Notes for "Question Answering with Subgraph Embeddings" paper

Question Answering with Subgraph Embeddings.md

Question Answering with Subgraph Embeddings

Introduction

- Open-domain Question Answering (Open QA) - efficiently querying large-scale knowledge base(KB) using natural language.
- Two main approaches:
  - Information Retrieval
    - Transform question (in natural language) into a valid query(in terms of KB) to get a broad set of candidate answers.
    - Perform fine-grained detection on candidate answers.
  - Semantic Parsing
    - Interpret the correct meaning of the question and convert it into an exact query.
- Limitations:
  - Human intervention to create lexicon, grammar, and schema.
- This work builds upon the previous work where an embedding model learns low dimensional vector representation of words and symbols.
- Link to the paper.

Task Definition

- Input - Training set of questions (paired with answers).
- KB providing a structure among the answers.
- Answers are entities in KB and questions are strings with one identified KB entity.
- The paper has used FREEBASE as the KB.

- Datasets
    - WebQuestions - Built using FREEBASE, Google Suggest API, and Mechanical Turk.
    - FREEBASE triplets transformed into questions.
    - Clue Web Extractions dataset with entities linked with FREEBASE triplets.
    - Dataset of paraphrased questions using WIKIANSWERS.

Embedding Questions and Answers

- Model learns low-dimensional vector embeddings of words in question entities and relation types of FREEBASE such that questions and their answers are represented close to each other in the joint embedding space.
- Scoring function $S(q, a)$, where q is a question and a is an answer, generates high score if a answers q.
    - $S(q, a) = f(q)^T.g(a)$
    - $f(q)$ maps question to embedding space.
    - $f(q) = W\varphi(q)$
    - W is a matrix of dimension $K * N$
    - K - dimension of embedding space (hyper parameter).
    - N - total number of words/entities/relation types.
    - $\psi(q)$ - Sparse Vector encoding the number of times a word appears in q.
- Similarly, $g(a) = W\psi(a)$ maps answer to embedding space.
- *&psi(a)* gives answer representation, as discussed below.

Possible Representations of Candidate Answers

- Answer represented as a **single entity** from FREEBASE and TBD is a one-of-N encoded vector.
- Answer represented as a **path** from question to answer. The paper considers only one or two hop paths resulting in 3-of-N or 4-of-N encoded vectors(middle entities are not recorded).
- Encode the above two representations using **subgraph representation** which represents both the path and the entire subgraph of entities connected to answer entity as a subgraph. Two embedding representations are used to differentiate between entities in path and entities in the subgraph.
- SubGraph approach is based on the hypothesis that including more information about the answers would improve results.

Training and Loss Function

- Minimize margin based ranking loss to learn matrix *W*.
- Stochastic Gradient Descent, multi-threaded with Hogwild.

Multitask Training of Embeddings

- To account for a large number of synthetically generated questions, the paper also multi-tasks the training of model with paraphrased prediction.
- Scoring function $S_{prp}(q1, q2) = f(q1)^T f(q2)$, where *f* uses the same weight matrix *W* as before.
- High score is assigned if *q1* and *q2* belong to same paraphrase cluster.
- Additionally, the model multitasks the task of mapping embeddings of FREEBASE entities (mids) to actual words.

Inference

- For each question, a candidate set is generated.
- The answer (from candidate set) with the highest set is reported as the correct answer.
- Candidate set generation strategy
    - **C_1** - All KB triplets containing the KB entity from the question forms a candidate set. Answers would be limited to 1-hop paths.
    - **C_2** - Rank all relation types and keep top 10 types and add only those 2-hop candidates where the selected relations appear in the path.

Results

- **C_2** strategy outperforms **C_1** approach supporting the hypothesis that a richer representation for answers can store more information.
- Proposed approach outperforms the baseline methods but is outperformed by an ensemble of proposed approach with semantic parsing via paraphrasing model.