CAsT - History (Q&A) Expansion

CFDA_CLIP & h2oloo

Presenters:

PARTI:

Sheng-Chieh Lin CITI, Academia Sinica

PARTII:

Jheng-Hong Yang CITI, Academia Sinica

Outline

- Our Conversational QA system
- Historical Query Expansion
- Historical Answer Expansion
- Experimental Results

Conversational QA System

Conversational QA System

Query1

Query2

. . .

Anserini

Document Retrieval

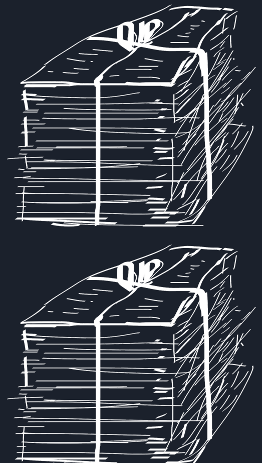
BERT Large

Document Reranking

Answer1

Answer2

. . .



Conversational QA System

Query1 +Session Title

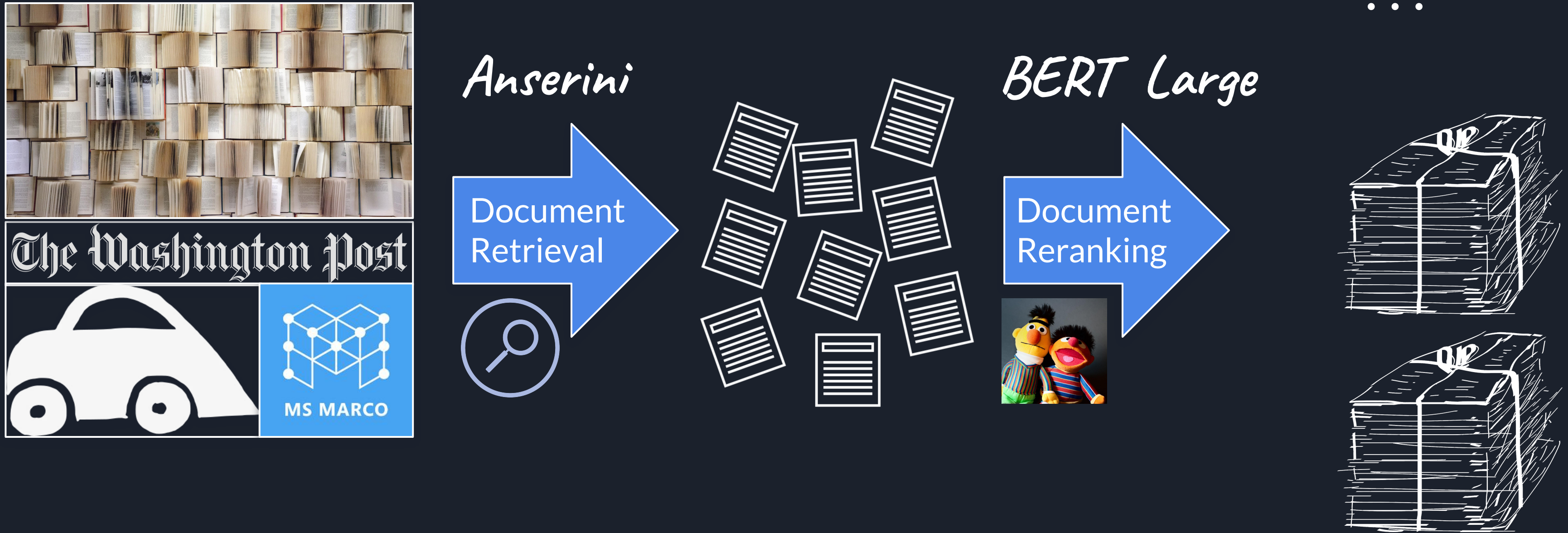
Query2 +Session Title

...

Answer1

Answer2

...



Our Strong Baseline! But...

