Project Proposal:
“Exploring the effect of dataset on chat bot performance”

Weiting Zhan and Hyein Jeong

Outline

- Goal of the project
- Motivation/previous work
- Pipeline
    - Data
    - sequence to sequence model
    - sequence to sequence +Dialogue act model
- Set of tasks and timeline

Goal of the project

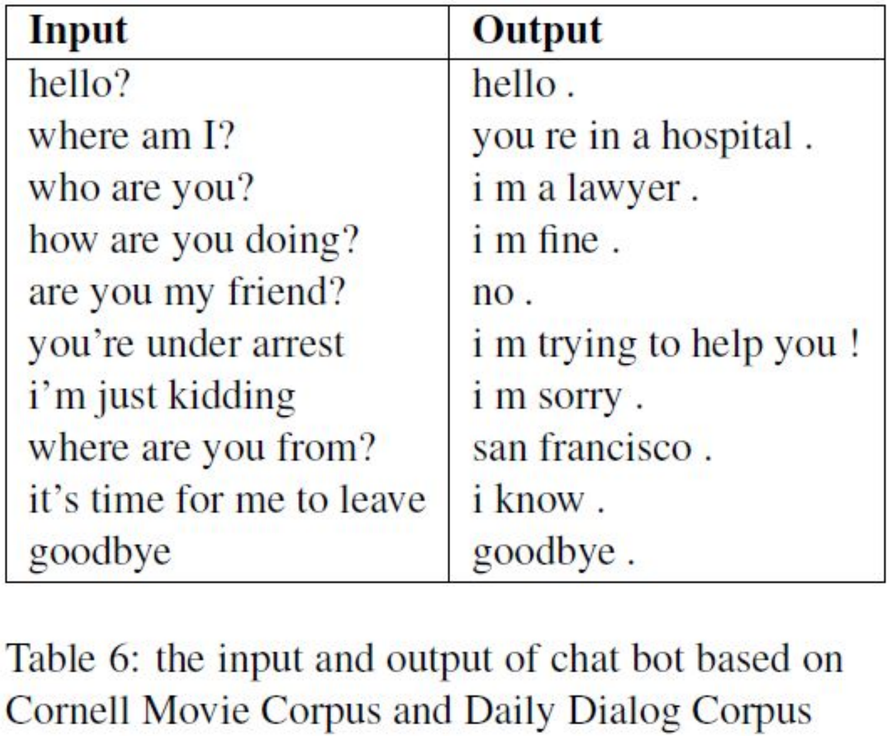
1. To learn how to implement sequence to sequence model to build a chat bot
2. To explore the effect of dataset on chat bot performance

   Cornell movie corpus

   Daily Dialogue

3. Use statistics method instead of template method, which is a typical way to build a chat bot

Input and output:



Input	Output
hello?	hello .
where am I?	you re in a hospital .
who are you?	i m a lawyer .
how are you doing?	i m fine .
are you my friend?	no .
you’re under arrest	i m trying to help you !
i’m just kidding	i m sorry .
where are you from?	san francisco .
it’s time for me to leave	i know .
goodbye	goodbye .

Table 6: the input and output of chat bot based on Cornell Movie Corpus and Daily Dialog Corpus

Dataset - Cornell movie dataset

Movie character dialogue

- 220,579 conversational exchanges between 10,292 pairs of movie characters
- 9,035 characters from 617 movies
- 304,713 total utterances
- There is a great variation of language formality, time periods, sentiment, which makes a model robust to generate many forms of inputs and queries. Link: https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html

Previous work

- ***"Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs"*** by Cristian Danescu-Niculescu-Mizil and Lillian Lee: focused on adjacent-utterance coordination with respect to classes of function words

Dataset - Cornell movie dataset

Original format of cornell-movie dialogue corpus: contains the actual text of each utterance (line ID, character ID, movie ID, character names, and text of the utterance)

```
b'L1045 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ They do not!\n'

b'L1044 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ They do to!\n'

b'L985 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ I hope so.\n'

b'L984 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ She okay?\n'

b"L925 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ Let's go.\n"

b'L924 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ Wow\n'
```

Dataset - Cornell movie dataset

Reformatting cornell-movie dialogue corpus: a tab-separated query sentence and a response sentence pair

```
b"Can we make this quick?  Roxanne Korrine and Andrew Barrett are having an incredibly horrendous public break- up on the quad.  Again.\tWell, I thought we'd start with pronunciation, if that's okay with you.\n"

b"Well, I thought we'd start with pronunciation, if that's okay with you.\tNot the hacking and gagging and spitting part.  Please.\n"

b"Not the hacking and gagging and spitting part.  Please.\tOkay... then how 'bout we try out some French cuisine.  Saturday?  Night?\n"

b"You're asking me out.  That's so cute. What's your name again?\tForget it.\n"
```

Preprocessing Dataset

Assembling vocabulary and query/response sentence pairs, and preprocessing: convert into lowercase, trim all non-letter characters/rarely used words, and limit the length of sentences

