University of Lorraine

Tutored Project

Dialogue Processing on Twitter

JAFARI Somaye and OLUOKUN Adedayo

supervised by
Prof. CERISARA Christophe

May 31, 2017

Contents

Chapter 1

Introduction

In order to understand dialogues, the ability to model and automatically understand discourse structure is essential. To achieve this, there is a need to describe discourse structure. The identification of dialogue acts (DAs) is a useful first step towards describing discourse structure.
A Dialogue Act is approximately the equivalent of the speech act of [Searle, 1969]

DAs can be thought of as a tag set that classifies utterances according to a combination of pragmatic, semantic, and syntactic criteria [Stolcke et al., 2000]. There exist DA labeling systems that are domain-independent, such as the Discourse Resource Initiatives DAMSL architecture [Core and Allen, 1997].

Social media and particularly Twitter have become a central data source for natural language processing methods and applications in recent years. The social and interactive nature of posts on twitter has not received much attention. Interestingly, up to 40% of all twitter messages are part of conversations [Scheffler, 2014].

The goal of this project is to analyze and annotate Twitter dialogues and build a neural network model that automatically classifies the Twitter dialogues into corresponding dialogue acts.

In chapter 2 of this report, related works on dialogue act classification/recognition are extensively described. Chapter 3 describes the annotation and modeling processes in full detail. The empirical evaluation is described in chapter 4 followed by the conclusion and future work in chapter 5.

Chapter 2

Literature review on dialogue act recognition

Alot of work has been done in the field of dialogue act recognition. In this chapter, some of the works done in the field will be reviewed.

[Vosoughi and Roy, 2016] explored speech act recognition by treating it as a multi-class identification system. [Zhao and Jiang, 2011] definitions for topic and type was used. A topic is a subject discussed in one or more tweets (e.g., Boston Marathon bombings, Red Sox, etc). The type characterizes the nature of the topic, these are: Entity-oriented, Event-oriented topics, and Long-standing topics. Two topics for each of the three topic types were selected. They collected a few thousand tweets from the Twitter public API for each of these topics using topic-specific queries (e.g., #fergusonriots, #redsox, etc).

For training, the labels that the majority of annotators agreed upon (7,563 total tweets) were used. The features used can be divided into two general categories: Semantic and Syntactic. Some of these features were motivated by various works on speech act classification, while others are novel features. Overall, 3313 binary features, composed of 1647 semantic and 1666 syntactic features were selected. Using these features they were able to achieve state-of-the-art performance for Twitter speech act classification, with an average F1 score of .70.

Four different classifiers were trained on 3,313 binary features using the following methods: naive bayes (NB), decision tree (DT), logistic regression (LR), SVM, and a baseline max classifier BL.

[Ang et al., 2005] used a MaxEnt classifier over a small set of 5 broad DA classes. Reports an overall classification accuracy of 81 percent on gold segments and based on lexical and prosodic features, which is only marginally improved by adding sequence information.

For social media data, [Forsyth and Martell, 2007] built a dialogue act recognizer for chat messages with a custom made schema of 15 dialogue acts. They consider each turn to correspond to only one DA, even though they note that several acts can appear within one turn in their data.

For Twitter data, the earliest work is [Ritter et al., 2010], who use unsupervised learning to extract dialogue act functions from Twitter data. Summary of Dialogue Act Recognition for Twitter Conversations. All the previous work on DA classification in social media assign exactly one DA to each post even though they might contain more than one DA.

[Scheffler and Zarisheva, 2016] introduced an approach to supervised dialogue act classification for German Twitter conversations where they viewed entire conversations as dialogues and classified individual segments within tweets.

Their work compares well to some previous DA recognition projects such as [Ang et al., 2005] on multiparty meetings, but stays far behind large efforts like [Stolcke et al., 2000], who report recognition accuracy of 0.71 on the Switchboard corpus with 42 DAs (their work: 0.37 for 51 DAs).

[Stolcke et al., 2000] used a statistical approach for modelling dialogue acts in conversational speech. The model detects and predicts the dialogue acts based on some parameters, such as, lexical, collocational, and prosodic cues, as well as on the discourse coherence of the dialogue act sequence. 42 dialogue act labels are defined on spontaneous telephone speech. The structure followed for this tag set was based on discourse structure annotation, the dialogue Act Markup in Several Layers (DAMSL) tag set [Core and Allen, 1997].

The tag set was defined based on DAMSL markup system, but later, modifications were made based on the corpus and task. The modifications were made in way that the tage set can be mapped back to DAMSL categories.

The domain which was chosen in this paper was the Switchboard corpus of human-human conversational telephone speech [Godfrey et al., 1992]. A large hand-labeled database of 1,155 conversation were produced for this work out of this corpus which was later used for training the model.

Dialogue act classification which was the main goal of the paper to be performed was done by probabilistic approach for combining multiple knowledge sources, and the ability to derive model parameters automatically from a corpus, using statistical inference techniques. In this paper, dialogue Act Decoding, is finally done using HMM representation, which allows computing aspects of dialogue modeling like the most probable DA sequence and the posterior probability of various DAs for a given utterance, after considering all the evidence. The prior probabilities of DA sequences are modeled in this paper by the statistical discourse grammar. Since a computationally convenient type of discourse grammar which also allows efficient decoding in the HMM framework is an n -gram model based on DA tags, the standard backoff n-gram models was used here for this purpose.

