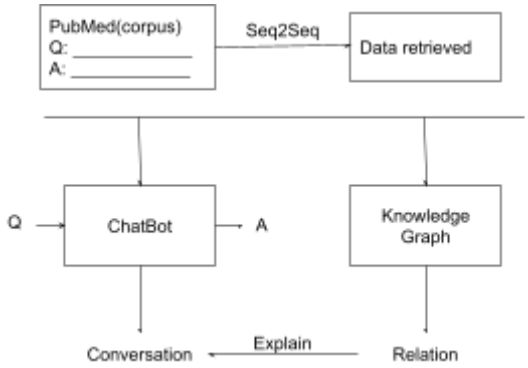
CS505 Final Project Milestone

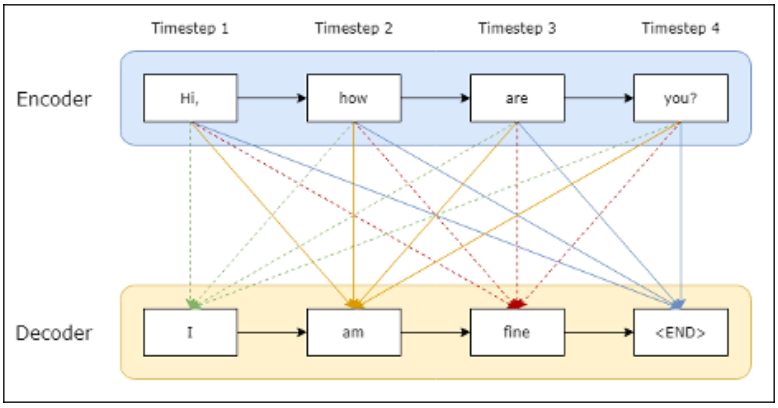
Members: Yuchen Lu, Zhengxu Wang, Melissa Zhen

Description:

We are going to make a chatbot using data from PubMed and other open source QA based data as our corpus. A summary of our idea is shown in the graph on the right. In doing so, we will make it easier for people to find the answers to their questions without having to manually look through the Q&A's which will increase efficiency. We chose to use medical data because it has high importance and is relevant to daily life. We decided on this project because we are interested in the topic of chatbots and knowledge graphs. After learning about neural networks in lecture, we would like to practice working more with LSTM and GRU models.

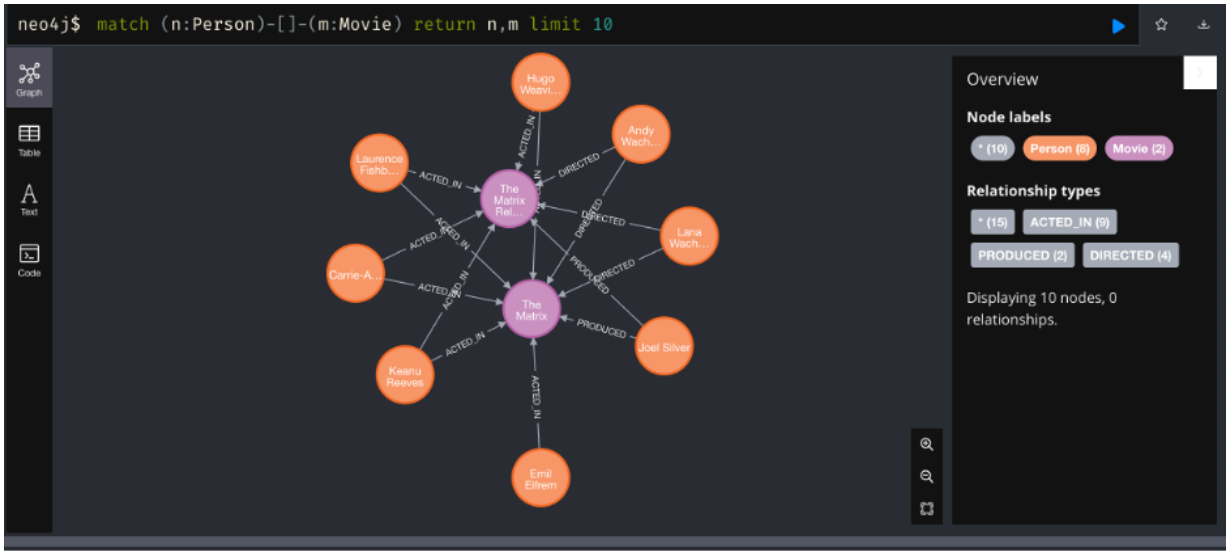


The Seq2Seq model mainly includes two basic blocks, Encoder and Decoder. Among them, the encoder is responsible for compressing the input sequence into a vector of specified length, and the network structure is a two-layer bidirectional GRU model; while the decoder is responsible for generating the specified sequence according to the semantic vector. The process, also known as decoding, is a two-layer unidirectional GRU model. And because in the Seq2seq model, the encoding process of the original encoded model will generate an intermediate vector C, which is used to save the semantic information of the original sequence. However, the length of this vector is fixed. When the original input sequence length is relatively long, vector C cannot save all the semantic information, and the context semantic information is limited, which also limits the understanding ability of the model. So use the Attention mechanism to break the limitation of the original encode model on the fixed vector. Or use the Tf-idf to simply keep the representative features.



As shown in the diagram above, it is an example of using the Seq2Seq model and the GRU/LSTM to generate the answering sentence based on training corpus.

For the knowledge graph part we plan to use the Neo4j graphical database, because it has a user-friendly interface of relationship searching sql and has a visualization tool called Bloom. Like the screenshot below.



We plan to use some pre-trained model or same seq2seq model to do Named Entity Recognition from the same QA corpus that we use for training the chatbot. Then we can get the entity and relation between these entities. Restore these data into Neo4j to get a knowledge graph. When chatting with the chatbot, we can use the graph to explain why our chatbot has these results for questions. This can be used to verify and evaluate whether the chatbot gets an answer based on correct knowledge.

Resources:

Graph Database Applications and Concepts with Neo4j - Scinapse. https://asset-pdf.scinapse.io/prod/776871969/776871969.pdf.

Hussain, Mustaffa. “Knowledge Graph-Based Chatbot.” Medium, TheCyPhy, 31 Oct. 2020, https://medium.com/thecyphy/knowledge-graph-based-chatbot-5416a79d7f17.

Shah, Dhruvil. “Generative Chatbots Using the seq2seq Model!” Medium, Towards Data Science, 28 July 2020, https://towardsdatascience.com/generative-chatbots-using-the-seq2seq-model-d411c8738ab5.

Manideep, Vvs. “Interview Chat Bot Using SEQ2SEQ Model.” Medium, Medium, 2 Aug. 2020, https://vvsmanideep.medium.com/interview-chat-bot-using-seq2seq-model-fe9059fffe6

Project Plan:

We are going to use open source data from PubMed that are grouped by questions and answers. Seq2seq model will be used to help retrieve features that we would like to use to feed into neo4j, a platform that helps generate knowledge graphs with processed data. The same dataset will be used to train both the chatbot and the NER model. The knowledge graph will be used to explain the answers given by the chatbot with a question as the input. The project will mostly be organized using github and run on SCC.

Here is our divide work plan. (Just temporarily the plan which will change during the process.)

A: Install the Neo4j platform and learn how to build a graph using triplet data prepare for KG constructing. Find a strategy for doing the NER job on QA corpus to generate triplets.

B: Build the Seq2seq model with GRU to make the chatbot work.

C: Help B work on chatbot first and try to use LSTM on Seq2seq. Help A work on NER and test the chatbot and knowledge graph result.