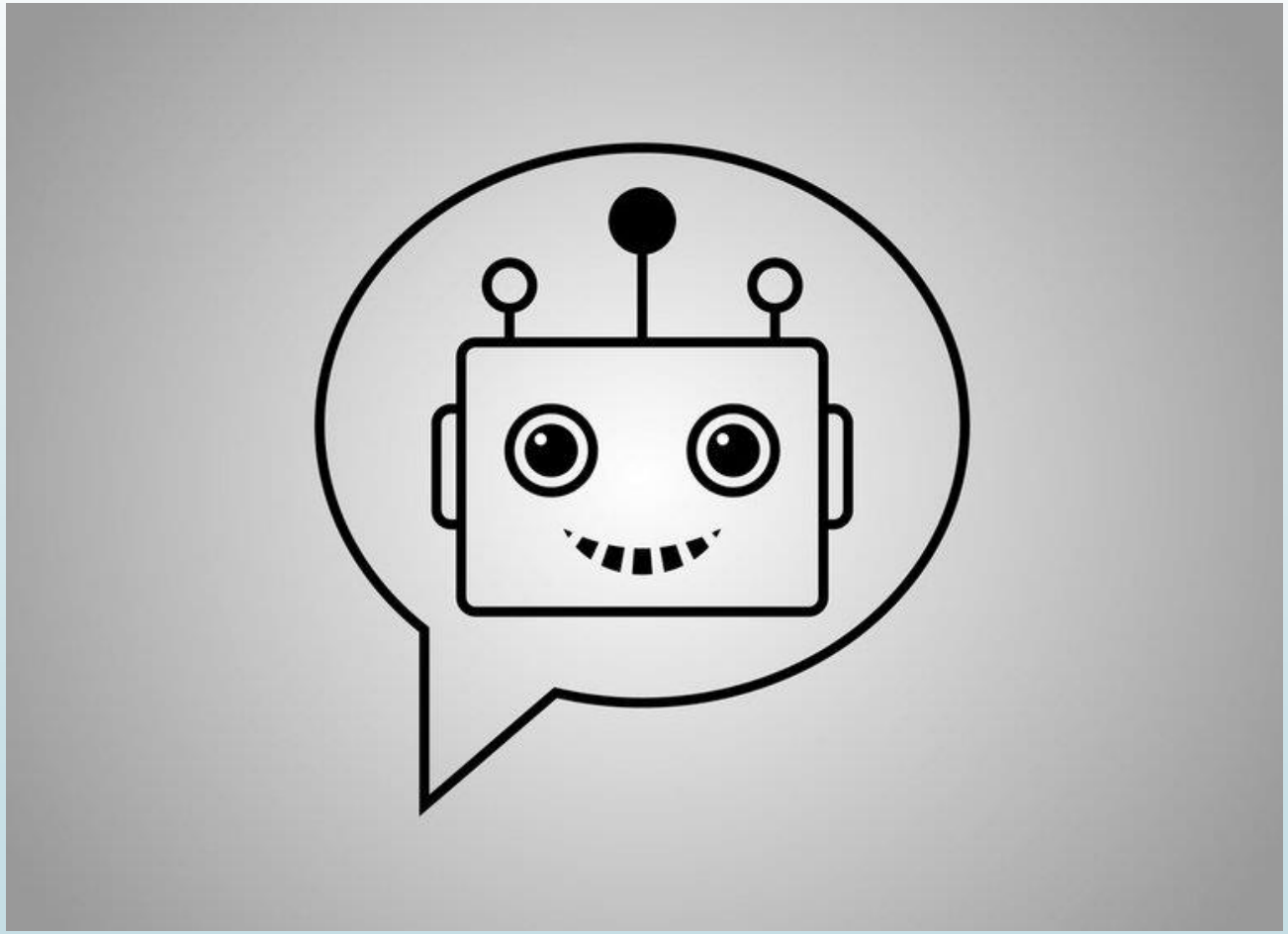
Retrieval based ChatBot



Team Members:

Kaustubh Daware 16BM6JP11

Ranjit Kumar Malloji 16BM6JP38

Sidharth Kumar 16BM6JP45

Prakhar Gupta 16BM6JP34

Ashwani Attri 10CS30010

Mentor: Abhishek Gangwar

Overview of ChatBots

A computer program designed to communicate with human users.
ChatBots are typically used in dialog system for various practical purposes including customer service or information acquisition.

