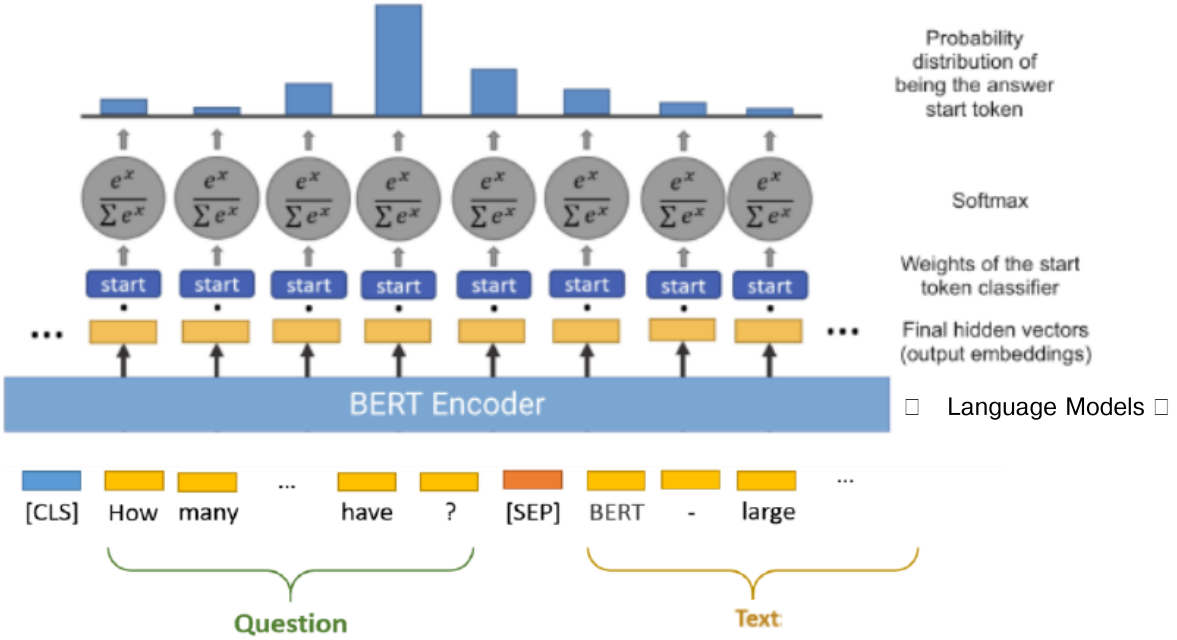


DeepAL, the combination of Deep Learning (DL) and Active Learning (AL), considers the complementary advantages of the two methods to achieve better results.

- DL achieved state-of-the-art results in QA tasks, but is limited by the high cost of labeling,
- AL maximize the value of labeling a small set of examples.

Deep Reader with BERT

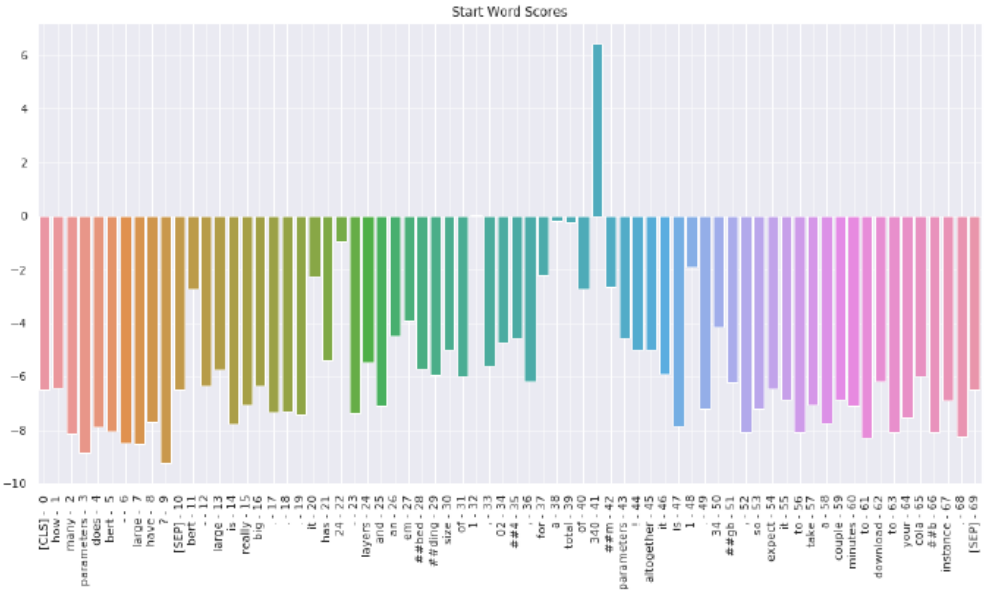
- Most modern textual QA systems has a deep reader model performing reading comprehension (RC) to extract an answer from given documents.
- A question-answering head is applied on top of the BERT model to produce probabilities over the tokens in the documents for being answer start and end token.