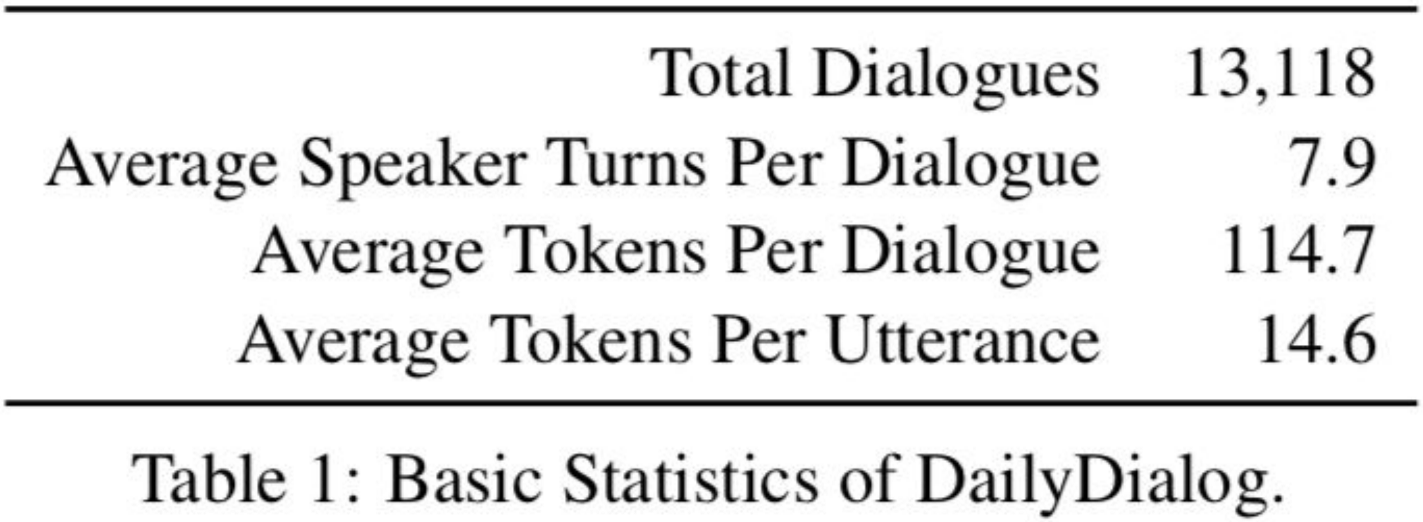
```
pairs:

['there .', 'where ?']

['you have my word . as a gentleman', 'you re sweet .']

['hi .', 'looks like things worked out tonight huh ?']

['you know chastity ?', 'i believe we share an art instructor']
```

Preprocessing Dataset

Changing sentence pairs into numerical index by using mini-batches

```
input_variable: tensor([[  34,    16,   147, 4556,    50],

        [   4, 5553,   379, 2528,     6],

        [  76, 3810,    45,     6,     2],

        [ 331,     4,   380,     2,     0],

        [ 117,     4, 1967,     0,     0],

        [  47,     4,     6,     0,     0],

        [   4,     2,     2,     0,     0],
```

Dataset - Daily Dialogue



Total Dialogues	13,118
Average Speaker Turns Per Dialogue	7.9
Average Tokens Per Dialogue	114.7
Average Tokens Per Utterance	14.6

Table 1: Basic Statistics of DailyDialog.

Dataset - Daily Dialogue

Original data

I made my little brother cry . __eou__ Shame on you ! __eou__

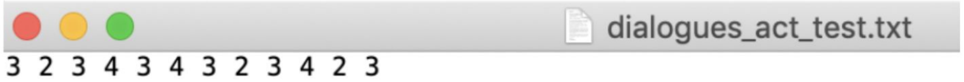
Can I have the check please ? __eou__ Right away , sir . __eou__ Is this amount correct ? __eou__ Yes , sir , it is . __eou__ Thank you , the service was good __eou__ Please come again . __eou__

Are you through with your meal ? __eou__ Yes , we are . Could we have the check please ? __eou__ Here you go . I can take care of it here when you are already . __eou__ Do you accept checks ? __eou__ No , I'm sorry we don't . We accept credit cards and cash . __eou__ Well , I don't have any cash with me , I'll have to put it on a credit .

Dataset - Features of Daily Dialogue

- Manually label each utterance with one of four dialogue act classes
    - **Inform:** Contains all statements and questions by which the speaker is providing information.
    - **Questions**: Labeled when the speaker wants to know something and seeks for information.
    - **Directives**: Contains dialogue acts like request, instruct, suggest, and accept/reject offer.
    - **Commissive**: About accept/reject request, suggestion and offer.
- Manually label the dialogue with emotions
    - six primary and universal emotions: {Anger, Disgust, Fear, Happiness, Sadness, Surprise}

Dataset - labeled with dialogue act



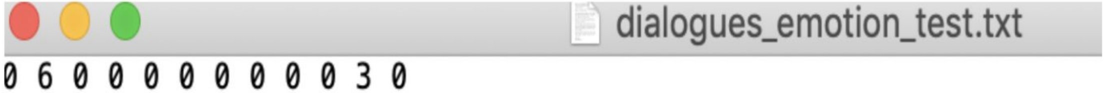
3 2 3 4 3 4 3 2 3 4 2 3

Dialog :{ 1: inform, 2: question, 3: directive, 4: commissive }

Hey man , you wanna buy some weed ? (3 directive)
__eou__ Some what ? (2 question)
__eou__ Weed ! You know ? Pot , Ganja , Mary Jane some chronic ! (3 directive)
__eou__ Oh , umm , no thanks . (4 commissive)
__eou__ I also have blow if you prefer to do a few lines . (3 directive)
__eou__ No , I am ok , really . (4 commissive)
__eou__ Come on man ! I even got dope and acid ! Try some ! (3 directive)
__eou__ Do you really have all of these drugs ? Where do you get them from ? (2 question)
__eou__ I got my connections ! Just tell me what you want and I ' ll even give you one ounce for free . (3 directive)
__eou__ Sounds good ! Let ' s see , I want . (4 commissive)
__eou__ Yeah ? (2 question)
__eou__ I want you to put your hands behind your head ! You are under arrest ! (3 directive)
__eou__

