Quick and (not so) Dirty: Unsupervised Selection of Justification Sentences for Multi-hop Question Answering

Vikas Yadav Steven Bethard Mihai Surdeanu

{vikasy, bethard, msurdeanu}@email.arizona.edu || University of Arizona, Tucson

Task and Contributions

Task:

- Explainable **Multi-hop Question Answering** (QA) - experiments on MultiRC and ARC.
- Justification selection for improving **interpretability** of a question answering system.

Contributions:

- An **unsupervised** justification selection approach which increases relevance (R) and coverage (C) of query terms in justifications with reduced redundancy (O)
- Unsupervised state-of-the-art results on justification selection task with similar or better performance over supervised BERT in various domains.
- State-of-the-art QA performance on MultiRC and ARC by coupling AutoROCC with BERT.

Example

Question: To which organ system do the esophagus, liver, pancreas, small intestine, and colon belong?
(A) reproductive system (B) excretory system (C) **digestive system** (D) endocrine system

ROCC-selected justification sentences:

- vertebrate digestive system has oral cavity, teeth and pharynx, *esophagus* and stomach, *small intestine, pancreas, liver* and the large intestine.
- digestive system consists liver, stomach, large intestine, *small intestine, colon*, rectum, anus

BM25-selected justification sentences:

- their digestive system consists of a stomach, *liver, pancreas, small intestine*, large intestine.
- the *liver pancreas* and gallbladder are the solid organ of the digestive system

Components

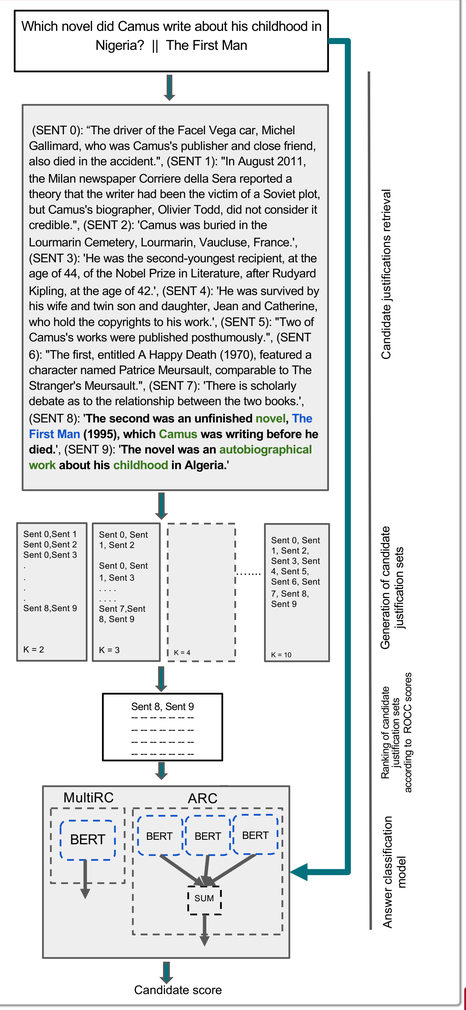
Relevance (R) - BM25 score, query = question + cand_ans

Coverage (C) - Lexical coverage of the question and the answer texts by the given set of justifications

$$C(X) = \frac{\sum_{t=1}^{|C_t(X)|} IDF[(\bigcup_{s_i \in S} t(X) \cap t(s_i))[t]]}{|t(X)|} \quad (1)$$

Overlap (O) - Ratio of common terms between individual justifications

Approach



Score

Goal is to maximize relevance(**R**) and coverage(**C**) while reducing the redundancy amongst justification sentences(**O**).

