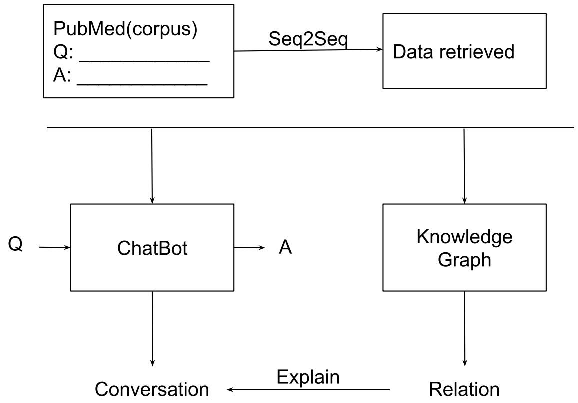
Knowledge Graph Based Chatbot

Zhengxu Wang, Yuchen Lu, Melissa Zhen

code @ github/google drive[1]

ABSTRACT

Chatbots (e.g. Buzz on BU websites) are used for giving automatic responses to the user's input questions on various websites. This will make it easier for people to find answers to their questions without having to manually look through the Q&A's, which will increase efficiency. A brief outline of our project is shown below.

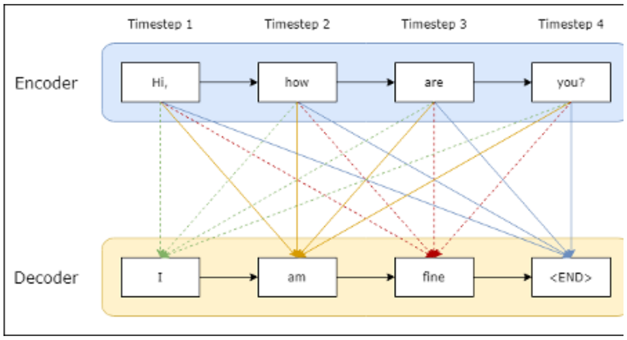


Our team decided to use medical data for our corpus to train a medical chatbot because it is a useful and highly relevant topic in today's society. We trained a chatbot using data from PubMed and other open-source Q&A based-data as our corpus. After comparing the accuracy of various pre-trained models for capturing named entity relations, we found that Bert-Based-NER, which utilizes transformers, performed better than other open-source libraries such as NLTK and Spacy. Thus, Bert-Based-NER was used to extract NERs from the corpus to generate a list of tuples to feed into Neo4j, which created a knowledge graph. Neo4j is a platform that helps generate knowledge graphs with processed data. The resulting graph can be used to explain the reasoning behind the corresponding conversation between the chatbot and the user.

1 INTRODUCTION

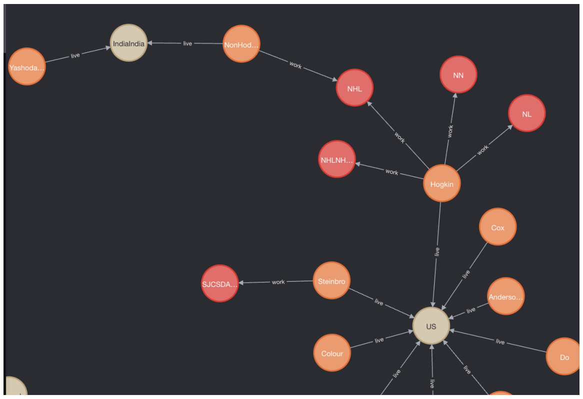
The Seq2Seq model mainly includes two basic blocks, Encoder and Decoder. Among them, the encoder is responsible for compressing the input sequence into a vector of specified length, and the network structure is a two-layer bidirectional GRU model; while the decoder is responsible for generating the specified sequence according to the semantic vector. The process, also known as decoding, is a two-layer unidirectional GRU model. 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 we can either use the Attention mechanism to break the limitation of the original encoded model on the fixed vector, or use Tf-idf to simply keep the representative features.

