Abstractive Summarization of Spoken and Written Conversation

Prakhar Ganesh
Junior Undergraduate
Department of Computer Science and Engineering
IIT Delhi
cs1150245@cse.iitd.ac.in

Saket Dingliwal
Junior Undergraduate
Department of Computer Science and Engineering
IIT Delhi
cs1150254@cse.iitd.ac.in

Abstract

Nowadays, lots of information is available in form of dialogues. We propose a novel abstractive summarization system for conversations. We use sequence tagging of utterances for identifying the discourse relations of the dialogue. After aptly capturing these relations in a paragraph, we feed it into an Attention-based pointer network to produce abstractive summaries. We obtain ROUGE-1, 2 F-scores similar to those of extractive summaries of various previous works. (See et al, 2017)

1 Introduction

There has been increasing amount of diffenrent ways for people to share and exchange information. Phone calls, e-mails, blogs and social networking applications are tools which have been in great use for communication. However, these are in form of dialogues containing spontaneous utterances with speech disfluencies. This makes them cumbersome, complex and time consuming to read.

And hence, there has been increasing demand for the creation of systems to automatically summarize both text and spoken conversations. So far, most summarization systems have applied extractive approaches to this problem. However, these traditional approaches do not seem to work effectively in this domain. Also, as pointed out in **(Murray et al., 2010)** and **(Oya et al., 2014)**, abstractive summaries are often preferred to extractive ones by human judges.

Work on abstractive conversation summarization systems have previously been done but the methods are mainly based on the extraction of lexical information **(Mehdad et al., 2013)** where the authors cluster conversation sentences/utterances into communities to identify most relevant ones and aggregate them using word-graph models.

We however propose a different method where we use discourse relations and lexical information to remove pauses, abandoned sentences, nonverbal cues etc. and replace acknowledgements, appreciations, agreements etc. with sentences having better context.

Most summarization systems that have been extensively studied so far are designed in a way that they work well on organized texts such as news articles, where all documents have few grammatical errors and less redundancy **(Oya et al., 2014)**. So instead of trying to convert conversations directly into summaries we experiment by trying to convert our conversation into a structured and organized text document. We later use a pointer-generator, coverage based, Attention model **(See et al., 2017)** to create abstractive summaries of our converted text.

2 Related Work

2.1 Abstractive Summarization of Conversations

Previous work has mostly been focused on extractive approaches for meeting summarization **(Garg et al. 2009; Murray et al. 2008).** These works focusses on learning a model to classify sentences as important/unimportant.

Recently many different approaches are taken towards abstractive summarization. **Banerjee et al. (2016)** uses the dialogue structure directly to generate summary sentences . **Mehdad et al. (2013)** used clustering of sentences and bulding entailment graphs followed by aggregation using word graph model. Query based ranking approach was used in **(Mehdad et al. 2014)**.

However our approach resembles more to that of **Stone et al. (2013)**. They try to simplify conversations using discourse relations. But they have a lot of complex rules that helps them remove less important information from the conversation. We

however don't want to do that since we use discourse relations only as a means of modelling the conversion instead of ranking them. So we instead use a very simple and intuitive set of rules for utterance to paragraph conversion using these relations and follow it up with a attention-based, pointer-generator and coverage network.

2.2 Pointer-generator networks and Coverage.

The pointer network **(Vinyals et al., 2015)** is a sequence-to-sequence model that uses the soft attention distribution of **Bahdanau et al. (2015)** to produce an output sequence consisting of elements from the input sequence. The pointer network has been used to create hybrid approaches for NMT **(Gulcehre et al., 2016)**, language modeling **(Merity et al., 2016)**, and summarization **(Gulcehre et al., 2016; See et al. 2017)**.

Originating from Statistical Machine Translation **(Koehn, 2009)**, coverage was adapted for NMT by **Tu et al. (2016)**, who used a GRU to update the coverage vector each step. However **See et al. (2017)** showed that a simpler approach summing the attention distributions to obtain the coverage vector suffices. We borrow his approach for the generating the abstractive summary of the discourse document.

3 Our Models

In this section we describe the complete pipeline of our model which includes (1) Sequence labelling of utterance tags , (2) Re-ordering of conversation to model discourse relations, and (3) Pointer-generator, coverage based model for abstractive summarization.

