Improving Question Answering with External Knowledge

Authors: Xiaoman Pan*, Kai Sun*, Dian Yu, Heng Ji and Dong Yu

Presented by: Shichen Liu
Background and Motivation

- Multiple-choice machine reading comprehension tasks require diverse types of knowledge. Recent works leverage rich world knowledge by pre-training deep neural models. However, this could be time-consuming and resource-extensive.
Contribution

- This work evaluates the effects of utilizing external knowledge during fine-tuning stage with:

  1. augmenting training data by using external in-domain question answering datasets

  2. enriching reference corpora by retrieving additional knowledge from external open-domain resources
Base Framework

• Given question Q, answer option O, and reference
document D, feed into BERT

\[\text{[CLS]} \ (D.+[A]) \ [SEP] \ (Q.+[B]) \ [SEP] \ (O.+[B]) \ [SEP]\]

where [A], [B], [CLS] and [SEP] are tokens from BERT

• The final prediction is obtained by a linear layer and a
softmax layer over the hidden states of input sequences.

• The model is first fine-tuned on RACE and then fine-tuned
on the specific dataset (ARC and OpenBookQA)
Using In-Domain Data

- The two step design of the framework naturally integrate the external in-domain data (i.e. RACE dataset).
- The paper investigate using multiple in-domain resources as external knowledge.
Using Open-Domain Data

• The paper propose to use **entity discovery and linking** (EDL) techniques to learn from open-domain resources.

• Entity discovery: simply consider all noun phrases as entity mentions.

• Entity linking: 1) generate an initial list of **candidate entities**; 2) rank them to select the candidate entity based on **salience**, **similarity**, and **coherence**.
Using Open-Domain Data

• Generate initial candidate list $E_m$ for entity mention $M$:
  \[ M = \{m_1, m_2, \ldots, m_n\} \quad E_m = \{e_1, e_2, \ldots, e_n\} \]

Based on commonness:

\[ F_{mention}(e|m) = \frac{A_{m,e}}{A_{m,*}} \]

Where $A_{m,e}$ is the subset in Wikipedia links where entity mention $m$ points to entity $e$. 
Using Open-Domain Data

• Salience

\[ F_{prior}(e) = \frac{A_{*,e}}{A_{*,*}} \]

• Similarity: context similarity between mention-entity pairs. Learned with neural networks using cosine similarity

\[ F_{sim}(m, e) \]
Coherence: if multiple entity mentions appear together within a sentence, their referent entities are more likely to be coherent.

First we construct a graph $G=(E, D)$, $E$ is all entities and $D$ is the Jaccard similarity with weight:

$$w_{ij} = \frac{|p_i \cap p_j|}{\max(|p_i|, |p_j|)}$$

Then we learn the graph embedding with these weights. The coherence of two entities $coh(e_i, e_j)$ is achieved by cosine similarity between their embeddings.
Using Open-Domain Data

• The coherence score of candidate entity e is:

\[ F_{coh}(e) = \frac{1}{|C_m|} \sum_{c \in C_m} coh(e, c) \]

where \( C_m \) is the union of all candidate entities for coherent mentions of \( m \)

• We combine the salience, similarity and coherence scores to compute the final score.
• Consistent improvements could be observed in accuracy across all tasks after we apply EDL to enrich the reference document for each question.
• Case study of EDL:

• Q: Which of the following statements best explains why magnets usually stick to a refrigerator door?
   A1: “The refrigerator door contains iron.”
   A2: “The refrigerator door is a good conductor.”
   A3: “The refrigerator door has electric wires in it.”
   A4: “The refrigerator door is smooth.”

• EDL links the mention “magnets” to its corresponding Wikipedia entry Magnet and attach its description in Wikipedia “A magnet is a material or object that produces a magnetic field. This magnetic field is invisible but is responsible for the most notable property of a magnet: a force that pulls on other ferromagnetic materials, such as iron, and attracts or repels other magnets.” after its reference document.
Conclusion

- In this work, we study improving question answering by utilizing in-domain external question answering datasets and utilizing open-domain external corpora to enrich the reference corpus.

- Preliminary experimental results on ARC and OpenBookQA datasets demonstrate the effectiveness of our proposed approaches.
Questions
Thanks!