[1] Average shared workload with Zhengxu guiding the work. Specifically, *A* built the Seq2seq model with GRU to make the chatbot work. *B* helped *A* work on chatbot first and try to use LSTM on Seq2seq and then helped *C* work on NER and tested the chatbot and knowledge graph result. *C* installed the Neo4j platform and learned how to build a graph using triplet data prepared for knowledge constructing. Find a strategy for doing the NER job on QA corpus to generate triplets.



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 as shown in the screenshot below.



We plan to use some pre-trained model or same seq2seq model to do Named Entity Recognition from the same Q&A 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.

2 APPROACH

We now describe the implementation of the chatbot, generating the triplets, and building the knowledge graph for evaluation.

2.1 Chatbot

The chatbot is a highly thematic system, which requires a professional corpus that has a high coverage of medical question and answer pairs. An appropriate amount of daily conversation corpus was added to make sure the smoothness of the conversation. In addition, in order to make sure that answers are returned in real-time, an optimizer was added to improve the efficiency of the algorithm.

Knowledge Graph Creation for Chatbot In order to create a training set, we produced a database, including jokes, stories, and some capitalized frequent words. These capitalized frequent words were used to avoid transforming all words into lower cases when tokenizing the inputs. Jokes and stories were stored in the form of name-to-content triplets.

Build the Training Sets Since there is no open source training set available that meets our needs, we implemented scripts to preprocess the raw data and retrieve words to achieve the same goal. Corpora used and their size are shown in the table below.

Weight	Corpus	Size (kb)
augment0	cornell_cleaned_new	5177
augment0	misc_new	15
augment0	reddit_cleaned_part	19136
augment0	scenarios_new	82
augment0	med_advice	103968
augment1	rule1 + papaya	88
augment2	rule 2-5	21

We used mini-batches and data augmentation to train different corpora in different files so that they can be assigned with different weights. For example, medical terms would be trained with the least weight. It is unrealistic for the training accuracy to reach 100% and thus errors will occur. However, due to the highly specialized nature of medical terminology, small errors can lead to completely different meanings for different semantics.

In contrast, corpora that were assigned large weights would not have such a problem since after several iterations of rearranging the order of words and training, sentences in everyday conversations that are not rigorous would not cause comprehension biases.

This kind of training technique allows the chatbot to accurately understand the question and thus provide a better answer. As we can see in the previous table, corpora that are highly specialized are assigned to files with lower weights and vice versa.

We used a corpus from PudMed–a certified professional medical corpus–to feed the specialized knowledge into the chatbot in order to make sure that the chatbot is able to provide professional medical advice after training.

It is worth mentioning that we used variables instead of specific contents when the training data involves information such as time, name and arithmetics. Such information usually appears in daily conversations frequently while the chatbot cannot learn how to answer through training. In order to make implementation easier, such information was replaced by variables started with underscores. In the training stage, if there is a variable in such format, the variable and its value will be stored. Then later in the prediction stage, the actual value of the corresponding variable will be retrieved and printed in the terminal instead of a fixed value. In this way, the chatbot is given a "short-term memory" to remember some of the conversation. For example, when being asked "what is 3 + 4", what is actually stored is "what is *_num1_* + *_num2_*" with *_num1_*=3 and *_num2_*=4. In order to answer this question, the chatbot calls a function that takes *_num1_* and *_num2_* as the parameters and does the addition.

Model The chatbot was trained with the Seq2Seq model and is able to predict according to the user-input questions. Seq2Seq model is a classic solution that applies to any task that involves vary-length output, including training chatbot.

Steps can be summarized as follows. First, create output tensors to store all the predictions. Then, create a source sequence as the input to the encoder to store vectors of contexts. Use the beginning-of-sentence tag as the first input to the decoder. Decode as needed until the end-of-sentence tag.

