MACHINE COMPREHENSION OF SPOKEN CONTENT

Akshay Mallipeddi **Anirudh Kulkarni** **Rushikesh Nalla**
112078311 112025249 111971011

Abstract

Huge amounts of multimedia data is available over the internet via YouTube, MOOCs like edX, Coursera, Udacity etc. The spoken part of these sources encodes very important information about the content. We can build a machine that listens to this audio data and extracts useful information to answer the questions posed to it. To tackle this task, we'll build models using two approches namely, End-to-end Memory Network (MEM2NN) and Hierarchical Attention Model (HAM) and compare their performances on TOEFL dataset.

1 Introduction

1.1 Objective

The objective of our project is to address the QA mechanism involving spoken data. In general QA is not just mere comparison among the text in the question and answer. For instance, in our task we are exploring TOEFL listening comprehension data that includes complex questions that ask for the conclusion of the theory, inferences etc presented to the listener. So, clear understanding of the text is required, for the system to answer these kinds of questions.

1.2 Task Description

Test of English as a Foreign Language (TOEFL) is a standardized test to measure the English language ability of non-native speakers wishing to enroll in English-speaking universities. In the listening section of the test, each question consists of an audio story, a question, and four answer choices. Among these choices, one or two of them are correct. Given the manual or automatic speech recognition (ASR) transcriptions of an audio story and a question, machine has to select the correct answer out of the four choices.

1.3 Applications

This ability of the machine to comprehend text will lead to a **better search**. Considering the journey with a traditional keyword-based search, we have a question in mind and hope the internet can answer. As a start, we formulate the question with some keywords and type them into the search box. The search engine returns a set of matched documents. Now, we pick top-ranked documents and quickly scan them through, as the answer to the question often hides somewhere in these documents. Finally, we marry multiple ideas from various texts and draw a conclusion about the answer. In this journey, the search engine only helps in filtering out irrelevant documents, leaving the hard tasks such as reading, comprehending and summarizing. In other words, the internet doesnt actually answer your question, we answer it by ourselves. Having a search system with the reading comprehension ability can make the user experience smoother and more efficient, as all time-consuming tasks such as retrieving, inferring and summarizing are left to the computers. For impatient users who dont have time to read text carefully, such system can be very useful. This system when integrated with voice assistants can significantly improve their performance.

1.4 Challenges

This task of machine comprehension is very challenging. Firstly, this system requires the spoken content to be first transcribed into text by ASR, and the machine will subsequently process the ASR output. This process is largely vulnerable to Out-of-vocabulary problem. Secondly, the amount of labelled data available at our disposal is very limited.

1.5 Baseline and Approaches

Strawman Solution(1) : A strawman solution to the problem in hand is to make use of pre-trained GloVe vectors to obtain the vector representation of each word. By averaging out the word vectors we can represent the question and option phrase by a fixed-length vector. We now evaluate the cosine similarities between question and each of the option vector. The option with the highest similarity is deemed correct. The main disadvantage of this approach is that it ignores the story completely.

Sliding Window(2) : This approach overcomes the shortcoming of the above approach by taking into account the story as well. In this approach we slide a window of fixed size and compare the cosine similarity of the window with the question word vector. The word vector of the window with maximum cosine similarity is compared with each of the choices. The option with the highest similarity is selected.

LSTM(3) : Using LSTM encoder for representing story and question together. Options were also encoded using the LSTM encoder. The option with the most similarity is selected.

The problem with general RNN(10) and LSTM models is that it fails to capture very long term dependencies.

We will be looking at more advanced models which are variations of RNN and LSTM (end-to-end memory networks and Tree-LSTM (9)) with attention modules. This would solve the above mentioned drawbacks of the existing methods.

1.6 Dataset and evaluation metrics

Dataset : The collected TOEFL dataset includes 963 examples in total (717 for training, 124 for validation, 122 for testing). Each example consists of a story, a question, and 4 choices.

