Dialogue Summarization

Prakhar Ganesh and Saket Dingliwal

The Problem and Challenges

The problem

From Here

Prakhar : I think it is time that we decide on topic for the NLP project . Mausam sir expects a paper from the batch
Saket :You are right .Ummm.. .I am too confused . How can we read a paper , code it and improve the baseline within a month ?This seems too much to ask for .
Prakhar : Haha. .do not worry. .We can look for code of a paper online and directly jump on improving baseline .I think finding a paper with code as well as dataset available will be a challenge . Why do not I try something on sentiment analysis ?
Saket : Yeah ! that seems cool . Are you familiar with generative language models ?

To Here

Prakhar and Saket discuss about the NLP Project. Saket was confused. Reading, Coding and Improving baseline seemed to much to ask. Prakhar suggests sentiment analysis while Saket jumps to Generative Models.

The problem

Formal Definition

We aim to develop a novel pipeline consisting of Discourse Relations extraction followed by an Attention based LSTM-RNN model which creates Abstractive Summaries of Discussions.

P.S. : The task is as complex as this definition :-)

The problem

Why This Problem ?

- **Well… In this busy world, Concise and Compact sell.**
- **There is overwhelming amount of text information available on web. Also, there is growing amount of it in form of dialogues. Online blogs and discussion forums are full of debates that can be cumbersome to read in full .**
- **Email threads and chat logs are another form of dialogues that can be summarized saving critical time.**

Challenges deep-dive

Challenge 1

Different from Paragraph based text

Dialogues do not have a simple line by line conversion to a paragraph. Various Question-Answers and Exclamations form a part of dialogue.

Ex - “Umm..”, “well okay”, “yes” , “Yeah ” etc are valid dialogues that can’t be converted into paragraph.

Challenge 2

Multiple Participants

There may be multiple speakers. Question and Answers between them can be separated by others in between. Ex-

A -> What are you doing C?
…….(multiple dialogues)
C -> I am doing fine.

Challenges deep-dive

Challenge 3

Lack of large amount of dataset for learning any Neural Model

There is no dataset with abstractive reference summary large enough to train any Neural Model for summarization. Only Annotated datasets are BC3 (40 email threads), IAC (45 arguments) .

Challenge 4

Abstractive Summarization is itself difficult

Attention based models for abstractive summarization have achieved limited success. Rouge-1 f-scores of state-of-the-art paper are close to 0.38 . Lots of data and training is required.

Challenges deep-dive

Challenge 5

Incorrect Factual Description

The summaries sometimes reproduce factual details inaccurately. This is common for rare or out-of-vocabulary words such as “3-2”. Ex->

Original: Germany beat Argentina 3-2

Generated: Germany beat Argentina 2-0

Challenge 6

Words Repeating Themselves

Encoder- decoder summaries tend to repeat themselves because of over reliance on previous summary word (Decoder input) causing endless repeating cycle. Ex-

“Germany beat Germany beat Germany…..”

Proposed Solutions

Pipeline

- Parsing, cleaning and Anaphora Resolution of the dialogue.
- CRF based Sequence Labelling of dialogue act tags like “qw” (wh question) , “nn” (negative answer) etc.
- Formation of paragraph like structure based on discourse relation derived from the act tags.
- Passing the paragraph as input to pointer-generator based abstractive summarization.

Anaphora Resolution

Replacing pronouns with the respective speaker names

S1 : Why do you think gay marriages are banned? I cannot accept your argument of ...

S2 : I think the reason for gay marriages being banned is not entirely political. But you cannot deny ...

S1 : Why do S2 think gay marriages are banned? S1 cannot accept S2's argument of ...

S2 : S2 think the reason for gay marriages being banned is not entirely political. But S1 cannot deny ...

What is Dialogue Act Modelling and Discourse ?

	name	act_tag	example	train_count	full_count
1	Statement-non-opinion	sd	Me, I'm in the legal department.	72824	75145
2	Acknowledge (Backchannel)	b	Uh-huh.	37096	38298
3	Statement-opinion	sv	I think it's great	25197	26428
4	Agree/Accept	aa	That's exactly it.	10820	11133
5	Abandoned or Turn-Exit	%	So, -	10569	15550
6	Appreciation	ba	I can imagine.	4633	4765
7	Yes-No-Question	qy	Do you have to have any special training?	4624	4727
8	Non-verbal	x	[Laughter], [Throat_clearing]	3548	3630
9	Yes answers	ny	Yes.	2934	3034

Using 43 such tags and learning a CRF based model to predict them

What is Dialogue Act Modelling and Discourse ?

