Machine Comprehension of Spoken Content

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

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. Also, until now large scale training and test datasets have been missing for this type of evaluation. In our work we make an attempt to try and resolve this bottleneck and provides large scale supervised reading comprehension data for spoken content.

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.

In this project we are also trying to tackle one of the main challenge posed by any QA task - lack of supervised data. We have made use of dailymail news articles and bAbI dataset present in the text format for training. Several preprocessing steps like generating question and choices, tagging sentences, dropping words and characters, identifying entities etc have been addressed.

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 Previous 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 Ideas Implemented

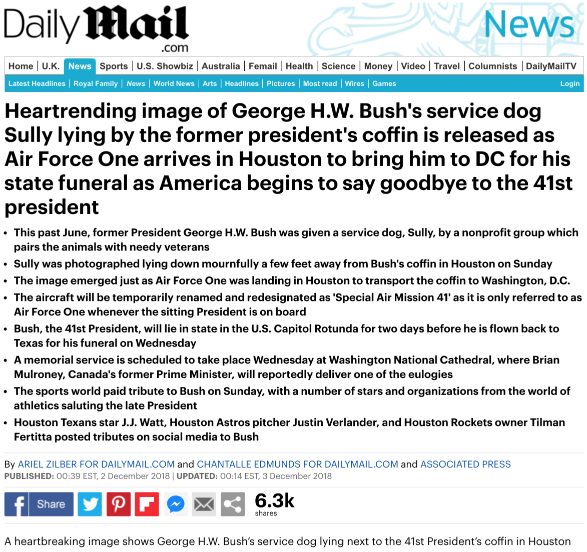
Generation of dataset :
One of the main challenge posed by the QA task is lack of supervised content during the training phase. We have made an attempt to generate large scale supervised reading comprehension data to address this issue. The main idea here is to modify the structured text data to replicate characteristics of the ASR transcriptions (usually prone to semantic and syntactic errors).
Sources of Text Data :
We have made use of two different sources of text data to train our model -

- Dailymail and CNN news articles
- bAbI dataset (Children Book Test)

Preprocessing News Data :
In the case of news data, we use a web scraper script to extract large amounts of data from dailymail websites whereby the content is extracted and saved to a local file. One important reason why we chose CNN and dailymail as our source of information is that these providers supplement their articles with a number of bullet points, summarizing aspects of the information contained in the article. These summary points are abstractive and are not simple copy of sentences from the documents. Training the model on such data makes sure that the model has a greater coverage and is mature enough to comprehend complex text and draw inferences.

Once we have the story and bullet points questions along with possible choices needs to be generated. We made use of 20 different combinations of the questions of the following form and randomly pick one question from this set.

- Which of the following is correct
- Which of the following is most relevant to the story
- Choose the correct alternative using the above context


Daily Mail
.com
News
Home | U.K. | News | Sports | U.S. Showbiz | Australia | Femail | Health | Science | Money | Video | Travel | Columnists | DailyMailTV
Latest Headlines | Royal Family | News | World News | Arts | Headlines | Pictures | Most read | Wires | Games
Login
Share
6.3k
shares

Heartrending image of George H.W. Bush's service dog Sully lying by the former president's coffin is released as Air Force One arrives in Houston to bring him to DC for his state funeral as America begins to say goodbye to the 41st president

- This past June, former President George H.W. Bush was given a service dog, Sully, by a nonprofit group which pairs the animals with needy veterans
- Sully was photographed lying down mournfully a few feet away from Bush's coffin in Houston on Sunday
- The image emerged just as Air Force One was landing in Houston to transport the coffin to Washington, D.C.
- The aircraft will be temporarily renamed and redesignated as 'Special Air Mission 41' as it is only referred to as Air Force One whenever the sitting President is on board
- Bush, the 41st President, will lie in state in the U.S. Capitol Rotunda for two days before he is flown back to Texas for his funeral on Wednesday
- A memorial service is scheduled to take place Wednesday at Washington National Cathedral, where Brian Mulroney, Canada's former Prime Minister, will reportedly deliver one of the eulogies
- The sports world paid tribute to Bush on Sunday, with a number of stars and organizations from the world of athletics saluting the late President
- Houston Texans star J.J. Watt, Houston Astros pitcher Justin Verlander, and Houston Rockets owner Tilman Fertitta posted tributes on social media to Bush