Example : Audio Story 🔊 (For few minutes)
uh , excuse me , professor thompson . i know your office hours are tomorrow , but i was wondering if you had a few minutes free now to discuss something . sure , john . what did you want to talk about ? well , i have some quick questions about how to write up the research project i did this semester about climate variations . oh , yes . you were looking at variations in climate in the grant city area.....
what does the professor offer to do for the man ?
help him collect more data in other areas of the state (0) .
submit his research findings for publication (0) .
give him the doctor's telephone number (0) .
review the first version of his report (1)

This example will be converted to :
SENTENCE uh , excuse me , professor thompson .
SENTENCE i know your office hours are tomorrow, but i
SENTENCE was wondering if you had a few minutes free
SENTENCE now to discuss something . sure, john .what
SENTENCE did you want to talk about ? well , i have
SENTENCE some quick questions about how to write up the
SENTENCE research project i did this semester about
SENTENCE climate variations. oh , yes. you were looking
SENTENCE at variations in climate in the grant city area .....
QUESTION what does the professor offer to do for the man?
OPTION help him collect more data in other areas of the state (0) .
OPTION submit his research findings for publication (0) .
OPTION give him the doctor's telephone number (0) .
OPTION review the first version of his report (1) .

A prefix is attached to each of the sentences present in the dataset which signifies the type of information. There are three possible values of prefix namely - SENTENCE, QUESTION and OPTION. The SENTENCE prefix indicates that the sentence is part of the spoken content, the QUESTION tag corresponds to the question and the OPTIONS tag corresponds to the possible answers to the question.

Evaluation Metrics :

- *Overall Accuracy:* This is calculated by checking number of questions that the model was able to answer correctly.
- *Conversation Accuracy:* This is calculated by checking the number of conversation related questions that the model is able to answer correctly.
- *Lecture Accuracy:* This is calculated by checking the number of lecture related questions that the model is able to answer correctly.

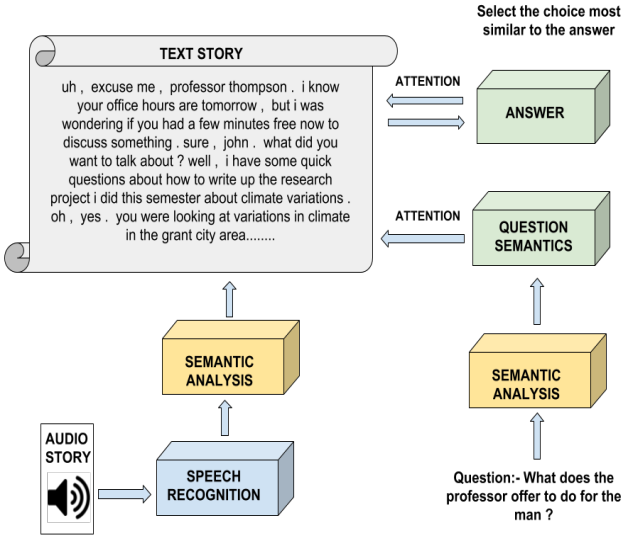

Figure 1: Overview of the system

1.7 Overview of the system

The basic system involves a sample of audio data being converted into text format, manually or using ASR(Automatic Speech Recognizer). This text would mainly be the story content presented to the listener. The story, questions and the options would then be fed to the model (that involves semantic analysis) which would return best answer for that particular question.

2 Current Progress

2.1 Preprocessing

- We have retrieved Stanford Parser, Stanford POS Tagger, GloVe (7) - Global Vectors for Word Representation and TOEFL data.
- The GloVe vectors have 2196017 words in total and the embedding size of each word is 300.
- For unknown words, we have assigned them the mean of all the embeddings of existing words.
- Building the vocabulary from TOEFL data which we have retrieved.
- We have split the TOEFL dataset into separate sentences, questions and options files.
- For each of these files we have built dependency parsers using Stanford Parser(8) considering each sentence at a time.