3.1 Sequence labelling of utterance tags

The first part of our pipeline is a sequence labeling task where we need to mark discourse labels for each utterance in the conversation. Since conversations are continuing across many utterances and are related to previous and next utterances, treating it as a sequence labeling task is clearly a better choice than using SVM or logistic regression.

We used a CRF model for our sequence labeling task with the following features

```
1. First Utterance of the
Conversation
2. Speaker
3. All the words in the
Utterance
4. POS Tags of all the words
in the Utterance etc.
```

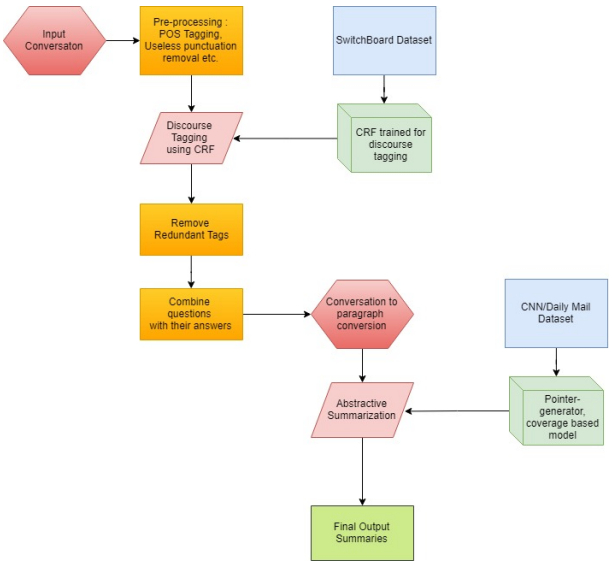

Figure 1: Proposed pipeline for Abstractive Summarization

3.2 Re-ordering of conversation to model discourse relations

We used an approach similar to **Stone et al. (2013)** but with a simpler set of rules. The basic idea behind the cleaning was to remove utterances with zero contribution to the conversation.

We removed utterances which had one of the following discourse tags

```
1. Formal Opening/closing
2. Non-Verbal
3. Abandoned Sentence
4. Self-talk etc.
```

We also identified questions and their answers by using the predicted labels and matching of similar words. The "yes/no" questions and their agreement/disagreement were also coupled together. Appreciation of some discussed topics by speakers was also identified.

3.3 Pointer-Generator, Coverage based model

Once the whole conversation is appropriately converted into a paragraph, we completed the pipeline with the pointer-generator, coverage based attention based model used by **See et al. (2017)**. A primitive code for the same was made available by

them online, however we adapted it to python 3 and tensorflow 1.2.1.

The trained model on CNN/DailyMail dataset was used because the conversation dataset did not have enough data to train a neural model by itself. Also once the conversations are converted into ordered text, we expect them to have similar behavior to that of a news article.

4 Dataset

4.1 The Switchboard Dialog Act Corpus

The Switchboard Dialog Act Corpus **(Jurafsky et al. 1997)** consist of 1115 conversations, containing 205,000 utterances and 43 different discourse tags was used to train the CRF.

4.2 Argumentative Dialogue Summary Corpus

The Argumentative Dialogue Summary Corpus **(Misra et al. 2015)** consist of 225 summaries, 5 different summaries produced by trained summarizers, of 45 dialogue excerpts on topics like gun control, gay marriage, the death penalty and abortion. This was used for experiments and evaluations.

4.3 CNN/Daily Mail Dataset

The dataset used by **See et al. (2017)** was the CNN/Daily Mail dataset **(Hermann et al., 2015; Nallapati et al., 2016)**, which contains online news articles (781 tokens on average) paired with multi-sentence summaries (3.75 sentences or 56 tokens on average).

5 Experiments

We experimented with modelling of dialogues into ordered form by sequence tagging and identification of discourse relations. We used the pointer-generator network at the end of every experimentation to obtain final summaries and compared them to the gold labeled extractive summaries using ROUGE-1 and ROUGE-2 scores. The results were than compared against the existing dialogue summarization methods for performance comparison. Some summaries were manually evaluated and compared. The features used for sequence labelling were chosen after rigrous expermentation. The best test accuracy for the task was obtained to be 0.7435. Also, We did the following variations

- Simple 'said that' baseline: To relate every utterance to it's speaker, speaker name followed by'said that' was used.
- Removing redundant utterances : We removed the sentences with particular tags that had no contribution to what the conversation was about.
- Realizing common actions : We found words like agreeing, denying, etc. and replaced utterances with sentences like 'Speaker1 agreed'.
- Joining Questions and Answers : Identification of question and their answers in dialgues by word matching.

