Multihop question answering for short and long narrative

Hoang Le

VinAI Residency

v.hoanglh88@vinai.io

August 5, 2021

Overview

1. Introduction of QA over multiple paragraphs
2. Paper 2: Simple and Effective

Introduction

- **Question answering** (QA) is the ability of reading text and getting the knowledge/insight about it.
- The crucial difference between single-hop QA and multihop QA is **the clues' distributing among passages**.

Example about a multihop QA dataset

A question - supporting documents - answer triple of **HotpotQA** - a typical multihop QA dataset.


```
{
  "_id": "5adf9ba1554299025d62a2db",
  "question": "What position on the Billboard Top 100 did Alison Moyet's late summer hit achieve?",
  "context": [
    [
      "The Other Side of Love",
      [
        "\"The Other Side of Love\" is a song by the British synthpop band Yazoo, released in 1982 as their
        " The single peaked at #13 on the UK Singles Chart, making it the band's least successful single an
        " The track was written by band members Vince Clarke and Alison Moyet, and was originally not inclu
        " It featured Stiff Records' all-girl band Sylvia and the Sapphires on backing vocals following a c
      ]
    ],
    [
      "All Cried Out (Alison Moyet song)",
      [
        "\"All Cried Out\" is a song by English singer-songwriter Alison Moyet.",
        " It was written by Moyet and producers Jolley & Swain for her debut studio album \"Alf\" (1984).",
        " Released as the album's second single in the autumn of 1984, the track peaked within the top ten
      ]
    ],
    [
      "The Vandals (UK band)",
      [
        "The Vandals were a late 1970s English rock band from Basildon in South East Essex.",
        " Playing in the punk rock style, they are mainly notable for featuring vocalist Alison Moyet; 'Alf
        " The other members were Robert Marlow, who during his tenure with the band was known as 'the guita
      ]
    ],
    [
      "Yazoo (band)",
      [
        "Yazoo (known as Yaz in North America for legal reasons involving Yazoo Records) were a British syn
        " Formed in late 1981 after Clarke responded to an advertisement Moyet placed in a UK music magazin
        " Yazoo enjoyed worldwide success, particularly in their home country where three of their four sin
        " In North America they are best known for the song \"Situation\", originally only a B-side in the
      ]
    ],
    [
      "Nick Morris",
      [
        "Nick Morris is a film maker who began writing and making amateur films at school, one of which was
        " His professional career began in the 1980s with music videos for \"The Final Countdown\" by Europ
        " After making approximately 100 promos for artists such as Elton John, Celine Dion, Alison Moyet,
        " He has also directed DVDs for comedians such as The Mighty Boosh, Mitchell and Webb and Bill Bail
        " Other work includes numerous trailers and music clips for West End shows such as \"The Producers\
        " in 2009 he captured Spandau Ballet's triumphant homecoming concert at the O2 Arena in London and
      ]
    ],
    [
      "Only You (Yazoo song)",
      [
        "\"Only You\" is a song written by English musician Vince Clarke.",
        " He wrote it while with Depeche Mode, but recorded it in 1982 after forming the duo Yazoo with Alis
```

Differences between multihop QA and single passage dataset

- Multihop QA datasets are given more than single passage (SQuAD) to search for the answer
- Answers of multihop QA datasets are keywords or their combination or free form (NarrativeQA) ; that of single QA datasets are keywords (SQuAD) or cloze-type (Children's Book Test) or multiple choices
- Length of supporting documents of multihop are varied vs usually short of single hop: Short paragraphs (HotpotQA, QAngaroo), enourmously long (NarrativeQA)

Reasoning is the key

In order to extract the answer, clues or something you can think of must be combined, i.e. **reasoned.**

My definition of Reasoning

Reasoning is a way of combining facts or clues from document(s) to get the final answer.

There are several tools used for reasoning: attention, graph.