Prakhar :

sv -Statement Opinion- I think it is time that we decide on topic for the NLP project .

sd -Statement non Opinion- Mausam sir expects a paper from the batch .

Saket :

b -Acknowledge- You are right .

x -Non Verbal- Ummm.. .

Prakhar :

h -Hedge- do not worry. .

sd -Statement non Opinion- We can look for code of a paper online and directly jump on improving baseline .

qh -Rhetorical Question- Why do not I try something on sentiment analysis ?

Discourse Relations

Removing unwanted sentences by removing certain discourse markers

Sentence	Discourse Tag
S1 : Well, it's been nice talking to you	fc
S1 : But, uh, yeah	%
S1 : I'm sorry	fa

All the above sentences will be removed from the conversation before moving forward

Discourse Relations

Joining one-word answers with the corresponding questions

S1 : Do you think he had any special training? Because if what I heard ...

S2 : Yes. Although what you heard is just a rumour ...

S1 asked do S2 think he had any special training and S2 agreed. Because if what S1 heard ..

Discourse Relations

Replacing sentences which indicate specific actions taken by the speaker with simpler sentences

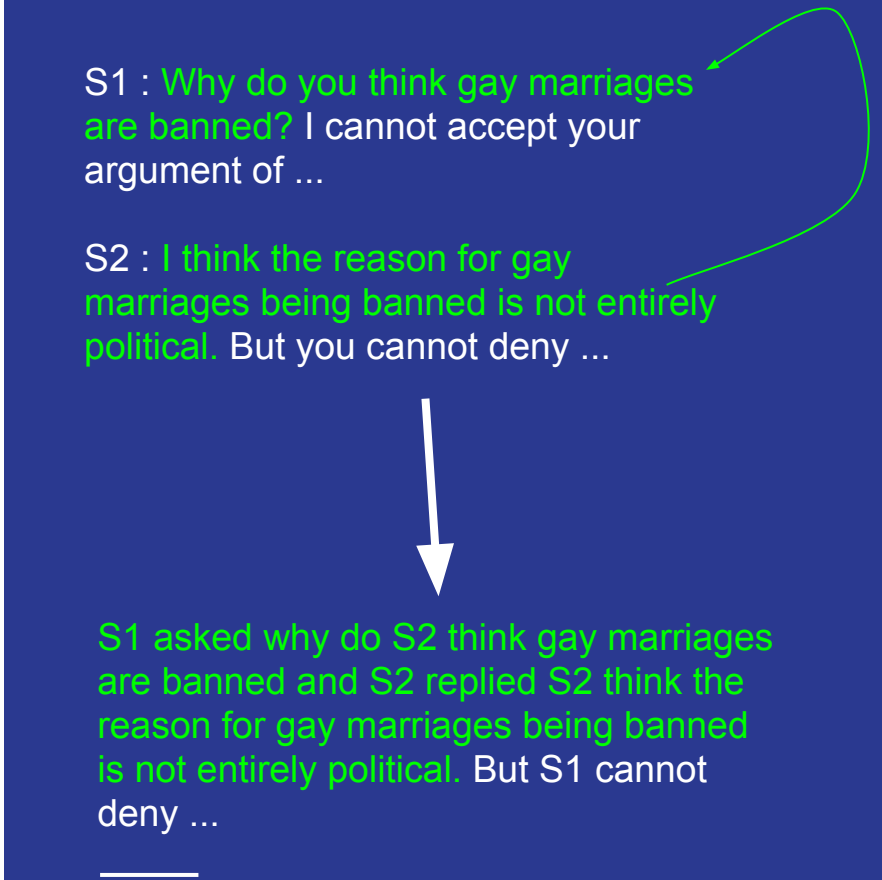
S1 : I am going for a walk. I will be back in a while.

S2 : Oh, Okay.

S1 is going for a walk. S1 will be back in a while. S2 acknowledged.

Discourse Relations

Looking for answers of wh-based questions in the dialogues of the subsequent speakers and matching them based on word count to form graph structure for the dialogue



Discourse Relations

How we did it?