Dataset - labeled with emotion

Emotion :{ 0: no emotion, 1: anger, 2: disgust, 3: fear, 4: happiness, 5: sadness, 6: surprise}



dialogues_emotion_test.txt

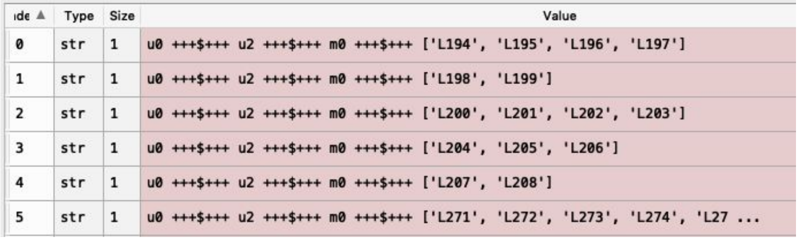
```
0 6 0 0 0 0 0 0 0 0 3 0
```

Hey man , you wanna buy some weed ? (0 no emotion)
__eou__ Some what ? (6 surprise)
__eou__ Weed ! You know ? Pot , Ganja , Mary Jane some chronic ! (0 no emotion)
__eou__ Oh , umm , no thanks . (0 no emotion)
__eou__ I also have blow if you prefer to do a few lines . (0 no emotion)
__eou__ No , I am ok , really . (0 no emotion)
__eou__ Come on man ! I even got dope and acid ! Try some ! (0 no emotion)
__eou__ Do you really have all of these drugs ? Where do you get them from ? (0 no emotion)
__eou__ I got my connections ! Just tell me what you want and I ' ll even give you one ounce for free . (0 no emotion)
__eou__ Sounds good ! Let ' s see , I want . (0 no emotion)
__eou__ Yeah ? (3 fear)
__eou__ I want you to put your hands behind your head ! You are under arrest ! (0 no emotion)
__eou__

Dataset - previous work using daily dialogue

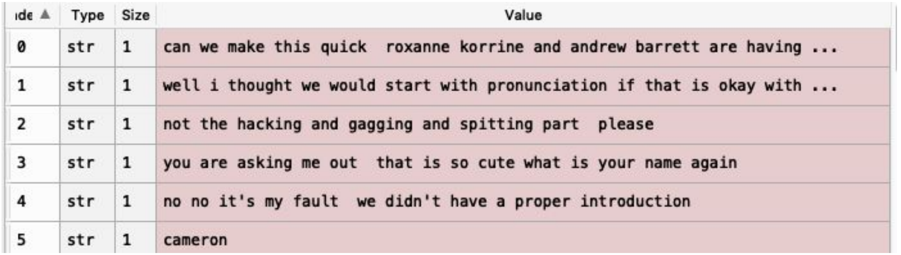
- **An Auto-Encoder Matching Model for Learning Utterance-Level Semantic Dependency in Dialogue Generation**: The paper said the Daily Dialogue is a high quality dataset. And their model aims to find utterance-level relations between conversations.
- **DIALOGWAE: MULTIMODAL RESPONSE GENERATION WITH CONDITIONAL WASSERSTEIN AUTO-ENCODER:** Uses Daily Dialogue as the benchmark.
- **Improving Variational Encoder-Decoders in Dialogue Generation:** Used the dataset as training corpus.
- **Chat More: Deepening and Widening the Chatting Topic via A Deep Model:** Used Daily Dialogue as a benchmark.

Cornell-Movie corpus

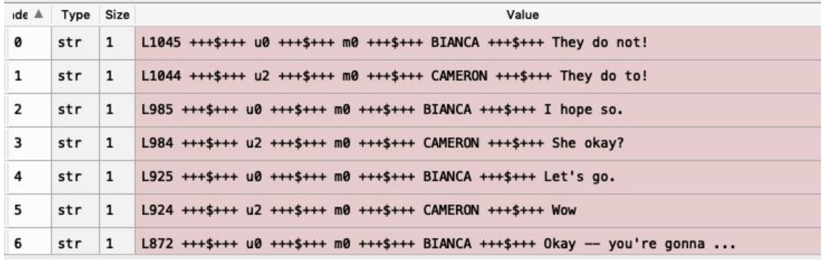


ıde ▲	Type	Size	Value
0	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L194', 'L195', 'L196', 'L197']
1	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L198', 'L199']
2	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L200', 'L201', 'L202', 'L203']
3	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L204', 'L205', 'L206']
4	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L207', 'L208']
5	str	1	u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L271', 'L272', 'L273', 'L274', 'L27 ...

Question:

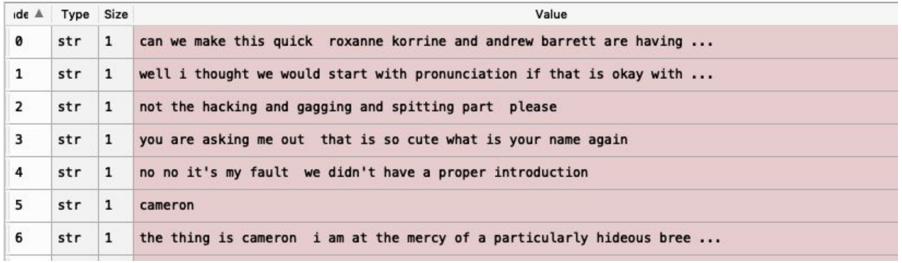


ıde ▲	Type	Size	Value
0	str	1	can we make this quick roxanne korrine and andrew barrett are having ...
1	str	1	well i thought we would start with pronunciation if that is okay with ...
2	str	1	not the hacking and gagging and spitting part please
3	str	1	you are asking me out that is so cute what is your name again
4	str	1	no no it's my fault we didn't have a proper introduction
5	str	1	cameron



ıde ▲	Type	Size	Value
0	str	1	L1045 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ They do not!
1	str	1	L1044 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ They do to!
2	str	1	L985 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ I hope so.
3	str	1	L984 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ She okay?
4	str	1	L925 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ Let's go.
5	str	1	L924 +++$+++ u2 +++$+++ m0 +++$+++ CAMERON +++$+++ Wow
6	str	1	L872 +++$+++ u0 +++$+++ m0 +++$+++ BIANCA +++$+++ Okay -- you're gonna ...