Simple and Effective Curriculum Pointer-Generator Networks for Reading Comprehension over Long Narratives

[1]Yi Tay, [2]Shuohang Wang, [3]Luu Anh Tuan, [4]Jie Fu, [5]Minh C. Phan
[6]Xingdi Yuan, [7]Jinfeng Rao, [8]Siu Cheung Hui, [9]Aston Zhang
[1,5,8]Nanyang Technological University [2]Singapore Management University [3]MIT CSAIL
[4]Mila, Polytechnique Montréal [6]Microsoft Research, Montréal [7]Facebook [8]Amazon AI

This work differs in previously presented one the followings:

- Deal with *NarrativeQA* - a multihop QA dataset - much much harder than *HotpotQA*
- It employs **IAL** as reasoning tool. This module is mainly inspired from attention.

Outline of this section:

1. Overview of system
2. IR block using Curriculum Learning
3. Attention-based IAL block for reasoning
4. Inferring answer with Pointer-Generator Network
5. Result

Fews about NarrativeQA

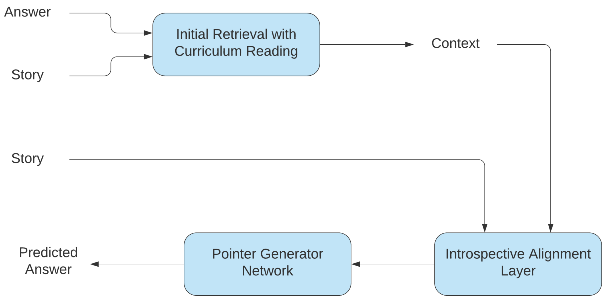
Few characteristics of **NarrativeQA** dataset are worth mentioning:

- Context is very long movie script or book and it is not decomposed into paragraphs beforehand
- Annotators are given summary of each context to produce question and answer and they are encouraged to produce answer in free form $\Rightarrow$ question and answers in many case do not share any common word with context

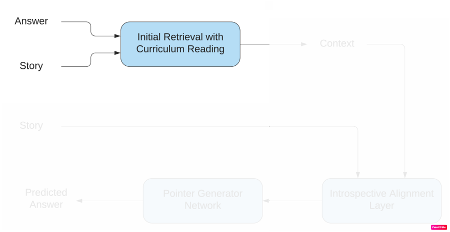
$\Rightarrow$ NarrativeQA is really challenging multihop QA problem.

Overview

This figure illustrates the model.



IR with Curriculum Learning



- This block selects which context is used to train the IAL model in next step
- Curriculum Learning is rather a training strategy than a model

IR with Curriculum Learning: Curriculum Learning

Idea of Curriculum Learning: Model which is trained with easy contexts first and hard contexts are provided gradually produces better result.

> Humans and animals learn much better when the examples are not randomly presented but organized in a meaningful order which illustrates gradually more concepts, and gradually more complex ones. Here, we formal-

In this paper, Curriculum Learning is used in selecting context to train. In other words, different contexts with different characteristics are provided to train the model.

IR with Curriculum Learning: Curriculum Learning

Characteristics of context:

- Where the context is derived from (from question or answer) (**Answerability**)
- How big the context is (50 words, 100 words...) (**Understandability**)

To combine 2 metrics, they use 3 term n, E_n and H_n, where:

n — a set containing several predefined lengths of context $k \in \{50, 100, 200, 500\}$

E_n — a set of questions and corresponding contexts
contexts are inferred from **answer** ; context's length are n words

H_n — a set of questions and corresponding contexts
contexts are inferred from **question** ; context's length are n words

For the details of how to create such sets, they use a *ranking function*.

Pseudo code

Algorithm 1 Curriculum Reading

```
1: chunk_list ← {50, 100, 200, 500}
2: n ← sample i in chunk_list
3: chunk_list ← chunk_list \ {n}
4: E_n ← F(Corpus, Answers, n)
5: H_n ← F(Corpus, Questions, n)
6: D ← E_n                                   ▷ initial training set
7: count ← 0              ▷ number of swaps within a chunk size
8: for i ← 1 to numEpochs do
9:     Train(D)
10:    score ← Evaluate(Dev_set)
11:    if score < bestDev then
12:        if count <= 1/δ then
13:            D ← Swap(D, E_n, H_n, δ)               ▷ Swap δ
   percent of easy set in D with the hard set
14:            count ← count + 1
15:        else
16:            Repeat step 3 to 8       ▷ Replace training set
   with new easy set of another chunk size
17:    else
18:        bestDev = score
```

