SDNet: Contextualized Attention-based Deep Network for Conversational Question Answering

Jan. 31, 2019

Chenguang Zhu
Microsoft Speech and Dialogue Research Group

Motivation

- Traditional MRC models target for single-turn QA scenario
- No correlation between different questions/answers
- Humans interact by conversation, with rich context
- Models with conversational AI better match realistic settings, like customer support chatbot, social chatbot

CoQA Dataset

- From Stanford NLP Lab
- First conversational question and answering datasets
- Context can be carried over, requiring coreference resolution
- Model should comprehend both the passage and questions/answers from previous rounds

Example

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Our formulation

- Passage (context): $\mathcal{C}$
- Previous rounds of questions and answers: $Q_1, A_1, \ldots, Q_{k-1}, A_{k-1}$
- Current question: Q_k

- We prepend the latest N rounds of QAs to the current question:
  - $\mathcal{Q}_k = (Q_{k-N}, A_{k-N}, \ldots, Q_{k-1}, A_{k-1}, Q_k)$

Encoding Layer

- For each word in passage and question
  - 300-dim GloVe word embedding
  - 1024-dim BERT contextual embedding

Using BERT

- We utilize BERT as a contextual embedder, with some adaptation
- BERT tokenizes text into **byte pair encoding** (**BPE**)
- It may not align with canonical tokenizers like spaCy

Using BERT--Alignment

- Tokenize text into word sequence $(w_1, w_2, \ldots, w_n)$
- Then use BERT to tokenize each word into BPEs:
  - $w_i \rightarrow (b_1, b_2, \ldots, b_{t_i})$
- After obtaining contextual embedding for each BPE, we average the BERT embeddings for all the BPEs corresponding to a word

Using BERT—Linear combination

- BERT has many layers and the paper recommends using only the last layer
- We linearly combine all layers of output
- This can greatly improve the results

Using BERT—Linear combination

$$\mathbf{BERT}_w = \sum_{l=1}^{L} \alpha_l \frac{\sum_{t=1}^{s} \mathbf{h}_t^l}{s}$$

Linear combination

Averaged BPE embedding for a word

Using BERT– Lock internal weights

- Locking weights of BERT (no backpropagation)
- In our case, finetuning does not work well

Representation

- Each word in passage is represented by $\boldsymbol{h}_i^C$
- Each word in question is represented by $\boldsymbol{h}_i^Q$

Integration Layer

- Word-level Inter-Attention
  - Attend from question to passage, i.e. $Attn(\{\boldsymbol{h}_i^C\}, \{\boldsymbol{h}_i^Q\}, \{\boldsymbol{h}_i^Q\})$:

$$S_{ij} = \mathrm{ReLU}(Uh_i^C)D\,\mathrm{ReLU}(Uh_j^Q),$$

$$\alpha_{ij} \propto exp(S_{ij}),$$

$$\hat{\boldsymbol{h}}_i^C = \sum_j \alpha_{ij}\boldsymbol{h}_j^Q,$$

- Passage words have several more features
  - POS, NER embedding
  - 3-dim exact matching vector
  - A normalized term frequency

RNN

- Passage (context) and question words both go through K layers of Bi-LSTMs

$$\boldsymbol{h}_1^{C,k}, ..., \boldsymbol{h}_m^{C,k} = \text{BiLSTM}\left(\boldsymbol{h}_1^{C,k-1}, ..., \boldsymbol{h}_m^{C,k-1}\right),$$
$$\boldsymbol{h}_1^{Q,k}, ..., \boldsymbol{h}_n^{Q,k} = \text{BiLSTM}\left(\boldsymbol{h}_1^{Q,k-1}, ..., \boldsymbol{h}_n^{Q,k-1}\right),$$
$$\boldsymbol{h}_i^{C,0} = \tilde{\boldsymbol{w}}_i^C, \boldsymbol{h}_i^{Q,0} = \tilde{\boldsymbol{w}}_i^Q,$$

- Question words go through one more RNN layer

$$\boldsymbol{h}_1^{Q,K+1}, ..., \boldsymbol{h}_n^{Q,K+1} = \text{BiLSTM}\left(\boldsymbol{h}_1^{Q}, ..., \boldsymbol{h}_n^{Q}\right),$$
$$\boldsymbol{h}_i^{Q} = [\boldsymbol{h}_i^{Q,1}; ...; \boldsymbol{h}_i^{Q,K}]$$

