Machine Comprehension of Spoken Content

{ rushikesh.nalla, anirudh.kulkarni and akshay.mallipeddi}@stonybrook.edu

Stony Brook University

Task Definition

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. We are making use of two architectures namely Hierarchical Attention Model (HAM) and End-to-end memory network (MemN2N) to train and test the model on TOEFL dataset.

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.

Transcribed Script :
uh , excuse me , professor thompson .
i know your office hours are tomorrow, but i
was wondering if you had few minutes
now to discuss . sure, john what ...
Question:
what does the professor offer to do for the man?
Choices:
A. help him collect more data in other areas of the state
B. submit his research findings for publication
C. give him the doctor's telephone number
D. review the first version of his report

Motivation

Our motivation comes from the fact that there is very less supervised data for machine comprehension and QA for the training of the machines. Our task which deals with TOEFL data is one such domain where systems couldn't achieve better accuracy due to less supervised data. We try to overcome this problem by integrating NEWS and Facebook's bAbI data to this available system and evaluate the performance.

The ability of the machine to comprehend text will lead to a better search which can read, comprehend and can make the user experience smoother and more efficient, as all time-consuming tasks such as inferring are left to the computers.

Evaluation

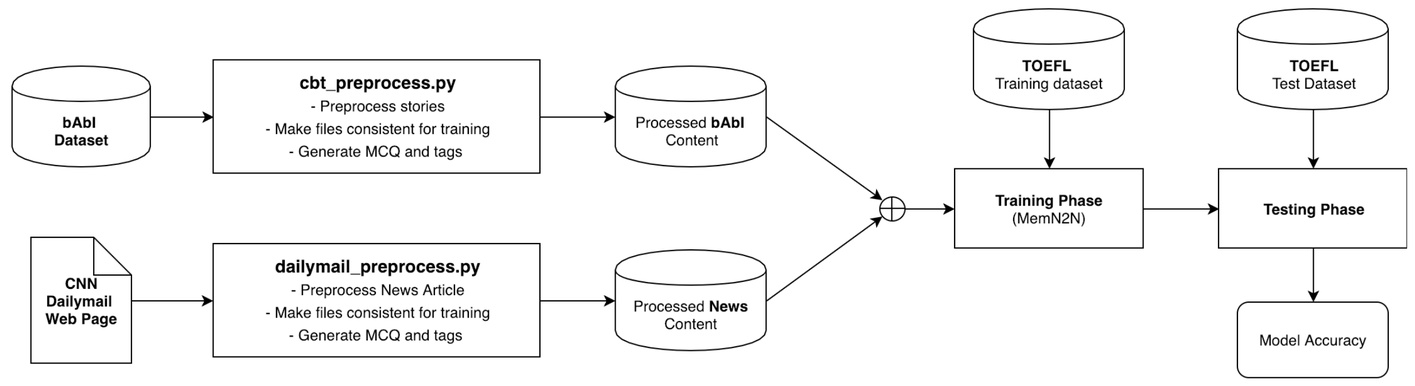
We are evaluating our model based on 3 aspects. In each of these cases, we use the TOEFL test dataset. **Accuracy** i.e the percentage of the questions answered correctly is used as the evaluation metric. The existing TOEFL data set contains 717 train, 124 dev and 122 test stories.

1. Varying the combinations of dataset :
- Constant Number of Hops : 1
- TOEFL Dataset v/s TOEFL Dataset + Generated Dataset
- 500, 1000 and 2000 stories

2. Varying the number of hops :
- Constant Dataset : TOEFL Dataset + Generated Dataset/TOEFL - Dataset
- Hops : 1 vs 3
- 500, 1000 and 2000 stories