Algorithm Used -> CRF - Sequence labeling task on sentence level

Features Used -> Combination of word level tokens and POS tags and other features. Intuitions taken from
Extractive Summarization and Dialogue Act Modeling on Email Threads: An Integrated Probabilistic Approach by Tatsuro Oya and Giuseppe Carenini

Dataset Used -> Switchboard Dataset - 1115 conversations, 43 different annotations

Accuracy Achieved -> 0.7435634

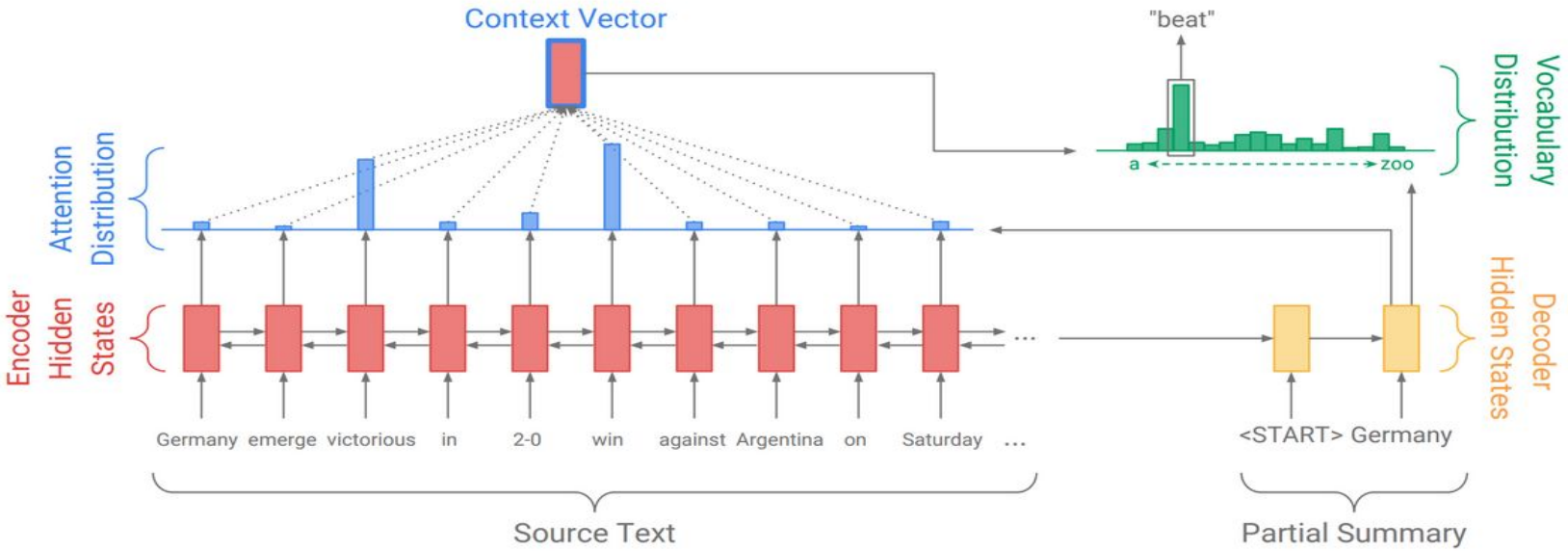
Attention based LSTM-RNN*

Seq2seq model to generate out of vocabulary words using LSTM-RNN

Germany emerged victorious in 2-0 win against Argentina.

Germany beat Argentina 2-0.

What is Attention based Summary Generator?



Source:
http://www.abigailsee.com/2017/04/16/taming-rnns-for-better-summarization.html

Pointer - Generator Networks*

Creating a balance between copying and generation of words to include only the correct factual information