Non n-gram discourse models, were also investigated in the approach of this paper such as decision trees, and neural networks which were used to model the idiosyncratic lexical and prosodic manifestations of each dialogue act. Dialogue act labeling accuracy achieved in this work was (65% based on errorful, automatically recognized words and prosody, and 71% based on word transcripts, compared to a chance baseline accuracy of 35% and human accuracy of 84%).

[Zarisheva and Scheffler, 2015] presents a dialogue act annotation for German Twitter conversations. The corpus was Twitter data that was collected within the BMBF project Analysis of Discourses in Social Media and it was extracted considering the following criteria, 1. filtering out non-German tweets using the langid [Lui and Baldwin, 2012] and Compact Language Detection, 2. using the 4 libraries for Python 2.7, with some manual correction.

The schema of the annotation was based on the DA annotation schema on the general-purpose DIT++ taxonomy for dialogue acts [Ide and Bunt, 2010], the choice was because most of the fact that DA taxonomies are suitable for task-oriented dialogues or human-machine dialogues, while, Twitter conversations are a type of human-human, non-task-oriented dialogue. The schema chosen is a full schema of 57 dialogue acts.

The schema of the DIT++ was adapted according to their needs, because even DIT++ has a very limited range of non-task-oriented acts and in order to reflect the type of interactions, they expected in their data, and to reduce the difficulty of the annotation task we needed to do the adaptation.

The annotation validation was done by Fleiss multi method, which measures how consistent the assigned labels are for each item, without regard to which annotator gave the label. The inter-annotator agreement for DA labels on the raw annotation data, using the same procedure. For this measure, they only included those tweets where all three annotators agreed on the segmentation. With a reduced set of 14 DAs, three annotators achieve multi-= 0.65.

In this paper, an attempt to annotate Twitter conversations with a detailed dialogue act schema was presented which is one of the few works done in this scope. They achieved only moderate inter- annotator agreement of = 0.56 between three annotators on the Dialogue act labeling task, in contrast with work in other domains that achieved good agreement, there are some ways suggested to improve annotation accuracy.

Chapter 3

Twitter dialogue act annotation and modeling

3.1 The dialogue Act Labeling Task

In order to understand and analyze dialogue or more specifically tweets forming dialogues, there is a need to model dialogue to detect their discourse structure. There are different methods to describe the discourse structure of the dialogues, one of which is detecting the dialogue acts (DAs). What is meant here by DA is the same as the speech act of [Searle, 1969], (nearly equivalent). To give a more precise definition, A DA represents the meaning of an utterance at the level of illocutionary force [Austin, 1962].

DA labeling seems to be a useful approach to reach the final goal of this work which is presenting a framework to classify DAs of English dialogue tweets automatically. Therefore, there was a need for defining tag set of dialogue acts and manually annotate English tweets, to produce the gold corpus for this work.

3.1.1 Corpus

We chose to model human-human conversations on twitter. Each conversation involved random people and is made up of at least five tweets. Two annotators labeled 148 dialogues consisting of a total of 1039 tweets.

The inter-annotator agreement was calculated using the S-coefficient. The inter annotator agreement of the two annotators on the first 162 tweets (33 dialogues) and initial list of labels is 85.12% and is shown in table 3.1. Table 3.2

shows confusion matrix for the inter annotation.

.	TagMatch	DiffTag	Total
TagMatch	243	–	243
DiffTag	–	42	42
Total	243	42	**285**

Table 3.1: Inter-Annotation Agreement, S coefficient **85.12%**

.	A	B	D	G	M	O	P	Q	R	S	T	U	W
A	**10**	1	0	0	0	0	1	0	0	1	0	0	0
B	0	**16**	0	0	1	1	1	0	1	1	0	0	0
D	0	0	**2**	0	0	0	0	0	0	1	0	0	0
G	1	0	0	**4**	0	2	1	0	1	0	0	1	0
M	0	0	0	0	**0**	0	0	0	0	0	0	0	0
O	0	0	0	0	0	**15**	0	1	0	1	0	0	0
P	0	0	0	0	0	0	**6**	0	0	1	0	0	0
Q	0	0	0	0	0	0	1	**15**	1	0	0	0	0
R	0	0	0	0	0	0	0	0	**16**	2	0	0	0
S	0	2	0	0	5	0	3	0	2	**78**	1	0	0
T	0	0	0	0	0	0	0	0	0	0	**6**	0	0
U	0	0	0	0	0	0	1	0	0	0	0	**0**	0
W	0	1	0	0	0	0	0	0	0	0	0	0	**10**

Table 3.2: Confusion matrix for manual labeling between 2 annotators

3.1.2 Tag set

The tag set is defined based on dialogue acts used in [Stolcke et al., 2000] , but later some modifications were made based on our corpus. However, the modified tags can be mapped to the original categories in [Stolcke et al., 2000]. The tag set was defined in two steps, the original tag set consisting of 17 dialogue acts and the modified tag set consisting of 13 dialogue acts. Table 3.3 shows the initial tag set and Table 3.4 shows the modified tag set. The last three tags in sets are ignored, because they are not participating in the classification process, in section 3.1.3 it is explained in more details. Therefore, the number of tags in former tag set is considered 14 and the number of tags in the latter tag set is considered 10.

The reason why tag set dialogue acts were later reduced to 10 categories is due to some dialogue acts appearing very low in the corpus. The low occurring dialogue acts were merged into another class.

Table 3.5 shows the comparison between the tag set in this work and the original dialogue acts used in [Stolcke et al., 2000]. Table 3.5 also shows the percentages corresponding to each tag in the corpus of each project.