Question self-attention

- Question contains conversation history
- We need to establish relationship between history and current question
- Self-attention:

$$\{\boldsymbol{u}_i^Q\} = Attn(\{\boldsymbol{h}_i^{Q,K+1}\}, \{\boldsymbol{h}_i^{Q,K+1}\}, \{\boldsymbol{h}_i^{Q,K+1}\})$$

- $\{\boldsymbol{u}_i^Q\}$ is the final representation of question words

Multi-level inter-attention

- Adapted from FusionNet (Huang et al. 2018)

$$\{\boldsymbol{m}_i^{(k),C}\}_{i=1}^m = \text{Attn}(\{\text{HoW}_i^C\}_{i=1}^m, \{\text{HoW}_i^Q\}_{i=1}^n, \{\boldsymbol{h}_i^{Q,k}\}_{i=1}^n), 1 \leq k \leq K+1$$

$$\text{HoW}_i^C = [\text{GloVe}(w_i^C); \text{BERT}_{w_i^C}; \boldsymbol{h}_i^{C,1}; ..., \boldsymbol{h}_i^{C,k}],$$
$$\text{HoW}_i^Q = [\text{GloVe}(w_i^Q); \text{BERT}_{w_i^Q}; \boldsymbol{h}_i^{Q,1}; ..., \boldsymbol{h}_i^{Q,k}].$$

- After multi-level inter-attention, we conduct RNN, self-attention and another RNN to obtain the final representation of context: $\{\boldsymbol{u}_i^C\}$

Generate answer – positional probability

- Firstly, we condense the question into one vector

$$\boldsymbol{u}^Q = \sum_i \beta_i \boldsymbol{u}_i^Q \,, \beta_i \propto \exp(\boldsymbol{w}^T \boldsymbol{u}_i^Q)$$

- Then we generate the probability that the answer starts and ends at each position:

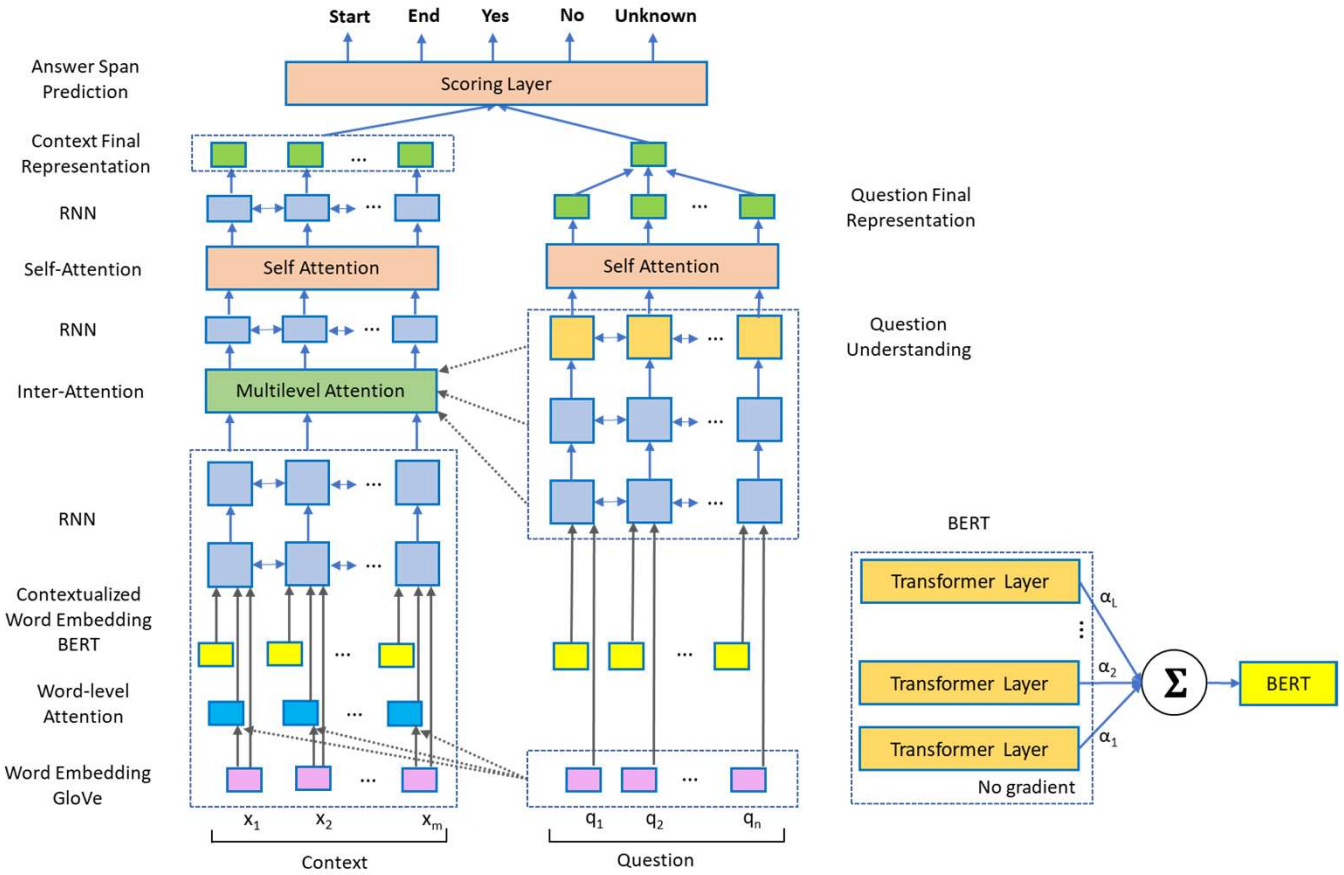
$$\text{Start probability: } \mathrm{P_i^S} \propto \exp\left(\left(\boldsymbol{u}^Q\right)^T W_S \boldsymbol{u}_i^C\right)$$