Application

1.Handling large number of clients/users concurrently.
2.Applications in Fields like: Automated Consultation, Assistance, Tracking consumer behaviour on company websites, Tutorials or Guides etc.
3.Major problems faced by users can be addressed to improve on customer satisfaction for example giving prompt reply to large number of customer queries.

Domains of ChatBot

Closed Domain:
Space of input and output in Closed Domain is somewhat limited and predictable since the system tries to achieve a very specific goal.

eg. Technical Customer Support, Automated Assistance etc.

Open Domain:
In Open Domain, conversations are not limited and not very predictable as there is not a well defined goal or intention.

eg. social media conversations on facebook, twitter etc.

ChatBot Models

Retrieval Based:
Retrieval Models use collection of predefined responses to obtain appropriate output using Input and context. So, they don't generate any new text.

Generative Models:
Generative models create new responses from scratch. such models usually use Machine Translation Approach, instead of translating it from one language to the other it's translated into the output from the input.

Objective

Our aim is to develop retrieval based chatBot on Ubuntu Dialogue Corpus. This chatbot will answer the user queries by choosing the best possible answer from the existing Q/A Corpus.

Ubuntu Dialogue Corpus

Requirements:

- Two-way (or dyadic) conversation, as opposed to multi-participant chat, preferably human-human.
- Large number of conversations; 10000 – 1000000 is typical of datasets used for neural-network learning in other areas of AI.
- Many conversations with several turns (more than 3).
- Task-specific domain, as opposed to chatbot systems.

All of these requirements are satisfied by the Ubuntu Dialogue Corpus.

# dialogues (human-human)	930,000
# utterances (in total)	7,100,000
# words (in total)	100,000,000
Min. # turns per dialogue	3
Avg. # turns per dialogue	7.71
Avg. # words per utterance	10.34
Median conversation length (min)	6

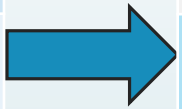
Filtering criteria for Dialogue Dataset creation

- We additionally discard conversations longer than five utterances where one user says more than 80% of the utterances.

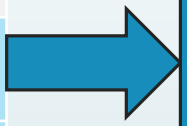
- We consider only extracted dialogues that consist of 3 turns or more to encourage the modeling of longer-term dependencies.

Dataset creation using Ubuntu chatroom Conversation

Time	User	Utterance
03:44	Old	I dont run graphical ubuntu, I run ubuntu server.
03:45	kuja	Taru: Haha sucker.
03:45	Taru	Kuja: ?
03:45	bur[n]er	Old: you can use "ps ax" and "kill (PID#)"
03:45	Kuja	Taru: Anyways, you made the changes right?
03:45	Taru	Kuja: Yes.
03:45	LiveCD	or killall speedlink
03:45	kuja	Taru: Then from the terminal type: sudo apt-get update
03:46	_pm	if i install the beta version, how can i update it when the final version comes out?
03:46	Taru	Kuja: I did.



Sender	Recipient	Utterance
Old		I dont run graphical ubuntu, I run ubuntu server.
bur[n]er	Old	you can use "ps ax" and "kill (PID#)"
Kuja	Taru	Haha sucker.
Taru	kuja	?
kuja	Taru	Anyways, you made the changes right?
Taru	kuja	Yes.
kuja	Taru	Then from the terminal type: sudo apt-get update
Taru	kuja	I did.



Train

Extract a pair of (context, response, flag) triples from each dialogue.

where one triple contains the correct response(i.e. the actual next utterance in the dialogue), and the other triple contains a false response, sampled randomly from elsewhere within the data set.

