An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge

Authors: Hao et al.

Presenter: Shivank Mishra

Link to complete paper: https://aclweb.org/anthology/P/P17/P17-1021.pdf
What is Knowledge base?

• It is a special type of database system

How is it special?

• It uses AI and data within it to give answers and not just some data
Question Answering

• We use it to build systems that automatically answer questions posed by humans in natural language [1]
  • Input: Natural Language Query
  • Output: Direct Answer

[1] https://en.wikipedia.org/wiki/Question_answering
Why QA when there are other ways to search?

• Keyword Search:
  • Simple information needs
  • Vocabulary redundancy

• Structured queries:
  • Demand for absolute precision
  • Small & centralized schema

• QA:
  • Specification of complex information needs
  • Schema-less data
Outline

• Introduction
  • High level view
• Existing Research
• Prior Issues
• Overview of KB-QA system
• Solution
• Model Analysis
• Results
• Error Analysis
• Conclusion

“First, I want to give you an overview of what I will tell you over and over again during the entire presentation.”
Introduction

- This paper presents:
  - A novel cross-attention based Neural Network model for Knowledge Base – Question Answering (KB-QA).
  - Reduces the Out Of Vocabulary problem by using Global Knowledge Base.
Introduction - High level view

• Design an end-to-end neural network model to represent the questions and their corresponding scores dynamically according to the various candidate answer aspects via cross-attention mechanism.
Existing Research

• Emphasis on learning representations of the answer end
  • Subgraph for candidate answer, Bordes et. al 2014a
  • Question -> single vector, bag-of-words, Bordes et. al 2014b
    • Relatedness of answer end has been neglected
  • Context and type of the answer, Dong et. al., 2015
Dong et al (2015)

• Use three CNNs for different answer aspects:
  • Answer path
  • Answer context
  • Answer type

• However, keeping only three independent CNNs has made the model mechanical and inflexible

• Therefore the authors decided to propose a cross-attention based neural network
Prior Issues

1) The global information of the KB is deficient
   • Entities and relations – KB resources are limited

2) out-of vocabulary (OOV) problem
   • Many entities in testing candidate have never been seen.
     • Attention of resources become same due to common OOV embedding
Overview of KB-QA system

- Identify topic entity of the question
- Generate candidate answer from Freebase
- Run a cross-attention based neural network to represent Question under the influence of Answer
- Rank the answers by score
- Highest score gets added to the set
Cross-attention based neural network architecture
Solution

• Incorporate Freebase KB itself as training data with Q&A pairs
  • Ensure that the global KB information acts as additional supervision, and the
    interconnections among the resources are fully considered.

• The Out Of Vocabulary problem is relieved.
Overall Approach

• Candidate Generation
• Neural Cross-Attention Model
  • Question Representation
  • Answer aspect representation
  • Cross-attention model
    • A-Q attention
    • Q-A attention
• Training
• Inference
  • Combining Global Knowledge
Candidate Generation

• Utilize Freebase API to identify topic of the question
  • Use top1 result (Yao and Van Durme, 2014) to get 86% correct results
  • Get topic entity connected with that one hop, called two hop.
Cross-Attention Model

“re-reading” mechanism to better understand the question.

• Judge candidate answer:
• Look at answer type
• re-read question
• Look where should the attention be
• Go the next aspect
• re-read question
• ..... 
• Read all answer aspects and get weighted sum of all scores
Cross Attention

- **Question-towards-answer attention**

- \( \beta_{e_i} = \) Attention of question towards answer aspects in one \((q, a)\) pair

\[
W \in \mathbb{R}^{2d \times d}
\]

\( W \) is the intermediate matrix for Q-A attention is pooling all the bi-directional LSTM hidden state sequence.

\( \overline{q} \) Result = vector that represents the question to determine which aspect of question should be more focused.
Cross Attention

- **Answer-towards-question attention**
  - Helps learn question-answer weight
  - Extent of attention can be measured by the relatedness between each word representation $h_j$
  - Answer aspect embedding $e_i$.
  - $\alpha_{ij}$ denotes the weight of attention from answer aspect $e_i$ to the $j$th word in the question, where $e_i \in \{e_e, e_r, e_t, e_c\}$.
  - $f(\cdot)$ is a non-linear activation function, such as hyperbolic tangent transformation here.
  - $n$ is the length of the question
  - $W$ is the intermediate matrix
  - $B$ is offset
  - $q$ is the question
Question Representation

- Question $q = (x_1, x_2, \ldots, x_n)$, $x_i$ is the $i$th word
- $E_w \in \mathbb{R}^{d \times V_w}$
  - Let $E_w$ be the word embedding matrix
  - $d =$ dimension of embeddings
  - $V_w =$ vocabulary size of natural language words