Training As mentioned in the previous section, the Seq2Seq model consists of two RNNs where RNN can learn the probability distribution and predict accordingly. The chatbot utilized the prediction ability of RNNs to predict the answer based on the user-input question. RNN uses *softmax* as an activation function in a neural network model to get the probability distribution, where *softmax* is defined as

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

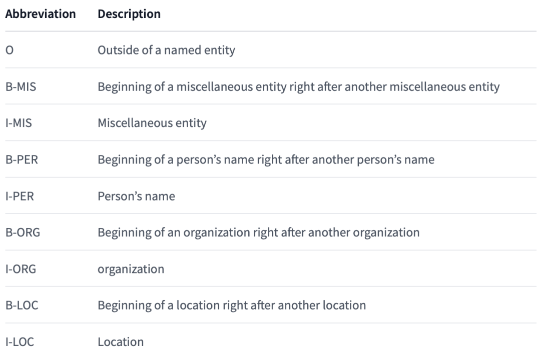
where e^{zi} is the standard exponential function for input vector and output vector, K is the number of classes in the multi-class classifier, and e^{zj} is the standard exponential function for output vector. Specifically, *softmax* function is used to normalize the outputs, converting them from weighted sum values into probabilities that sum to one for easier understanding and comparison.

Optimizer The chatbot used *Adam* optimizer, directly applying *adamOptimizer()* API of tensorflow. We picked this optimizer since *Adam* is computationally efficient and requires less memory, which makes it easier to run the script with limited resources. Different from classical stochastic gradient descent, *Adam* maintains a single learning rate alpha for all weight updates and the learning rate does not change during training. *Adam* provides an optimization algorithm that can handle sparse gradients on noisy problems.

2.2 Generating the Knowledge Graph

We first generated csv files from the medical data corpus where each line is a triplet in the form of (word, named entity recognition, index of sentence that the word belongs to). Then, the file is processed and then fed into Neo4j to help better visualize the relations.

Triplets Generation In order to create the list of triplets, we compared and contrasted four different pretrained NER models: NLTK, Spacy with en_core_web_sm, Spacy with en_core_web_lg and Bert-based-NER. NLTK has the worst performance such that it classified most of the names as organizations. This ruled out NLTK. Spacy is not able to capture lots of the NERs, while Bert-based-NER can. Thus, Bert-based-NER clearly has better performance compared to the others, correctly recognizing most of the entities and not missing most of the NERs. Each NER was classified as one of the following:



Abbreviation	Description
O	Outside of a named entity
B-MIS	Beginning of a miscellaneous entity right after another miscellaneous entity
I-MIS	Miscellaneous entity
B-PER	Beginning of a person's name right after another person's name
I-PER	Person's name
B-ORG	Beginning of an organization right after another organization
I-ORG	organization
B-LOC	Beginning of a location right after another location
I-LOC	Location

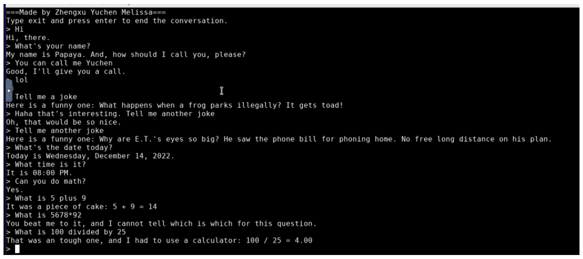
Neo4j We set 3 rules to generate the Knowledge Graph. We extract the Location, Person and Organization from each QA data. Set Organization locate in Location, Person lives in Location and Person works in Location to create triplets. And use py2neo to connect to Neo4j database and save triplets for generate Knowledge Graph.

3 RESULTS OVERVIEW

Experiment Setting

Software: Linux, python/3.7, tensorflow-gpu/1.13.1
Hardware: 1 Tesla V100 GPU, 12 core vcpu (SCC)