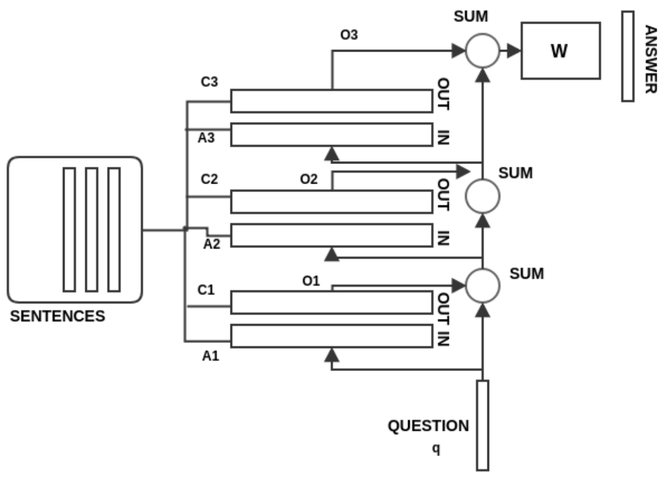
3. Varying the transcription mechanism :
- ASR v/s Manual Transcripts

Task Overview



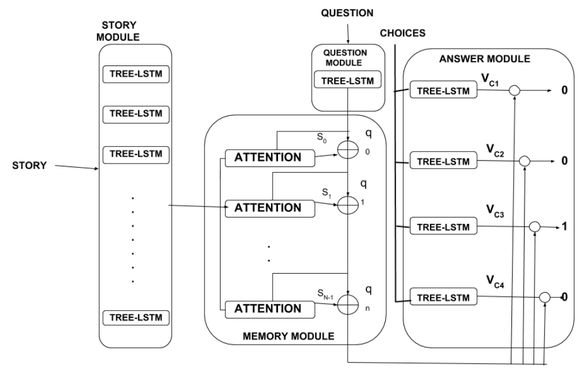
The above block diagram summarizes the overall picture of the project implementation. This flow can be divided into two phases - data generation and model training/predictions. In the data generation phase, two python scripts transform the textual content into a format consistent for model training. This data along with domain specific data is fed to the model for training.

MEMN2N Model



The architecture is a form of Memory Network with a recurrent attention model over a possibly large external memory.

HAM Model



Hierarchical Attention Model (HAM) constructs multi-hopped attention mechanism using tree-structured input rather than sequential representations of the sentences.

Our Accomplishments

- Achieved **2%** increase in the model prediction accuracy on the test data for MEMN2N and **0.1%** increase for HAM via judicious hyperparameter tuning.

- Generated supervised training data using news webpages and bAbI dataset to tackle the main challenge posed by this task.

- Successfully integrated the generated dataset with the existing architecture to train the model.

Analysis

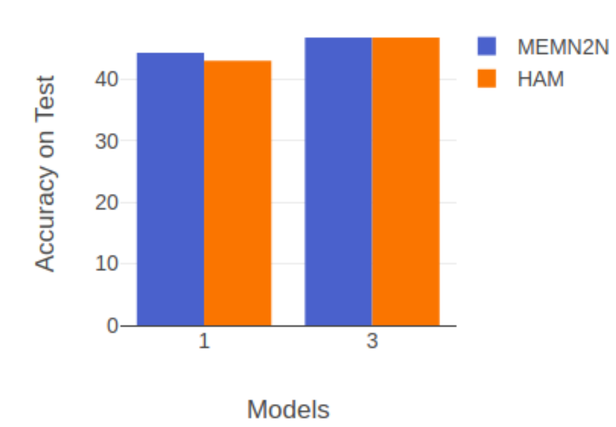
- Firstly, we observe that the overall accuracy of the model increases with the increase in the number of hops for both HAM and MEMN2N. This shows that attention over the stories actually extract better representation for the story.

- Secondly, 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 was unidirectional.

- Finally, we observe that the best configuration of HAM performs way better than MemN2N. This may be because of in MemN2N architecture sentences are represented using the bag-of-words representations and hence fail to take into account the word order. On the other hand, HAM leverages hierarchical structures to encode the syntactic and semantic structure of the sentence.

Results

Comparison between MEMN2N and HAM



Comparison between dataset sizes for MEMN2N

Conclusion

- Using multihops helped for better representation of the sentences for TOEFL dataset.
- The external dataset which are not compatible with conversation/lecture did not perform well.
- Character-level dropping can be an alternative to word-level dropping in preprocessing.

References

[1] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espe- holt, W. Kay, Teaching Machines to Read and Comprehend In *NIPS '15*