Figure: Pseudo code of normal training process

Algorithm 1 Curriculum Reading

```
1: chunk_list ← {50, 100, 200, 500}
2: n ← sample i in chunk_list                Set up:
3: chunk_list ← chunk_list \ {n}              - En
4: E_n ← F(Corpus, Answers, n)                - Hn
5: H_n ← F(Corpus, Questions, n)
6: D ← E_n                                   ▷ initial training set
7: count ← 0              ▷ number of swaps within a chunk size
8: for i ← 1 to numEpochs do
9:     Train(D)
10:    score ← Evaluate(Dev_set)              Increase
11:    if score < bestDev then                Answerability
12:        if count <= 1/δ then
13:            D ← Swap(D, E_n, H_n, δ)               ▷ Swap δ
   percent of easy set in D with the hard set
14:            count ← count + 1
15:        else
16:            Repeat step 3 to 8       ▷ Replace training set
   with new easy set of another chunk size
17:    else                                   Increase
18:        bestDev = score                    Understandability
```

Figure: Pseudo code of training process embedding Curriculum Reader

IR with Curriculum Learning: Curriculum Learning

Pseudo code (cont.)

- Swap: In training set D, remove several *easy context* belonging to E_n with *hard context* belonging to H_n
- If swapping time exceeds $1/\delta$, stop swapping and replace with new training dataset containing only *easy context* (i.e. redo step $2 \rightarrow 7$)

Algorithm 1 Curriculum Reading

```
1:  chunk_list ← {50, 100, 200, 500}
2:  n ← sample i in chunk_list
3:  chunk_list ← chunk_list \ {n}
4:  E_n ← F(Corpus, Answers, n)
5:  H_n ← F(Corpus, Questions, n)
6:  D ← E_n                                  ▷ initial training set
7:  count ← 0          ▷ number of swaps within a chunk size
8:  for i ← 1 to numEpochs do
9:      Train(D)
10:     score ← Evaluate(Dev_set)
11:     if score < bestDev then
12:         if count <= 1/δ then
13:             D ← Swap(D, E_n, H_n, δ)          ▷ Swap δ
    percent of easy set in D with the hard set
14:             count ← count + 1
15:         else
16:             Repeat step 3 to 8     ▷ Replace training set
    with new easy set of another chunk size
17:     else
18:         bestDev = score
```

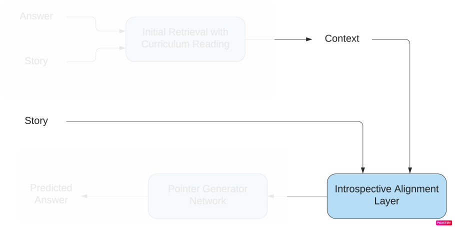
Set up:
- En
- Hn

Increase Answerability

Increase Understandability

Figure: Pseudo code of training process embedding Curriculum Reader

Introspective Alignment Reader



This block aims to find contextual representation of context given raw context and question. Reasoning step occurs within this.

Introspective Alignment Reader

Given:

- Raw context
- Raw question

Output:

- Context matrix $Y \in \mathbb{R}^{l_c \times 2d}$

How to do:

- Input and Context embedding
- Introspective Alignment
- Reasoning over alignments
- Local block-based self-attention

IAL-Reader: Input embedding

It passes context C and question Q in raw form into same BiLSTM layer

$$H^C = BiLSTM(C); H^Q = BiLSTM(Q)$$

where

$$H^C \in \mathbb{R}^{l_c \times d}; H^Q \in \mathbb{R}^{l_q \times d}$$

IAL-Reader: Introspective Alignment

This step creates alignments from question embd and context embd.
Step 1: Co-attention between H^C and H_Q

$$E_{ij} = F(h_i^C)^\top F(h_j^Q)$$

Matrix $E \in \mathbb{R}^{l_c \times l_q}$ is soft matching matrix ; i.e. Affinity matrix in some materials.

Step 2: Learn alignments between context and question

$$A = softmax(E)H^Q$$

Rows of matrix $A \in \mathbb{R}^{l_c \times d}$ are aligned representation of H^C.