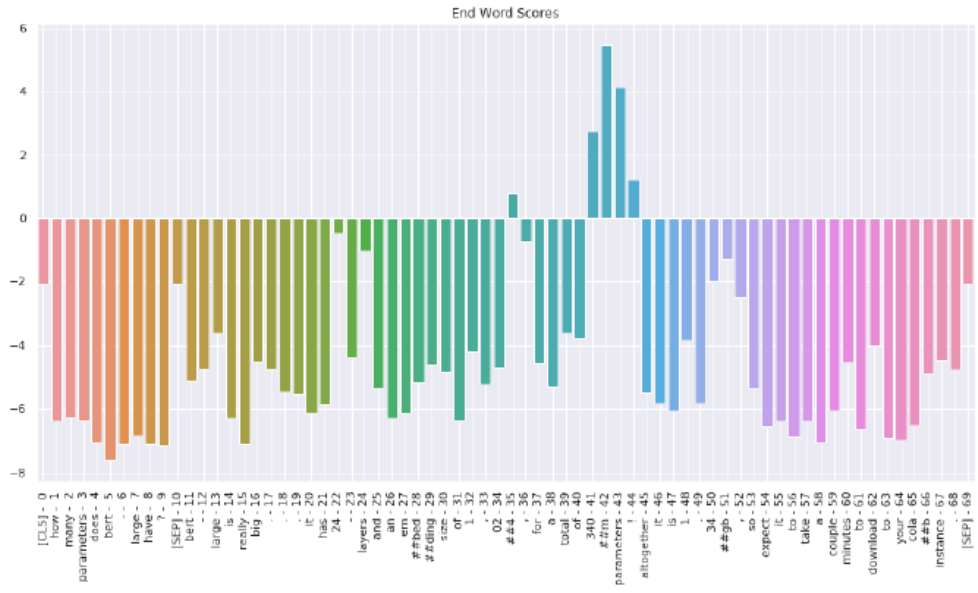


Question: How many parameters does BERT-large have?

Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

```
question = "How many parameters does BERT-large have?"
Context = "BERT-large is really big... it has 24-layers and an embedding size of 1,024,
for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple
minutes to download to your Colab instance."
```



