Automatic Generation of Introduction and Abstract for Research Papers

What

Why

How

- Generate Abstract & Introduction of research papers (computational Linguistic domain)
- To make authors’ life easier
- Abstractive text summarization approach

Backlog

Data gathering

Downloading arxiv papers frequently

Data formatting

Unzipped and then converted to JSON Line

Evaluating existing solutions

Main focus: Cohan et al.

Developing the model

Sequence to Sequence based summarization model

Data gathering

Data gathering

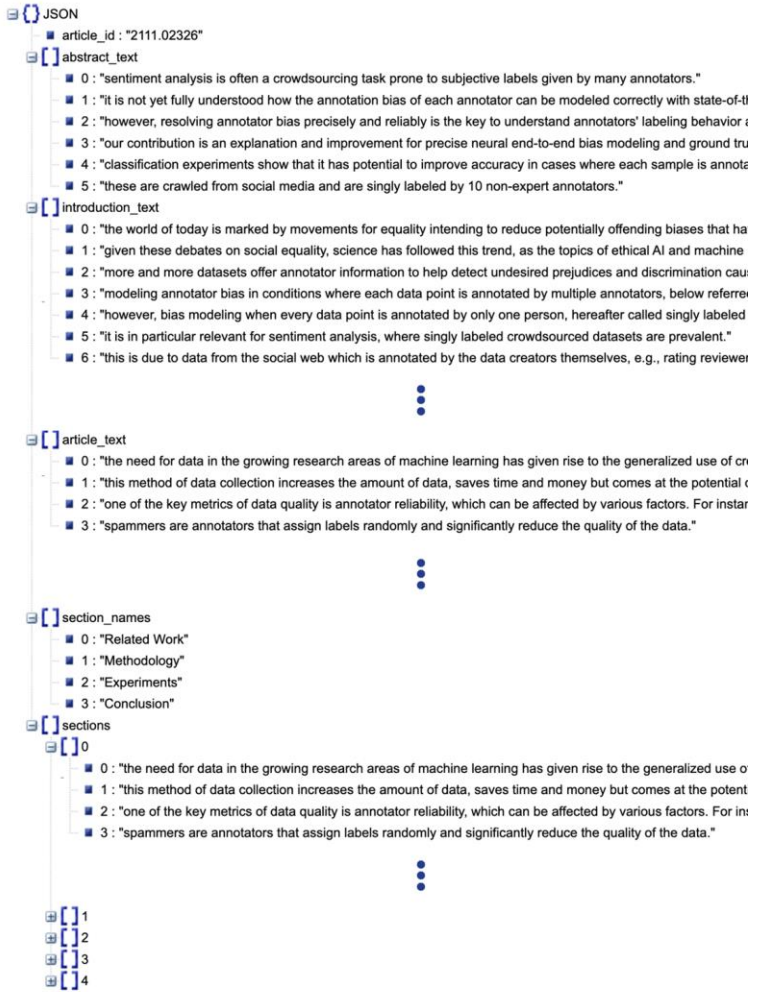
- At the moment **1048** papers from arxiv.org have been collected.

- Weekly running the downloader for the url, https://arxiv.org/list/cs.CL/recent

Data formatting

Data formatting

- Converting latex papers into **JSON**
- Or **JSON Lines** (Focusing existing solution has choose this way)



Evaluating existing solutions

Evaluating existing solutions

- Main focus is on Paper
- Code

A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents

Arman Cohan† Franck Dernoncourt* Doo Soon Kim* Trung Bui*
Seokhwan Kim* Walter Chang* Nazli Goharian†

†IRLab, Georgetown University, Washington, DC
{arman,nazli}@ir.cs.georgetown.edu

*Adobe Research, San Jose, CA
{dernonco,dkim,bui,seokim,wachang}@adobe.com

Abstract

Neural abstractive summarization models have led to promising results in summarizing relatively short documents. We propose the first model for abstractive summarization of single, longer-form documents (e.g., research papers). Our approach consists of a new hierarchical encoder that models the discourse structure of a document, and an attentive discourse-aware decoder to generate the summary. Empirical results on two large-scale datasets of scientific papers show that our model significantly outperforms state-of-the-art models.