IAL-Reader: Reasoning over alignments

To do so, it computes self-attentive reasoning over alignments.

$$G_{ij} = F_s([A_i; H_i^c; A_i - H_i^c, A_i \odot H_i^c])^\top \cdot \\ F_s([A_j; H_j^c; A_j - H_j^c, A_j \odot H_j^c])$$

Note that this calculation is done with index i and j satisfying the following condition:

$$|i - j| \leq b$$

where b is hyperparameter.

The appearance of this condition is to ensure the computation doesn't become prohibitive as $l_c > 2000$.

IAL-Reader: Local Block-based Self-attention

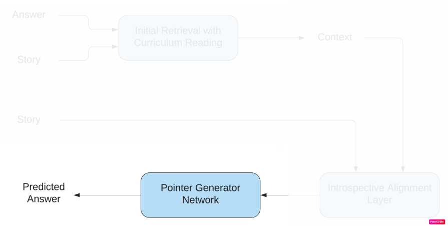
To calculate introspective alignment representation, it uses the following equation:

$$B = \mathrm{Softmax}(G)\,[A; H^c; A - H^c; A \odot H^c]$$

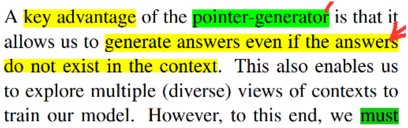
Matrix B above is then passes through *BiLSTM* layer to aggregate final representation of context, say Y:

$$Y = \mathrm{BiLSTM}([B; A; H^c; A - H^c; A \odot H^c])$$

Inferring answer with Pointer Generator Network



Recall that, NarrativeQA dataset encourages generating answer in free from and many questions have no answer in plain text in context. As such, Pointer Generator Network is suitable for this.



A key advantage of the pointer-generator is that it allows us to generate answers even if the answers do not exist in the context. This also enables us to explore multiple (diverse) views of contexts to train our model. However, to this end, we must

For more information about how to apply Pointer Generator Network into this problem, please visit the paper.

Result

Model	ℓ	Dev Set BLEU-1	Dev Set BLEU-4	Dev Set Meteor	Dev Set Rouge	Test Set BLEU-1	Test Set BLEU-4	Test Set Meteor	Test Set Rouge
IR (BLEU)	-	6.73	0.30	3.58	6.73	6.52	0.34	3.35	6.45
IR (ROUGE)	-	5.78	0.25	3.71	6.36	5.69	0.32	3.64	6.26
IR (Cosine)	-	6.40	0.28	3.54	6.50	6.33	0.29	3.28	6.43
BiDAF	-	5.82	0.22	3.84	6.33	5.68	0.25	3.72	6.22
ASR	200	16.95	1.26	3.84	1.12	16.08	1.08	3.56	11.94
ASR	400	18.54	0.00	4.2	13.5	17.76	1.10	4.01	12.83
ASR	1K	18.91	1.37	4.48	14.47	18.36	1.64	4.24	13.4
ASR	2K	20.00	2.23	4.45	14.47	19.09	1.81	4.29	14.03
ASR	4K	19.79	1.79	4.60	14.86	19.06	2.11	4.37	14.02
ASR (Ours)	4K	12.03	1.06	3.10	8.87	11.26	0.65	2.66	8.68
R^3	-	16.40	0.50	3.52	11.40	15.70	0.49	3.47	11.90
RNET-PG	4K	17.74	0.00	3.95	14.56	16.89	0.00	3.84	14.35
RNET-CPG	4K	19.71	2.05	4.91	15.05	19.27	1.45	4.87	15.50
IAL-CPG	4K	**23.31**	**2.70**	**5.68**	**17.33**	**22.92**	**2.47**	**5.59**	**17.67**
Rel. Gain	-	+31%	+51%	+23%	+17%	+20%	+17%	+28%	+26%

Figure: Result of model

Conclusion

- Multihop RC requires more than just search answer keywords from text, it needs reasoning
- Many available reasoning strategies: graph-based, attention-based
- Some datasets (like NarrativeQA) is really difficult

The End