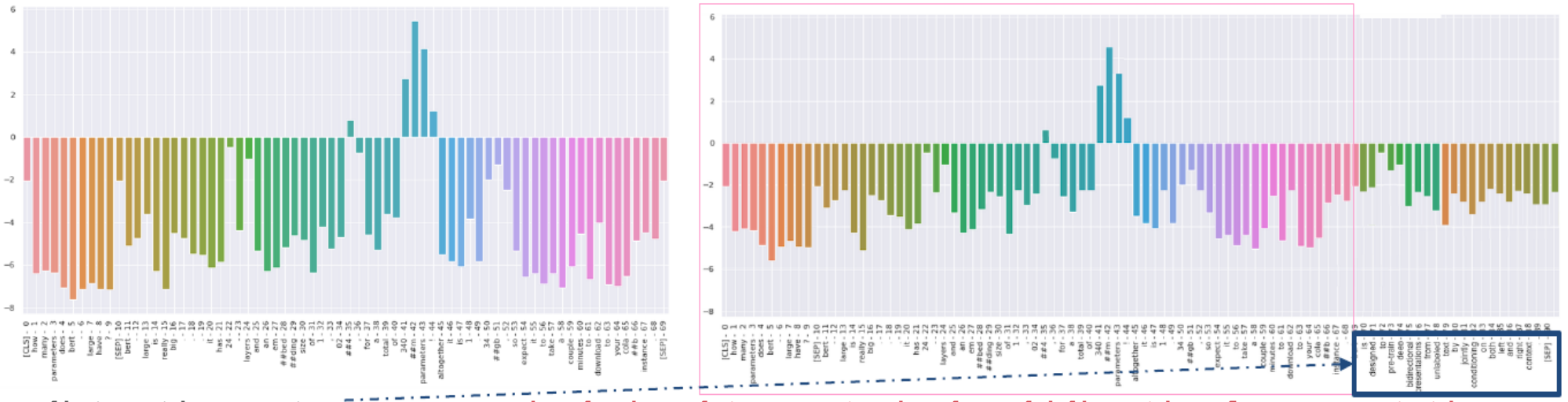


Answer: "340 ##m"

https://mccormickml.com/2020/03/10/question-answering-with-a-fine-tuned-BERT/

Key Idea

We hypothesize that a robust model should produce similar probability distributions on the **original context's part** <u>after perturbation of the context with an additional distracting sentence.</u>



```
distracting sentence = "BERT is designed to pre-train deep bidirectional representations
from unlabeled text by jointly conditioning on both left and right context."
```

Our approach: Perturbation-based AL

1 Creating perturbation for unlabeled candidates

The first step of our PAL acquisition strategy **finds the distracting sentence** from the context of the most similar labeled questions using the **embeddings** of the fine-tuned model.

A perturbed instance is generated by appending the distractor sentence to the original context.

2 Scoring the robustness to perturbation

We compute the **Kullback-Leibler divergences** in the model predictive probabilities between each candidate unlabeled question and its corresponding perturbed question as the **perturbation sensitive score.**

3 Select candidates to query

We then **rank** the unlabeled candidates according to their perturbation sensitive scores.

PAL **selects** top n unlabeled questions that have the highest score to improve the robustness of the current model.

Experiments

We experiment with the pre-trained BERT-BASE model in combination with AL query strategies for selecting the next informative question examples to be evolve the QA reader model.

The model predicts the start and the end position of the answer, and calculates the cross-entropy loss.

In each iteration, we continue fine-tuning the model on the newly labelled 10% of the rest of the unlabeled dataset selected by the active learning acquisition functions.

We used SQuAD as the benchmark dataset for the QA answer extraction task.

Algorithm : Active Learning Approach

Input: Unlabeled data pool $\mathcal{U}$, AL acquisition function $\phi(\cdot,\cdot,\cdot)$, AL selected unlabeled samples $\mathcal{D}_{(t)}al$

while $|\mathcal{D}_{(t)}u| > 0$ **do**

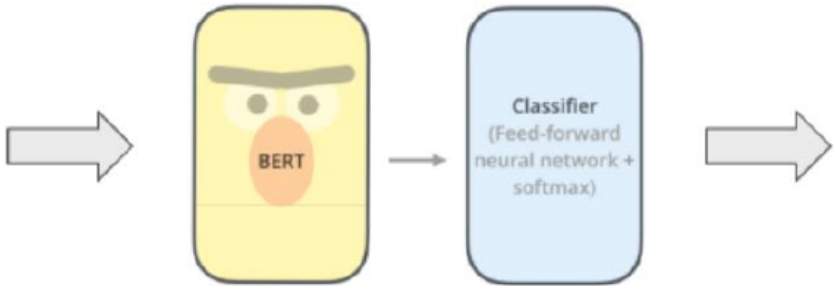
  $M_t \leftarrow$ Continue fine-tuning M_{t-1} with $\mathcal{D}_{(t)}al$;