Retrieval	Origin	Title
Re-ranking	Origin	Title
R@1000	0.440	0.774
mAP	0.069	0.187
mRR@10	0.120	0.273

Historical Query Expansion

Motivation:
Why some queries get better mAP scores?

Session1:

1 What is a physician's assistant?

2 What are the educational requirements required to become one?

3 What does it cost?

4 What's the average starting salary in the UK?

5 What about in the US?

6 What school subjects are needed to become a registered nurse?

7 What is the PA average salary vs an RN?

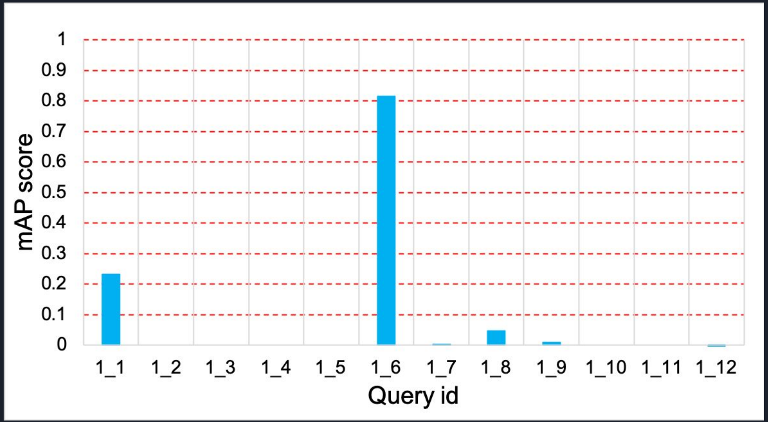
8 What the difference between a PA and a nurse practitioner?

9 Do NPs or PAs make more?

10 Is a PA above a NP?

11 What is the fastest way to become a NP?

12 How much longer does it take to become a doctor after being an NP?



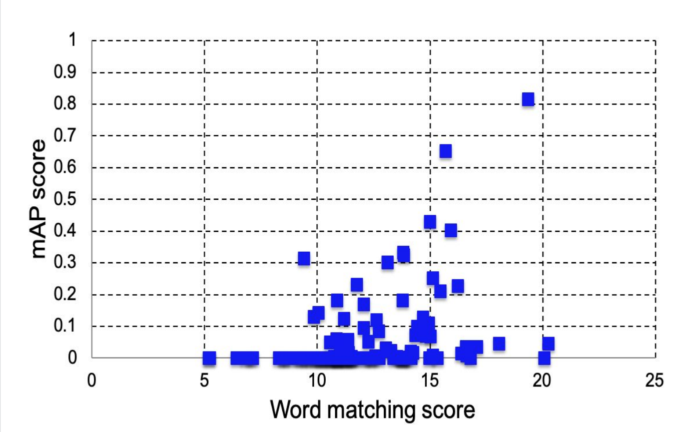
Observation 1:
Ambiguous queries can be detected automatically

Score	Query
11.73	What is a physician's assistant?
14.96	What are the educational requirements required to become one?
9.29	What does it cost?

Query	Doc1	Score1
	Doc2	Score2
	...	...
	DocN	ScoreN

Observation 1:
Ambiguous queries can be detected automatically

Score	Query
11.73	What is a physician's assistant?
14.96	What are the educational requirements required to become one?
9.29	What does it cost?



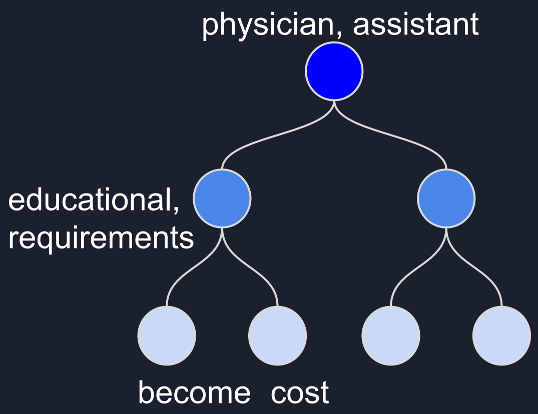
Observation 2:
Keywords can be extracted from queries