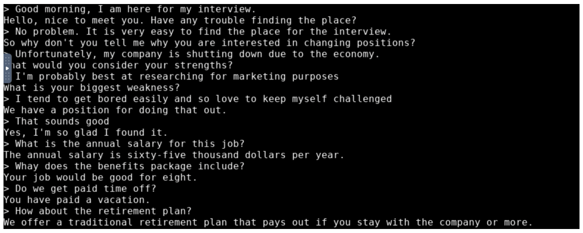
FIGURE 1



```
===Made by Zhengxu Yuchen Melissa===
Type exit and press enter to end the conversation.
> Hi
Hi, there.
> What's your name?
My name is Papaya. And, how should I call you, please?
> You can call me Yuchen
Good, I'll give you a call.
> lol

  Tell me a joke
Here is a funny one: What happens when a frog parks illegally? It gets toad!
> Haha that's interesting. Tell me another joke
Oh, that would be so nice.
> Tell me another joke
Here is a funny one: Why are E.T.'s eyes so big? He saw the phone bill for phoning home. No free long distance on his plan.
> What's the date today?
Today is Wednesday, December 14, 2022.
> What time is it?
It is 08:00 PM.
> Can you do math?
Yes.
> What is 5 plus 9
It was a piece of cake: 5 + 9 = 14
> What is 5678*92
You beat me to it, and I cannot tell which is which for this question.
> What is 100 divided by 25
That was an tough one, and I had to use a calculator: 100 / 25 = 4.00
> 
```

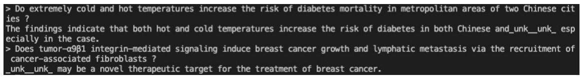
FIGURE 2



```
> Good morning, I am here for my interview.
Hello, nice to meet you. Have any trouble finding the place?
> No problem. It is very easy to find the place for the interview.
So why don't you tell me why you are interested in changing positions?
> Unfortunately, my company is shutting down due to the economy.
hat would you consider your strengths?
  I'm probably best at researching for marketing purposes
What is your biggest weakness?
> I tend to get bored easily and so love to keep myself challenged
We have a position for doing that out.
> That sounds good
Yes, I'm so glad I found it.
> What is the annual salary for this job?
The annual salary is sixty-five thousand dollars per year.
> Whay does the benefits package include?
Your job would be good for eight.
> Do we get paid time off?
You have paid a vacation.
> How about the retirement plan?
We offer a traditional retirement plan that pays out if you stay with the company or more.
```

As shown in Figures 1 and 2, the chatbot we configured can respond to our user's input message. The output responses are highly correlated to the inputs. For example, when the user inputs "Tell me a joke", the chatbot responds back with "Here is a funny one: What happens when a frog parks illegally? It gets toad!", which is indeed a joke. At the same time, if the chatbot is asked to give another joke, the same joke will not be given again, showing that the chatbot has a "short-term" memory and has the ability to collaborate with the context of the conversation.

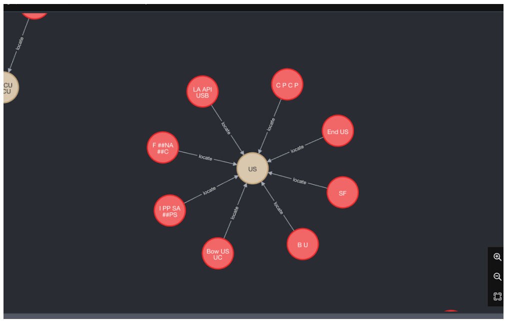
FiGURE 3



```
> Do extremely cold and hot temperatures increase the risk of diabetes mortality in metropolitan areas of two Chinese cit
ies ?
The findings indicate that both hot and cold temperatures increase the risk of diabetes in both Chinese and_unk__unk_ esp
ecially in the case.
> Does tumor-α9β1 integrin-mediated signaling induce breast cancer growth and lymphatic metastasis via the recruitment of
 cancer-associated fibroblasts ?
_unk__unk_ may be a novel therapeutic target for the treatment of breast cancer.
```

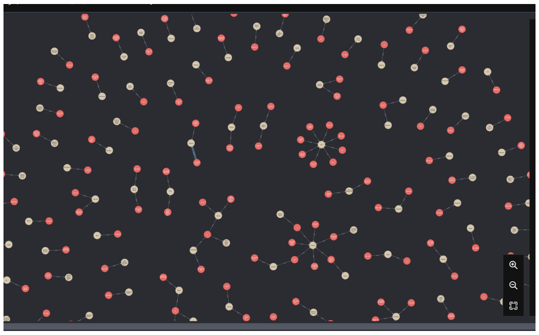
As seen above in figure 3, we tried testing out inputs with medical-related inquiries for our chatbot. We still got responses that were somewhat relevant to the question, however, they were not as comprehensible as the conversations from Figures 1 and 2. This is likely due to the corpus size being too small. Since the corpus for training was not large enough, some of the medical terms were not recognized by the model. Therefore, the variable *_unk__unk_* is used to indicate that the word is outside of the corpus.