1 Introduction

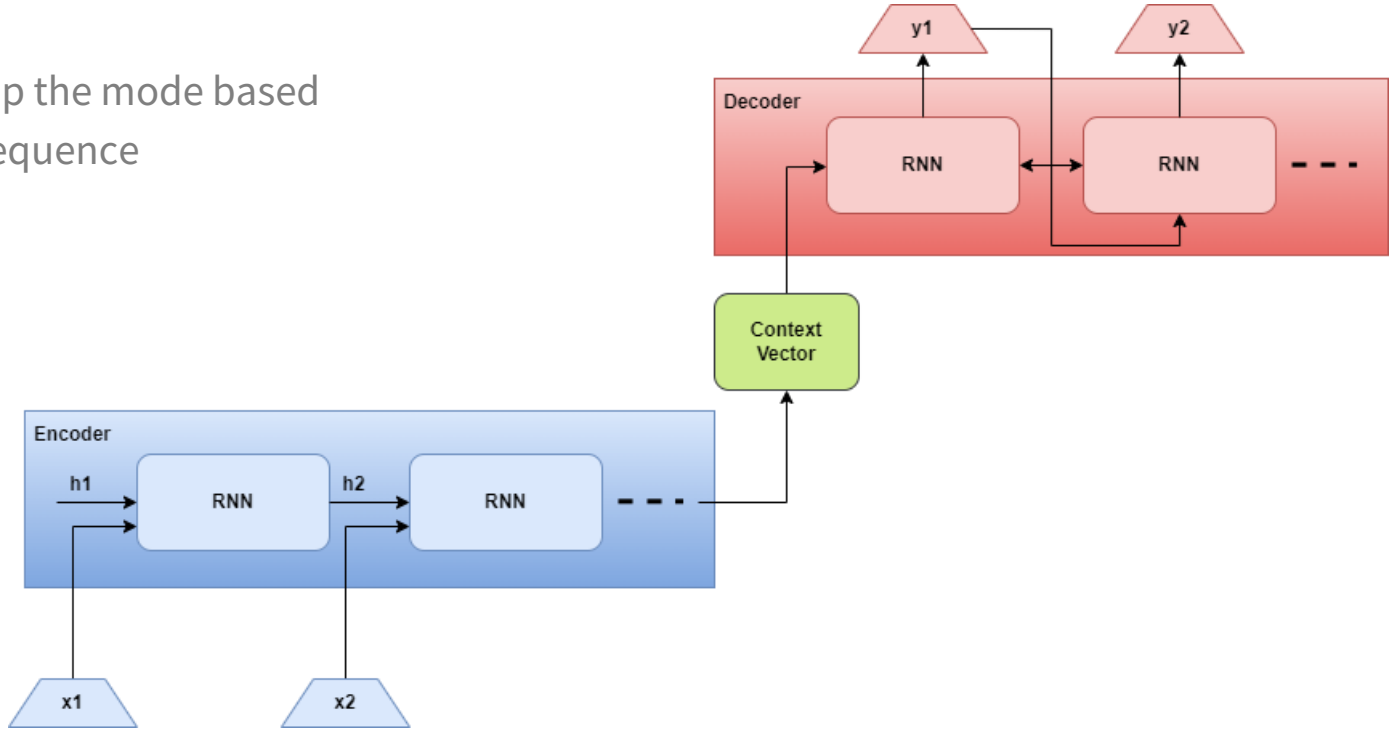
Existing large-scale summarization datasets consist of relatively short documents. For example, articles in the CNN/Daily Mail dataset (Hermann et al., 2015) are on average about 600 words long. Similarly, existing neural summarization models have focused on summarizing sentences and short documents. In this work, we propose a model for effective abstractive summarization of longer documents. Scientific papers are an example of documents that are significantly longer than news articles (see Table 1). They also follow a standard discourse structure describing the abstractive summarization (Nallapati et al., 2016; See et al., 2017; Paulus et al., 2017; Li et al., 2017). These approaches employ a general framework of sequence-to-sequence (seq2seq) models (Sutskever et al., 2014) where the document is fed to an encoder network and another (recurrent) network learns to decode the summary. While promising, these methods focus on summarizing news articles which are relatively short. Many other document types, however, are longer and structured. Seq2seq models tend to struggle with longer sequences because at each decoding step, the decoder needs to learn to construct a context vector capturing relevant information from all the tokens in the source sequence (Shao et al., 2017).

Our main contribution is an abstractive model for summarizing scientific papers which are an example of long-form structured document types. Our model includes a hierarchical encoder, capturing the discourse structure of the document and a discourse-aware decoder that generates the summary. Our decoder attends to different discourse sections and allows the model to more accurately represent important information from the source resulting in a better context vector. We also introduce two large-scale datasets of long and struc-

Developing the model

Developing the model

- Planned to develop the mode based on Sequence to Sequence architecture



Questions

Suggestions