Index	Label	Abbr
1	Statement	S
2	Request (Recommendation)	R
3	Rhetorical Question	Q
4	Open Question	O
5	Y/N Question	I
6	Agreement	A
7	Reject (Disagreement)	D
8	Y answer	Y
9	N answer	N
10	Thanking	T
11	Opinion	P
12	Greet	G
13	Open answer	W
14	Acknowledgement	C
15	Miscellaneous	M
16	UNK	U
17	Bug	B

Table 3.3: 17 labeled tag set and their acronyms,

Index	Label	Abbr
1	Statement	S
2	Request (Recommendation)	R
3	Rhetorical Question and Y/N Question	Q
4	Open Question	O
5	Agreement	A
6	Reject (Disagreement)	D
7	Thanking	T
8	Opinion	P
9	Greet and Acknowledgement	G
10	Open answer, N answer and Y answer	W
11	Miscellaneous	M
12	UNK	U
13	Bug	B

Table 3.4: 13 labeled tag set (Merged tag set) and their acronyms

Some tags in the tag set defined by [Stolcke et al., 2000] are not the same as defined in our tag set. An example is N and Y^* answer tag in [Stolcke et al., 2000].

no	Tweet Tag labels	%	[Stolcke et al., 2000] Tag labels	%
1	S (Statement)	43.82	Statement	36
2	Request (Recommendation)	8.99	(used in some later models)	-
3	Q (Rhetorical and Y/N Question)	8.43	Rhetorical and Y/N Question	2+0.2= 2.2
4	O (Open Question)	8.43	Open Question	0.3
5	A (Agreement)	5.62	Agreement/ Accept	5
6	D (Disagreement)	1.12	Reject	0.2
7	T (Thanking)	3.37	Thanking	$\tilde{0}.1$
8	P (Opinion)	3.37	Opinion	13
9	G (Greet and Acknowledgement)	2.25	Acknowledgement	19
10	W (Open, N and Y answer)	5.62	N and Y^* answer	1+1= 2

Table 3.5: Comparing Tweet_tag_% to [Stolcke et al., 2000]_tag_%

3.1.3 Dialogue Act Types

1. Statement and Opinion: The most common types of dialogue acts were STATEMENTS. ‘Descriptive, narrative, or personal’ statements are classified as STATEMENTS while other-directed opinion statements are classified as OPINION.

2. Request (Recommendation): Different types of requests and recommendations are classified as REQUEST. Table 3.6 shows examples of Requests and Recommendations from our corpus.

3. Rhetorical Question and Y/N Question: Different types of questions which do NOT need answers are classified as RHETORICAL QUESTIONS and Different types of questions which need yes/ no answers are classified as YES/ NO QUESTIONS.

4. Open Question: Different types of questions which need answers, but not yes/no answers are classified as OPEN QUESTIONS.

5. Agreement: Different types of sentences showing one speaker agrees with another one in a conversations are classified as ”AGREEMENT”.

6. Reject (Disagreement): Different types of phrases showing one speaker disagrees with another one in a conversations or rejects someone or something are classified as ”REJECT (DISAGREEMENT)”.

7. Y-answer, N-answer and Open answer: All different types of answers are classified in this category.

8. Thanking: Different types of phrases showing someone is appreciating some thing is classifies as ”THANKING”.

9. Greet and Acknowledgement: Different types of phrases used for greetings are classified as ”GREETING” and Different types of phrases showing one speaker understands what the other one says are classified as ”ACHKNOWLEDGMENT”.

10. Miscellaneous: This tag refers to the phrases roughly understood, but none of the tags in the short-list refers to them.

    This group of tag can either be kept as it is, because it’s a limit of the annotation or it can be removed from the corpus. In this work, they were removed of the corpus.

11. UNK: This tag refer to the phrases which are (roughly) understood, but the decision needs to be made between the two tags A or B.

    The decision which is made about the this group of tag is to solve the ambiguity by concerting with other annotators.

12. Bug: This tag refer to the phrases which are not understood at all. The decision which is made about this group of tag is to filter them out from corpus, because they are a bug in the corpus.

    Table 3.6 shows some examples from our corpus for all tags in the tag set.

Dialogue Act	**Example**
Statement	It'll be a month on august 1st
Statement	She has a boyfriend
Opinion	I think I shoulder some of the 'blame' in that. :d
Opinion	We need 1 more imo, better as a 5 premade :d
Request (Recommendation)	tell that shrimp hurry up
Request **(Recommendation)**	let's get married now
Request **(Recommendation)**	Why not hanging hindu fanatics
Request (Recommendation)	don't put him on the list
Rhetorical question	omg can you believe this
Y/N Question	is this your tank top
Open Question	how did you get it started?
Agreement	it is definitely going to be an adventure
Reject (Disagreement)	no no noooo I don't know what to say to the hot dwarf.
Y , N and Open answer	yeah it was good thanks
Y, **N** and Open answer	no sir it's locked in already
Y, N and **Open** answer	because it is easy for whites to ignore this genocide.
Y, N and **Open** answer	I leave friday.
Thanking	Thanks for the shout out again.
Miscellaneous	I'm sorry.
Bug	http://t.co/cALJWgg8eY

Table 3.6: Corpus examples for all the tags

3.2 Dialogue Act Model

The goal of this project is to build a model that takes as input a sentence in a dialog and outputs a dialogue act for that sentence. The following information is relevant for the model to accomplish this task.

1. Lexical information: The words play an important role in deciding the dialogue act

2. Dialogue history information: The dialogue history plans an important role. For instance, a question is always followed by an answer.

