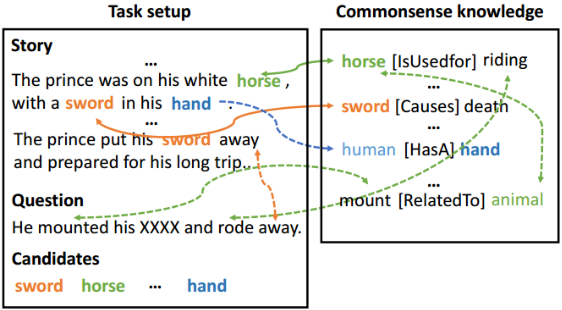
Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge

Guo Tianyi

2018.10.18

Summary

Background knowledge helps a lot in reading comprehension.



Many reading comprehension models don't take external knowledge into account.

This model:

- attends to relevant selected external knowledge
- combines this knowledge with the context representation before inferring the answer

Summary

Symbols

$f_{1...p}$:facts relevant form connecting story, question and candidate answers.
$d_{1...m}$:the story context tokens.
$q_{1...n}$:the question tokens.
$a_{1...k}$:the answer candidates.

Knowledge Retrieval

This model takes Open Mind Common Sence(OMCS, singh et al.) part of ConceptNet as knowledge base. It is a a crowd-sourced resource of commonsense knowledge with a total of ∼630k facts.
Knowledge is represented as a triple:

$$f_i = (\text{subject}, \text{relation}, \text{object})$$

Example:

$$([\text{bow}]_{subj}, [\text{IsUsedFor}]_{rel}, [\text{hunt, animals}]_{obj})$$

The model experiment with three set-ups:

- all facts from OMCS that pertain to ConceptNet, referred to as CN5All
- all facts from CN5All excluding some WordNet relations referred to as CN5Sel(ected)
- facts from OMCS that have *source* set to WordNet (CN5WN3)

Retrieving relevant knowledge

Each sample is like $(D, Q, A_{1\ldots10})$, with ten answer candidates.

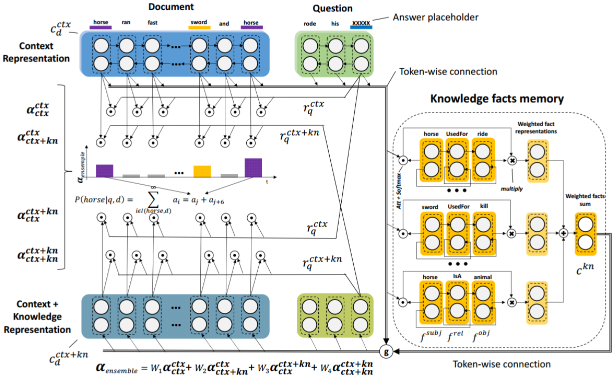
Look up facts from tokens contained in any document, question and answer candidates. Score the fact with some conditions:

- 4 if it contains a lemma of a candidate token from A
- 3 if it contains a lemma from the tokens of Q
- 2 if it contains a lemma from the tokens of D

Weight over the facts by summing the weights of the subject and object arguments. Keep first N facts.

Extending the Attention Sum Reader with a Knowledge Memory

The paper takes Attention-Sum Reader(Kadlec et al.) as their base model and extend it with a knowledge fact memory that is filled with pre-selected facts.



Word Embedding Layer:
looking up the embeddings of document and question tokens:

$$e_{d_{1\dots n}} = Emb(d_{1\dots n})$$
$$e_{q_{1\dots m}} = Emb(q_{1\dots m})$$

Set dropout rate at 0.2
Context Representations:
First encode the tokens with a Bi-directional GRU:

$$c^{\text{ctx}}_{d_{1\dots n}} = \text{BiGRU}(e_{d_{1\dots n}}) \in \mathbb{R}^{n\times 2h}$$
$$c^{\text{ctx}}_{q_{1\dots m}} = \text{BiGRU}(e_{q_{1\dots m}}) \in \mathbb{R}^{m\times 2h}$$

h is the output hidden size of a single GRU unit.