Score	Query
11.73	['What', 'is', 'a', 'physician', "'s", 'assistant', '?'] [3.13, -1, -1, 5.87, 3.58, 4.24, -1]
14.96	['What', 'are', 'the', 'educational', 'requirements', 'required', 'to', 'become', 'one', '?'] [3.13, -1, -1, 3.92, 3.75, 3.75, -1, 3.48, 2.09, -1]
9.29	['What', 'does', 'it', 'cost', '?'] [3.13, 3.90, -1, 3.42, -1]

Word		
	Doc1	Score1
	Doc2	Score2
	...	...
	DocN	ScoreN

Observation 3:
Tree structure of keywords

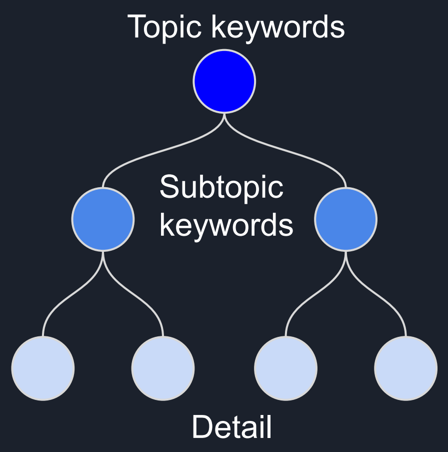
Score	Query
11.73	['What', 'is', 'a', 'physician', "'s", 'assistant', '?']
	[3.13, -1, -1, 5.87, 3.58, 4.24, -1]
14.96	['What', 'are', 'the', 'educational', 'requirements', 'required', 'to', 'become', 'one', '?']
	[3.13, -1, -1, 3.92, 3.75, 3.75, -1, 3.48, 2.09, -1]
9.29	['What', 'does', 'it', 'cost', '?']
	[3.13, 3.90, -1, 3.42, -1]



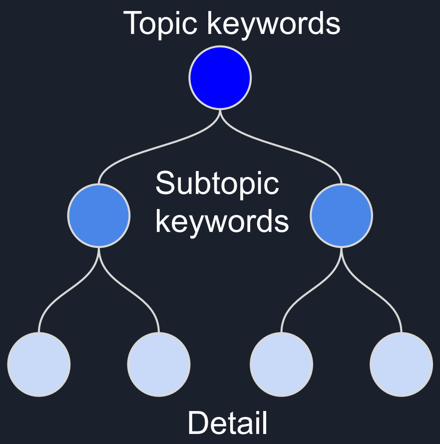
Assumption

A clear query requires:

1. Topic keywords: last along the whole session
2. Subtopic keywords: last along several turns
3. Detail: last only one turn



Methodology

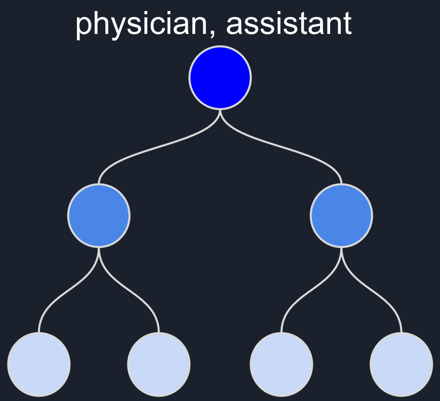


A clear query requires:

1. Topic keywords: last along the whole session
2. Subtopic keywords: last along several turns
3. Detail: last only one turn

For each query:

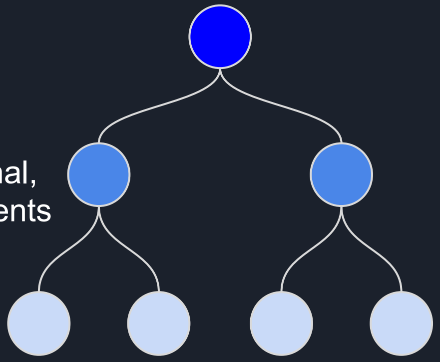
1. Extract topic keywords (R_1=4)
2. Extract subtopic keywords (R_2=3.5)
3. Check if a query is ambiguous (θ=10)
   a. Yes: Add topic keywords except for the first query
   b. No: Add topic keywords + subtopic keywords extracted from previous N turns