Test

Similar to train set but instead it has one correct and Nine false responses

Dataset:

- Training set:

	Context	Utterance	Label
0	i think we could import the old comment via rs...	basic each xfree86 upload will not forc user t...	1
1	i 'm not suggest all - onli the one you modifi...	sorri __eou__ i think it be ubuntu relat . __e...	0
2	afternoon all __eou__ not entir relat to warti...	yep . __eou__ oh , okay . i wonder what happen...	0

- Test set:

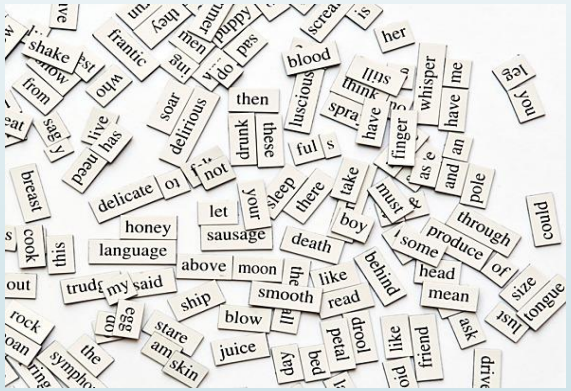
	Context	Ground Truth Utterance	Distractor_0	Distractor_1	Distractor_2	Distractor_3	Distractor_4	Distractor_5	Distractor_6	Distractor_7	Distra
0	anyon know whi my stock oneir export env var u...	nice thank ! __eou__	wrong channel for it , but check efnet.org , u...	everi time the kernel chang , you will lose vi...	ok __eou__	! nomodeset > acer __eou__ i 'm assum it be a ...	http : //www.ubuntu.com /project/about-ubuntu/d...	thx __eou__ unfortun the program be n't instal...	how can i check ? by do a recoveri for test ? ...	my humbl apolog __eou__	# ubur offtop __eou

Approach

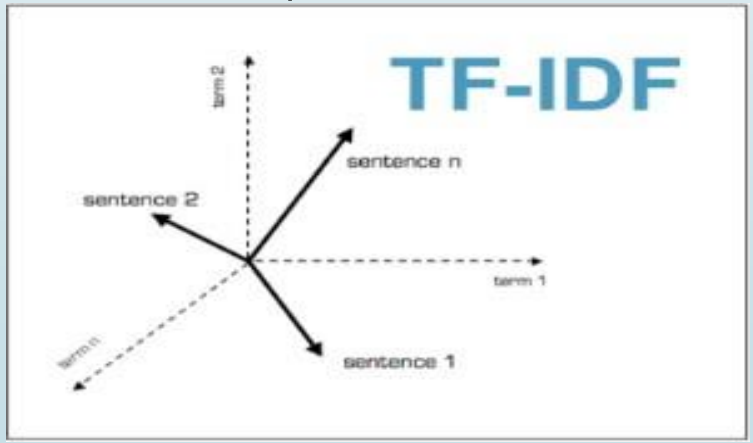
- Cosine Model
- RNN with LSTM Model

Traditional Chat Bot : Cosine Model

- Bag of Words: Using Context and Utterance in the train set.



- Vector Space model: Using term-frequency and Inverse Document Frequency
  - 4 lakhs approx features created
- Mapping of test set in this feature space.



$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

erm x within document y

- Ranking the responses using cosine Similarity.
- Returning the Top 1, Top 2, Top 5 responses.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Traditional model Issues and N-Grams Approach

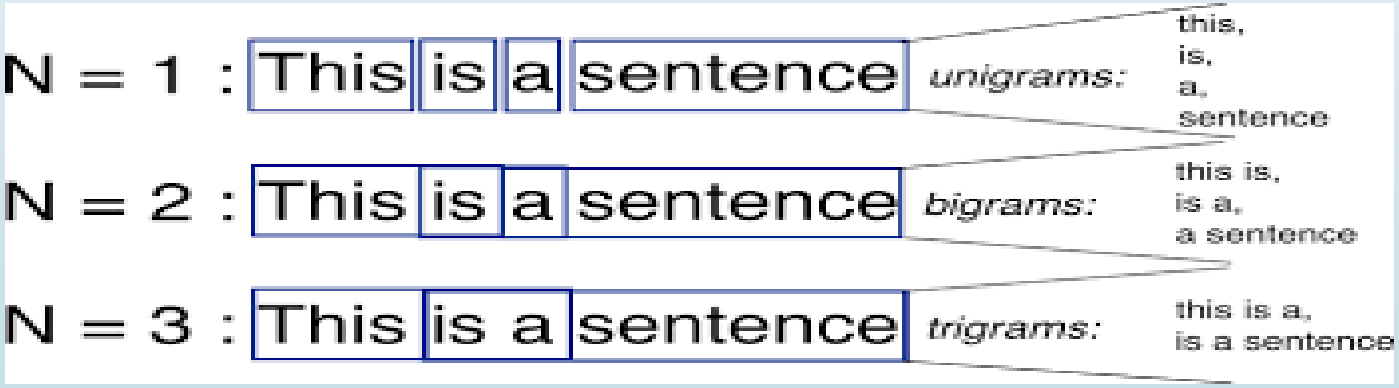
- Unable to capture the semantics of the Context.