$$\text{GRU: } \boldsymbol{t}^Q = GRU(\boldsymbol{u}^Q, \textstyle\sum_i \mathrm{P_i^S} \boldsymbol{u}_i^C)$$

$$\text{End probability: } \mathrm{P_i^E} \propto \exp(\left(\boldsymbol{t}^Q\right)^T W_E \boldsymbol{u}_i^C)$$

- We generate the probability of “yes”, “no”, “no answer” similarly

Overall structure -- SDNet



Training and prediction

- Loss is cross-entropy
- Batch size: 1 dialogue
- 30 epochs
- ~2 hours per epoch on one Volta-100 GPU

- Prediction of span is done by choosing maximum (start x end) probability, with length <= 15

- Ensemble model consists of 12 best single models with different random seeds

Experiments

- CoQA dataset
- Dev set: 100 passages per in-domain topics
- Test set: 100 passages per in-domain and out-of-domain topics

Table 1: Domain distribution in CoQA dataset.

Domain	#Passage	#QA turn
Child Story	750	14.0
Literature	1,815	15.6
Mid/High Sc.	1,911	15.0
News	1,902	15.1
Wikipedia	1,821	15.4
Out of domain		
Science	100	15.3
Reddit	100	16.6
Total	8,399	15.2

Leaderboard (As of 12/11/2018)

Table 2: Model and human performance (% in F1 score) on the CoQA test set.

	Child.	Liter.	Mid-High.	News	Wiki	Reddit	Science	Overall
PGNet	49.0	43.3	47.5	47.5	45.1	38.6	38.1	44.1
DrQA	46.7	53.9	54.1	57.8	59.4	45.0	51.0	52.6
DrQA+PGNet	64.2	63.7	67.1	68.3	71.4	57.8	63.1	65.1
BiDAF++	66.5	65.7	70.2	71.6	72.6	60.8	67.1	67.8
FlowQA	73.7	71.6	76.8	79.0	80.2	67.8	76.1	75.0
SDNet (single)	**75.4**	**73.9**	**77.1**	**80.3**	**83.1**	**69.8**	**76.8**	**76.6**
SDNet (ensemble)	**78.7**	**77.1**	**80.2**	**81.9**	**85.2**	**72.3**	**79.7**	**79.3**
Human	90.2	88.4	89.8	88.6	89.9	86.7	88.1	88.8

