Contextualized Query Embeddings for Conversational Search

Presenter: Sheng-Chieh Lin
carving

/ˈkærvɪŋ/

See definitions in:

Art  Skiing

noun

the action of carving.
• an object or design cut from a hard material as an artistic work.
plural noun: carvings
"a carving of an Asian elephant"

Similar: sculpture  model  statue  

Definitions from Oxford Languages

Translations and more definitions

Carving

Hobbies  Videos  Types
TREC: Conversational Assistant Track

CAsT Example

$q_1$ What is throat cancer?
$q_2$ Is it treatable?
$q_3$ Tell me about lung cancer.
$q_4$ What are its symptoms?
TL;DR
Successfully Apply DPR to Conversational Search

Interpret how DPR Understand Conversational Query
WHY
Previous Solution

CAst Example

$q_1$ What is throat cancer?
$q_2$ Is it treatable?
$q_3$ Tell me about lung cancer.
$q_4$ What are its symptoms?


What are lung cancer's symptoms?

$q_{<4}; q_4$

CQR

BERT Re-ranker

BM25
Our Idea: Conversational DPR

CAst Example

$q_1$ What is throat cancer?
$q_2$ Is it treatable?
$q_3$ Tell me about lung cancer.
$q_4$ What are its symptoms?


What are lung cancer's symptoms?

$q_{<4}; q_4$ CQR DPR BERT Re-ranker

8
Our Idea: Conversational DPR

What are lung cancer's symptoms?

$q_{<4}; q_4 \rightarrow \text{CQR} \rightarrow \text{DPR} \rightarrow \text{BERT Re-ranker}
## Challenge

Table 2: CAst dataset statistics.

<table>
<thead>
<tr>
<th></th>
<th>CAst19 Training</th>
<th>CAst19 Eval</th>
<th>CAst20 Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td># Queries</td>
<td>108</td>
<td>173</td>
<td>208</td>
</tr>
<tr>
<td># Dialogues</td>
<td>13</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td># Passages</td>
<td>38M</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What are lung cancer's symptoms?

$q_{<4}; q_4$ → CQR → DPR → BERT Re-ranker
Successfully Apply DPR to Conversational Search

Interpret how DPR Understand Conversational Query
Fine-tuning with Pseudo Relevance Judgement

Table 1: CANARD dataset statistics.

<table>
<thead>
<tr>
<th>CANARD</th>
<th>Training</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td># Queries</td>
<td>31,526</td>
<td>3,430</td>
<td>5,571</td>
</tr>
<tr>
<td># Dialogues</td>
<td>4,383</td>
<td>490</td>
<td>771</td>
</tr>
</tbody>
</table>

Fine-tuning with Pseudo Relevance Judgement

Q1: What happened in 1983?
A1: In May 1983, she marries Nikos Karvelas, a composer

Q2: Did they have any children?
A2: In November, she gave birth to her daughter Sofia

Q3: Did she have any other children?

Did Anna Vissi and Nikos Karvelas have any children together?

Did Anna Vissi have any other children than her daughter Sofia?


Fine-tuning with Pseudo Relevance Judgement

**Question Rewriting**

Q1: What happened in 1983? → What happened to Anna Vissi in 1983?

A1: In May 1983, she marries Nikos Karvelas, a composer

Q2: Did they have any children? → Did Anna Vissi and Nikos Karvelas have any children together?

A2: In November, she gave birth to her daughter Sofia

Q3: Did she have any other children? → Did Anna Vissi have any other children than her daughter Sofia?

Contextualized Query Embeddings

What was the neolithic revolution? When did it start and end? [Q] Why did it start?

Conversational Query Encoder

Context

What was the neolithic revolution? When did it start and end? [Q] Why did it start?

Query

What was the neolithic revolution? When did it start and end? [Q] Why did it start?
Successfully Apply DPR to Conversational Search
○ Reformulate conversational query directly in dense space
○ Create training data with pseudo relevance judgement

Interpret how DPR Understand Conversational Query
○ Text compression (or filtering)
Successfully Apply DPR to Conversational Search

- Reformulate conversational query directly in dense space
- Create training data with pseudo relevance judgement

Interpret how DPR Understand Conversational Query

- Text compression (or filtering)
What was the Neolithic Revolution? When did it start and end?

[Q] Why did it start?
Interpretation

What was the Neolithic revolution? When did it start and end? [Q] Why did it start?
The Neolithic Revolution

What was the Neolithic Revolution? When did it start and end? Why did it start?

Interpretation

Before fine-tuning

After fine-tuning

Dense index

ANN search

AvgPool

CQE

Conversational Query Encoder

Context

Query

Bag of words

Neolithic Revolution? Why did it start?