  $\mathcal{D}_{(t)}al \longleftarrow \arg\max_{*x \in \mathcal{D}_{(t)}u} \phi(M_t, x, 10\%)$;

  $\mathcal{D}_{(t+1)}l \longleftarrow \mathcal{D}_{(t)}l \cup label(\mathcal{D}_{(t)}al)$;

  $\mathcal{D}_{(t+1)}u \longleftarrow \mathcal{D}_{(t)}u \setminus \mathcal{D}_{(t)}al$

end



Stanford Question Answering Dataset (SQuAD)

- 107,785 question-answer pairs on 536 articles.
- The text passages are taken from Wikipedia across a wide range of topics

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

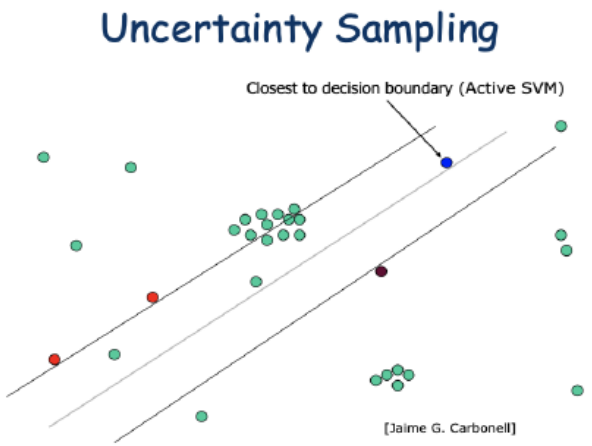
What causes precipitation to fall?
gravity

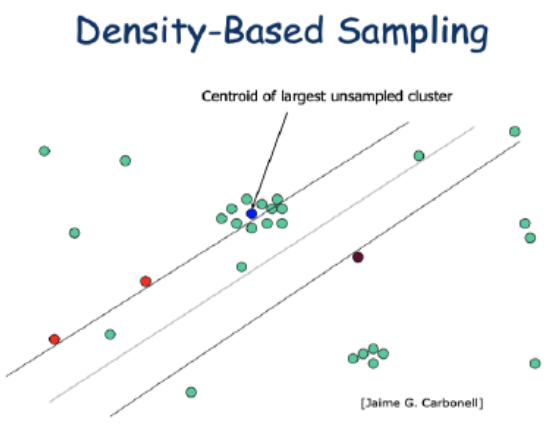
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

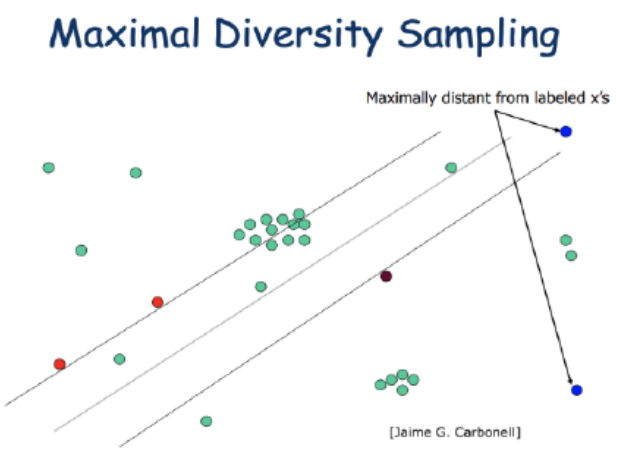
Where do water droplets collide with ice crystals to form precipitation?
within a cloud

"SQuAD: 100,000+ Questions for Machine Comprehension of Text"

Common Active Learning Strategies







- Uncertainty: select the unlabeled data samples with least confidence (largest uncertainty), measured based on the output predictions.
- Density/Clustering-based: finds representative data samples by clustering data in the embedding space and selected ones close to centroids.
- Maximal Diversity: query the unlabeled ones that have maximally distant from the labeled ones

Results

Fine-tuning BERT-base model with various AL acquisition strategies for the QA task.