By ARIEL ZILBER FOR DAILYMAIL.COM and CHANTALLE EDMUNDS FOR DAILYMAIL.COM and ASSOCIATED PRESS
PUBLISHED: 00:39 EST, 2 December 2018 | **UPDATED:** 00:14 EST, 3 December 2018

A heartbreaking image shows George H.W. Bush's service dog lying next to the 41st President's coffin in Houston

Figure 1: Sample news article (from daily mail website)

Next step is to generate four possible choices one of which is the actual answer and the rest are false. For this, we made use of Named Entity Recognition. We pick one the bullet points from the scraped data as the correct choice or answer. Now, we have to generate three wrong statement. For this, we first identified the named entity in the correct choice and replaced it with top 3 most frequently occurring entities in the story other than the correct. We are replacing named entities as it is usually the most significant term that drives the meaning of the sentence.

We are employing one-word dropping mechanism to make the text data consistent with the spoken transcriptions. The text data that we are extracting is structured and usually conforms to the syntactic and semantic structure of the language. On the other hand, the spoken data is usually informal and unstructured. To normalize the text data consistent with the spoken content, we randomly drop one word from each sentence of the story.

Preprocessing of Children Book Test(CBT) data : In the case of CBT, we extract stories from the books from various authors. Each story is presented with a question (which is in Cloze format where some part of text is replaced with blank) and few options to choose an answer from. The questions were generated by replacing the Named Entities with various possible options among which only one answer is appropriate. We have used this dataset since the stories align to the lecture format questions in TOEFL and the level of semantics involved within the text in these stories are of low level which do not require deeper understanding of the text to answer the questions. We made use of 20 different combinations of the questions as mentioned above. Next we generate answers by replacing the blanks with the possible options given to us. We are employing one-word dropping mechanism to make the text data consistent with the spoken transcriptions as explained above.

Example:

Story:

With almost everything else to make them happy, they wanted one thing they had no children . This vexed the king even more than the queen .....

Question:

replied the XXXXX ; for the king 's aunts were old-fashioned , and did not approve of her , and she knew it .

Answer: queen

Options: baby|boy|mother|portrait|queen|time|wall

Converted Format:

Question: Pick from the random 20 as explained above.

Answer: replied the **queen** ; for the king 's aunts were old-fashioned , a...

Other Options:

replied the boy ; for the king 's aunts were old-fashioned , a...

replied the mother ; for the king 's aunts were old-fashioned ,a...

replied the time ; for the king 's aunts were old-fashioned , a...

Intuition : As we are increasing the size of the dataset and training the model extensively on a different types of text, the accuracy of the model should increase.

1.7 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.

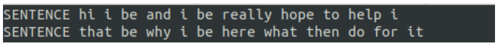
1.8 Outcomes

- We have written scripts to create labelled CNN/Daily News and Children Book Test dataset. The dataset is generated by performing several preprocessing steps like generating question and choices, tagging sentences, dropping words and characters and identifying entities.

- Next, we integrated the newly geenrated CNN/Daily news and Children Book Test dataset to evaluate how the existing models (HAM and MEMN2N) perform.

- Additionally, we have also implemented a Google Speech Recognition script that converts an audio into text, and then compared the text that was generated by the ASR system mentioned in (5). The exact ASR system has not been mentioned but the error rate by using the Google speech Recognition API was very less compared to the ASR system used by (5). So, the voice from the listening comprehension can be sent to the Google speech recognition API to increase the performance of the system.

Sample ASR output:-



```
SENTENCE hi i be and i be really hope to help i
SENTENCE that be why i be here what then do for it
```

Figure 2: Sample ASR output

We can see that the ASR output is far from readable. So, there is a need for better Audio to Text converter, we could use Google speech Recognition API.

1.9 Issues

- Our evaluation shows that with the increase in the size of the external dataset, the overall accuracy decreased. This is probably because the listening section of the toefl has audio that is two-way or conversational in nature and the dataset that we chose is unidirectional.

- We implemented word level dropping and could not see much effect. Character level dropping or a combination of word-character dropping could be effective since the ASR creates numerous spelling mistakes and grammatical errors. These techniques for character level dropping can be tried as future work.