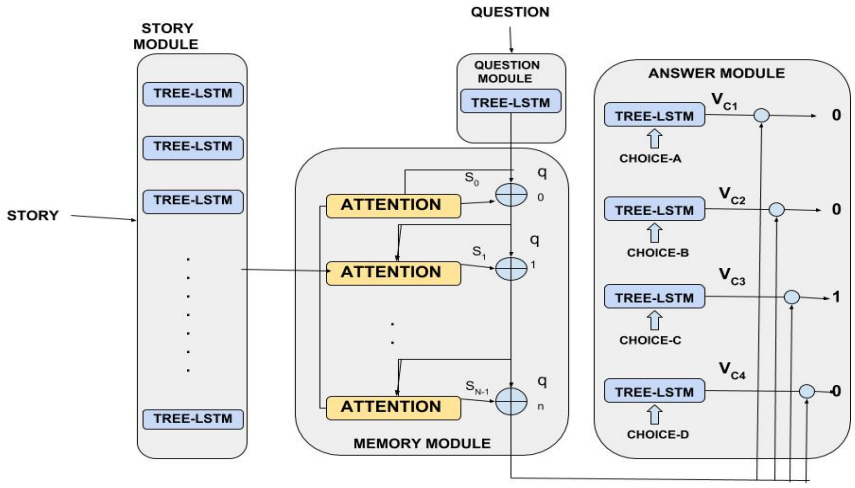

Figure 2: Overview of the HAM system

Model	Conversation Accuracy	Lecture Accuracy	Overall Accuracy
MEM2NN			
HAM			

Table 1: Comparison between MEM2NN and HAM

2.2 Implementation

We would be implementing the following two systems.

MEM2NN(4) : In this model, the embeddings of the input are converted to input memory vectors and we compute its inner product with the embeddings of the query. We take the softmax of the result and take the product with output memory units to find the correct option by evaluating the similarity.

HAM(5) : This model is one of the most mature one to carry out this task. The proposed system is a Hierarchical Attention Model (HAM) which contains Tree-LSTM modules to represent the sentences/phrases in the story, questions and choices. Tree structured models help in representing the syntactic structure of the sentence and understanding the semantics. Attention and memory modules are essential when dealing with long sentences because for answering a question we are interested in finding out the most relevant words in the text and this can be achieved by having access to previous states/words of the story/question. Evaluation metrics mentioned in 1.6 will be used to compare both of the above mentioned methods.

3 Expected Results

We would be filling table 1 with the results that we obtain for both the models. We would analyze the system in detail to generate plots for final report.

4 Questions

- One of the major bottlenecks that the model will suffer from is the lack of efficient ASR model. A huge error rate is noticed even in the state-of-the art ASR technique.
- Can the model that is being trained on TOEFL data easily used be for other domains?
- As the model we are looking to implement is already advanced, what are the standard techniques to tweak the architecture?

References

[1] B.-H. Tseng, S.-S. Shen, H.-Y. Lee, and L.-S. Lee, Towards machine learning comprehension of spoken content: Initial toefl listening comprehension test by machine, in INTERSPEECH, 2016.

[2] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun,and S. Fidler, Movieqa: Understanding stories in movies through question-answering, CoRR, vol.abs/1512.02902, 2015.

[3] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espe- holt, W. Kay, M. Suleyman, and P. Blunsom, Teaching machines to read and comprehend, CoRR, vol.abs/1506.03340, 2015.

[4] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, Weakly supervised memory networks, CoRR, vol. abs/1503.08895, 2015.

[5] Wei Fang, Jui-Yang Hsu, Hung-yi Lee, Lin-Shan Lee "Hierarchical Attention Model for Improved Machine Comprehension of Spoken Content".

[6] Chung, Yu-An and Lee, Hung-Yi and Glass, James "Supervised and unsupervised transfer learning for question answering", NAACL HLT, 2018.

[7] J. Pennington, R. Socher, and C. D. Manning, Glove: Global vectors for word representation., EMNLP, vol.14, pp. 153243, 2014.

[8] D. Chen and C. D. Manning, A fast and accurate dependency parser using neural networks., in EMNLP, 2014, pp. 740750.

[9] K. S. Tai, R. Socher, and C. D. Manning, Improved semantic representations from tree-structured long shortterm memory networks, CoRR, vol. abs/1503.00075, 2015.

[10] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, Movieqa: Understanding stories in movies through question-answering, CoRR, vol. abs/1512.02902, 2015.