Answer:



ıde ▲	Type	Size	Value
0	str	1	can we make this quick roxanne korrine and andrew barrett are having ...
1	str	1	well i thought we would start with pronunciation if that is okay with ...
2	str	1	not the hacking and gagging and spitting part please
3	str	1	you are asking me out that is so cute what is your name again
4	str	1	no no it's my fault we didn't have a proper introduction
5	str	1	cameron
6	str	1	the thing is cameron i am at the mercy of a particularly hideous bree ...

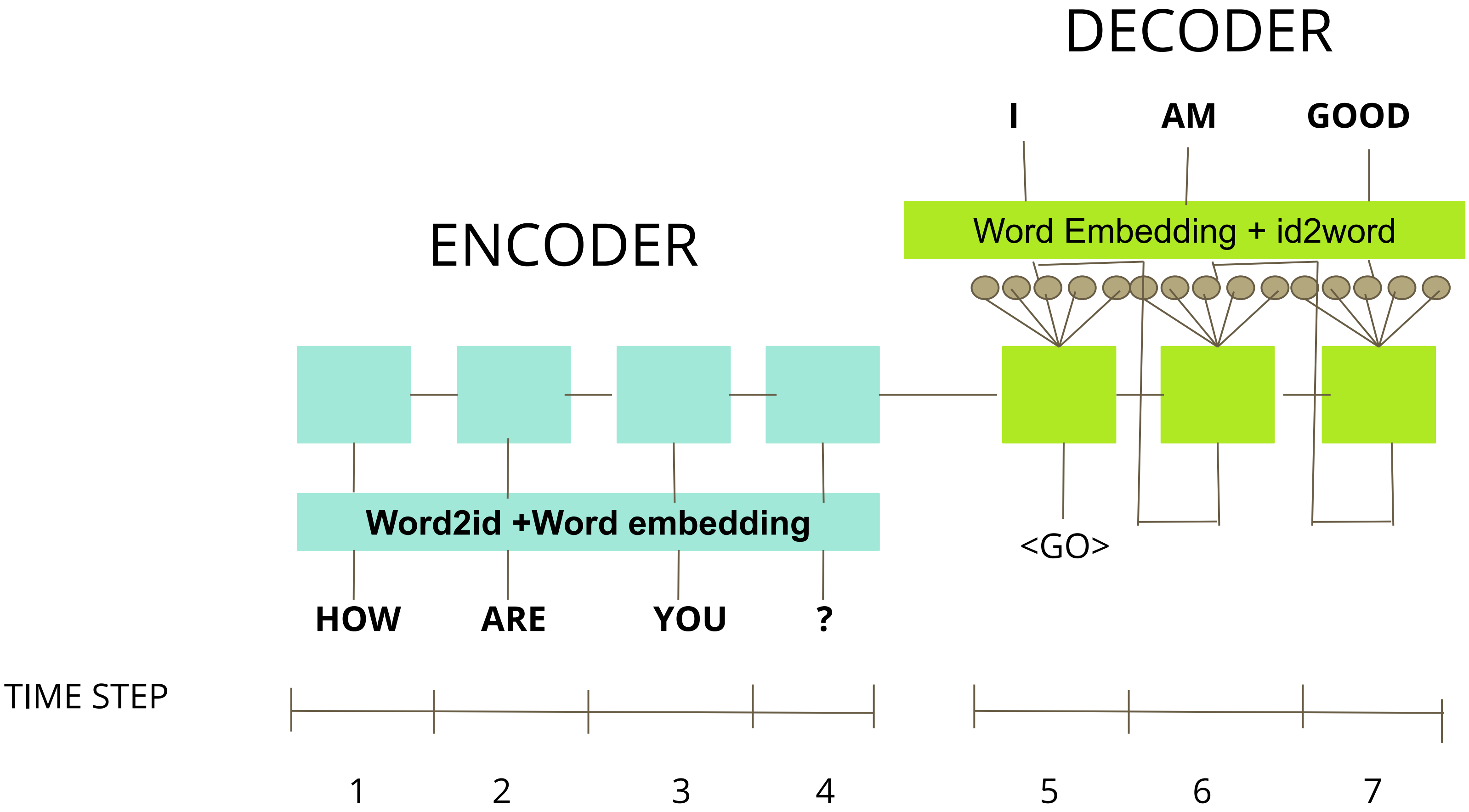
Clean data

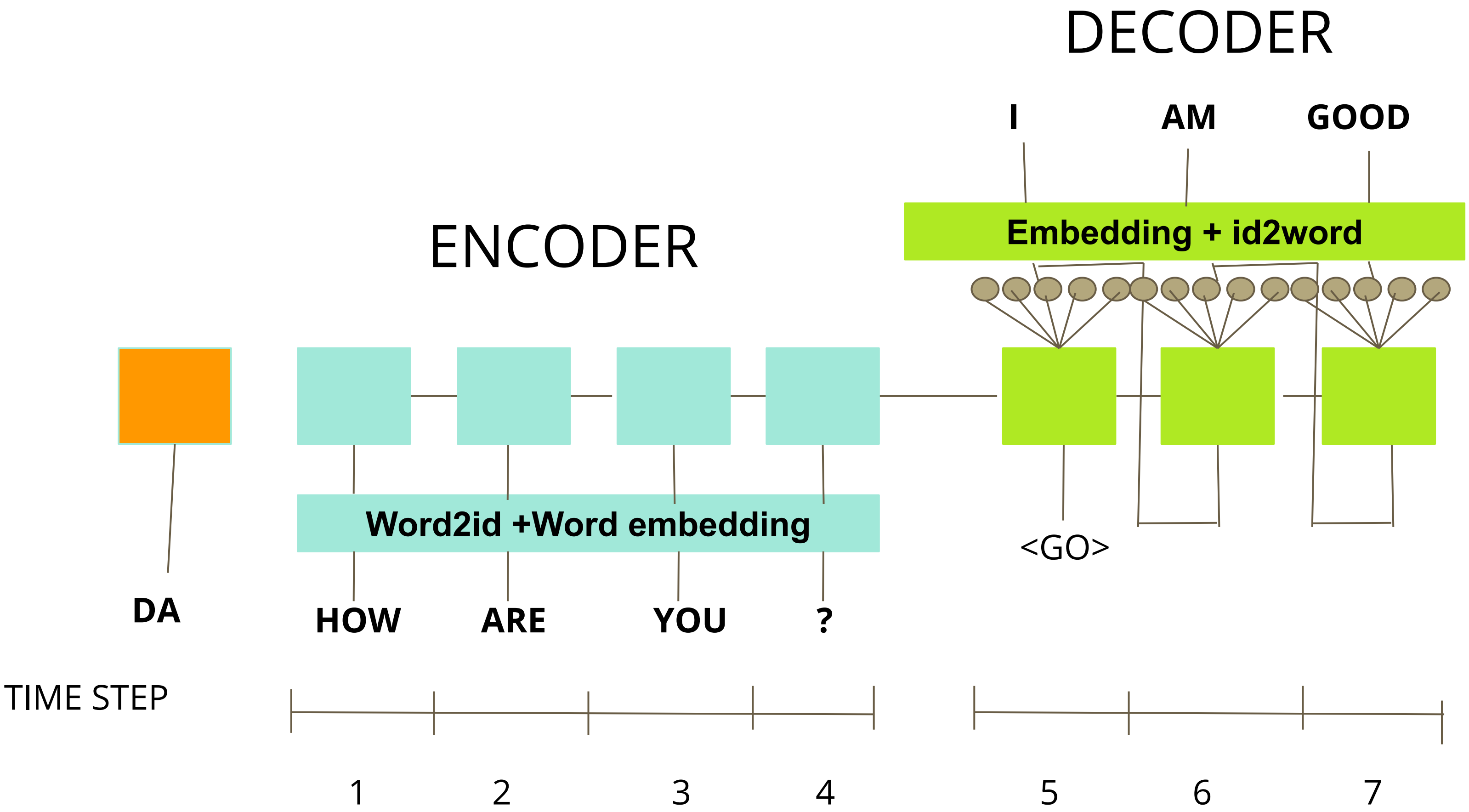
1. Lower case all the words.
2. Replacement

```
text = re.sub(r"i'm", "i am", text)
text = re.sub(r"he's", "he is", text)
text = re.sub(r"she's", "she is", text)
text = re.sub(r"that's", "that is", text)
text = re.sub(r"what's", "what is", text)
text = re.sub(r"where's", "where is", text)
text = re.sub(r"\'ll", " will", text)