Methodology

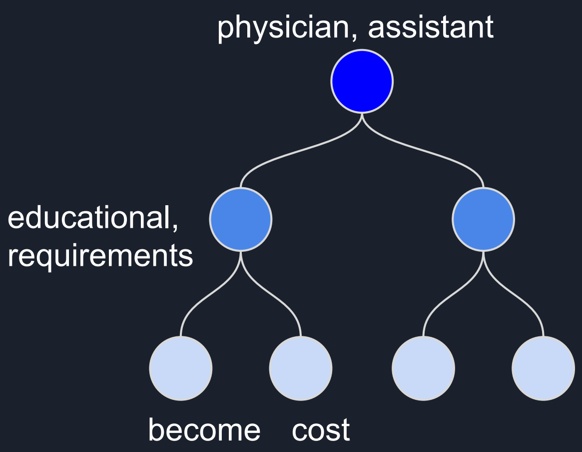
Score	Query
11.73	['What', 'is', 'a', 'physician', "'s", 'assistant', '?'] [3.13, -1, -1, 5.87, 3.58, 4.24, -1]

Methodology



Score	Query
11.73	['What', 'is', 'a', 'physician', "'s", 'assistant', '?'] [3.13, -1, -1, 5.87, 3.58, 4.24, -1]
14.96	['What', 'are', 'the', 'educational', 'requirements', 'required', 'to', 'become', 'one', '?'] [3.13, -1, -1, 3.92, 3.75, 3.75, -1, 3.48, 2.09, -1] + ['physician', 'assistant']

Methodology



Score	Query
11.73	['What', 'is', 'a', 'physician', "'s", 'assistant', '?'] [3.13, -1, -1, 5.87, 3.58, 4.24, -1]
14.96	['What', 'are', 'the', 'educational', 'requirements', 'required', 'to', 'become', 'one', '?'] [3.13, -1, -1, 3.92, 3.75, 3.75, -1, 3.48, 2.09, -1] + ['physician', 'assistant']
9.29	['What', 'does', 'it', 'cost', '?'] [3.13, 3.90, -1, 3.42, -1] + ['physician', 'assistant'] + ['educational', 'requirements', 'required']

Result

Retrieval	Origin	Title	HQExp
Re-ranking	Origin	Title	HQExp
R@1000	0.440	0.774	0.818
mAP	0.069	0.187	0.192
mRR@10	0.120	0.273	0.264

Historical Answer Expansion

Motivation

- A rule-based model for history selection [1]
- Given (p, q_k, a_k, H_k), where H_k stands for the history of (q_k, a_k) pairs
- A subset of history turns: $H_k' = H_{k-T}$ is considered useful, where T = # of lagged turns

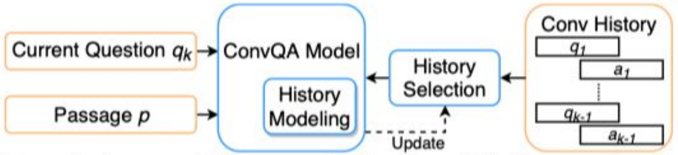


Figure 1: A general framework for ConvQA. Orange denotes model input and blue denotes model components.

Reference:
[1] Qu, Chen, et al. BERT with History Answer Embedding for Conversational Question Answering. 2019. In SIGIR 1133-1136.