Question Query Representation:
Select the state at the placeholder index pl:

$$r^{\text{ctx}}_{q} = c^{\text{ctx}}_{q_{1\dots m}}[pl] \in \mathbb{R}^{2h}$$

Answer Prediction:
Rank the given answer candidates $a_1 \ldots a_L$ according to the **normalized attention sum score** between the context representation of **the question placeholder** and the representation of the **candidate tokens** in the document:

$$P(a_i|q, d) = \mathrm{softmax}(\sum \alpha_{i_j})$$
$$\alpha_{i_j} = Att(r_q^{\mathrm{ctx}}, c_{d_{1\ldots n}}^{\mathrm{ctx}}), i \in [1, L]$$

where j is an index pointer from the list of indices that point to the candidate occurrences in the document.

Enriching Context Representations with Knowledge (Context+Knowledge).

Knowledge Encoding:
A number of relevant facts are retrieved and represented as a triple:

$$f = (w^{subj}_{1 \ldots L_{subj}}, w^{rel}_0, w^{obj}_{1 \ldots L_{obj}})$$

Using *BiGRU* to encode the triple argument tokens into the following cotext-encoded representations:

$$\begin{aligned} f^{subj}_{last} &= BiGRU(Emb(w^{subj}_{1 \ldots L_{subj}}), 0) \\ f^{rel}_{last} &= BiGRU(Emb(w^{rel}_0), f^{subj}_{last}) \\ f^{obj}_{last} &= BiGRU(Emb(w^{obj}_{1 \ldots L_{obj}}), f^{rel}_{last}) \end{aligned}$$

Querying the Knowledge Memory:
Use Key-Value retrieval(Miller et al.) to query the knowledge memory:

$$c_{s_i}^{kn} = \sum softmax(Att(c_{s_i}^{ctx}, M_{1 \dots P}^{k}))^T M_{1 \dots P}^{v}$$

s in $c_{s_i}^{kn}$ represents d and q.

Combine Context and Knowledge $(ctx + kn)$:
Combine the original context token representation $c_{s_i}^{ctx}$, with the acquired knowledge representation $c_{s_i}^{kn}$ to obtain $c_{s_i}^{ctx+kn}$:

$$c_{s_i}^{ctx+kn} = \gamma c_{s_i}^{ctx} + (1 - \gamma) c_{s_i}^{kn}$$

where $\gamma = 0.5$, and can be replaced with a gating function.

Answer Prediction: Similar to base model, the paper construct normal attention sum score between the context and the candidates:

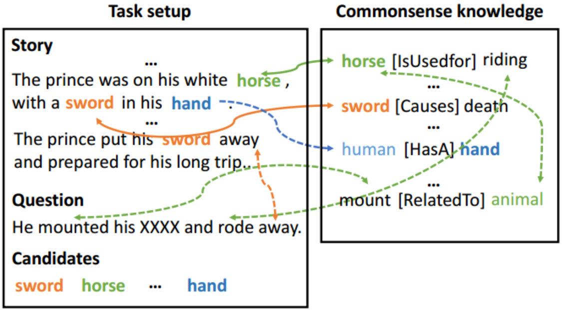
$$
\begin{aligned}
P(a_i|q,d) &= softmax(\sum \alpha_{i_j}^{ensemble}) \\
\alpha_{i_j}^{ensemble} &= W_1 Att(r_q^{ctx}, c_{d_j}^{ctx}) \\
&+ W_2 Att(r_q^{ctx}, c_{d_j}^{ctx+kn}) \\
&+ W_3 Att(r_q^{ctx+kn}, c_{d_j}^{ctx}) \\
&+ W_4 Att(r_q^{ctx+kn}, c_{d_j}^{ctx+kn})
\end{aligned}
$$

j is the same as before. $W_{1...4}$ are scalar weights initialized with 1.0. Paper propose the combination of ctx and $ctx + kn$ attentions because the task does not provide supervision whether the knowledge is needed or not.

Summary

Dataset

Use the *Common Nouns and Named Entities partitions* of the Children's Book Test (CBT) dataset (Hill et al., 2015) .



Hill et al. (2015) show that the human performance when given the full context is at 81.6% since the best neural model (Munkhdalai and Yu, 2016) achieves only 72.0% on the task.