- Word embedding are fed into LSTM (good for harnessing long sentences)
  - Use bidirectional LSTM to forward and backward of a word $x_i$
    - Read question Left -> Right
    - Read question Right -> Left
Answer Retention

- Use KB embedding matrix $E_k \in \mathbb{R}^{d \times v_k}$
- $V_k =$ vocabulary size; $d =$ dimension
- $a_e =$ answer entity
- $a_r =$ answer relation
- $a_t =$ answer type
- $a_c =$ answer context (can contain multiple KB resources)
- Similarly we have embedding aspects

\[
\left( e_{c_1}, e_{c_2}, \ldots, e_{c_m} \right)
\]

Average embedding:
\[
e_c = \frac{1}{m} \sum_{i=1}^{m} e_{c_i}
\]
Training

Training Loss, hinge loss

\[ L_{q,a,a'} = [\gamma + S(q,a') - S(q,a)]_+ \]

Objective function

\[
\min \sum_q \frac{1}{|P_q|} \sum_{a \in P_q} \sum_{a' \in N_q} L_{q,a,a'}
\]

SGD to minimize loss, with mini-batch sizes

Inference

• We need to get maximum similarity, \( S_{\text{max}} \)

• \( S(q,a) \) for each \( a \) that is part of candidate answer set \( C_q \)

\[
S_{\text{max}} = \arg \max_{a \in C_q} \{ S(q,a) \}
\]

• Use margin \( \gamma \) if there is more than 1 answer

• If the score of candidate answer is within margin \( \gamma \) v/s \( S_{\text{max}} \)
  • Add to the final answer set
Combining Global knowledge

• Adopt the TransE model (translation in embedding space) (like Bordes et al., 2013)
• Train both KB-QA and TransE models together
• e.g. Facts are subject-predicate-object triples \((s, p, o)\)
  • (/m/0f8l9c, location.country.capital,/m/05qtj)
  • France, relation, Paris
• \((s', p, o')\) are the negative examples
• Completely unrelated facts are deleted
• Training loss (S is set of KB & S’ is set of corrupted facts)

\[
L_k = \sum_{(s, p, o) \in S} \sum_{(s', p, o') \in S'} [\gamma_k + d(s + p, o) - d(s' + p, o')]_+
\]
Experiments

• Use WebQuestions (Google Suggest API)
  • 3778 QA pairs for training
  • 2032 pairs for testing
• Answers (from Freebase) are labeled manually by AMT
• Training data: ¾ training set, rest – validation set
• F1 score is used as the evaluation metric
• Average result is computed by script from Berant et al. (2013)
Settings

• KB-QA training:
  • Mini-batch SGD to reduce pairwise training loss
    • Mini-batch = 100
    • Learning rate = 0.01
    • $E_w$ (word embedding matrix) $E_v$ (KB embedding matrix are normalized after every epoch)
    • Embedding size $d = 512$
    • Hidden unit size = 256
    • Margin $\gamma = 0.6$
Model Analysis

| Methods               | Avg $F_1$ |
|-----------------------|-----------|
| LSTM                  | 38.2      |
| Bi.LSTM               | 39.1      |
| Bi.LSTM+A-Q-ATT       | 41.6      |
| Bi.LSTM+C-ATT         | 41.8      |
| Bi.LSTM+GKI           | 40.4      |
| Bi.LSTM+A-Q-ATT+GKI   | 42.6      |
| Bi.LSTM+C-ATT+GKI     | 42.9      |

Table 2: The ablation results of our models.

Figure 3: The visualized attention heat map. Answer entity: /m/06npd(Slovakia), answer relation: partially_containedby, answer type: /location/country, answer context: (/m/04d9kf, /m/01mp, ...
Results

Comparison of our method with state-of-the-art end-to-end NN-based methods

| Methods         | Avg $F_1$ |
|-----------------|-----------|
| Bordes et al., 2014b | 29.7      |
| Bordes et al., 2014a | 39.2      |
| Yang et al., 2014 | 41.3      |
| Dong et al., 2015 | 40.8      |
| Bordes et al., 2015 | 42.2      |
| **our approach** | **42.9**  |

The evaluation results on WebQuestions
Error Analysis

• Wrong attention
  • Q: “What are the songs that Justin Bieber wrote?”
  • A: answer type /music/composition pays the most attention on “What” rather than “songs”.

• Complex questions
  • Complex Q: “When was the last time Arsenal won the championship?”
  • A: Prints all championships. - model did not train with “last”

• Label Error:
  • Q: “What college did John Nash teach at?”
  • A: prints Princeton University, but misses Massachusetts Institute of Technology
Conclusion

• Proposed a novel cross-attention model for KB-QA
  • Utilized Q-A and A-Q attention
  • Leveraged the global KB information to alleviate the OOV problem for the attention model

• The experimental results proved to give better performances than the current state of the art end-to-end methods
Thank you