2 Dataset Details

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.We have used CNN/Daily News and Children Book Test data for data generation step and this data will be preprocessed before integrating to the system. We have created 500,1000 and 2000 stories to integrate with the TOEFL data and test how the system works with this integrated data. We will be comparing with the existing MemN2N and HAM models with the accumulated dataset as mentioned above. We are calculating the accuracy of the model by matching the predicted answers along with the truth values associated with the story, question pairs.

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.

3 Baseline Model : End-to-End Memory Networks

Key ideas : The idea of improving the performance of the model by performing multiple hops over long-term memory is what drives its architecture.

Approach :

- The model takes a discrete set of sentences $x_1, .., x_n$ and a question q as inputs and returns a as output. Each of these *x*, *q* and *a* are composed of symbols coming from dictionary *V*.
- The model writes all *x* to the memory up to a fixed buffer size, and then finds continuous representation for *x* and *q*.
- This continuous representation is then processed using multiple hops to output the answer *a*.

Single Memory Hop Operation
Input memory Representation : We have $x_1, .., x_n$ to be stored in the memory. The entire set $\{x_i\}$ are converted to memory vectors $\{m_i\}$ of dimension *d* computed via embedding matrix *A* of dimension *(d x V)*. In a similar fashion query q is embedded to obtain obtain *u* via embedding matrix *B*. The match between the question and the sentences is computed by taking the inner product of *u* and *m* followed by softmax:

$$p_i = Softmax(u^T m_i)$$

Output memory Representation : Each x_i has an output vector c_i. The response from the memory *o* is the sum over the transformed inputs c_i and probability vector from the input.

$$o = \Sigma p_i c_i$$

Generating Final Prediction : The sum of the output vector and the input embedding is passed through a final weight matrix W of dimension *(V x d)* and a softmax to produce the predicted answer

$$\hat{a} = Softmax(W(o + u)))$$

Multiple Memory Hop Operation

We can stack the memory layers of the single layer described above to build multiple memory hop operations. The memory layers are stacked in the following fashion :

- The input to the layer *k+1* is the sum of output and input at layer k i.e. $u^{k+1} = u^k + o^k$
- Each layer has its own embedding matrices A^k, C^k, used to embed inputs.

There are multiple ways in which the weight tying can be done within the model, namely - Adjacent and Layer-wise.
Sentence Representation : There are two different ways in which the sentences are represented. Firstly, the bag-of-words (BoW) representation that takes the input sentence $x_i = \{x_{i1}, x_{i2}, .., x_{in}\}$, embeds each word and sums the resulting vector. The **drawback** of this technique is that it cannot capture the order of words. To overcome this, a second representation that encodes the word position is used.
Temporal Encoding : In the QA task temporal information is of atmost importance. The model should be mature enough to identify the state of an entity at a given point of time. To enable this the memory vector is modified as, $m_i = \Sigma\ Ax_{ij}$ + $T_A(i)$. Here $T_A(i)$ is the ith row of a special matrix T_A that encodes temporal information.
Learning time invariance by injecting random noise : To regularize the temporal data, an approach of random noise (RN) is introduced that adds 10% of empty memories to the stories.