Knowledge Source

CN5ALL:*Open Mind Common Sense* part of ConceptNet 5.0 that contains 630k fact triples.
CN5WN3:the WordNet 3 part of CN5All (213k triples).
CN5Sel:CN5ALL excludes the following WordNet relations: *RelatedTo*, *IsA*, *Synonym*, *SimilarTo*, *HasContext*.

Summary

Model Parameters

Number of facts: 50 by default. 100, 200, 500 also evaluated.
Key-Value Selection Strategy:*Subj/Obj* and *Obj/Obj*, *Subj/Obj* by default.
Answer Selection Components:Use ensemble attention $\alpha_{ensemble}$ by default, referred as *Full model*.
Hyper Parameters:Pretrained GloVe, BiGRU with hidden size 256, batch size of 64 and learning rate of 0.001.

Empirical Results

Knowledge Sources:

Source	**Dev**	**Test**
CN5All	71.40	66.72
CN5WN3(WN3)	70.70	68.48
CN5Sel(ected)	**71.85**	67.64

CN5Sel works best on the *Dev* set but *CN5WN3* works much better on *Test*. Further experiments use the *CN5Sel* setup.

Number of facts:

facts	**50**	**100**	**200**	**500**
Dev	**71.85**	71.35	71.40	71.20
Test	67.64	67.44	**68.12**	67.24

The best result on the *Dev* set is for 50 facts.

Empirical Results

Component ablations:

	NE		CN	
D_{repr} **to** Q_{repr} **interaction**	**Dev**	**Test**	**Dev**	**Test**
D_{ctx}, Q_{ctx} (w/o know)	75.50	70.30	68.20	64.80
D_{ctx+kn}, Q_{ctx+kn}	76.45	69.68	70.85	66.32
D_{ctx}, Q_{ctx+kn}	**77.10**	69.72	70.80	66.32
D_{ctx+kn}, Q_{ctx}	75.65	**70.88**	71.20	**67.96**
Full model	76.80	70.24	**71.85**	67.64
w/o D_{ctx}, Q_{ctx}	75.95	70.24	70.65	67.12
w/o D_{ctx+kn}, Q_{ctx+kn}	76.20	69.80	70.75	67.00
w/o D_{ctx}, Q_{ctx+kn}	76.55	70.52	71.75	66.32
w/o D_{ctx+kn}, Q_{ctx}	76.05	70.84	70.80	66.80

The combination of different interactions between *ctx* and *ctx+kn* representations leads to clear improvement over the w/o knowledge setup.

Empirical Results

Key Value Selection Strategy:

	NE		CN	
Key/Value	**Dev**	**Test**	**Dev**	**Test**
Subj/Obj	76.65	75.52	71.85	67.64
Obj/Obj	76.70	71.28	71.25	67.48

Empirical Results

Comparison to Previous Work:

Models	NE dev	NE test	CN dev	CN test
Human (ctx + q)	-	81.6	-	81.6
Single interaction				
LSTMs (ctx + q) (Hill et al., 2015)	51.2	41.8	62.6	56.0
AS Reader	73.8	68.6	68.8	63.4
AS Reader (our impl)	75.5	70.3	68.2	64.8
KnReader (ours)	77.4	71.4	71.8	67.6
Multiple interactions				
MemNNs (Weston et al., 2015)	70.4	66.6	64.2	63.0
EpiReader (Trischler et al., 2016)	74.9	69.0	71.5	67.4
GA Reader (Dhingra et al., 2017)	77.2	71.4	71.6	68.0
IAA Reader (Sordoni et al., 2016)	75.3	69.7	72.1	69.2
AoA Reader (Cui et al., 2017)	75.2	68.6	72.2	69.4
GA Reader (+feat)	77.8	72.0	74.4	70.7
NSE (Munkhdalai and Yu, 2016)	77.0	71.4	74.3	71.9

Summary

- The paper propose a neural cloze-style reading comprehension model that incorporates external commonsense knowledge, building on a single-turn neural model.
- Incorporating external knowledge improves its results with a relative error rate reduction of 9% on Common Nouns, thus the model is able to compete with more complex RC models.
- it shows that the types of knowledge contained in ConceptNet are useful.
- The model can be adapted to very different task settings.