Microsoft

Microsoft Speech and Dialogue Research Group

Learning to Attend On Essential Terms: An Enhanced Retriever-Reader Model for Open-domain Question Answering

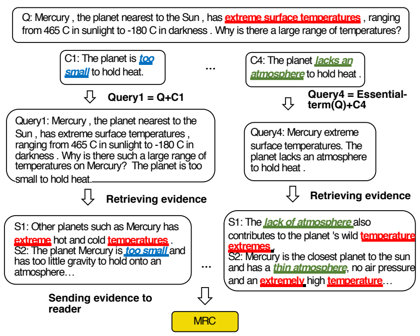
UC San Diego
JACOBS SCHOOL OF ENGINEERING
Computer Science and Engineering

Jianmo Ni, Chenguang Zhu, Weizhu Chen, Julian McAuley

UC San Diego, Microsoft Speech and Dialogue Research Group

Introduction

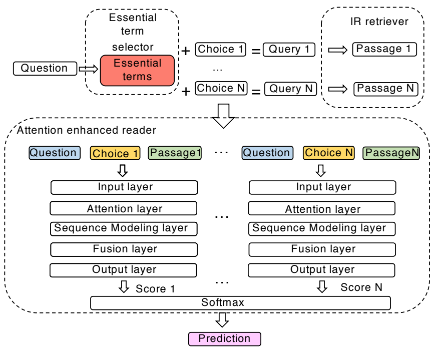
- Open-domain QA on large-scale corpus is framed as machine reading comprehension based on evidence retrieved from corpus by search engines.
- Answer quality highly depends on retrieval results.
- Previous work uses question+choice as search query, which may return irrelevant results.
- Finding essential terms within questions can help improve quality.



Our Contributions

- Propose a retriever-reader model that learns to attend on essential terms
- *Essential term selector*
  - Identifies important words in a question
  - Reformulates the query
  - Searches for related evidence
- *Enhanced reader*
  - Distinguishes between essential terms and other words
  - Use machine reading comprehension to predict the answer

Our Model

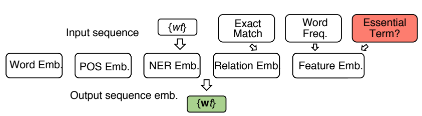


Essential Term Selector

- Given a question Q and K answer choices $C_1, \ldots, C_K$
- Predict a binary variable y_i for each word Q_i in the question Q, where y_i=1 means Q_i is an essential term and 0 otherwise.
- Use BiLSTM to learn each term's y_i based on a supervised dataset

Enhanced Reader

- Input layer
  - Question Q $= \{w_t^Q\}_{t=1}^q$
  - K choices $C_k = \{w_t^C\}_{t=1}^c$
  - K retrieved passages $P_k = \{w_t^P\}_{t=1}^p$
  - Use a mix of embeddings including output from essential term selector



- Attention Layer
  - Cross attention between choices and retrieved passages
- Sequence Modeling Layer

$$\mathbf{h}^q = \mathrm{BiLSTM}\,\mathbf{w}_Q$$
$$\mathbf{h}^c = \mathrm{BiLSTM}[\mathbf{w}_C; \mathbf{w}_C^P; \mathbf{w}_C^Q]$$
$$\mathbf{h}^p = \mathrm{BiLSTM}[\mathbf{w}_P; \mathbf{w}_P^Q],$$

- Fusion Layer
  - Attention over $\boldsymbol{h}^q, \boldsymbol{h}^c, \boldsymbol{h}^p$
  - Inter-choice attention
  - $\boldsymbol{c}'' = \boldsymbol{Maxpool}\left(\boldsymbol{h}^{c_k} - \frac{1}{K-1}\sum_{i \neq k} \boldsymbol{h}^{c_i}\right)$
- Output Layer
  - Generate scores for each (question, passage, choice) tuple

Results

Essential Term Selector: ET-Net

- We use the public dataset from Khashabi et al. (2017) which contains 2,223 questions, each accompanied by four answer choices.
- Labels about essential terms are available

Model	Precision	Recall	F1
MaxPMI	0.88	0.65	0.75
SumPMI	0.88	0.65	0.75
PropSurf	0.68	0.64	0.66
PropLem	0.76	0.64	0.69
ET Classifier	0.91	0.71	0.80
ET-Net	0.74	**0.90**	**0.81**

Dataset	Example questions
ARC	The best way to separate salt from water is with the use of Which geologic process most likely caused the formation of the Mount St. Helens Volcano?
RACE-Open	According to the article, what does the band Four Square hope to do in the future? According to the article we know it is _ to prevent the forests from slowly disappearing.
Amazon-QA	For anyone with small ears, do these fit comfortably or do they feel like they are always going to fall out, not in correctly, etc. Does it remove easily and does it leave any sticky residue behind? thanks in advance.

ET-RR

We compare ET-RR with existing retrieve-and-read methods on four datasets:

- ARC: Multiple-choice scientific QA dataset.
- A corpus of relevant knowledge (14M sentences).

??? Jianmo, add more datasets here

Process/causal
What is the first step of the process in the formation of sedimentary rocks? (A) erosion (B) deposition (C) compaction (D) cementation

Basic facts
Which element makes up most of the air we breathe? (A) carbon (B) nitrogen (C) oxygen (D) argon

Multihop reasoning
Which property of a mineral can be determined just by looking at it? (A) luster (B) mass (C) weight (D) hardness

Comparison
Compared to the Sun, a red star most likely has a greater (A) volume. (B) rate of rotation. (C) surface temperature. (D) number of orbiting planets

ET-RR Performance

Model	ARC Test	RACE-Open Test	MCScript-Open Test
IR solver	20.26	30.70	60.46
Random	25.02	25.01	50.02
BiDAF	26.54	26.89	50.81
BiLSTM Max-out	33.87	/	/
ET-RR (Concat)	35.33	36.87	66.46
ET-RR	**36.61**	**38.61**	**67.71**

Model	Amazon -Patio	Amazon -Auto	Aamzon -Cell
IR solver	72.80	73.60	70.50
Moqa	84.80	86.30	88.60
ET-RR (Concat)	96.19	95.21	93.26
ET-RR	**96.61**	**95.96**	**93.81**

Training Corpus	model	ARC
ARC	Reading Strategies	35.0
	ET-RR	36.6
ARC+RACE	Reading Strategies	40.7

Impact of Essential term selection

Model	ET-RR (Concat)		ET-RR (TF-IDF)		ET-RR	
Top K	Dev	Test	Dev	Test	Dev	Test
5	39.26	33.36	39.93	34.73	39.93	35.59
10	38.93	35.33	39.43	35.24	**43.96**	**36.61**
20	41.28	34.56	38.59	33.88	42.28	35.67

- ET-RR outperforms other baselines using concatenation and TF-IDF, given different numbers of retrievals K.

Conclusion

- Retriever-reader model ET-RR for open-domain QA. Strong performance on various datasets.
- For future work, we plan to explore multi-hop query and end-to-end retriever-reader model via reinforcement learning.

Acknowledgement

The authors thank Jade Huang, Liang Wang and Daniel Khashabi. This work is partly supported by NSF#1750063