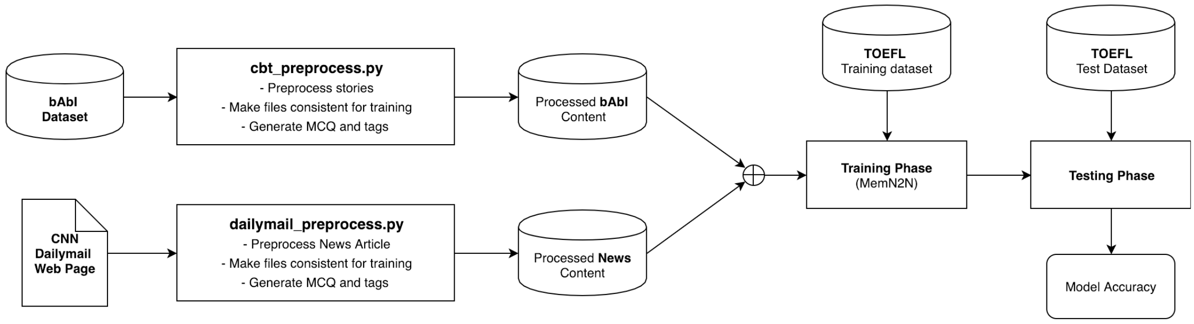


Figure 3: Overview of the system

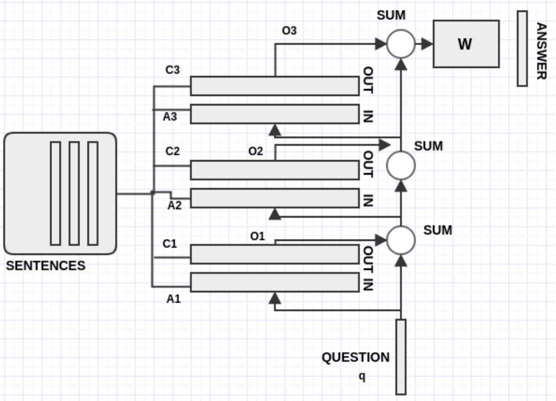


Figure 4: Multilayer MEMN2N architecture

3.1 Issues

MemN2N model represents the input sentences as Bag of words(BOW) which fails to capture the word order of the sentence. This is the major issue. This representation doesnt take into account the semantic relation between words. In most tasks the neighbor words in a sentence is useful for predicting the middle word. Consider the following sentences -
1) i had cleaned my car, which means the person has cleaned his own car.
2) i had my car cleaned, which means the person's car was cleaned by someone else.
The above two sentences have the same BOW representation (frequency) but have different meanings. We implement a model with tackles this issue by using tree representations (HAM) for the input sentences.

4 Approach

4.1 Overall Architecture

The Fig 3 summarizes the overall picture of the project implementation. This flow can be divided into two phases - data generation and model training/predictions. In the data generation phase, two python scripts transform the textual content into a format consistent for model training. This data along with domain specific data is fed to the model for training. The model that is trained is based on MemN2N and HAM architectures.

4.2 Hierarchical Attention Model (HAM)

4.2.1 Details

Key ideas : The architecture described here is named Hierarchical Attention Model (HAM). The name comes intuitively as it makes use of a hierarchical structure (i.e TreeLSTM) and attention mechanism to comprehend and answer. TreeLSTM improves the encoding of sentences by using their hierarchical structure and hence successfully improves semantic representation. In addition, a memory module with attention mechanism is used to match the question-context sentences.
TreeLSTM : There are two types of TreeLSTM(9) - Child-Sum TreeLSTM and N-ary TreeLSTM. In this architecture, a Child-Sum TreeLSTM over the dependency tree of a sentence is used to represent a sentence. Just like an LSTM, the TreeLSTM has input(i_j) gates, output gates(o_j) for memory cells, a set of memory cells(c_j) and forget gates(f_{jk}) that controls the information flowing in from the child nodes *k*.
A hidden state h_j at node *j* is the representation for a phrase consisting of words in the subtree rooted at node j.
Story and Question Modules : We make use of the above steps to represent the story and question. For story, the hidden representations of each sentence is store for future use. Contrastingly, we sum up the hidden states of the root nodes of the Tree-LSTM of each sentence for the questions to obtain question vector V_Q.
Memory Module : The memory module has two main aspects. Firstly, the attention mechanism to obtain the question context match and the multi-hopping to refine the representations.
Attention Mechanism - There are two different ways to obtain vector representation O=$\{o_1,o_2,..,o_T\}$ for story - (a) Phrase-level (b) Sentence Level. In Phrase-level, each o_t represents a phrase where as in sentence level representation, each o_t represents a story sentence. These vectors are first transformed into memory vectors M=$\{m_1,m_2,..,m_T\}$ and evidence vectors C=$\{c_1,c_2,..,c_T\}$ using embedding matrices as below :

$$m_t = W^{(m)}o_t$$
$$c_t = W^{(c)}o_t$$

The question vector V_Q obtained from question module is transformed into an initial query q_o using another embedding matrix :

$$q_o = W^{(q)} V_Q$$

Now, cosine similarity is used to compute the attention score η_t between query vector q_o and each memory vector m_t and a softmax score is evaluated to give attention score :

$$\eta_t = q_o \odot m_t$$
$$\alpha_t = Softmax(\eta_t)$$

The story vector s_0 is then the weighted sum of the evidence vectors c_t with the attention weights :