Windows is Better than Ubuntu

Ubuntu is better than Windows

- We tried Bi grams and tri grams features.

N-Grams	# Features
1	4,32,656
1+2	51,21,875
1+2+3	2,48,60,976



Sublinear term frequency Scaling

- The possibility of X occurrences of a term in a document actually carry X times the significance of a single occurrence is very unlikely.

$$\mathrm{wf}_{t,d} = \begin{cases} 1 + \log \mathrm{tf}_{t,d} & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

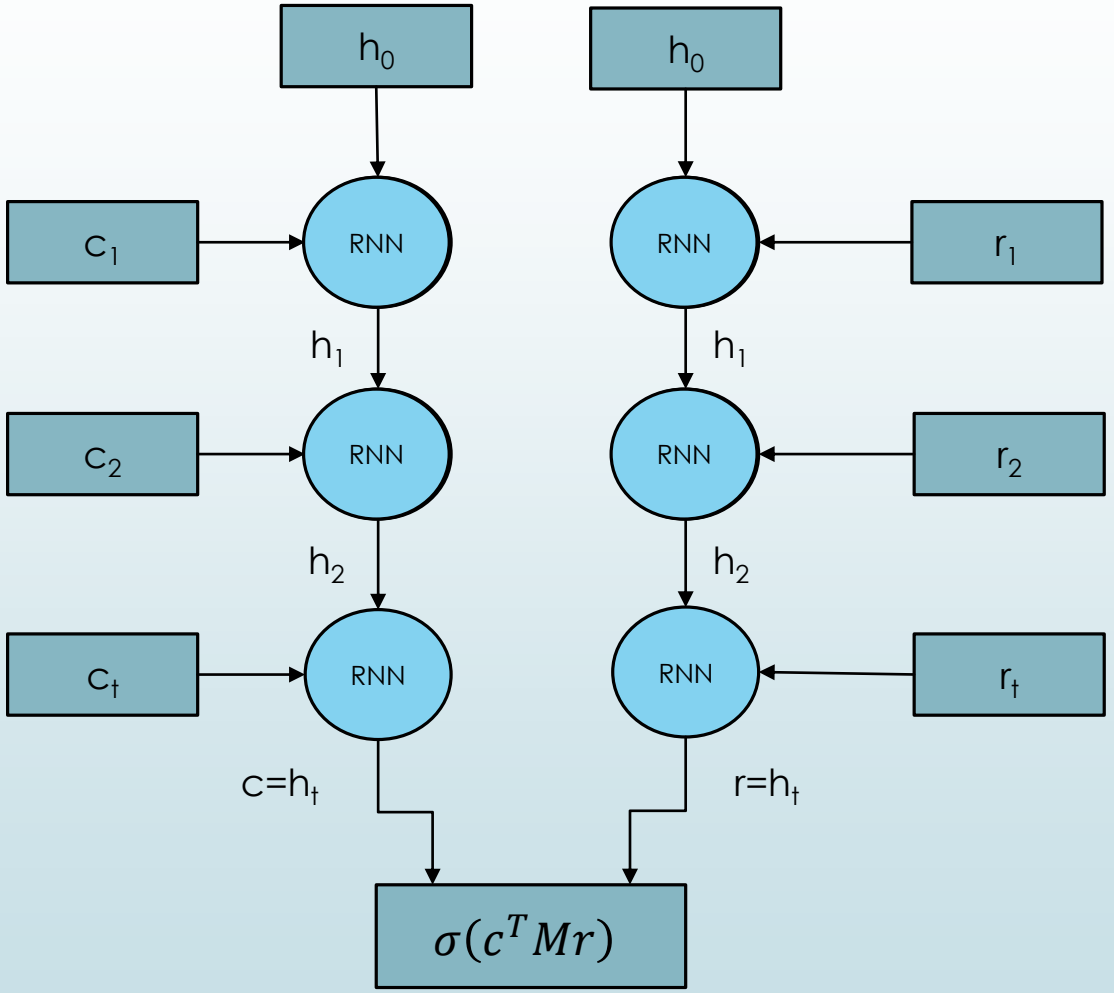
$$\text{wf-idf}_{t,d} = \mathrm{wf}_{t,d} \times \mathrm{idf}_t$$