3.2.1 Lexical model

Lexical models can be limited to a bag-of-words model, which assumes that the sequence of words is not relevant for DA recognition. But we know this is a too poor hypothesis. In this project, we use a sequence model that further considers the sequential ordering of words in the sentence. There are several types of sequence models: generative ones, like HMMs, but they require more data to be learnt efficiently; discriminative ones, like CRFs, but they usually have less performance when compared to neural networks.

3.2.2 Recurrent Neural Networks

Simple RNN

We consider simple recurrent neural networks (RNN), which are composed of a repeating neural cell, one for each word in the sentence. This cell computes the following function, given the input word x_t

$$ht = f(Wx_t + Uh_{t-1} + b) \tag{3.1}$$

where h_t is a hidden state in the cell that encodes the information from the current word and from the past hidden state. At the end of the sentence, the last cell outputs the last hidden state h_T. This last state is transformed, it summarizes all the information from the sentence about the dialogue act, into an actual decision, i.e., the choice of a DA. This is done by adding one feed-forward layer with as many output than the number of possible DAs.

$$y = g(Vh_T + c) \tag{3.2}$$

where the dimensions of y is the number of DAs: the chosen DA is the one corresponding to the dimension of y with the largest value. In order to get a differentiable function, and also to be able to interpret y as a probability distribution over DAs, the softmax function is used.

Word representation

Each word is encoded by its index in a fixed vocabulary: ‘a’ is 0, ‘at’ is 1, etc. But these indexes can not be directly given to the RNN, because the RNN would interpret them as real values, which makes no sense: the distance between a and at would be 1, while it would be 1000 between a and the. So they must be encoded into a vector (categorical) representation that makes no difference between words: a = [0100...] and at = [0010...] which is called the one-hot encoding, where the vector has the size of the vocabulary and a single dimension is activated (with 1) for each word. But the vocabulary may be quite large, and this representation takes too much space. So we rather associate to each word a smaller embedding vector full of real numbers. These embeddings were trained by the model.

LSTM

LSTM networks have been shown to learn long-term dependencies more easily than the simple recurrent architectures. We consider LSTM networks which are composed of a repeating neural cell, one for each word in the sentence. This cell computes the following function, given the input word x_t

$$h_i^{(t)} = tanh(s_i^{(t)})q_i^{(t)} \tag{3.3}$$

$$q_i^{(t)} = \sigma(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o x_j^{(t-1)}) \tag{3.4}$$

Which has parameters b^o, U^o, W^o for its biases, input weights and recurrent weights, respectively. At the end of the sentence, the last cell outputs the last hidden state $h_i^{(t)}$. This last state is transformed, it summarizes all the information from the sentence about the dialogue act, into an actual decision, i.e., the choice of a DA. This is done by adding one feed-forward layer with as many output than the number of possible DAs.

$$y = g(Vh_i^{(t)} + c) \tag{3.5}$$

where the dimensions of y is the number of DAs: the chosen DA is the one corresponding to the dimension of y with the largest value. In order to get a differentiable function, and also to be able to interpret y as a probability distribution over DAs, the softmax function is used.

Chapter 4

Empirical evaluation

4.1 Experimental setup

4.1.1 Tweets preprocessing

The tweets were tokenized using the NLTK [Loper and Bird, 2002] tweet tokenizer. Words occuring just once in the corpus were seen as OOV(Out of vocabulary words). Stops words were not removed from the corpus.

4.1.2 Deep learning library

Keras [Chollet et al., 2015] and Tensorflow [Abadi et al., 2015] were used for implementing the deep learning models

4.1.3 Manipulation of DA categories

As shown in Table 3.3, 43% of the tweets in the corpus are mainly Statements. This means that our model might be biased towards tweets that are Statements. In order eliminate this bias, we decided to create a more balanced corpus by augmenting the occurrence of other tags.

This was done using different approaches such as merging the tags into new categories such as:

1. **Statement** and **Non Statement**: Here all dialogue acts which are not Statements are categorized as Non Statements.

2. **Statement**, **Question**, **Opinion**, **Answer**: The tags were merged to create the new categories below. The merging was done based on the syntactic relationship between the dialogue acts and the number of occurrence of the dialogue acts.

   - Statement = S
   - Open Question (O) + Question (Q) = $Q^{'}$
   - Thanking (T) + Greeting (G) + Opinion (P) + Agreement (A) = $P^{'}$
   - Disagreement (D) + Answer (W) + Request(Recommendation) (R) = $W^{'}$

4.2 Experimental result

In this section, we describe the results of our implementations. Each section is divided based on the number of tweets, dialogues and tags used.

For calculating the average of Precision, Recall and f-measure, the results of tag ”S”, was ignored.

4.2.1 162 tweets, 33 dialogues, 10 tags

- Description of model: LSTM, history of the tweets were not taken into consideration.
- Dialogue acts: The dialogue acts considered here are the ones defined in Table 3.3.
- Parameters of model:

Parameters	Values
Vocabulary size	789
Max length of tweet	20
No of epochs	50
Embedding dimension	32
No of neurons	16
Activation	Softmax
Optimizer	Adam
Loss	Categorical cross entropy
LSTM dropOut	0.5
Dense layer dropout	0.2
Development set	0.2%

Table 4.1: Parameters of the model for 33 dialogues and 10 tags

Figure 4.1 shows the development and training loss of model at 50 epochs.