text = re.sub(r"\'ve", " have", text)
text = re.sub(r"\'re", " are", text)
text = re.sub(r"\'d", " would", text)
text = re.sub(r"won't", "will not", text)
text = re.sub(r"can't", "cannot", text)
text = re.sub(r"[-()\"#/@;:<>{}+=~|.?,]", "", text)
```

Sequence to Sequence model

- Build dictionary of the dataset
- Build encoder model
- Build decoder model
- Greedy decoding and Beam search decoding


DECODER
I
AM
GOOD
Word Embedding + id2word
ENCODER
Word2id +Word embedding
<GO>
HOW
ARE
YOU
?
TIME STEP
1
2
3
4
5
6
7



DECODER

I

AM

GOOD

Embedding + id2word

ENCODER

Word2id +Word embedding

<GO>

DA

HOW

ARE

YOU

?

TIME STEP

1

2

3

4

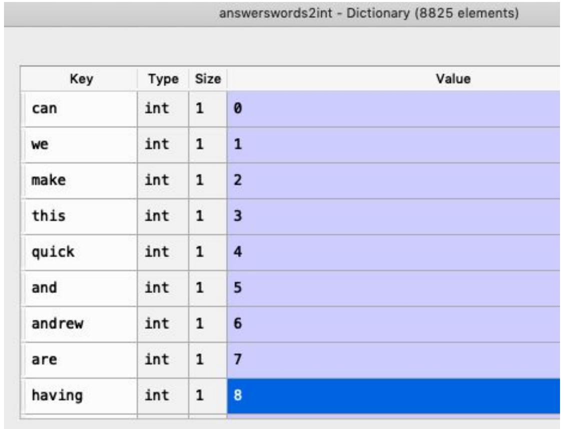
5

6

7

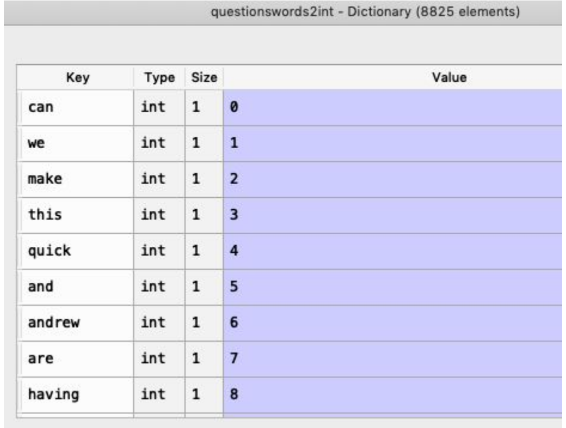
Building Dictionary: Word2id and id2Word

- Creating two dictionaries that map the questions words and the answers words to a unique integer
- Translating all the questions and the answers into integers.



