Yale

Memory Networks for Multi-Document Summarization

Alexander R. Fabbri, Irene Li, Tianwei She, Dragomir R. Radev PhD

Department of Computer Science, Yale University

LILY Lab

Introduction

Neural network-based methods for abstractive summarization have largely focused on single-document summarization (SDS). Multi-document summarization (MDS), on the other hand, does not benefit from large datasets as in SDS. Neural abstractive summarization for single document summarization uses datasets such as the CNN/Daily Mail dataset (Hermann:15), Gigaword (Graft:03), the NYT dataset (NYT) and the Newsroom corpus (Grusky:18).

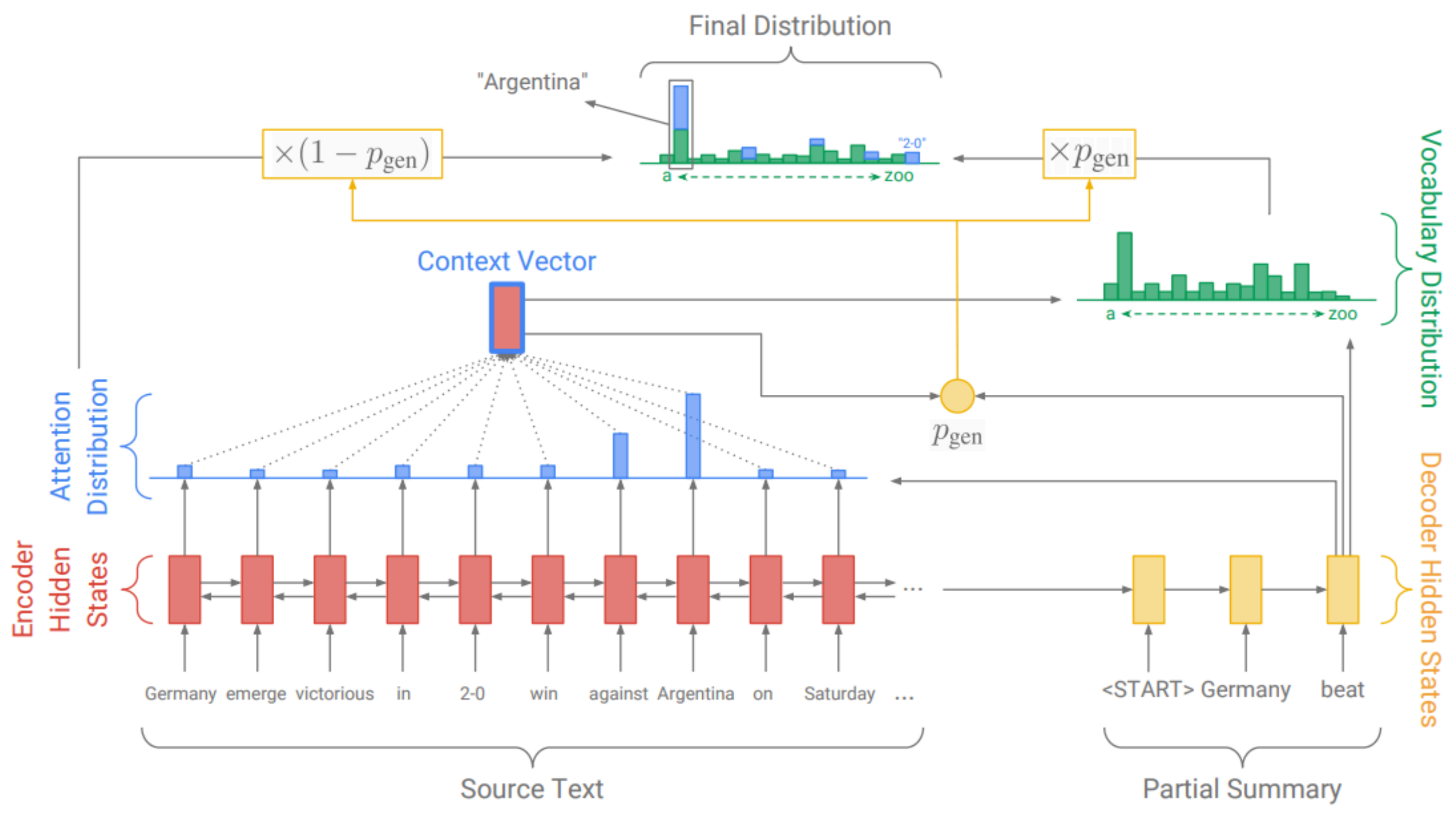
However, multi-document summarization (MDS), which aims to output summaries from document clusters on the same topic, has largely been performed on datasets with less than 100 clusters such has the DUC 2004 (Paul:04) and TAC 2011 (Dang:08) As a results, neural encoder decoder models for multi-document summarization have not received as much attention as their single document counterparts. In this paper, we propose to adapt neural methods trained on SDS datasets to MDS through fine-tuning. This is related to some recent attempts to adapt neural encoder decoder models trained on single document summarization datasets to MDS. To do so, we use memory networks, which aim to encode context via multi-hop attention over memories, or previous input/output and have not been previously used in summarization. We use memory networks combined with pointer networks common in summarization models, inspired by recent work in dialogue systems.