Results on TF-IDF Models : Recall@K

Approach	Recall @ 1	Recall @ 2	Recall @ 5
Random	0.0937632	0.194503	0.49297
Unigram	0.495032	0.596882	0.766121
Bigram	0.4804556	0.601344	0.775422
Trigram	0.4814974	0.602344	0.772134
Unigram with Sublinear tf Scaling	0.5134245	0.603767	0.769298

RNN with LSTM Model

Training

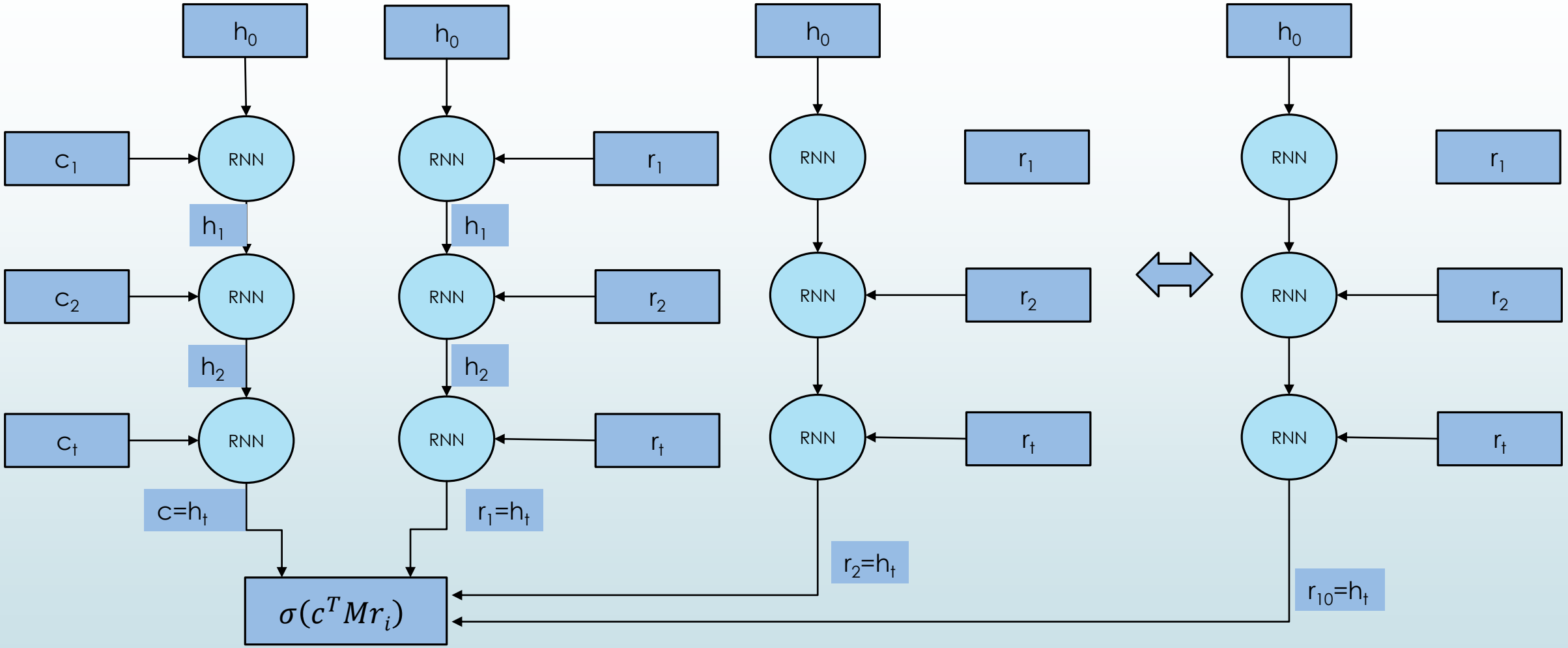


Loss Function:

Binary Cross Entropy :

$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right],$$

Test



Here we pass 10 utterances of which 1 is ground truth and 9 are distractors.