FIGURE 4



A close-up view of the knowledge graph

FIGURE 5

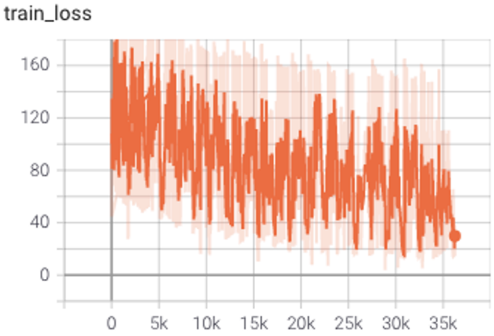


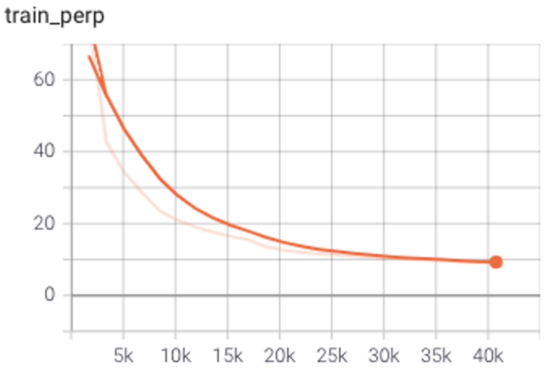
In Figure 4, we can see what each word is and what it is connected to as well as what their relationship towards each other is. Figure 5 shows an expanded view of the knowledge graph as a whole. This shows the many relationships that were captured after training our chatbot with the PubMed corpus.

4 EVALUATION

Open source data from PubMed that are grouped by questions and answers is used to train both the chatbot and the NER model. The knowledge graph generated from Neo4j that was fed in triplets generated by the trained NER model is used to evaluate the answers given by the chatbot with a question as the input. At the same time, subjective judgements were also involved.

The loss and the perplexity of the training is as shown below.





The chatbot has the ability to have simple daily conversations and is able to answer questions taking the previous conversations into account (see joke part). In addition, the chatbot is able to give some professional advice in terms of questions related to illness and diseases. However, as already mentioned, since the corpus for training is not big enough, the chatbot is not able to recognize some of the medical terms.

5 FUTURE WORK

There is definitely more to improve on the entire project, which can be divided into two parts: chatbot part and knowledge graph part.

For the chatbot part, two things can be improved in the future: training efficiency and accuracy. Based on our current training model, the epoch time for training is around 1 hour per epoch, which is not efficient. In order to have better performance, we trained the model with 60 epochs for the illustration of this paper, which takes more than 3 days. We believe that by increasing the parallelism of the implementation for training will increase the efficiency. In terms of accuracy of the chatbot, as shown in the evaluation section, the chatbot is good for daily conversations but not enough for giving professional suggestions. Thus, a larger corpus will be necessary to train the chatbot.

For the knowledge graph part, when retrieving triplets, the accuracy of recognizing NERs is not as high. Specifically, if there is a typo in the corpus, the word will then not be recognized correctly. For example, if “LA” is typed as “La”, it is recognized as MISC instead of LOC. Thus, more work needs to be done to improve the accuracy of extracting NERs.

6 RESOURCES

Graph Database Applications and Concepts with Neo4j - Scinapse.https://asset-pdf.scinapse.io/prod/776871969/77687199.pdf.

Hussain, Mustaffa. “Knowledge Graph-Based Chatbot.” Medium, TheCyPhy, 31 Oct. 2020, https://medium.com/thecyphy/knowledge-graph-based-chatbot-541679d7f17.

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

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.