Generating Natural Answers by Incorporating Copying and Retrieving Mechanisms in Sequence-to-Sequence Learning

Guo Tianyi

2017.12.21

Outline

1 Introduction

- Question answering (QA) systems devote to providing exact answers, often in the form of phrases and entities for natural language questions (Woods, 1977; Ferrucci et al., 2010; Lopez et al., 2011; Yih et al., 2015)
- However, in real-world environments, most people prefer the correct answer replied with a more natural way.
- This paper proposed a new and practical question answering task which devotes to generating natural answers for information inquired questions.
- Proposed a neural network based model, named as COREQA, by incorporating copying and retrieving mechanism in Seq2Seq learning.

Outline

2.1 RNN Encoder-Decoder

- Recurrent Neural Network(RNN) based Encoder-Decoder is the backbone of Seq2Seq learning (Cho et al., 2014)
- An encoding RNN first transform a source sequential object $X = [x_1, \ldots, x_{L_X}]$ into an encoded representation $\boldsymbol{c}$.
- Once the source sequence is encoded, another decoding RNN model is to generate a target sequence $Y = [y_1, \ldots, y_{L_Y}]$ through the following prediction model:

$$\boldsymbol{s}_t = f(y_{t-1}, \boldsymbol{s}_{t-1}, \boldsymbol{c}); \quad p(y_t | y_{<t}, X) = g(y_{t-1}, \boldsymbol{s}_t, \boldsymbol{c})$$

2.2 The Attention Mechanism

- The prediction model of classical decoders for each target word y_t share the same context vector $\boldsymbol{c}$.
- However, a fixed vector is not enough to obtain a better result on generating a long targets.
- The attention mechanism in the decoding can dynamically choose context $\boldsymbol{c}_t$ at each time step.

$$\boldsymbol{c}_t = \sum_{i=1}^{L_X} \alpha_{ti} \boldsymbol{h_i}; \quad \alpha_{ti} = \frac{e^{\rho(\boldsymbol{s}_{t-1}, \boldsymbol{h}_i)}}{\sum_{i'} e^{\rho(\boldsymbol{s}_{t-1}, \boldsymbol{h}'_i)}}$$

where the function ρ use to compute the attentive strength with each source state, which usually adopts a neural network such as multi-layer perceptron (MLP).

2.3 The Copying Mechanism

Seq2Seq learning heavily rely on the "meaning" for each word in source and target sequences, however, some words in sequences are "no-meaning" symbols and it is improper to encode them in encoding and decoding processes.

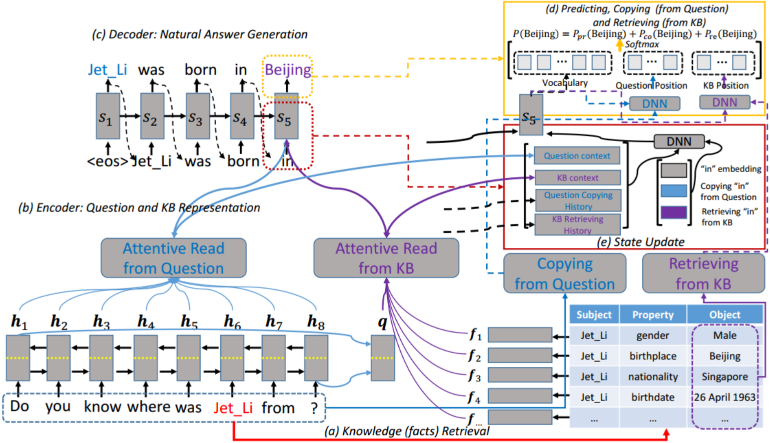
For example, generating the response "*Of course, read*" for replying the message "*Can you read the word 'read'?*" should not consider the meaning of the second "*read*".

By incorporating the copying mechanism, the decoder could directly copy the sub-sequences of source into the target (Vinyals et al., 2015). The basic approach is to jointly predict the indexes of the target word in the fixed vocabulary and/or matched positions in the source sequences (Gu et al., 2016; Gulcehre et al., 2016).

Outline

3.1 Model Overview

The overall diagram of COREQA

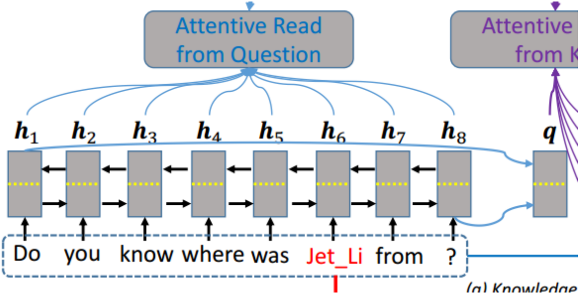


3.2 Knowledge (facts) Retrieval

- We mainly focus on answering the information inquired questions. This paper utilizes the gold topic entities for simplifying our design.
- Given the topic entities, we retrieve the related facts from the corresponding KB.
- KB consists of many relational data, which usually are sets of inter-linked subject-property-object (*SPO*) triple statements.
- Usually, question contains the information used to match the *subject* and *property* parts in a fact triple, and answer incorporates the *object* part information.