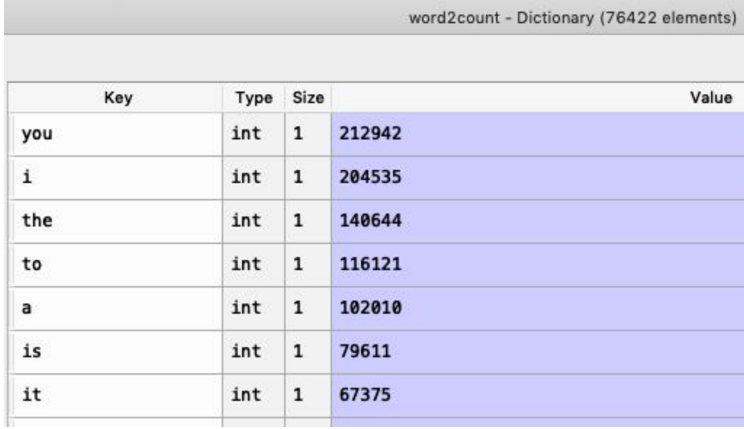
answerswords2int - Dictionary (8825 elements)

Key	Type	Size	Value
can	int	1	0
we	int	1	1
make	int	1	2
this	int	1	3
quick	int	1	4
and	int	1	5
andrew	int	1	6
are	int	1	7
having	int	1	8



questionswords2int - Dictionary (8825 elements)

Key	Type	Size	Value
can	int	1	0
we	int	1	1
make	int	1	2
this	int	1	3
quick	int	1	4
and	int	1	5
andrew	int	1	6
are	int	1	7
having	int	1	8

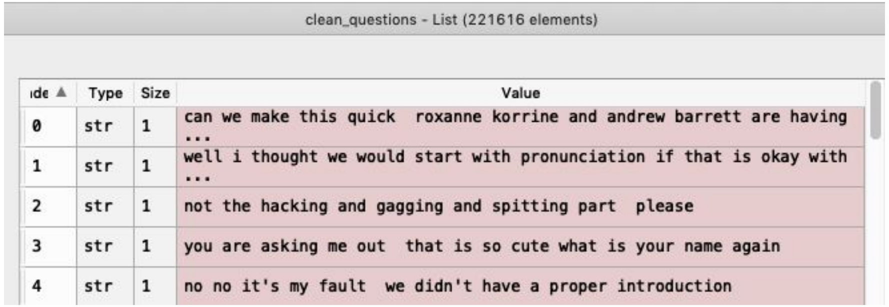


word2count - Dictionary (76422 elements)

Key	Type	Size	Value
you	int	1	212942
i	int	1	204535
the	int	1	140644
to	int	1	116121
a	int	1	102010
is	int	1	79611
it	int	1	67375

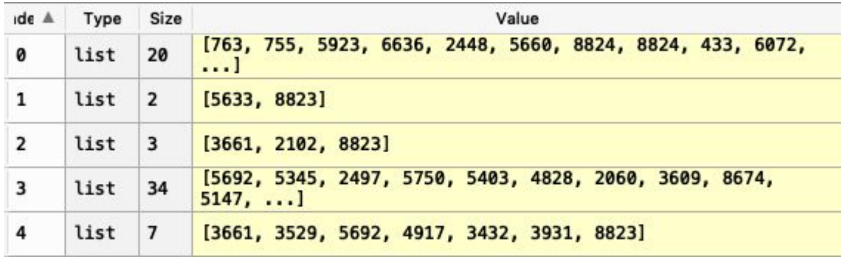
Word2id

Question:



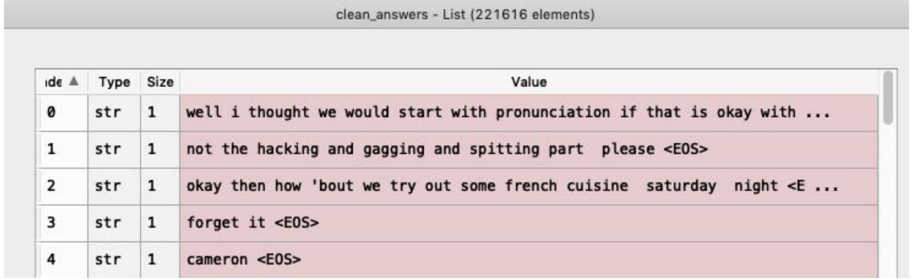
clean_questions - List (221616 elements)

ıde ▲	Type	Size	Value
0	str	1	can we make this quick roxanne korrine and andrew barrett are having ...
1	str	1	well i thought we would start with pronunciation if that is okay with ...
2	str	1	not the hacking and gagging and spitting part please
3	str	1	you are asking me out that is so cute what is your name again
4	str	1	no no it's my fault we didn't have a proper introduction