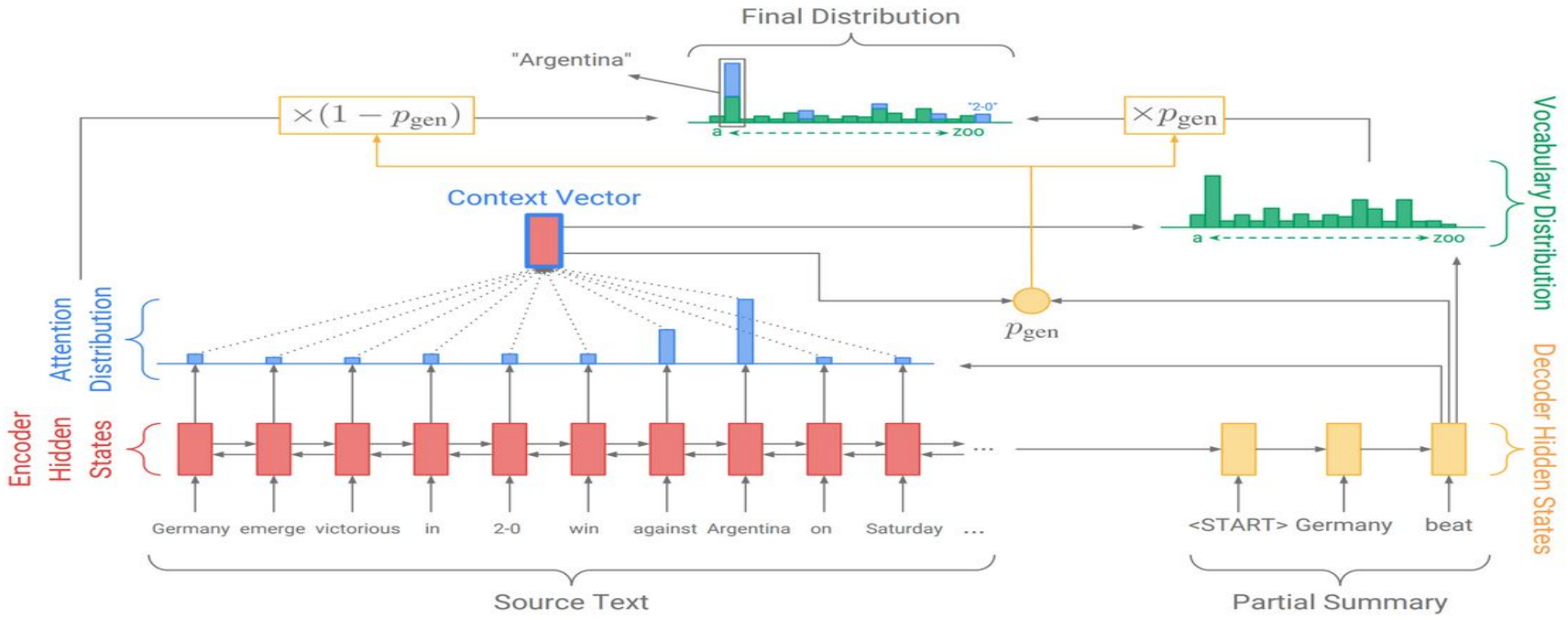
Germany emerged victorious in 2-0 win against Argentina.

Germany beat Argentina 2-0. ✓

Germany beat Argentina 3-2. ✗

2-0 is an out of vocabulary word. Trying to generate it would be really hard for the LSTM-RNN alone. It would instead generate a similar seen vocabulary word 3-2.

What is Pointer Generator?



Source:
http://www.abigailsee.com/2017/04/16/taming-rnns-for-better-summarization.html

Coverage*

Penalising the model for using generating/pointing the same words again, thus removing repetitions.

Germany emerged victorious in 2-0 win against Argentina.

Germany beat Argentina 2-0. 

Germany beat Germany beat Germany beat .., 

The model may get stuck in repetition due to high probability of a particular word in the given context. To prevent the same we can penalise repetition of words.

*Pointer-Generator Network Summarization

How we did it?

Algorithm Used -> Attention based LSTM-RNN along with pointer-generator and coverage penalty

Code Used -> Parameters were tuned for dialogue system from the following#
Get To The Point : Summarization with Pointer-Generator Networks by Abigail See , Peter J. Liu, Christopher D. Manning

Dataset Used -> DeepMind Q&A Dataset : Consisting of CNN/Daily-Mail News Articles with a total of around 3 lakh stories and corresponding abstractive summaries available.

code adapted to tensorflow 1.5 and python 3 from https://github.com/abisee/pointer-generator

Abstractive Nature of Summaries

Text -> The law that will bar S1's family from legal protections . It will not protect her marriage but will bar S1 and S1's people from being full citizens .
Summary -> the law that will not protect her marriage but will bar s1 and s1 's family from legal protections.