6 Results

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	0.38934	0.25894
Attention + CRF	0.40354	0.27467
Attention + CRF + Discourse Relations Tree	**0.43885**	**0.30748**
Attention + CRF + Discourse Relations Tree + Wh-questions	0.42461	0.28951

Table 1: F1-Scores

Comapring ROUGE results of extractive and abstrative summarization tells us very little about the quality of summaries generated. ROUGE measure favours extractive summaries although they might not be any better.

Also on manual evaluation the summaries produced by our model revealed they had abstractive properties present in it and hence provided more readability.

7 Conclusion

The research in the field of summarization of organized data has grown a lot in the recent years. The paper adds to current work on abstractive dialogue summarization by bringing up a new pipeline based on discourse relations and modelling. This paper achieves results comparable to those of extractive summarizers in terms of ROUGE scores.

References

[Jurafsky et al.1997] Dan Jurafsky, Liz Shriberg, and Debra Biasca 1997. *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation.*

[Murray et al.2008] Gabriel Murray and Giuseppe Carenini 2008. *Summarizing Spoken and Written Conversations.* In Proceeding of EMNLP, Waikiki, Hawaii.

[Koehn2009] Philipp Koehn 2009. *Statistical machine translation.* Cambridge University Press.

[Garg et al.2009] Nikhil Garg, Benoit Favre, Korbinian Reidhammer, and Dilek Hakkani-Tur 2009. *ClusterRank: a graph based method for meeting summarization.* In proceeding of Interspeech, 2009.

[Murray et al2010] Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010. *Generating and validating abstracts of meeting conversations: a user study.* In Proceedings of the 6th International Natural Language Generation Conference. Association for Computational Linguistics, pages 105-113.

[Mehdad et al2013] Yashar Mehdad, Giuseppe Carenini, Frank W. Tompa and Raymond T. Ng. 2013. *Abstractive meeting summarization with entailment and fusion.* In Proc. of European Natural Language Generation Workshop (ENLG). pages 136–146.

[Stone et al.2013] Matthew B Stone, Una Stojnic, and Ernest Lepore 2013. *Situated Utterances and Discourse Relations.* In Proceeding of IWCS.

[Oya et al2014] Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. 2014. *A template-based abstractive meeting summarization: Leveraging summary and source text relationships.* In Proc. of the 8th International Natural Language Generation Conference (INLG 2014). pages 45-53.

[Mehdad et al.2014] Yashar Mehdad, Giuseppe Carenini and Raymond T. Ng 2014. *Abstractive Summarization of Spoken and Written Conversations Based on Phrasal Queries.* Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

[Hermann et al.2015] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom 2015. *Teaching machines to read and comprehend.* In Neural Information Processing Systems.

[Misra et al.2015] Amita Misra, Pranav Anand, Jean E. Fox Tree, Marilyn Walker. 2015. *Using Summarization to Discover Argument Facets in Online Idealogical Dialog.* In The North American Chapter of the Association for Computational Linguistics (NAACL), Denver, Colorado.

[Vinyals et al2015] Oriol Vinyals, Meire Fortunato and Navdeep Jaitly 2015. *Pointer networks.* In Neural Information Processing Systems.

[Bahdanau et al2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio 2015. *Neural machine translation by jointly learning to align and translate.* In International Conference on Learning Representations.

[Gulcehre et al2016] Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio 2016. *Pointing the unknown words.* In Association for Computational Linguistics.

[Merity et al2016] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher 2016. *Pointer sentinel mixture models.* In NIPS 2016 Workshop on Multi-class and Multi-label Learning in Extremely Large Label Spaces.

[Tu et al.2016] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li 2016. *Modeling coverage for neural machine translation.* In Association for Computational Linguistics.

[Banerjee et al.2016] Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama 2016. *Abstractive Meeting Summarization Using Dependency Graph Fusion.* arXiv preprint arXiv:1609.07035

[Nallapati et al.2016] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, C aglar Gulcehre, and Bing Xiang 2016. *Abstractive text summarization using sequence-to-sequence RNNs and beyond.* In Computational Natural Language Learning.

[See et al2017] Abigail See, Peter J. Liu and Christopher D. Manning 2017. *Get To The Point: Summarization with Pointer-Generator Networks.* arXiv preprint arXiv:1704.04368