Sparse index

BM25 search

21
Experiments
# Experiments

## Models

<table>
<thead>
<tr>
<th>CQE</th>
<th>DPR</th>
<th>CQR</th>
</tr>
</thead>
</table>

## Training Data

Table 1: CANARD dataset statistics.

<table>
<thead>
<tr>
<th>CANARD</th>
<th>Training</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td># Queries</td>
<td>31,526</td>
<td>3,430</td>
<td>5,571</td>
</tr>
<tr>
<td># Dialogues</td>
<td>4,383</td>
<td>490</td>
<td>771</td>
</tr>
</tbody>
</table>

## Eval Data

Table 2: CAsT dataset statistics.

<table>
<thead>
<tr>
<th></th>
<th>CAsT19</th>
<th>CAsT20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Eval</td>
</tr>
<tr>
<td># Queries</td>
<td>108</td>
<td>173</td>
</tr>
<tr>
<td># Dialogues</td>
<td>13</td>
<td>20</td>
</tr>
<tr>
<td># Passages</td>
<td>38M</td>
<td></td>
</tr>
</tbody>
</table>
Dense Retrieval

CQE (BERT-base)  
DPR  

CQR  
DPR  

CQE  
DPR  

NDCG  

Human  
GPT2  
BERT-large  
T5-base  
CQE (BERT-base)  

NDCG@3  

Human  
GPT2  
BERT-large  
T5-base  
CQE (BERT-base)  

[9] Lin et al. 2020
Comparison to Multi-stage Pipeline

<table>
<thead>
<tr>
<th>CAsT19 Eval</th>
<th>nDCG@3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BERT-base</strong>: latency = 314 ms</td>
<td></td>
</tr>
<tr>
<td>CQE</td>
<td>.499</td>
</tr>
<tr>
<td>CQE-hybrid</td>
<td>.515</td>
</tr>
</tbody>
</table>

CQE

DPR
Comparison to Multi-stage Pipeline

<table>
<thead>
<tr>
<th>CAsT19 Eval</th>
<th>nDCG@3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BERT-base</strong> latency = 314 ms</td>
<td></td>
</tr>
<tr>
<td>CQE</td>
<td>.499</td>
</tr>
<tr>
<td>CQE-hybrid</td>
<td>.515</td>
</tr>
<tr>
<td><strong>CQR + BM25 + BERT-base</strong> latency = 5,350 ms</td>
<td></td>
</tr>
<tr>
<td>QuReTec (Voskarides et al., 2020)</td>
<td>.476</td>
</tr>
<tr>
<td>Few-Shot Rewriter (Yu et al., 2020)</td>
<td>.492</td>
</tr>
<tr>
<td><strong>3CQR + BM25 + BERT-base</strong> latency = 8,025 ms (est.)</td>
<td></td>
</tr>
<tr>
<td>MVR (Kumar and Callan, 2020)</td>
<td>.565</td>
</tr>
</tbody>
</table>
Comparison to Multi-stage Pipeline

CQE

DPR

<table>
<thead>
<tr>
<th></th>
<th>CAsT19 Eval</th>
<th>nDCG@3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BERT-base:</strong> latency = 314 ms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CQE</td>
<td>.499</td>
<td></td>
</tr>
<tr>
<td>CQE-hybrid</td>
<td>.515</td>
<td></td>
</tr>
<tr>
<td><strong>CQR + BM25 + BERT-base:</strong> latency = 5,350 ms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QuReTec (Voskarides et al., 2020)</td>
<td>.476</td>
<td></td>
</tr>
<tr>
<td>Few-Shot Rewriter (Yu et al., 2020)</td>
<td>.492</td>
<td></td>
</tr>
<tr>
<td><strong>3CQR + BM25 + BERT-base:</strong> latency = 8,025 ms (est.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MVR (Kumar and Callan, 2020)</td>
<td>.565</td>
<td></td>
</tr>
<tr>
<td><strong>CQR + BM25 + BERT-large:</strong> latency = 16,450 ms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformer++ (Vakulenko et al., 2020)</td>
<td>.529</td>
<td></td>
</tr>
<tr>
<td>NTR (T5) (Lin et al., 2021c)</td>
<td>.556</td>
<td></td>
</tr>
<tr>
<td>HQE + NTR (T5) (Lin et al., 2021c)</td>
<td>.565</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions and Future Work

• Create training data with pseudo relevance judgement
  ○ Reformulate conversational query directly in dense space
  ○ Create training data with pseudo relevance judgement

• Explain how DPR reformulates queries in embedding space
  ○ Text compression (or filtering)

• Future Work:
  ○ Add system responses as context
[1] https://blog.google/products/search/the-new-conversational-search-experience-were-thankful-for/