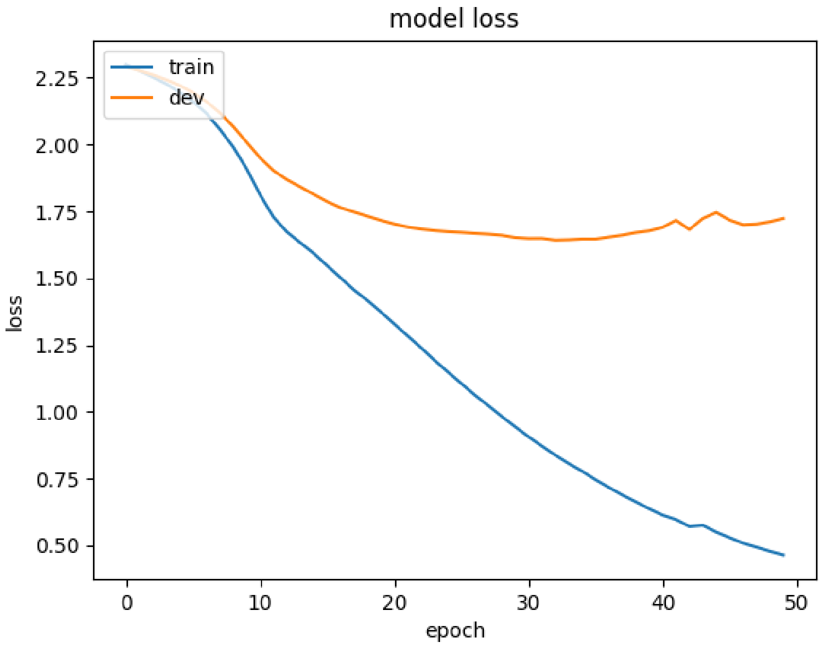

Figure 4.1: Model loss for LSTM with 33 dialogues and 10 tags

- Precision, Recall and F-measure

Tag	Precision	Recall	F-measure
Statement	0.51	0.96	0.67
Request (Recommendation)	0	0	0
Rhetorical Question and Y/N Question	0	0	0
Open Question	0	0	0
Agreement	0	0	0
Reject (Disagreement)	0	0	0
Thanking	0	0	0
Opinion	0	0	0
Greet and Acknowledgement	0	0	0
Open answer, N answer and Y answer	0	0	0
Average	0	0	0

Table 4.2: Precision, Recall and F-measure for 33 dialogues and 10 tags

4.2.2 650 tweets, 84 dialogues, 10 tags

- Description of model: LSTM, history of the tweets were not taken into consideration.

- Parameters of model: Used the parameters in Table 4.1

- Precision, Recall and F-measure

Tag	Precision	Recall	F-measure
Statement	0.5	0.64	0.56
Request (Recommendation)	0.16	0.15	0.15
Rhetorical Question and Y/N Question	0.27	0.18	0.22
Open Question	0.53	0.52	0.52
Agreement	0.18	0.06	0.09
Reject (Disagreement)	0	0	0
Thanking	0	0	0
Opinion	0.53	0.18	0.27
Greet and Acknowledgement	0.18	0.09	0.12
Open answer, N answer and Y answer	0.14	0.2	0.17
Average	0.22	0.15	0.17

Table 4.3: Precision, Recall and F-measure for 84 dialogues and 10 tags

4.2.3 1039 tweets, 148 dialogues, 10 tags

- Description of model: LSTM, history of the tweets were not taken into consideration.

- Dialogue acts: The dialogue acts considered here are the ones defined in Table 3.3.

- Parameters of the model

Parameters	**Values**
Vocabulary size	2884
Max length of tweet	20
No of epochs	50
Embedding dimension	100
No of neurons	16
Activation	Softmax
Optimizer	Adam
Loss	Categorical cross entropy
LSTM dropOut	-
Dense layer dropout	-
Development set	0.2%

Table 4.4: Parameters of the model for 148 dialogues and 10 tags

- Precision, Recall and F-measure

Tag	Precision	Recall	F-measure
Statement	0.54	0.61	0.57
Request (Recommendation)	0.22	0.2	0.2
Rhetorical Question and Y/N Question	0.43	0.29	0.35
Open Question	0.55	0.55	0.55
Agreement	0.18	0.19	0.18
Reject (Disagreement)	0	0	0
Thanking	0.7	0.37	0.18
Opinion	0.24	0.21	0.22
Greet and Acknowledgement	0.26	0.26	0.26
Open answer, N answer and Y answer	0.13	0.13	0.13
Average	0.3	0.24	0.23

Table 4.5: Precision, Recall and F-measure for 148 dialogues and 10 tags

4.2.4 891 tweets, 148 dialogues, 4 tags

- Description of model:

  1. LSTM without dialogue history
  2. Bidirectional LSTM with dialogue history: Here, 2 Bidirectional LSTMs containing the previous and next tweets were merged in order to model the dialogue. Figure 4.2 shows the architecture of the bi-directional LSTM.