Assumption

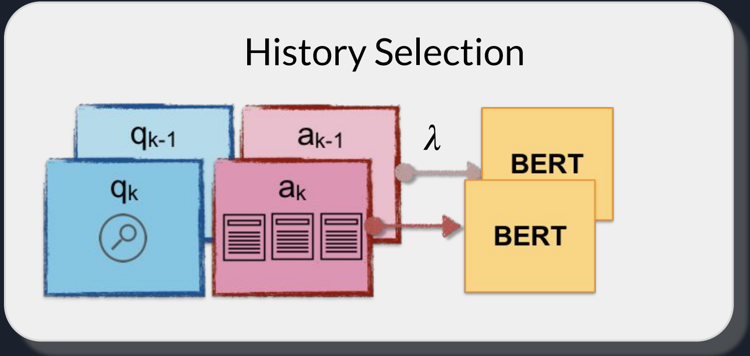
- Semantic of q_k changed smoothly within a conversation: $\varphi(q_{k-1}) \sim= \varphi(q_k)$, where φ stands for semantic mapping function
- Historical answer candidates are less important

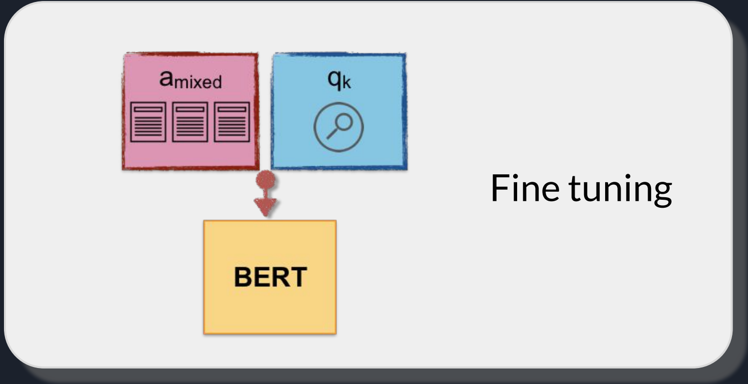
Example

```
Session1:
1 What is a physician's assistant?
2 What are the educational requirements required to become one?
3 What does it cost?
4 What's the average starting salary in the UK?
5 What about in the US?
6 ...
```

Methodology

- Consider previous one turn only: $H_k' = H_{k-1}$
- Just apply pretrained BERT as our query-passage likelihood function
- Log-likelihood of BERT(q_{k-1}, a_{k-1}) is modulated by a hyperparameter λ
- Take top-1000 by sorted list with [BERT(q_k, a_k), λ * BERT(q_{k-1}, a_{k-1})]
- Fine-tuning with BERT(q_k, set(a_k, a_{k-1})$_{top\text{-}1000}$)





Results - Training Set of CAsT

- HAE improves recall in all cases
  - Subset of history is useful
- HAE + full HQE in both stages have the best performance in recall and ranking
  - 3rd fine-tuning procedure is helpful in this combination

Retrieval	Title	Title	HQExp	HQExp
Re-ranking	Title	HQExp	Title	HQExp
R@1000	0.755	0.755	0.818	0.818
mAP	0.188	0.194	0.189	0.192
mRR@10	0.274	0.281	0.257	0.264
		+HAExp		
R@1000	**0.772**	**0.772**	**0.844**	**0.844**
mAP	0.187	0.191	0.193	**0.197**
mRR@10	0.273	0.277	0.268	**0.277**

Other Methods

Query (Corpus) Expansion

- RM3 (Query)
  - Relevance feedback model implemented in Anserini [2]
  - We use RM3 to improve recall in the first stage
- Doc2Query (Corpus)
  - Seq2seq query generation model pre-trained on MS MARCO [3]
  - Use predicted top-5 queries from D2Q to expand only one of the three corpus (MS MARCO) in CAsT

Reference:

[2] Victor Lavrenko and W. Bruce Croft□. 2001. Relevance Based Language Models. In SIGIR. 120–127

[3] Nogueira, Rodrigo, et al. Document Expansion by Query Prediction. 2019. In arXiv preprint arXiv:1904.08375.