The F1 scores are evaluated at every n-th training step (with batch size of 12) on the SQuAD dataset.

AL Strategies	200	300	400	500	600	700	800	AUC
Confidence	52.4	66.5	71.7	**74.8**	**77.2**	**77.5**	79	71.3
Clustering	53.6	67.1	68.4	73.6	76.3	76.5	77.7	70.5
Diversity	50.4	65.7	71.4	71.6	75.6	74.7	78.2	70
PAL	**57.7**	**70.1**	**72.6**	74.1	76.2	78.5	**79.9**	**72.7**

- In general, uncertainty-based strategy outperforms the two other common sampling strategies as it always searches for the “valuable” samples around the current decision boundary.
- Clustering sampling strategy performs better when the number of labeled samples is very small, while uncertainty-based criterion usually overtakes the clustering strategy afterwards.
- Our PAL acquisition method, utilize both the input feature and model’s output predictions to select the most informative instances.

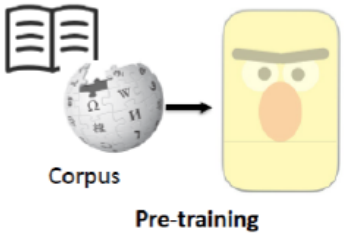
Future works

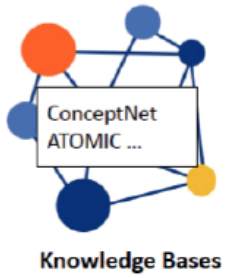
Knowledge Base as external resources

- KBQA (Knowledge-Base Question Answering) uses structured knowledge graphs as the knowledge sources.
  - Pros: High precision
  - Cons: Low coverage, expensive to obtain an extensive and high-quality KB

- TextQA (often referred to as ODQA) leverages text (e.g. Wikipedia articles).
  - Pros: Vast amount of data; can readily use SoTA transformer models
  - Cons: Neglects valuable knowledge sources such as KBs and tables.

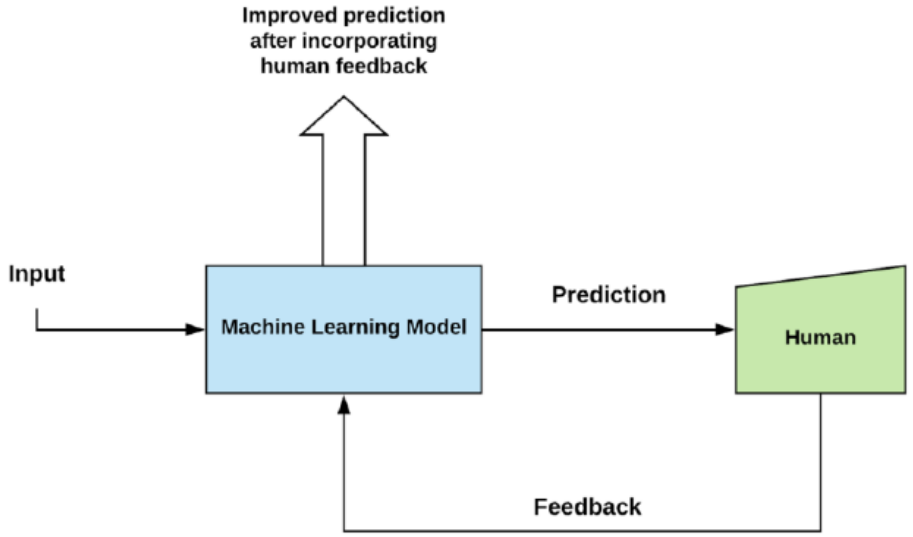
Unifying KBQA and TextQA has proven challenging