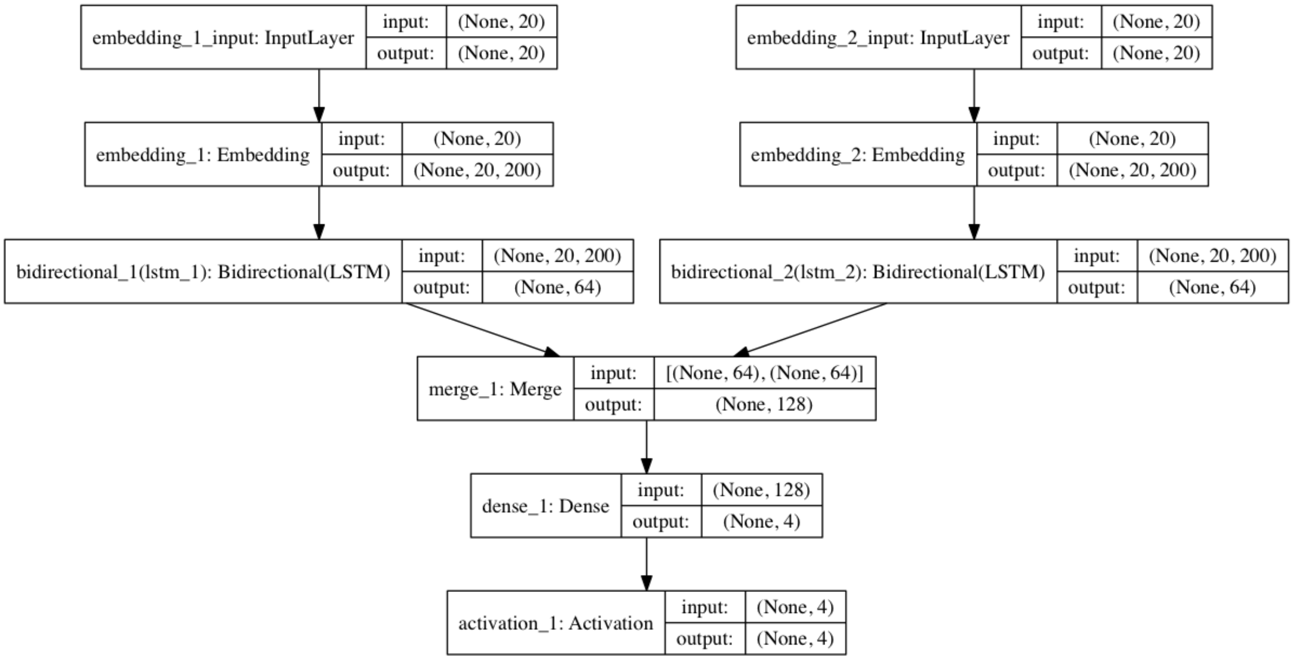

Figure 4.2: Model for bi-directional LSTM

- Dialogue acts: The dialogue acts considered here are Statements, Opinions, Answers and Questions.

- Parameters of model:

Parameters	**LSTM**	**Bi-LSTM**
Vocabulary size	2884	2884
Max length of tweet	20	20
No of epochs	50	5
Embedding dimension	32	200
No of neurons	16	64*2
Activation	Softmax	Softmax
Optimizer	Adam	Adamax
Loss	Categorical cross entropy	Categorical cross entropy
LSTM dropuOut	0.5	-
Dense layer dropout	0.2	-
Development set	0.2%	0.2%

Table 4.6: Parameters of the model for 148 dialogues and 4 tags

Fig 4.3 shows the loss of the training and development set of the bi-directional LSTM with dialogue history taken into consideration. The best development loss is at 5 epochs. After 5 epochs, the development loss increases while the training loss decreases. This clearly shows that at epochs greater than 5, the model over-fits.

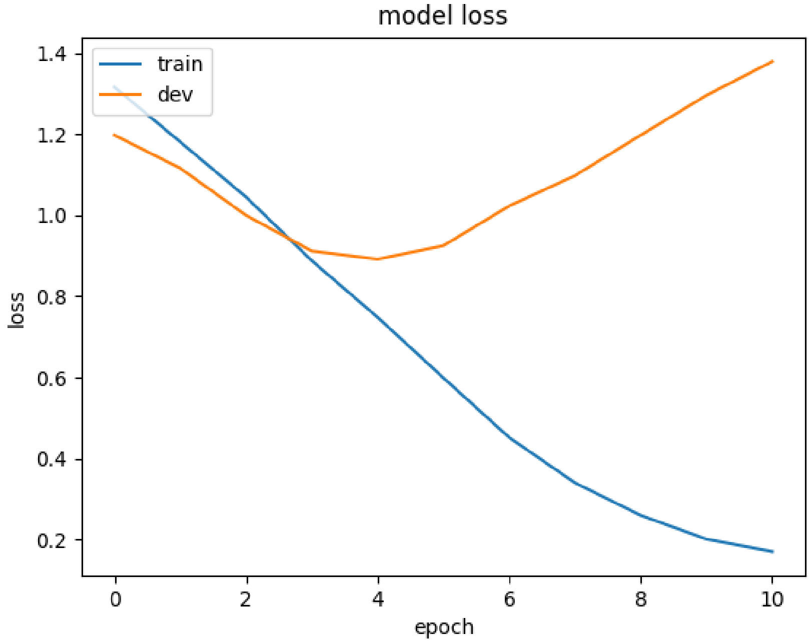

Figure 4.3: Model loss for bi-directional LSTM

- Precision, Recall and F-measure

Tag	Precision	Recall	F-measure
S	0.59	0.88	0.71
$P^{'}$	0.53	0.08	0.13
$Q^{'}$	0.81	0.75	0.78
$W^{'}$	0.53	0.31	0.39
Average	0.62	0.38	0.43

Table 4.7: Precision, Recall and F-measure for Bi-directional LSTM for 148 dialogues and 4 tags

Tag	Precision	Recall	F-measure
S	0.53	0.76	0.62
$P^{'}$	0.36	0.19	0.25
$Q^{'}$	0.79	0.73	0.76
$W^{'}$	0.23	0.13	0.17
Average	0.46	0.35	0.39

Table 4.8: Precision, Recall and F-measure for LSTM for 148 dialogues and 4 tags

4.2.5 1039 tweets, 148 dialogues, 2 tags

- Description of model: LSTM, history of the tweets were not taken into consideration.