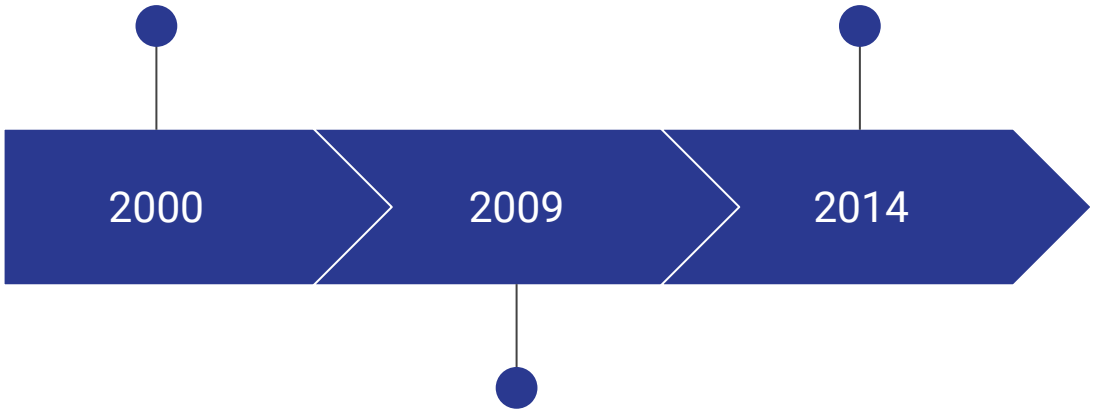
Text -> However , S2 are for gay marriage , ergo , it makes more sense S2 would be bothered by that inconsistency ..... Again , S2 do not see an inconsistency in that position .
Summary -> s2 don't see an inconsistency in that position of gay marriage , ergo .

Text -> If a law that the majority favors is unjust in an individual's life then it must be overruled for the individuals sake .... If the law effects even one person adversely then it is unjust and needs to be overturned .
Summary -> if the law effects even one person adversely then it must be overruled for the individuals sake.

Experiments and Baseline

DiaSumm: Flexible Summarization of spontaneous dialogues in Unrestricted Domains - K Zechner (turn linking, topic segmentation)

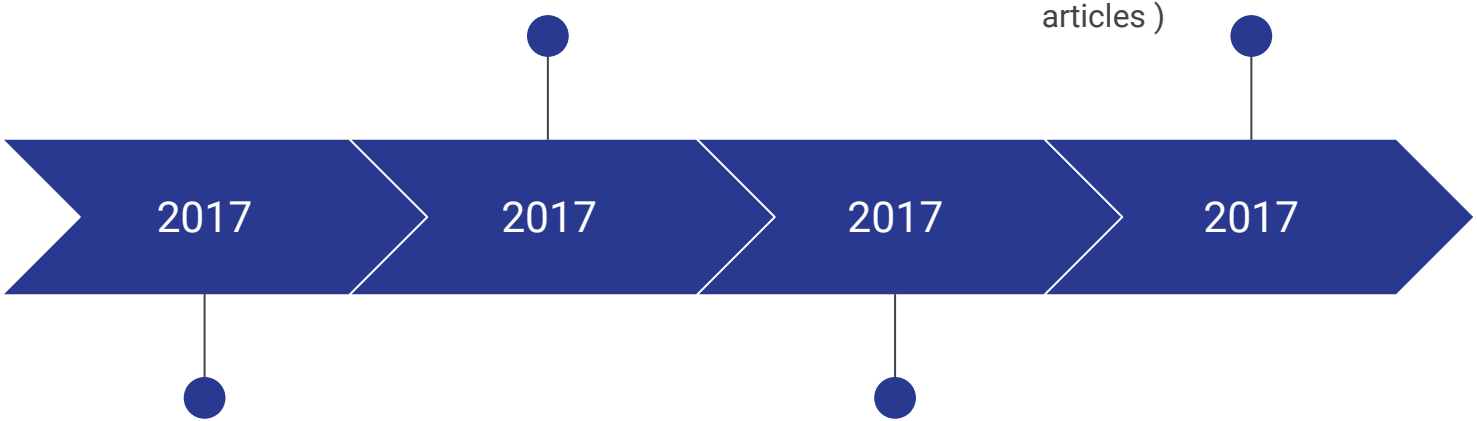
Extractive Summarization and Dialogue Act Modeling on Email Threads: An Integrated Probabilistic Approach - Tatsuro Oya (modelled dialogue acts and discourse relation in emails)



Regression-Based Summarization of Email Conversation- Jan Ulrich (use SVM based features to classify important / unimportant classifiers)

Toward Extractive Summarization of Online Forum Discussions via Hierarchical Attention Networks - Sansiri Tarnpradab

Get To The Point: Summarization with Pointer-Generator Networks - A.See (abstractive summarization of news articles)



Joint Modeling of Content and Discourse Relations in Dialogues - K Quin (used MCMC for sampling discourse distribution and phrase labels)

Summarizing Dialogic Arguments from Social Media - Amita Misra
(important sentences marked using ML Classifiers)

Evaluation Metric

ROUGE-Scores

ROUGE-Scores -> We will be using ROUGE-1 and ROUGE-2 scores to measure our models performance against widely used dialogue summarization (extractive) baselines.