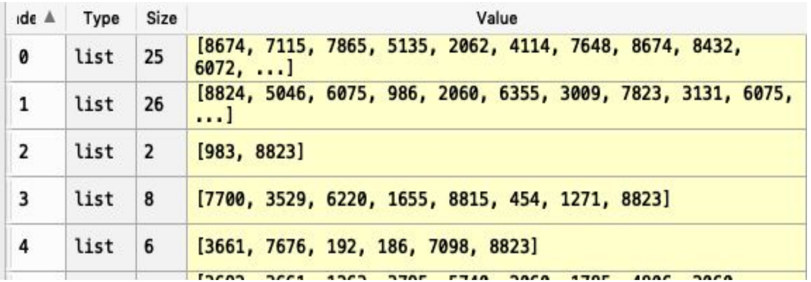
ıde ▲	Type	Size	Value
0	list	20	[763, 755, 5923, 6636, 2448, 5660, 8824, 8824, 433, 6072, ...]
1	list	2	[5633, 8823]
2	list	3	[3661, 2102, 8823]
3	list	34	[5692, 5345, 2497, 5750, 5403, 4828, 2060, 3609, 8674, 5147, ...]
4	list	7	[3661, 3529, 5692, 4917, 3432, 3931, 8823]

Answer:



clean_answers - List (221616 elements)

ıde ▲	Type	Size	Value
0	str	1	well i thought we would start with pronunciation if that is okay with ...
1	str	1	not the hacking and gagging and spitting part please <EOS>
2	str	1	okay then how 'bout we try out some french cuisine saturday night <E ...
3	str	1	forget it <EOS>
4	str	1	cameron <EOS>



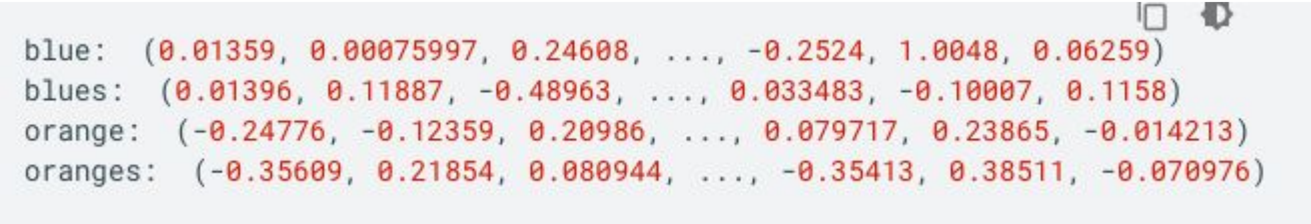
ıde ▲	Type	Size	Value
0	list	25	[8674, 7115, 7865, 5135, 2062, 4114, 7648, 8674, 8432, 6072, ...]
1	list	26	[8824, 5046, 6075, 986, 2060, 6355, 3009, 7823, 3131, 6075, ...]
2	list	2	[983, 8823]
3	list	8	[7700, 3529, 6220, 1655, 8815, 454, 1271, 8823]
4	list	6	[3661, 7676, 192, 186, 7098, 8823]

Word embedding

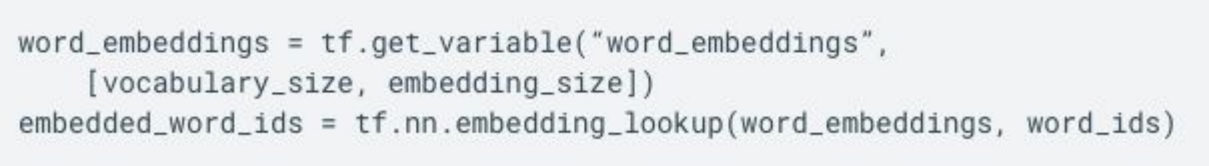
It takes integers as input, it looks up these integers in an internal dictionary,

and it returns the associated vectors.

Use Tensorflow embedding function:https://www.tensorflow.org/guide/embedding



```
blue:  (0.01359, 0.00075997, 0.24608, ..., -0.2524, 1.0048, 0.06259)
blues:  (0.01396, 0.11887, -0.48963, ..., 0.033483, -0.10007, 0.1158)
orange:  (-0.24776, -0.12359, 0.20986, ..., 0.079717, 0.23865, -0.014213)
oranges:  (-0.35609, 0.21854, 0.080944, ..., -0.35413, 0.38511, -0.070976)
```



```
word_embeddings = tf.get_variable("word_embeddings",
    [vocabulary_size, embedding_size])
embedded_word_ids = tf.nn.embedding_lookup(word_embeddings, word_ids)
```

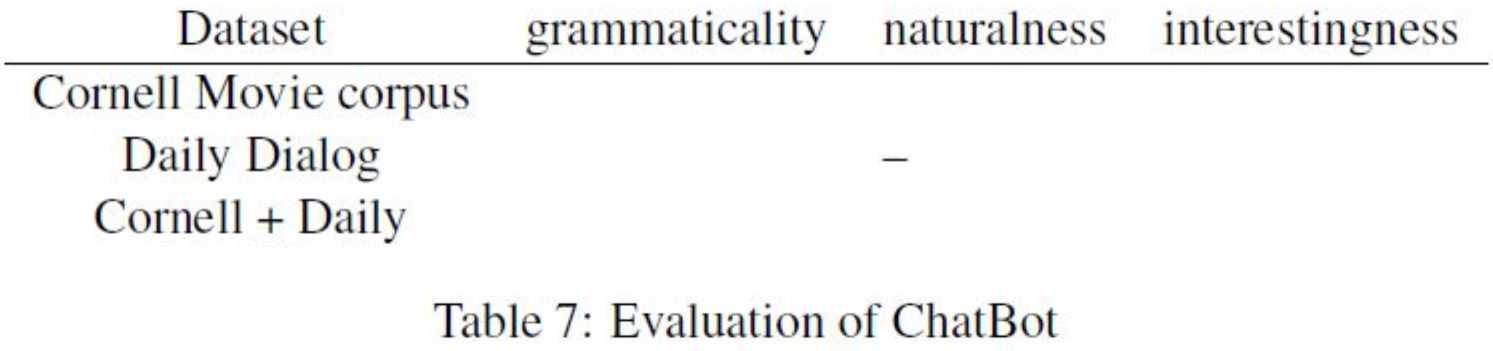
RNN Encoder

```
def encoder_rnn(rnn_inputs, rnn_size, num_layers, keep_prob, sequence_length):
        lstm = tf.contrib.rnn.BasicLSTMCell(rnn_size)
        lstm_dropout = tf.contrib.rnn.DropoutWrapper(lstm, input_keep_prob = keep_prob)
        encoder_cell = tf.contrib.rnn.MultiRNNCell([lstm_dropout] * num_layers)
        encoder_output, encoder_state = tf.nn.bidirectional_dynamic_rnn(cell_fw = encoder_cell,
cell_bw = encoder_cell, sequence_length = sequence_length,inputs = rnn_inputs, dtype =
tf.float32)
        return encoder_state
```