- Dialogue acts: The two dialogue acts considered here are the Statements and Non Statements.

- Parameters of model:

Parameters	**Values**
Vocabulary size	2884
Max length of tweet	20
No of epochs	50
Embedding dimension	100
No of neurons	32
Activation	Softmax
Optimizer	Rmsprop
Loss	Categorical cross entropy
Development set	0.2%

Table 4.9: Parameters of the model for LSTM with 148 dialogues and 2 tags

- Precision, Recall and F-measure

Tag	Precision	Recall	F-measure
S	0.55	0.59	0.5
N	0.66	0.63	0.64
Average	0.66	0.63	0.64

Table 4.10: Precision, Recall and F-measure for LSTM with 148 dialogues and 2 tags

4.3 Discussion

4.3.1 Comparison of the models

Model	Tags	No of tweets	Accuracy	Chance Accuracy
LSTM	2	1039	52.0	53.0
LSTM	4	1039	**53.4**	40.0
LSTM	10	1039	46.0	47.5
LSTM	10	162	46.0	47.5
Bi-LSTM (dialogue history)	4	891	**60.0**	40.0

Table 4.11: Comparison of the results of different models

Table 4.11 shows the accuracy of the different models experimented with.

The Accuracy was calculated by dividing the total number of tags which were predicted correctly by the model by the total number of tags.

An accuracy of 52.0% was gotten on the LSTM with 2 tags(Statement and Non-statement). This is poor compared to the chance accuracy which is 53.0%.

Here we can see that it is difficult for the model to differentiate between both tags, this can be due to reasons such as little data or no explicit distinction between the two tags in reality.

The model with 10 tags did not show significant improvement against the chance accuracy when the data increased. This can be due to reasons such as large number of classes which in turn makes it difficult for the model to distinguish between them.

When the tags were merged into 4, the accuracy of the model increased significantly. The accuracy beat the chance accuracy by 13%. This shows that the model learns better with the merged tags.

The best accuracy was gotten when the dialogue history was taken into consideration and the tags were merged into 4. The accuracy beat the chance accuracy by 20%. This shows that dialogue history plays a major role in knowing which dialogue act a tweet represents.
For instance, we know that a question has a high probability of being followed by an answer.

Analysis of the results of model considering dialogue history vs model with no history

	P	Q	S	W
P	**34**	9	35	17
Q	12	**152**	17	12
S	120	40	**344**	139
W	16	8	59	**25**

Table 4.12: Confusion matrix of model with 4 tags and no dialogue history

The confusion matrix of the results of the model with 4 tags with no dialogue history is shown in Table 4.12 As shown in Table 4.12, the model confuses the two tags S' and W' the most. After analyzing the structure of the tweets having these two tags, it was discovered that since "O" is a sub-tag of W' and the structure of statements are in most cases very similar to open answers, it is difficult for the system to distinguish them.

	P	Q	S	W
P	**14**	3	6	3
Q	9	**133**	15	7
S	126	35	**330**	108
W	16	7	25	**54**

Table 4.13: Confusion matrix of model with 4 tags with dialogue history

In order to improve the learning, the model with 4 tags was trained taking the history of the dialogue into account. The confusion matrix of the results is shown Table 4.13. As shown in Table 4.13 the accuracy of detecting W' increased but not as expected. The reason can be due to sub-tags consisting W'.

As illustrated in the confusion matrix of the model with 4 tags and history consideration, P' is also confused with S' in many places. This can be due to the fact that sub-tags of P' are not consistent with each other neither structurally, nor semantically. Therefore, it may be difficult for the system to learn.

Thanking + Greeting + Opinion + Agreement = P'

Disagreement(D) + Answer(W) + Request and Recommendation(R) = W' Apart from ‘O’, W' consists of tags, ‘R’ and ‘D’. Structurally speaking, ‘R’ is not similar to ‘D’ and ‘W’ . This can be one of the reasons why the system can not learn as expected considering dialogue history.

4.3.2 Modeling with syntactic features

In annotating the tweets in the corpus, the syntactic structure of the DA labels were taken into consideration to the highest extent possible and it was agreed between annotators to follow the rules extracted for each DA label, for instance for the label ”Recommendation”, some syntactic structures symbolizing this label were: *”How about verb + ing ...”*, *”Let’s verb +...”*, *”you’d better verb + ...”*.

While applying syntactic structure for modeling is precise, it is not comprehensive enough, and it does not capture the semantic and pragmatic aspects of dialogues extensively. It can be one one of the reasons why the model does not learn very well.

4.3.3 Volume of data

Models in this work are trained and evaluated using a hand-labeled database of 148 conversations, 1039 tweets from spontaneous human-to-human English tweets of the Tweeter. However this database is not considered as a large one, and the more data is available for the training the better results is gained at the end.

The precision of the model using different volumes of data is shown in Figure 4.1.