$$s_o = \sum_{t=1}^{T} \alpha_t c_t$$

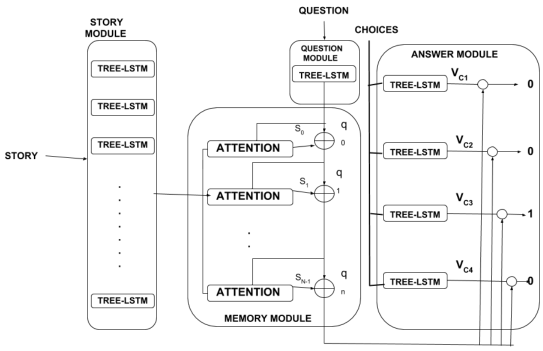


Figure 5: HAM architecture

Answer Module : We calculate the cosine similarity between the memory module output and the choice vector followed by softmax to estimate the correct option.

5 Evaluation

5.1 Hyperparameter Tuning

We started with the default set of parameters and changed each parameter at a time to see its effect. Depending on the accuracy values we modified the other set of hyper parameters. Using this approach we were able to get the best set of hyperparameters which resulted in increase in accuracy as compared to the original paper (2% increment for MEMN2N and 0.1% for HAM).

max epochs = 5, 10
learning rate = 1e-03, 2e-03, 5e-03
mini batch size = 10
regularization = 1e-04, 1e-03
hops = 1, 2, 3
attention level = phrase, sentence
attention similarity = cosine
optimizer = adagrad

5.2 Results and Analysis

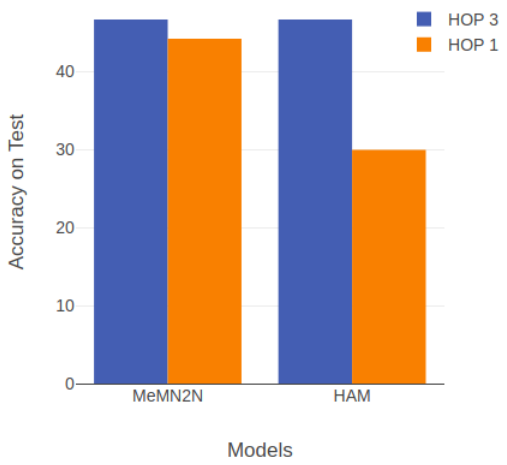


Figure 6: Comparison between MEMN2N and HAM

Using both MEMN2N and HAM models on TOEFL dataset we observe an increase in accuracy on increasing the attention hops. This is intuitive as additional memory mechanism helps the model in developing the relation between passage and the question to answer it correctly as seen in figure 6.

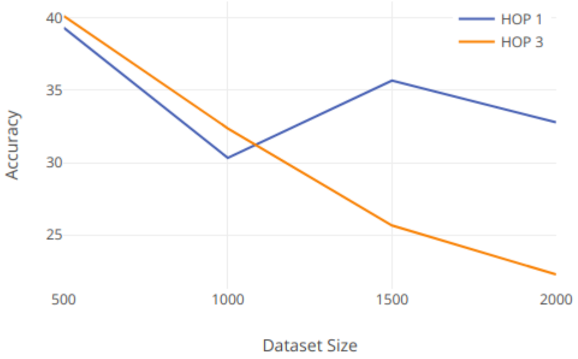


Figure 7: Comparison between datasizes for MEMN2N

Seeing an increase in accuracy on increasing the dataset size from 1000 to 1500 while using 1 hop we expected to see a similar trend on changing dataset size to 2000 but the accuracy decreased. Refer figure 7. This might be because the excess data from daily mail has confused the model (as it is not conversational in nature unlike TOEFL).

From figure 7 we can see that on increasing the daily mail data-set from 1000 to 1500 passages the accuracy in the case of 1 hop attention model increased. As observed in figure 6 we thought increasing the hops might

Table 1: Accuracies according to story type for MEMN2N

memn2n	Lecture	Conversation	Overall
TOEFL	47.16	45.53	46.72
TOEFL+500	42.68	33.36	40.16
TOEFL+1000	35.06	25.11	32.37

Table 2: Accuracies according to story type for HAM

HAM	Lecture	Conversation	Overall
TOEFL	48.26	42.56	46.72
TOEFL+500	44.69	35.96	42.33
TOEFL+1000	38.27	35.24	37.45

improve the results but the accuracy decreased drastically. One possible explanation is that hop 1 has learnt incorrect relations and that propagated to hops 2 and 3. This way the predictions took a hit.