RNN -decoder

```
def decoder_rnn(decoder_embedded_input, decoder_embeddings_matrix, encoder_state, num_words,
sequence_length, rnn_size, num_layers, word2int, keep_prob, batch_size):
        with tf.variable_scope("decoding") as decoding_scope:
        lstm = tf.contrib.rnn.BasicLSTMCell(rnn_size):
        lstm_dropout = tf.contrib.rnn.DropoutWrapper(lstm, input_keep_prob = keep_prob)
        decoder_cell = tf.contrib.rnn.MultiRNNCell([lstm_dropout] * num_layers)
        weights = tf.truncated_normal_initializer(stddev = 0.1)
        biases = tf.zeros_initializer()
        output_function = lambda x: tf.contrib.layers.fully_connected(x,num_words,None,scope = decoding_scope,
weights_initializer = weights, biases_initializer = biases)
        training_predictions = decode_training_set(encoder_state,decoder_cel,
decoder_embedded_input,sequence_length,decoding_scope,output_function, keep_prob,batch_size)
        decoding_scope.reuse_variables()
        test_predictions = decode_test_set(encoder_state,decoder_cell,decoder_embeddings_matrix,
word2int['<SOS>'],word2int['<EOS>'],sequence_length -
1,num_words,decoding_scope,output_function,keep_prob,batch_size)
        return training_predictions, test_predictions
```

Human Evaluation



Dataset	grammaticality	naturalness	interestingness
Cornell Movie corpus			
Daily Dialog		–	
Cornell + Daily			

Table 7: Evaluation of ChatBot

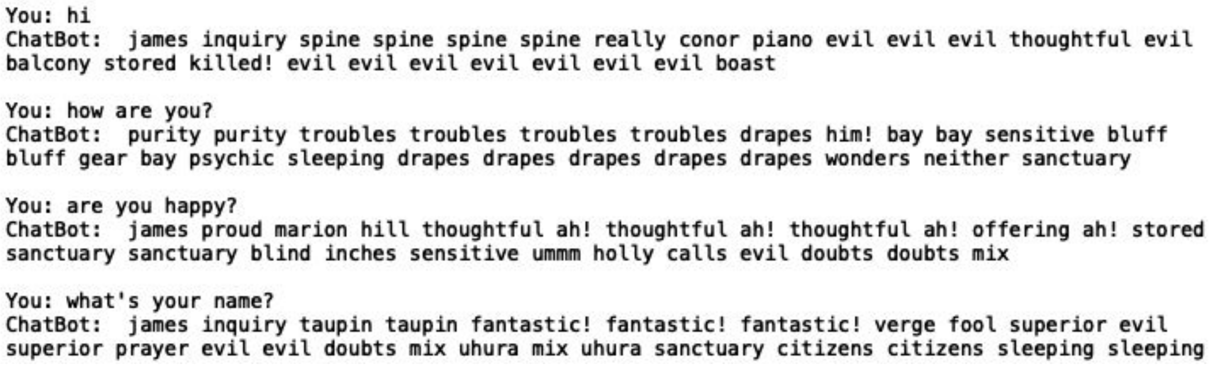
Timeline

May 10th	[Question, Answer] pairs format for Cornell Corpus [Question, answer] paris format for Daily Dialogue
May 15th	Sequence to sequence chat bot model (take at least 20 hours to train)
May 20th	Sequence to sequence chat bot model + dialog act
May 25th	Improve the model
May 30th	Evaluation
June 5th	Writing Report

Initial result

20 hours training .

```
# Setting the Hyperparameters
epochs = 10   ----1000
batch_size = 32
rnn_size = 102
num_layers = 3
encoding_embedding_size = 102 ----1000
decoding_embedding_size = 102 ----1000
learning_rate = 0.1 ------0.001
learning_rate_decay = 0.9
min_learning_rate = 0.1
keep_probability = 0.5
```



```
You: hi
ChatBot:  james inquiry spine spine spine spine really conor piano evil evil evil thoughtful evil
balcony stored killed! evil evil evil evil evil evil evil boast

You: how are you?
ChatBot:  purity purity troubles troubles troubles troubles drapes him! bay bay sensitive bluff
bluff gear bay psychic sleeping drapes drapes drapes drapes drapes wonders neither sanctuary

You: are you happy?
ChatBot:  james proud marion hill thoughtful ah! thoughtful ah! thoughtful ah! offering ah! stored
sanctuary sanctuary blind inches sensitive ummm holly calls evil doubts doubts mix

You: what's your name?
ChatBot:  james inquiry taupin taupin fantastic! fantastic! fantastic! verge fool superior evil
superior prayer evil evil doubts mix uhura mix uhura sanctuary citizens citizens sleeping sleeping
```