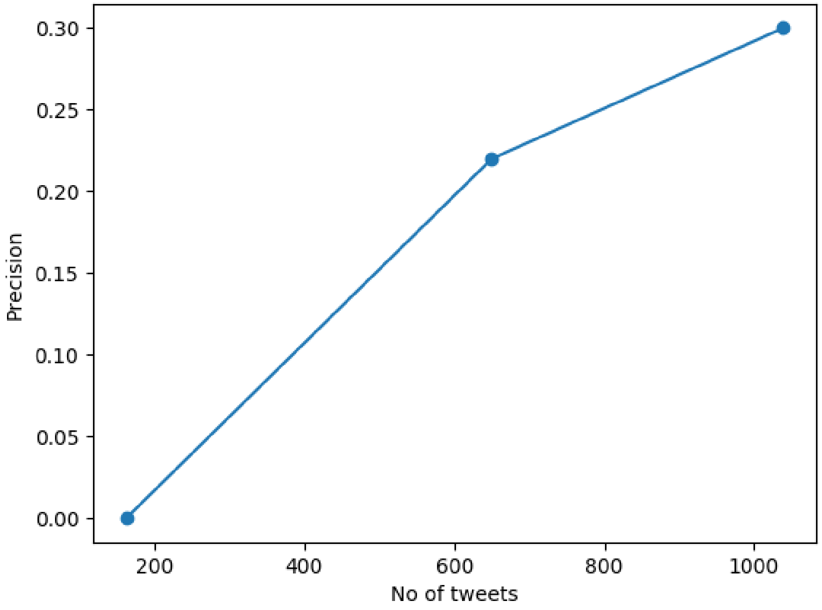

Figure 4.4: Volume of data vs precision

4.3.4 Structure of the language of Tweets

It is known that the language of tweets has specific characteristics, such as compactness , informality, users' unique linguistically styles of developing tweets, users' different usage of temporal references.

Taking all differences into consideration, twitter language is less structured compared to standard form of language. This might make it hard to develop a model that truly captures the uniqueness in this language as shown in the accuracy of the different models described in Table 4.10.

4.3.5 Comparison with [Stolcke et al., 2000]

Dialogue act modeling in [Stolcke et al., 2000] reached the accuracy of 71% based on word transcripts, compared to a chance baseline accuracy of 35% and human accuracy of 84%.
Compared to our model, [Stolcke et al., 2000] uses a structured dialogue of human-

human conversation. Our model uses twitter dialogues which is unstructured and difficult to model. Also, our model is trained using 148 dialogues which is low when compared to 1,155 dialogues used by [Stolcke et al., 2000]

Chapter 5

Conclusion and Future work

We have developed a neural network based model for classifying dialogue acts for human-human conversation on twitter. The approach combines lexical and dialogue history information. Classification accuracies achieved so far are highly encouraging, relative to the inherent difficulty of the task as measured by human labeler performance.

We implemented different models for the dialogue act recognition and found that performance depends on the number of dialogue acts and dialogue history. The best performance was achieved by implementing a bidirectional LSTM with dialogue history.

In future, in order to achieve better accuracy, the following items are considered to be done.

- Incorporating semantic and pragmatic information into the modeling the dialogues
- Increasing the amount of training data (adding more hand-labeled data to the corpus).
- Balancing the weight of data corresponding to different by statistical methods of reducing the weight of frequent tags like ”S”.
- Re-merging the labels considering semantic and syntactic structure of the labels, in order to improve the model and resolve the challenge mentioned in section 4.3.1.

Bibliography

[Abadi et al., 2015] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[Ang et al., 2005] Ang, J., Liu, Y., and Shriberg, E. (2005). Automatic dialog act segmentation and classification in multiparty meetings. In *ICASSP*, pages 1061–1064.

[Austin, 1962] Austin, J. L. (1962). *How to do Things with Words*. Clarendon Press, Oxford.

[Chollet et al., 2015] Chollet, F. et al. (2015). Keras. `https://github.com/fchollet/keras`.

[Core and Allen, 1997] Core, M. and Allen, J. (1997). Coding dialogs with the damsl annotation scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35.

[Forsyth and Martell, 2007] Forsyth, E. N. and Martell, C. H. (2007). Lexical and discourse analysis of online chat dialog. pages 19–26.

[Godfrey et al., 1992] Godfrey, J. J., Holliman, E. C., and McDaniel., J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520.

[Ide and Bunt, 2010] Ide, N. and Bunt, H. (2010). Anatomy of annotation schemes: mapping to graf. pages 247–255.

[Loper and Bird, 2002] Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

[Lui and Baldwin, 2012] Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. pages 25–30.

[Ritter et al., 2010] Ritter, A., Cherry, C., and Dolan, B. (2010). Unsupervised modeling of twitter conversations. In *Proceedings of NAACL.*

[Scheffler, 2014] Scheffler, T. (2014). A german twitter snapshot. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).*

[Scheffler and Zarisheva, 2016] Scheffler, T. and Zarisheva, E. (2016). Dialog act recognition for twitter conversations. In *Proceedings of the Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*, pages 31–38.

[Searle, 1969] Searle, J. R. (1969). *Speech Acts*. Cambridge University Press, London-New York.

[Stolcke et al., 2000] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

[Vosoughi and Roy, 2016] Vosoughi, S. and Roy, D. (2016). Tweet acts: A speech act classifier for twitter. In *10th International AAAI Conference on Weblogs and Social Media (ICWSM 2016).*

[Zarisheva and Scheffler, 2015] Zarisheva, E. and Scheffler, T. (2015). Dialog act annotation for twitter conversations. In *Proceedings of the SIGDIAL 2015 Conference*, pages 114–123.

[Zhao and Jiang, 2011] Zhao, X. and Jiang, J. (2011). An empirical comparison of topics in twitter and traditional media. In *Singapore Management University School of Information Systems Technical paper series. Retrieved November 10:2011.*