Cornell Notes

20000512

Multi-hop question answering over knowledge graphs

1. What is the problem expected to be solved?	* The paper aims to address the limited information grasp of the questions in the SOTA models for multi-hop question answering which results in reducing the accuracy of answers, by using a novel approach for multihop question answering over KG mainly utilizing question and path embedding strategy.
2. What are knowledge graphs?	* Knowledge graphs are a way of organizing knowledge with entitres as nodes and their relationships as edges
3. What's a hop? Where are multi-hops needed?	* Hop is the distance between 2 entities in KGs. Complex queries needs to traverse multiple hops joining entities to derive at answers
4. What are the steps followed in the approach? (High level)	* In the approach used, question is decomposed into subject-verb-object triples, optimal subgraphs are extracted, question and subgraph path embeddings are created and the best matching answer is selected by comparing the embeddings.
5. Previous approaches for the steps included in the methodology.	* Question decomposition, Subgraph extraction, Generation of path and sentence embeddings, Selecting the answer from potential options criteria is reviewed in detail with regards to the previous researches and findings in the paper.
6. How KG is structured and the intention?	* Knowledge graph contains SVO tiples (h, r, t). The object entity e_t is to be extracted from question q, that contains head entity e_h.
7. How the steps in the methodology are carried out?	* Question q in natural language is decomposed into SVO triples to identify head entity (e_h) and relationship (R). * Then the subgraph is extracted from the KG using these entities as key identifiers. * Sentence embedding (S_q) and path embeddings $P_{h_i}; i = 0..n$ is obtained for paths in the extracted subgraph. These embeddings are fixed size vectors capturing the semantic meanings utilizing pretrained BERT language model embeddings and a multihead attention encoder. * Out of all path embeddings (P_{h_i}) in the subgraph extracted, the nearest match is selected using L_1 distance metric with the sentence embedding (S_q). For multi hop questions the selected path leads answering to the question. * A sigmoid function is applied to the L1 layer, and a threshold is set to identify viable answers.
8. What is the used objective function?	* Binary cross entropy is used as the objective function, and the final layer of the model uses the sigmoid function in combination with another vector and the multi hot encoding scheme to compute the prediction and target set binary cross entropy.
9. Key facts of the experimental procedure	* The MetaQA dataset which contains a knowledge graph dataset and a QA dataset was trained using ComplexSQ model. * The QA dataset reading section is categorized into 3 types as MetaQA 1Hop, MetaQA 2Hop and MetaQA 3Hop, following a consistent head-body-tail structure. * Path dictionaries were used to map head entities to their outgoing links in subgraph extraction * The tail is transformed into one-hot representation of size 43,234. * Utilized batching to manage huge amount of entities.
10. What are the results of the experiments and the areas of improvement?	* Experiments were performed on WebQSP and MetaQA datasets and for the evaluation Hits at N (H@N) metric was employed. * The model outperformed SOTA models with high scores of 83.8%, 98.0%, 62.0% for 1-3 hop questions. * For 50% of the data intentianally removed, the model scored 91.4%, 90.8% and 63.9% for 1-3 hop categories. * As future work, WebQuestionSP dataset can be used to improve and generalize the model.
11. Summary	This research focuses on improving the accuracy of Multi-hop question answering over knowledge graphs using a five step approach wit which contains question decomposition, subgraph extraction, question embedding creation, path embedding creation and optimal path seletion using L1 distance metric. With the techniques used, the model was able to outperform SOTA models in terms of accuracy. Due to time and resource limitations, the author provides insights to further improve and generalize the model using WebQuestionSP dataset