ROCC score - The final score to rank the justification set.

$$S(P_i) = \frac{R}{\epsilon + O(P_i)} \cdot (\epsilon + C(A)) \cdot (\epsilon + C(Q)) \quad (2)$$

Results

#	External resource?	Supervised justifications?	Method	$F1_m$	$F1_a$	EM0	Justification P	Justification R	Justification F1
			DEVELOPMENT DATASET						
			Previous work						
1	Yes	Yes	EER_{DPL} + FT (Wang et al. 2019)	70.5	67.8	13.3	–		
2	No	Yes	Multee (ELMo) (Trivedi et al. 2019)	***70.3***	67.3	***22.8***	–		
3	Yes	Yes	Multee (ELMo) (Trivedi et al. 2019)	73.0	69.6	22.8	–		
4	No	Yes	RS (Sun et al. 2018)	69.7	***67.9***	16.9	–		
5	Yes	Yes	RS (Sun et al. 2018)	**73.1**⋆	**70.5**⋆	21.8	–		
			BERT + IR baselines						
6	No	No	BERT + entire passage	65.7	62.7	17.0	17.4	100.0	29.6
7	No	No	BERT + BM25 (k = 3 sentences)	69.1	65.7	21.6	42.6	56.1	48.4
8	No	No	BERT + BM25 (k = 4 sentences)	70.05	66.7	22.3	36.9	64.6	47.0
9	No	No	BERT + BM25 (k = 5 sentences)	71.2	67.7	23.4	32.7	71.1	44.8
			BERT + parametric ROCC						
10	No	No	BERT + ROCC (k = 4 sentences)	72.2	69.0	25.0	40.6	71.0	51.6
11	No	No	BERT + ROCC (k = 5 sentences)	71.6	68.7	22.7	35.0	**76.5**	48.1
			BERT + non-parametric ROCC						
12	No	No	BERT + AutoROCC ($k \in \{2,3,4,5,6\}$)	72.1	69.2	**25.3**	48.2	68.2	**56.4**
13	No	No	BERT + BM25 (k from best AutoROCC)	71.1	67.4	23.1	43.8	61.2	51.0
14	No	No	BERT + AutoROCC ($k \in \{2,3,4,5,6\}$, pre-trained)	**72.9**	69.6	24.7	48.2	68.2	**56.4**
			Ceiling systems with gold justifications						
15	No	Yes	BERT + Gold knowledge	79.1	75.4	37.6	100.0	100.0	100.0
16	-	-	Human	86.4	83.8	56.6	–		
			TEST DATASET						
17	No	No	SurfaceLR Khashabi et al. 2018	66.9	63.5	12.8			
18	Yes	Yes	Multee (ELMo) Trivedi et al. 2019	73.8	70.4	24.5	–		
19	No	No	BERT + AutoROCC ($k \in \{2,3,4,5,6\}$, pre-trained)	**73.8**	**70.6**	**26.1**			

Table: Performance on the MultiRC dataset, under various configurations.

Analysis

train/test	Science textbook	Fiction	News	Wiki articles	Society, Law and Justice	All
AutoROCC	54.57	53.88	54.32	60.49	61.06	56.44
BERT+All passages	55.15	55.46	68.77	65.14	58.79	60.90
BERT+Science textbook	55.67	41.01	51.45	50.06	48.84	50.79
BERT+Fiction	45.16	57.60	63.05	63.13	50.94	58.31
BERT+News	44.11	50.77	68.82	65.45	58.30	59.30
GPT-2 Wang et al. 2019	-	-	-	-	-	60.7

Table: Domain robustness of the non-parametric ROCC vs. a supervised sentence selection model, evaluated on the gold justification sentences from MultiRC.

#	Ablations	ARC	MultiRC EM0	MultiRC Justification F1
0	Full AutoROCC	**56.09**	**25.29**	**56.44**
1	– IDF	54.11	24.65	54.19
2	– $C(A)$	54.90	21.82	52.93
3	– $C(Q)$	54.66	23.61	52.09
4	– O	55.88	24.03	55.97
5	R⋆	53.90	23.40	44.81

Table: Ablation study, removing different components of ROCC. The scores are reported on the ARC test set and MultiRC dev set.

Codes - https://github.com/vikas95/AutoROCC