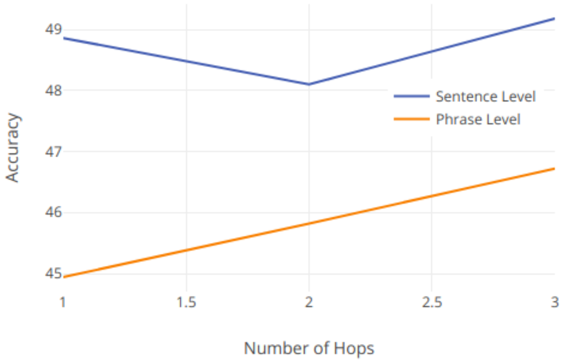


Figure 8: Comparison between sentence and phrase level from HAM

Overall the sentence level attention performed better than phrase level attention mechanism. In figure 8 we can observe that there is a drop in accuracy for 2 hops in sentence level attention. The paper authors might have observed this and didn't experiment with 3 hops but we observed an increase in accuracy (best model for HAM).

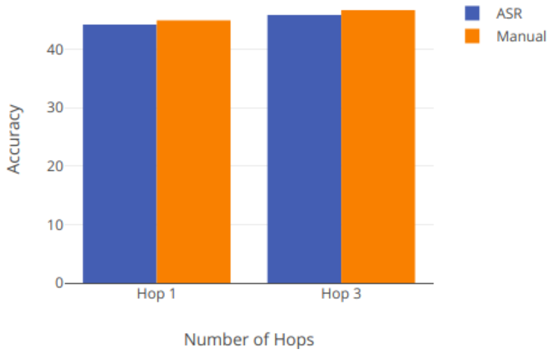


Figure 9: Comparison between ASR and Manual datasets for HAM (Phrase level)

We tried to compare both ASR and manual data to check the accuracy with respect to hops. We observe that manual data gave better results overall as seen in 9. One reason might be because the model was trained with manual data so testing on it produced better results. Also, ASR data is in general error prone compared to manual so it is difficult for the model to predict the correct answer.

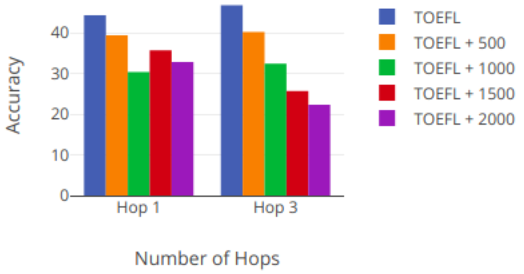


Figure 10: Comparison between TOEFL and Additional datasets for MEMN2N (Phrase level)

The model performed the best with only the TOEFL dataset. On adding Daily mail and Children Stories data there was a drop in accuracy. One of the major reasons is the type of stories which are available in new datasets (unidirectional and domain specific information) which didn't match up with the TOEFL dataset.

In TOEFL dataset the questions are divided into 3 types -
Type 1 : Basic Comprehension of the story
Type 2 : Understanding the attitude of the speaker
Type 3 : Making conclusions, inferences and forming generalizations

Our model was able to answer most of the Type 1 questions correctly as the external data-sets that we are integrating require only basic understanding of the story to answer the questions. Some of the Type 2 questions were also answered correctly but the models suffered when it was given Type 3 questions. Examples for each type of question can be found in the website demo. (**Drive link given below**)

5.3 Code

Google drive link-
https://drive.google.com/drive/folders/1SCV-e9gn2VObIu7ksQWFkLgebPzqc_em?usp=sharing

6 Conclusions

- Using Tree-LSTM and multi-hop attention mechanism helped for better representation of the sentences and hence better predictions on TOEFL dataset.
- The external datasets (Dailymail and Facebook bAbI) were not compatible with TOEFL dataset as they were unidirectional and domain specific. So the model didn't perform well when the datasets were integrated.
- In general the external data was more aligned to lectures (in TOEFL) and hence produced better results than on conversations (in TOEFL).
- Character-level dropping (to replicate data similar to ASR) from external datasets can be used an alternative to word-level dropping in preprocessing which might improve the model.
- A better speech recognizer could also convert the audio much better as shown above with Google Speech Recognizer API.

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.