3.3 Encoder
(1) Question Encoding

$$\{\overrightarrow{\boldsymbol{h}_1}, \ldots, \overrightarrow{\boldsymbol{h}_{L_X}}\} = \overrightarrow{RNN}(\boldsymbol{X})$$
$$\{\overleftarrow{\boldsymbol{h}_1}, \ldots, \overleftarrow{\boldsymbol{h}_{L_X}}\} = \overleftarrow{RNN}(\boldsymbol{X})$$
$$q = [\overrightarrow{\boldsymbol{h}_{L_X}}, \overleftarrow{\boldsymbol{h}_1}]$$
$$\boldsymbol{M}_Q = \{\boldsymbol{h}_t\}$$
$$\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_{L_X-t+1}}]$$



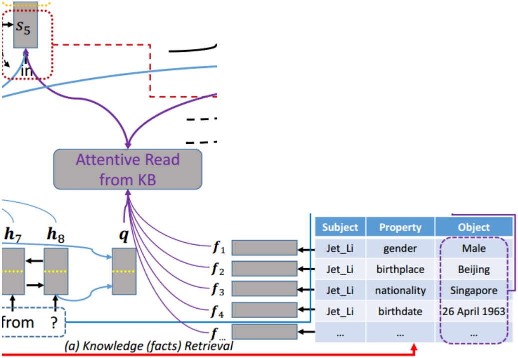
3.3 Encoder
(2) Knowledge Base Encoding

$\boldsymbol{f} = [\boldsymbol{e}_s, \boldsymbol{e}_p, \boldsymbol{e}_o]$

$\{\boldsymbol{f}\} = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_{L_F}\}$

$S(q, f_j) = DNN_1(\boldsymbol{q}, \boldsymbol{f}_j) = \tanh(\boldsymbol{W}_2 \cdot \tanh(\boldsymbol{W}_1 \cdot [\boldsymbol{q}, \boldsymbol{f}_j] + \boldsymbol{b}_1) + \boldsymbol{b}_2)$

$S(q, s_t, f_j) = DNN_1(\boldsymbol{q}, \boldsymbol{s}_t, \boldsymbol{f}_j)$



Subject	Property	Object
Jet_Li	gender	Male
Jet_Li	birthplace	Beijing
Jet_Li	nationality	Singapore
Jet_Li	birthdate	26 April 1963
...	...	...

3.4 Decoder

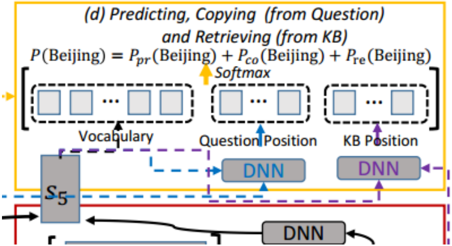
The decoding process of COREQA have the following differences compared with the conventional decoder:

- **Answer words prediction**
  COREQA predicts SUs based on a mixed probabilistic model of three modes, namely the predict-mode, the copy-mode and the retrieve-mode;
- **State Update**
  The predicted word at step $t{-}1$ is used to update s_t, but COREQA uses not only its word embedding but also its corresponding positional attention informations in M_Q and M_{KB};
- **Reading short-Memory M_Q and M_{KB}**
  M_Q and M_{KB} are fed into COREQA with two ways, the first one is the "meaning" with embeddings and the second one is the positions of different words (properties' values).

3.4 Decoder
(1) Answer Words Prediction

- Predicted word vocabulary $\mathcal{V} = \{v_1, \ldots, v_N\} \cup \{\text{UNK}\}$, UNK indicates any out-of-vocabulary (OOV) words.
- Adopt two set of SUs $\mathcal{X}_Q$ and $\mathcal{X}_{KB}$ which cover words/entities in the source question and the partial KB.
- Adopt the instance-specific vocabulary $\mathcal{V} \cup \mathcal{X}_Q \cup \mathcal{X}_{KB}$ for each question.