Ablation Studies

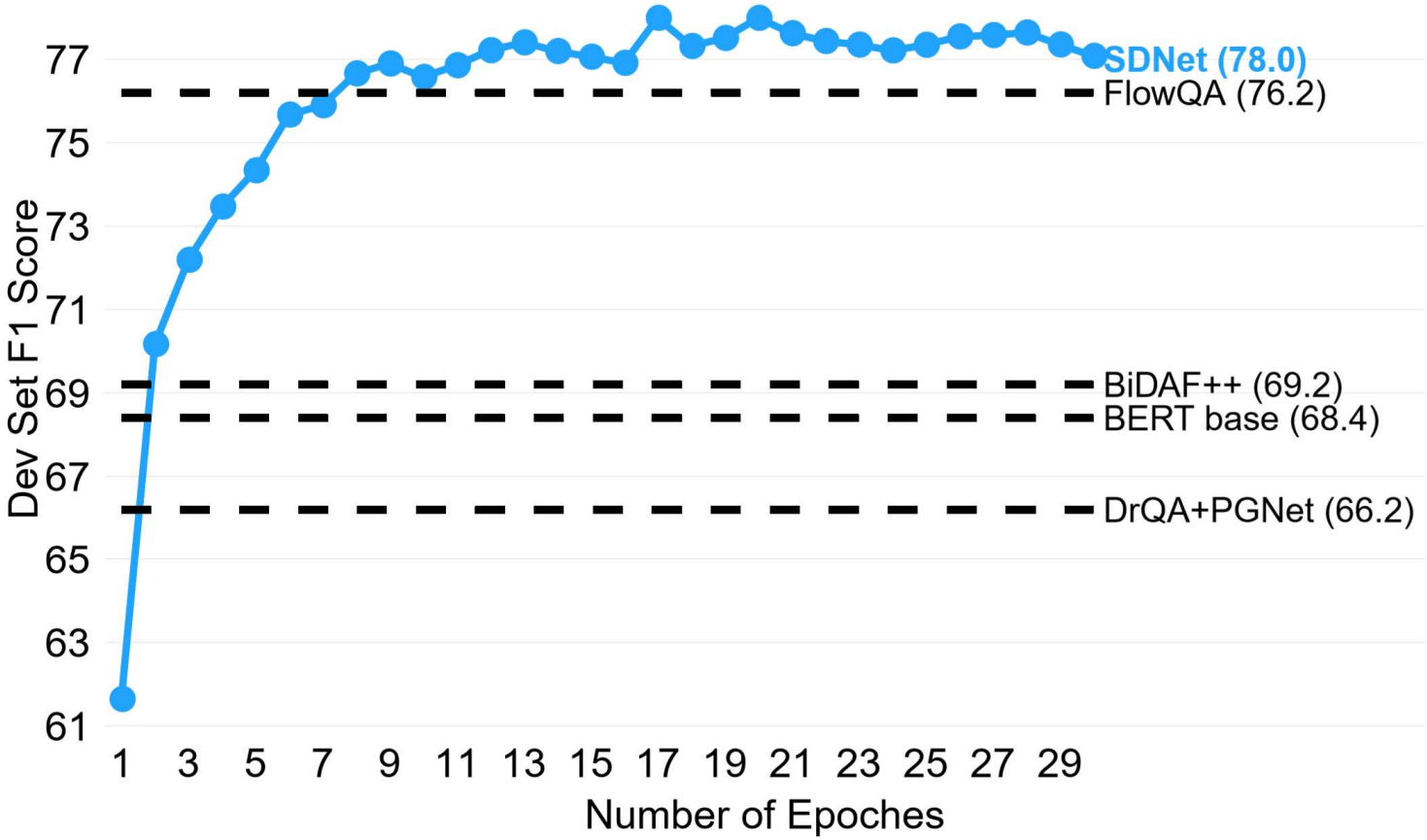
Table 3: Ablation study of SDNet on CoQA development dataset.

Model	F_1
SDNet	77.99
–Variational dropout	77.75
–Question self attention	77.24
Using last layer of BERT output (no weighted sum)	76.24
BERT-base	75.38
ELMo	73.37
–BERT	70.84

Effect of history

#previous QA rounds N	F_1
0	69.43
1	76.70
2	**77.99**
3	77.39

Training


77
75
73
71
69
67
65
63
61
Dev Set F1 Score
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29
Number of Epoches
SDNet (78.0)
FlowQA (76.2)
BiDAF++ (69.2)
BERT base (68.4)
DrQA+PGNet (66.2)

Summary

- SDNet leverages MRC, BERT and conversational history
- BERT is important, but adaptation is required to use it as a module
- Next step is open-domain conversational QA, with no passage given, and model may also ask questions

Introduction to SDRG

- Microsoft **S**peech and **D**ialogue Research Group
- Conduct research on speech recognition and conversational AI
- We're from Stanford, Princeton, Cambridge, IBM, …
- We're **now hiring** researchers and machine learning engineers
- We have world-class research environment
  - DGX-1 machine and large clusters with tons of GPUs
- Welcome to learn more!

Q&A

- Thank you!

- My Email: chezhu@microsoft.com

- More details in our paper:
  - SDNet: Contextualized Attention-based Deep Network for Conversational Question Answering – Chenguang Zhu, Michael Zeng, Xuedong Huang
  - https://arxiv.org/abs/1812.03593