Human-in-the-loop Interactive learning

Can human supervision and intervention in the learning process of the model help it learn faster and make better predictions and explanations?



https://hub.packtpub.com/what-is-interactive-machine-learning/

- Besides the answers as direct supervision, would extra information (feedback) provided by humans provide rich guidance to the model? (User input: corrections, rankings, or evaluations)
- How to incorporate the user feedback into an existing QA model?

Combination of active learning and self-learning

- **Self-learning**

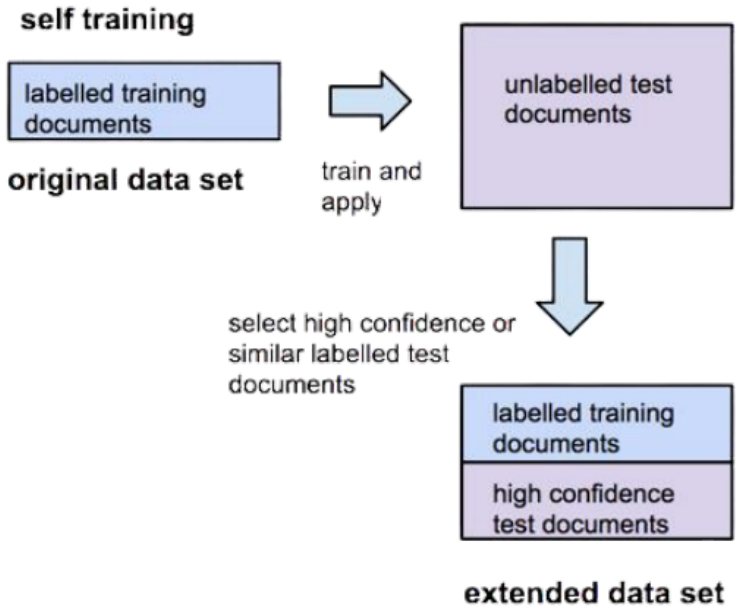
  Discovers highly reliable instances based on its own predictions to teach itself.

- **Active Learning**

  Select the most informative instances.

- **Hybrid**

  Update the model with the most informative and highly reliable instances.



More Challenging QA Tasks

- Conversational intelligence supported by QA
  - No longer an independent task
  - Integrated naturally in a conversational system

- Multi-modal interaction
  - Visual question answering
  - Virtual tour guide

Numerical Reasoning over Text

- A challenging subtask in Question Answering (QA)
- Discrete Reasoning Over Text (DROP[1]) Dataset

(Question)

"How many yards did Kasay kick in total?"

(Passage)

"John Kasay hitting a 45-yard field goal … with Kasay again hitting a 49-yard field goal…"

(Answer)
94

An example of the DROP data instance

Xih, Wen-tau, and Hao Ma. "Question answering with knowledge base, Web and beyond."

QUESTIONS & DISCUSSION

Acknowledgements

I would like to express my heartfelt gratitude to the following people for helping me realize my dream (in no particular order).

- My advisors.
- My committee.
- My colleagues.
- Most important of all, my family!

Thank You!