Datasets

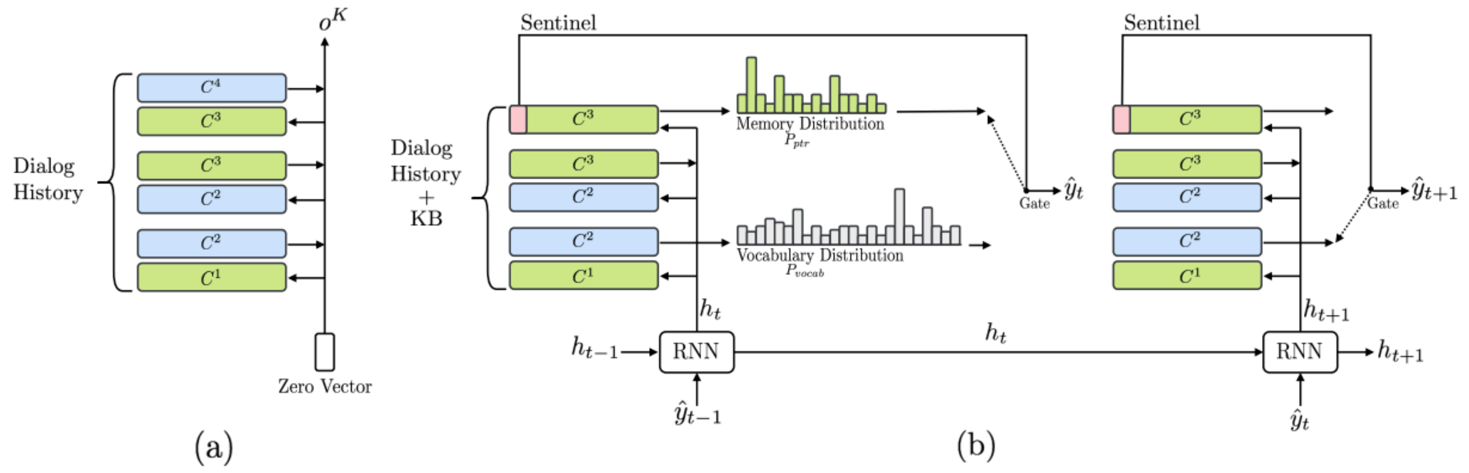
We use the CNN/Daily Mail dataset for pre-training single-document summarization systems. This corpus contains online news articles paired with multi-sentence summaries (781 tokens/article and 56 tokens/summary on average). The dataset consists of 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs. We use DUC data from 2001-2003 as training and test on DUC 2004 data.

	DUC04	TAC11	CNN	Dailymail
DUC03	1.2784	1.5584	1.5209	1.6385
DUC04		1.6705	1.3978	1.4995
TAC11			1.4564	1.6235
CNN				1.2480

Table 1: Proxy-A Distance scores between datasets



Pointer network from See et al. 2017, where words are generated from a fixed vocabulary as well as from the source sentence



Memory Networks with external knowledge bases for sequence to sequence tasks, specifically dialogue, from Madotto:18. The model forms a distribution over output vocabulary and memories.

DUC2001-2003 each contain 60 document clusters of 10 documents each. DUC 2004 contains 100 clusters of 10 documents each.

We considered using Wikisum dataset (Liu:18), although were unable initially to reproduce the dataset with sufficient coverage for our tests, although we plan to pursue this in future experiments. This dataset consists of 1865750, 233252, and 232998 training, validation and test examples, respectively.

Proposed Models

To alleviate the OOV problem, many models enable the ability to "copy" tokens from the source input while decoding. This allows for more abstractive summarization. Typically, the model upon which this copy mechanism is built is a seq2seq model of LSTM's with attention. However, for this project we are the first to propose using memory networks for summarization.

This model uses MemNN as an encoder (Sukhbaatar:15). The memories are represented with trainable embedding matrices, which are looped over for a specified number of "hops." At each hop, the model computes the attention weights for each memory through a softmax function. Then, the model reads out the memory through a weighted sum. The decoder then uses a RNN and MemNN to generate text. The probability distribution over the vocab is generated by concatenating the attention read from the first hop and the current query vector. The probability distribution of generating from the input history is calculated by using the attention weights at the last hop, which is motivated that the first hop detects a more general distribution while the last hop defines a sharper distribution

Difficulties in Neural MDS

Initial experiments of training on simply the DUC (01-03) and testing on DUC04 resulted in extreme overfitting. This was to be expected due to the size of the datasets. We have begun training on the CNN/Daily Mail dataset, but are waiting on training to finish and to fine tune the models. Fine-tuning brings to light the subject of transfer learning, which is not a simple problem, so we will have to explore creative ways to fine-tune our models

Conclusion and Future Work

We plan to fine-tune our models trained on single-document summarization data. We are curious as to the results on the single document summarization data, as this method has not previously been tried for summarization. Additionally, we would like to incorporate external knowledge bases into summarization through this memory network, as the initial paper for Mem2Seq did. Finally, we would like to explore smarter ways to incorporate memory networks in seq2seq tasks. There as been much work on gated reading networks and Dynamic Memory Networks which allow for more efficient reading and writing of memories which can scale to the task of summarization.

Acknowledgement

Thank you to Professor Radev for advising this project and for providing the foundation of work in AAN to build upon. For all references please see the associated PowerPoint Presentation for this project on the LILY website.