3.4 Decoder
(1) Answer Words Prediction



$$p(y_t|\boldsymbol{s}_t, y_{t-1}, \boldsymbol{M}_Q, \boldsymbol{M}_{KB}) = p_{pr}(y_t|\boldsymbol{s}_t, y_{t-1}, \boldsymbol{c}_t) \cdot p_m(pr|\boldsymbol{s}_t, y_{t-1})$$
$$+p_{co}(y_t|\boldsymbol{s}_t, y_{t-1}, \boldsymbol{M}_Q) \cdot p_m(co|\boldsymbol{s}_t, y_{t-1}) + p_{re}(y_t|\boldsymbol{s}_t, y_{t-1}, \boldsymbol{M}_{KB}) \cdot p_m(re|\boldsymbol{s}_t, y_{t-1})$$

pr, co, re stand for the predict-mode, the copy-mode and the retrieve-mode.

$p_m(.|.)$ indicates the probability model for choosing different modes.

3.4 Decoder
(1) Answer Words Prediction

$p_m(.|.)$ is a *softmax* classifier with two-layer MLP.

The probability of the three modes are given by

$$p_{pr}(y_t|.) = \frac{1}{Z} e^{\psi_{pr}(y_t)}$$
$$p_{co}(y_t|.) = \frac{1}{Z} \sum\nolimits_{j:Q_j=y_t} e^{\psi_{co}(y_t)}$$
$$p_{re}(y_t|.) = \frac{1}{Z} \sum\nolimits_{j:KB_j=y_t} e^{\psi_{re}(y_t)}$$
$$Z = e^{\psi_{pr}(y_t)} + \sum\nolimits_{j:Q_j=v} e^{\psi_{co}(v)} + \sum\nolimits_{j:KB_j=v} e^{\psi_{re}(v)}$$

3.4 Decoder
(1) Answer Words Prediction

The scoring functions are defined as follows:

$$
\begin{aligned}
\psi_{pr}(y_t = v_i) &= \boldsymbol{V}_i^T \boldsymbol{W}_{pr}[\boldsymbol{s}_t, \boldsymbol{c}_{q_t}, \boldsymbol{c}_{kb_t}] \\
\psi_{co}(y_t = x_j) &= DNN_2(\boldsymbol{h}_j, \boldsymbol{s}_t, \mathbf{hist}_Q) \\
\psi_{re}(y_t = v_j) &= DNN_3(\boldsymbol{f}_j, \boldsymbol{s}_t, \mathbf{hist}_{KB})
\end{aligned}
$$

3.4 Decoder
(2) State Update

y_{t-1} may not come from vocabulary $\mathcal{V}$ and not owns a word vector, so modify the state update process in COREQA.
y_{t-1} will be represented as follow:

$$y_{t-1} = [\boldsymbol{e}(y_{t-1}), \boldsymbol{r}_{q_{t-1}}, \boldsymbol{r}_{kb_{t-1}}]$$

$\boldsymbol{e}(y_{t-1})$ is the word embedding associated with y_{t-1}, $\boldsymbol{r}_{q_{t-1}}$ and $\boldsymbol{r}_{kb_{t-1}}$ are the weighted sum of hidden states in $\boldsymbol{M}_Q$ and $\boldsymbol{M}_{KB}$

3.4 Decoder
(2) State Update

$$\boldsymbol{r}_{q_t} = \sum_{j=1}^{L_X} \rho_{tj} \boldsymbol{h}_j, \boldsymbol{r}_{kb_t} = \sum_{j=1}^{L_F} \delta_{tj} \boldsymbol{f}_j$$

$$\rho_{tj} = \begin{cases} \frac{1}{K_1} p_{co}(x_j|.) & x_j = y_t \\ \mathbf{0} & otherwise \end{cases}$$

$$\delta_{tj} = \begin{cases} \frac{1}{K_2} p_{re}(f_j|.) & object(f_j) = y_t \\ \mathbf{0} & otherwise \end{cases}$$

K_1 and K_2 are the normalization terms which equals $\sum_{j':x_j'=y_t} p_{co}(x_j'|.)$ and $\sum_{j':object(f_j')=y_t} p_{re}(f_j'|.)$

3.4 Decoder
(3) Reading short-Memory $\boldsymbol{M}_Q$ and $\boldsymbol{M}_{KB}$

- COREQA employ the attention mechanism at decoding process.
- At each decoder time t, we selective read the context vector $\boldsymbol{c}_{q_t}$ and $\boldsymbol{c}_{kb_t}$ from the short-term memory of question $\boldsymbol{M}_Q$ and retrieval facts $\boldsymbol{M}_{KB}$.
- In addition, the accumulated attentive vectors $\mathbf{hist}_Q$ and $\mathbf{hist}_{KB}$ are able to record the positional information of SUs in the source question and retrieved facts.

3.5 Training

For the batches of the source questions $\{X\}_M$ and target answers $\{Y\}_M$ both expressed with natural language, the objective function is to minimize the negative log-likelihood:

$$\mathcal{L} = -\frac{1}{N}\sum_{k=1}^{M}\sum_{t=1}^{L_Y}\log[p(y_t^{(k)})|y_{<t}^{(k)}, X^{(k)}]$$

where the quperscript (k) indicates the index of one question-answer (Q-A) pair.