References

- Abdalla, Muhammad Anwar, and Sameh Basha. Active Learning on Graph Neural Network for Enzymes Classification. Diss. Cairo University, 2021.
- Yih, Wen-tau, and Hao Ma. "Question answering with knowledge base, Web and beyond." Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. 2016.
- Abbasiantaeb, Z. and S. Momtazi (2021). Text-based question answering from information retrieval and deep neural network perspectives: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(6), p. e1412.
- Allam, A. M. N. and M. H. Haggag (2012). The question answering systems: A survey. International Journal of Research and Reviews in Information Sciences (IJRRIS), 2(3).
- Beltagy, I., M. E. Peters, and A. Cohan (2020). Longformer: The Long-Document Transformer.
- Chen, D., A. Fisch, J. Weston, and A. Bordes (2017). Reading wikipedia to answer opendomain questions. arXiv preprint arXiv:1704.00051
- Clark, C. and M. Gardner (2017). Simple and effective multi-paragraph reading comprehension. arXiv preprint arXiv:1710.10723.
- Dasgupta, S. (2011). Two faces of active learning. Theoretical computer science, 412(19), pp. 1767–1781.
- De Cao, N., W. Aziz, and I. Titov (2018). Question answering by reasoning across documents with graph convolutional networks. arXiv preprint arXiv:1808.09920.
- Esposito, M., E. Damiano, A. Minutolo, G. De Pietro, and H. Fujita (2020). Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. Information Sciences, 514, pp. 88–105.
- Fu, R., H. Wang, X. Zhang, J. Zhou, and Y. Yan (2021). Decomposing complex questions makes multi-hop QA easier and more interpretable. arXiv preprint arXiv:2110.13472.
- Fu, Y., X. Zhu, and B. Li (2013). A survey on instance selection for active learning. Knowledge and information systems, 35(2), pp. 249–283.
- Guo, L., X. Su, L. Zhang, G. Huang, X. Gao, and Z. Ding (2018). Query expansion based on semantic related network. In Pacific Rim International Conference on Artificial Intelligence, pp. 19–28. Springer.
- Guu, K., K. Lee, Z. Tung, P. Pasupat, and M. Chang (2020a). Retrieval augmented language model pre-training. In International Conference on Machine Learning, pp. 3929–3938. PMLR.

Backup Slides

RELATED WORK

Question Decomposition

- (Min et al. 2019) proposed DecompRC, a system that learns to break compositional multi-hop questions into simpler, singlehop sub-questions.

- (Jiang and Bansal 2019) designed four types of language reasoning modules, and proposed a controller RNN which decomposes the multi-hop question into multiple single-hop sub-questions, and dynamically infers a series of reasoning modules.

Multi-step (iterative) retrievers

- (Feldman and El-Yaniv 2019)
  - A joint vector representation of both a question and a paragraph.
  - In each retrieval iteration, reformulate the search vector

- GOLDEN Retriever introduced by (Qi et al. 2019)
  - Generates queries given the question and available context for two steps to search documents for HotpotQA full wiki.

- (Asai et al. 2019)
  - Iteratively retrieve a subsequent passage in the reasoning chain with RNN, until the end-of-evidence symbol is selected.
  - Beam search outputs the top reasoning paths with the highest scores and passes them to the reader model.

Graph-based models

- Recent studies build **entity graphs** from multiple paragraphs, and apply graph neural networks to conduct reasoning across documents over the graphs (De Cao, Aziz, and Titov 2019; Xiao et al. 2019).
- DFGN(Qiu et al. 2019) also constructed an entity graph, and predicted a **dynamic mask** to select a subgraph, so that in each reasoning step irrelevant entities are softly masked out.
- CogQA (Ding et al. 2019) iteratively extracted entities and answer candidate spans for each hop and organized them as a **cognitive graph**.