Result of RNN model

Recall @ 1	0.447251585624
Recall @ 2	0.64011627907
Recall @ 5	0.887843551797
Recall @ 10	1.0

Experiment: Similar utterances

Cosine Model:

- Recall @ 1: 0.49
- Recall @ 2: 0.606
- Recall @ 5: 0.789

Context	Ground truth	9 Utterances Retrieved using cosine similarity

LSTM Model:

- Recall @ 1: 0.15
- Recall @ 2: 0.25
- Recall @ 5: 0.42

Context	Ground truth	99 Utterances Retrieved using cosine similarity.

Context	Top 10 utterances

- Observation: LSTM model shows poor performance than earlier test data as utterances are more similar to the ground truth as opposed to random utterances in earlier test data. Because of this, ranking by LSTM has become more difficult.

Experiment: Candidate Retrieval

We have defined a metric for candidate retrieval:

$$1/100 \sum_{i=1}^{100} \frac{number\ of\ relevant\ retrieved\ responses}{Total\ number\ of\ retrieved\ responses}$$

Comparing Cosine similarity and LSTM approach:

- Using Cosine similarity (for 100 samples):
  - Average number of responses greater than threshold, 0.0871
  - Number of cases in which at least one answer is above threshold, 67
- Using Cosine Similarity with LSTM (for 100 samples):
  - Average number of responses greater than threshold, 0.0820
  - Number of cases in which at least one answer is above threshold, 52

Sample Output: Cosine Model

```
Query:
i know i 'm probabl do someth stupid here , but i ca n't figur out how to instal ubuntu to sdb . all the instal show
be sda . gpart can see sdb __eou__ ani idea ? __eou__ __eot__ use the somthign else/custom ' option and make your par
tit on sdb as you want . ie : sdb1 = / sdb2 = /home/ sdb3 = swap . __eou__ __eot__ yeah when i choos `` someth els ''
i 'm not see sdb in there either . i 'm not sure whi __eou__ __eot__
--------------------------------------------------------
Actual Response
you can partion the hd with gpart from the live cd , then start the installer.. perhap . __eou__ that how i tend to d
o it . __eou__ the instal partion manag tool be a bite . annoy . __eou__ i also notic the instal do not have a instal
to a specif drive use the whold drive ' option.. __eou__ sort of annoy it will autom other things.. but not a fair co
mmon case of a seper hd just for linux . __eou__
--------------------------------------------------------
Generated Response
you can partion the hd with gpart from the live cd , then start the installer.. perhap . __eou__ that how i tend to d
o it . __eou__ the instal partion manag tool be a bite . annoy . __eou__ i also notic the instal do not have a instal
to a specif drive use the whold drive ' option.. __eou__ sort of annoy it will autom other things.. but not a fair co
mmon case of a seper hd just for linux . __eou__
```

Sample Output: LSTM Model

```
Input Context :
i know i 'm probabl do someth stupid here , but i ca n't figur out how to instal ubuntu to sdb . all the instal show be
sda . gpart can see sdb __eou__ ani idea ? __eou__ __eot__ use the somthign else/custom ' option and make your partit on
 sdb as you want . ie : sdb1 = / sdb2 = /home/ sdb3 = swap . __eou__ __eot__ yeah when i choos `` someth els '' i 'm not
 see sdb in there either . i 'm not sure whi __eou__ __eot__
--------------------------------------------------------------------------------------------------------------
Response given by the model and actual answer are same:
you can partion the hd with gpart from the live cd , then start the installer.. perhap . __eou__ that how i tend to do i
t . __eou__ the instal partion manag tool be a bite . annoy . __eou__ i also notic the instal do not have a instal to a
specif drive use the whold drive ' option.. __eou__ sort of annoy it will autom other things.. but not a fair common cas
e of a seper hd just for linux . __eou__
```

References:

- Ryan Lowe, Nissan Pow, Iulian Serban, Joelle Pineau, "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems", 2015
- Zhao Yan , Nan Duan , Junwei Bao , Peng Chen , Ming Zhou , Zhoujun Li , Jianshe Zhou, "DocChat: An Information Retrieval Approach for Chatbot Engines Using Unstructured Documents", 2016
- Introduction to information retrieval, Textbook by Christopher D. Manning, Hinrich Schütze, and Prabhakar Raghavan

Thank you