Results - Training Set of CAsT

Corpus	CAsT			CAsT + D2Q (MARCO)		
Retrieval	Title	Title + RM3	HQExp	Title	Title + RM3	HQExp
Re-ranking	Title	Title	Title	Title	Title	Title
R@1000	0.755	0.774	0.818	0.759	0.769	0.805
mAP	0.188	0.187	0.189	0.189	0.181	0.188
mRR@10	0.274	0.273	0.257	0.281	0.279	0.262

- RM3 improves recall
  - Treat queries in each turn independently still works
- D2Q only improves the case involving title in both stages
  - May due to the mismatch of queries between CAsT and pure MS MARCO
  - Predicted queries for CAR and WAPO are not considered
- HQE provides best performance with/without corpus modification

Coreference Resolution

Results - Annotated Subset of CAsT

- Coreference resolution from CAsT host
- Evaluated on annotated subset of CAsT training set only
- HQE is enhanced by Coref in the re-ranking stage
- HAE still improves recall, but it hurts ranking metrics in the fine-tuning stage

Retrieval	Title+RM3	Title+RM3	HQExp	HQExp
Re-ranking	Title	Coref	HQExp	Coref
R@1000	0.897	0.897	0.859	0.859
mAP	0.258	0.397	0.274	0.374
mRR@10	0.358	0.552	0.433	0.544
		+HAExp		
R@1000	**0.910**	**0.910**	**0.863**	**0.863**
mAP	0.257	0.388	0.272	**0.371**
mRR@10	0.358	0.524	0.431	**0.520**

Submissions

Overall Results - Training Set of CAsT

- Due to superiority of coreference resolution on the annotated training subset, we choose to submit heavily on combinations with Coref involved flow

Table 1: Training set

Condition	1	2	3	4	5	6
Retrieval	Title	Title	Title	HQExp	HQExp	HQExp
Re-ranking	Title	HQExp	Coref	Title	HQExp	Coref
R@1000	0.774	0.774		**0.818**	**0.818**	
mAP	0.187	0.194	-	0.189	0.192	-
mRR@10	0.273	0.282		0.257	0.264	
+HAExp						
R@1000	0.790	0.790		**0.844**	**0.844**	
mAP	0.187	0.192	-	0.193	**0.197**	-
mRR@10	0.273	**0.279**		0.268	0.277	

Table 2: Co-reference effect on annotated subset

Condition	1	2	3	4	5	6
Retrieval	Title	Title	Title	HQExp	HQExp	HQExp
Re-ranking	Title	HQExp	**Coref**	Title	HQExp	**Coref**
R@1000	0.897	0.897	0.897	0.859	0.859	0.859
mAP	0.258	0.291	**0.392**	0.261	0.274	**0.374**
mRR@10	0.358	0.442	**0.525**	0.377	0.433	**0.544**
+HAExp						
R@1000	0.910	0.910	0.910	0.863	0.863	0.863
mAP	0.257	0.285	0.388	0.261	0.272	0.371
mRR@10	0.358	0.440	0.524	0.377	0.431	0.520

Results on Evaluation Set of CAsT

Automatic Runs

RUN_TAG	CFDA_CLIP_1	H2OLOO_2	H2OLOO_3	H2OLOO_4	H2OLOO_5	CFDA_CLIP_6	CFDA_CLIP_7	CFDA_CLIP_8
Indexed	MARCO	CAsT	CAsT	CAsT	CAsT	CAsT	CAsT+D2Q	CAsT+D2Q
Retrieval	Title	Title	Title	Title+RM3	HQExp	Coref+RM3	Title	HQExp
Re-ranking	Coref	HQExp	Coref	Coref	Coref	Coref	HQExp	Coref
+HAE					V			V
R@1000	0.412	0.632	0.632	0.639	**0.689**	**0.812**	0.611	0.695
mAP	0.226	0.274	0.324	0.321	0.354	**0.395**	0.269	0.363
mAP@5	0.071	0.066	0.082	0.081	0.096	**0.101**	0.068	0.099
NDCG@5	0.459	0.427	0.530	0.532	0.564	**0.576**	0.427	0.568

Conclusions

- Proposed two ad-hoc methods for conversational-information-seeking problem defined in Conversational Assistant Track (CAsT) in TREC 2019
- Two proposed methods are suitable for the dataset with few labels
- Coreference resolution is not addressed in the current setting
- Need a detailed analysis on the full combinations of corpus expansion/query expansion/history (Q&A) expansion

Thank you