Outline

4.1 Natural QA in Restricted Domain

Q-A Patterns	**Examples**(e.g. KB facts (e2,year,1987);(e2,month,6); (e2,day,20);(e2,gender,male))
When is %*e* birthday? He was born in %*m* %*d* %*d*th	When is e2 birthday? He was born in June 20th.
What year were %*e* born? %*e* is born in %*y* year.	What year were e2 born? e2 is born in 1987 year.

Table 1: Sample KB facts, patterns and their generated Q-A pairs.

4.1 Natural QA in Restricted Domain

Models	P_g	P_y	P_m	P_d	P_A	R_A	$F1_A$
RNN	72.2	0	1.1	0.2	0	27.5	0
RNN+atten	55.8	1.1	11.3	9.5	1.7	34	3.2
CopyNet	75.2	8.7	28.3	5.8	3.7	32.5	6.7
GenQA	73.4	0	0	0	0	27.1	0
COREQA	**100**	**84.8**	**93.4**	**81**	**87.4**	**94**	**90.6**

Table 2: The AE results (%) on synthetic test data.

4.1 Natural QA in Restricted Domain

Entities	P_g	P_y	P_m	P_d	P_A	R_A	$F1_A$
Seen	100	84.8	93.4	81	87.4	94	90.6
Unseen	75.1	84.5	93.5	81.2	63.8	85.1	73.1

Table 3: The AE (%) for seen and unseen entities.

4.2 Natural QA in Open Domain

Models	Single	Multi	Mixed
CopyNet	9.7	0.8	8.7
GenQA	47.2	28.9	45.1
COREQA	**58.4**	**42.7**	**56.6**

Table 4: The AE accuracies (%) on real world test data.

Models	Correctness	Fluency	Coherence
CopyNet	0	13.3	3.3
GenQA	26.7	33.3	20
COREQA	**46.7**	**50**	**60**

Table 5: The ME results (%) on sampled mixed test data.

4.2 Natural QA in Open Domain

ID	Question	Gold Answer	Generated Natural Answer
1	拜仁的哈格里夫斯是哪国人？ Which country did *Hargreaves* of Bayern comes from?	英国人！加拿大出生 British! born in Canada	加拿大人，英国人 Canadians, British
2	沙祖康是哪个族的？ Which ethnic groups is *Sha Zukang* from?	汉族现在是联合国副秘书长 Han, now he is the DSG of the UN	联合国副秘书长，是汉族 DSG of the UN, Han
3	罗伯特·舒曼是谁？ Who is *Robert Schumann*?	是一个作家、音乐家 a writer, musician	**罗伯特·舒曼**个作家 哲学家 **Robert Schuhmann** is a writer and philosopher.
4	马斯切拉诺是谁？ Who is *Mascherano*?	他是一名足球运动员 He is a football player	一个优秀的足球 运动员 An excellent football player
5	巴尔扎克与小裁缝的导演是谁？ Who is the director of *The Little Chinese Seamstress*?	戴思杰 Dai Sijie	**导演**是戴思杰 **Director** Dai Sijie
6	电影铁三角是谁拍的？ Who shot the movie *The Iron Triangle*?	徐克、林岭东、杜琪峰 Tsui Hark, Johnny To, Ringo Lam	冯小刚 Feng Xiaogang
7	谁知道席慕蓉这个人的一些详细资料？ Who knows some details of *Xi Murong*?	女诗人 poetess	是是是 yes, yes, yes
8	泡泡堂是哪个公司开发的？ Which company developed the game *Crazy Arcade*?	应该是上海盛大 should be the Shanda Group	玩上海盛大 playing Shanda Group

Table 6: Examples of the generated natural answers by COREQA.

Outline

5 Related Work

- Seq2Seq learning is to maximize the likelihood of predicting the target sequence Y conditioned on the observed source sequence X (Sutskever et al., 2014)
- Recent work of QA and Dialogue which have adopted Seq2Seq learning. CopyNet (Gu et al., 2016) and Pointer Networks (Vinyals et al., 2015; Gulcehre et al., 2016)
- GenQA (Yin et al., 2016)
- Neural Abstract Machine (Graves et al., 2016; Yin et al., 2015; Liang et al., 2016) which could retrieve facts from KBs with neural models.

Outline

6 Conclusion and Future Work

The future work includes:

- lots of questions cannot be answered directly by facts in a KB, we plan to learn QA system with latent knowledge.
- we plan to adopt memory networks (Sukhbaatar et al., 2015) to encode the temporary KB for each question.