Dialogue Summarization

Prakhar Ganesh and Saket Dingliwal

Problem Statement

More and more of the information available on the web is dialogic. We aim to develop a novel pipeline consisting of Content and Discourse Relations extraction followed by an Attention based LSTM-RNN model which creates Abstractive Summaries of Discussions.

Dataset

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

2. Argumentative Dialogue Summary Corpus : Consisting 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. The dialogues are from an other corpora, the Internet Argument Corpus(IAC).

Baseline

Get To The Point: Summarization with Pointer-Generator Networks.

The code of the following paper was adapted to newer version of tensorflow and python. Then using trained model, evaluation was done on the test set of the CNN database. The ROUGE scores as mentioned by paper are achieved.

Pipeline	ROUGE-1 (AVG F-score)	ROUGE (AVG RECALL)
Pointer-generator + Coverage	0.3882	0.4146

Experimentation

Numerous experiments were conducted with the Dialogue datasets. Particularly, Internet Argument Corpus and BC3 were cleaned,parsed and chunked so that they are converted into the input format required by the pointer generator summarization code.

Simple intuition regarding converting a dialogue into paragraph was used. { Speaker1 -> “ …… “} was converted to {Speaker1 said that …….. }

The ROUGE scores were calculated and are as mentioned.

Results

Dataset	ROUGE - l Score		
	Average Recall	Average Precision	Average F-score
CNN/Daily-mail News Articles (155 Articles)	0.4139462581	0.3446227097	0.3635005806
British Columbia Conversation Corpora (40 Conversations)	0.1945895	0.339817	0.233186
Argumentative Dialogue Summary Corpus (45 Conversations)	0.1679982222	0.4017931111	0.231714

Further Innovations

Our aim is to do domain adaptation of the Attention based LSTM-RNN summarization model into the domain of conversations.

To do so, we will find content and discourse relations between different sentences in the conversation and try and reduce it into the form of an article. Then we will input it into the pre-trained Attention Model to get the summary.

Timeline

Adaptation

Adapting the available code into tensorflow 1.5 and python 3

Chunking

Chunking our input files and creating .stories files as required.

Re-Training

Training again on the pre-trained Attention Model to adapt to the dialogue domain

Done

To Do

Article Format

Converting our input from conversation format into article format.

Content and Discourse Relations

Extracting content and discourse relations from the conversations to make a better corresponding article.