Learning with Limited Annotations

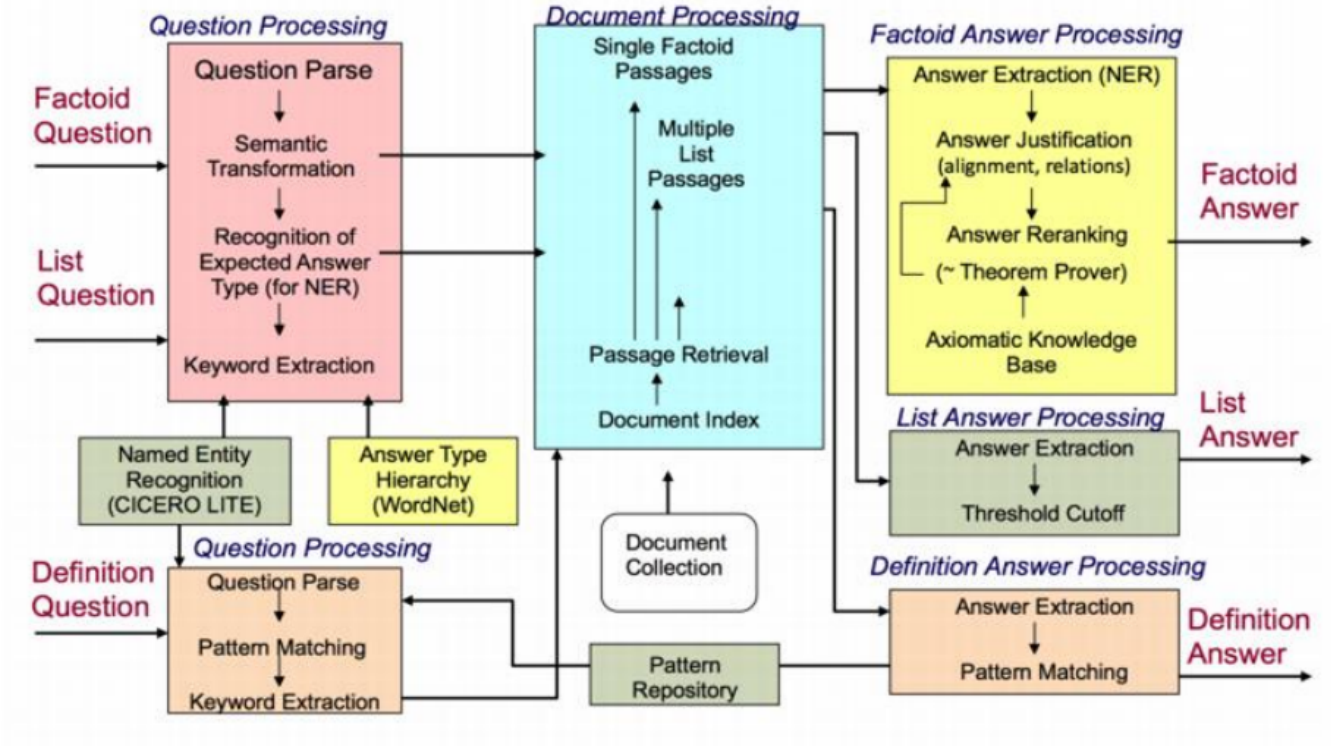
- (Celikyilmaz, Thint, and Huang 2009) implemented a SSL approach by creating a graph for labeled and unlabeled data using match-scores of textual entailment features as similarity weights between data points, and demonstrated that **utilization of more unlabeled data points** can improve the answer-ranking task of QA.

- (Dhingra, Danish, and Rajagopal 2018) showed that **fine-tuning** the pre-trained QA models **on the small set of labeled QA pairs** improves the performance of the models significantly.

- (Zhou, Chen, and Wang 2010) applied **active learning** in the semi-supervised learning framework to identify reviews that should be labeled as training data for **review sentiment classification.**

Early QA systems

[architecture of LCC (Harabagiu/Moldovan) QA system, circa 2003]
Complex systems but they did work fairly well on "factoid" questions



Modern QA systems - Deep Neural Models

- **Representation-based models**
  - Encode Q and A into fixed vectors (using BiLSTM and CNN) + similarity of these vectors
- **Interaction-based models**
  - Capture the interaction between individual words in Q and A usually using attention mechanisms (e.g., Transformer models)

Comparison against KB-based QA

- KB-based QA also use graph-based representations, but
- Our approach uses a *dynamically-constructed* graph that is built on-the-fly from the documents relevant for a query
  - More relevant information than a static KB
  - Smaller search space than a static KB