Is ROUGE-Score a good enough measure? -> ROUGE-Scores are calculated based on the number of overlapping n-grams present in the reference and generated summaries. This clearly means extractive summaries are by default better performers than abstractive summaries in terms of ROUGE-Scores.

Dataset

1. **The Switchboard Dialogue Act Corpus - 1115 conversations, 205,000 utterances, 43 different annotations**

2. **BC3 Email Corpus - 40 Email threads, 3222 sentences**

Existing Baselines

Dialogue Extractive Summarization

1. MaxLength: Selects the longest sentences of every participant. This is a proven strong baseline for conversation summarization.
2. HAL: A sentence extraction based system , where the sentences are selected based on the words score in the semantic space built using a lexical co-occurrence model.
3. FirstSent: Selects the first sentence from each message in the conversation sequence (Position Hypothesis).
4. DiaSumm: This system creates a summary by extracting an interconnected structure of segments that quoted and responded to each other.

ROUGE - F-Scores

Model	ROUGE-1	ROUGE-2
MaxLength	0.27755	0.14909
HAL	0.37527	0.24820
FirstSent	0.41933	0.26926
DiaSumm	0.43815	0.30736

Attention Based LSTM-RNN

S1 : I don't think this is a ..

Is converted to

S1 said that S1 don't think this is a ..

The final article created is fed into the Attention LSTM-RNN model for summarization.

ROUGE - F-Scores

Model	**ROUGE-1**	**ROUGE-2**
MaxLength	0.27755	0.14909
HAL	0.37527	0.24820
FirstSent	0.41933	0.26926
DiaSumm	0.43815	0.30736
Attention based LSTM-RNN	**0.38934**	**0.25894**

CRF + Attention

Sentences which have a certain tag given by the CRF are removed.

The remaining conversation is converted into an article by linearly joining all utterances without adding “said that”.

The final article created is fed into the Attention LSTM-RNN model for summarization.

ROUGE - F-Scores

Model	**ROUGE-1**	**ROUGE-2**
MaxLength	0.27755	0.14909
HAL	0.37527	0.24820
FirstSent	0.41933	0.26926
DiaSumm	0.43815	0.30736
Attention based LSTM-RNN	0.38934	0.25894
CRF + Attention	**0.40354**	**0.27467**

CRF + Discourse Relation Tree + Attention

Sentences which have a certain tag given by the CRF are removed.

Sentences which have tags that suggest particular actions by the speaker are appropriately converted.

Eg -> *S1: Okay* becomes *S1 acknowledged*.

The conversation now formed is converted into an article by linearly joining all utterances.

The final article created is fed into the Attention LSTM-RNN model for summarization.

ROUGE - F-Scores

Model	**ROUGE-1**	**ROUGE-2**
MaxLength	0.27755	0.14909
HAL	0.37527	0.24820
FirstSent	0.41933	0.26926
DiaSumm	0.43815	0.30736
Attention based LSTM-RNN	0.38934	0.25894
CRF + Attention	0.40354	0.27467
CRF + Discourse Relation Tree + Attention	**0.43885**	**0.30748**

CRF + Discourse Relation Tree + Wh-questions + Attention

Sentences which have a certain tag given by the CRF are removed.

Sentences which have tags that suggest particular actions by the speaker are appropriately converted.

Questions and their corresponding answers are bought together.

The conversation now formed is converted into an article by linearly joining all utterances.

The final article created is fed into the Attention LSTM-RNN model for summarization.

ROUGE - F-Scores

Model	ROUGE-1	ROUGE-2
MaxLength	0.27755	0.14909
HAL	0.37527	0.24820
FirstSent	0.41933	0.26926
DiaSumm	0.43815	0.30736
Attention based LSTM-RNN	0.38934	0.25894
CRF + Attention	0.40354	0.27467
CRF + Discourse Relation Tree + Attention	0.43885	0.30748
CRF + Discourse Relation Tree + Wh-Questions + Attention	**0.42461**	**0.28951**

ROUGE - F-Scores

Model	ROUGE-1	ROUGE-2
MaxLength	0.27755	0.14909
HAL	0.37527	0.24820
FirstSent	0.41933	0.26926
DiaSumm	0.43815	0.30736
Attention based LSTM-RNN	0.38934	0.25894
CRF + Attention	0.40354	0.27467
CRF + Discourse Relation Tree + Attention	**0.43885**	**0.30748**
CRF + Discourse Relation Tree + Wh-Questions + Attention	0.42461	0.28951

Thank You