

Automatic Generation of Introduction and Abstract for Research Papers



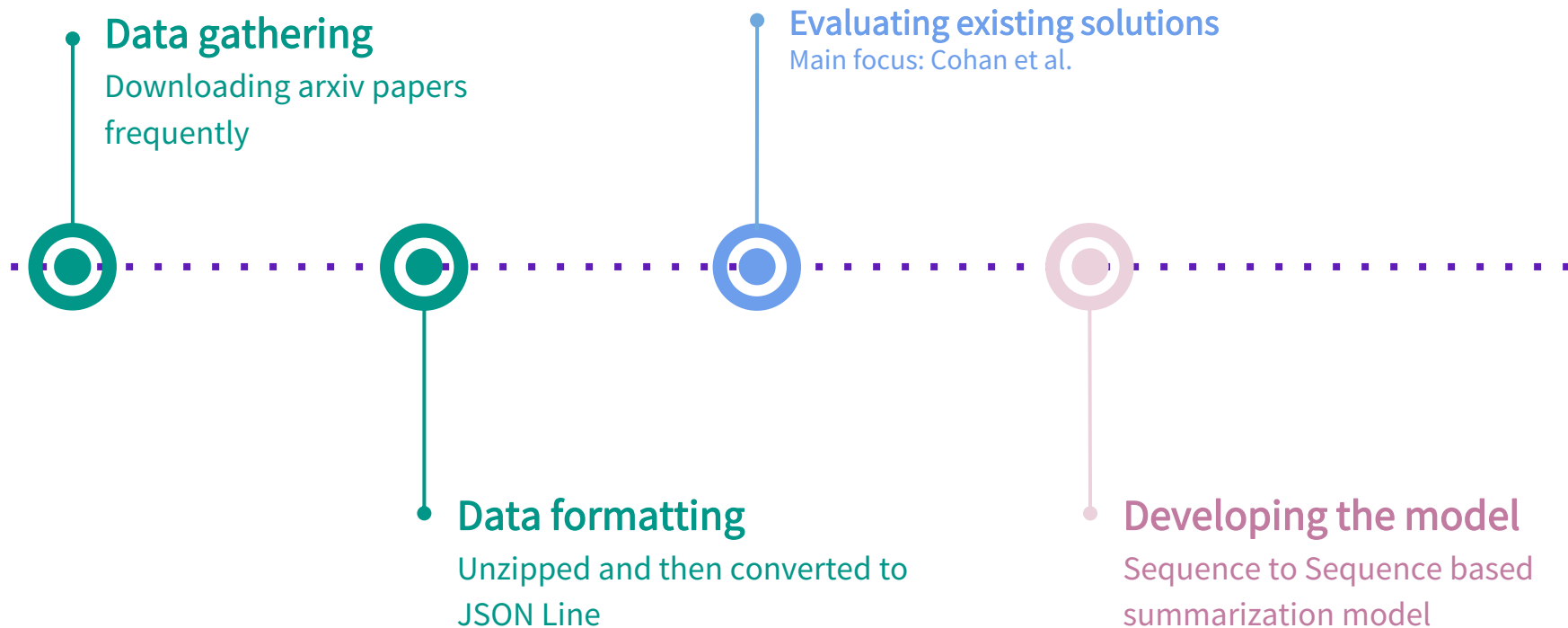
What

Why

How

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

Backlog



Data gathering

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)

```
{
  "article_id": "2111.02326",
  "abstract_text": [
    "0: \"sentiment analysis is often a crowdsourcing task prone to subjective labels given by many annotators.\"",
    "1: \"it is not yet fully understood how the annotation bias of each annotator can be modeled correctly with state-of-the-art methods.\"",
    "2: \"however, resolving annotator bias precisely and reliably is the key to understand annotators' labeling behavior.\"",
    "3: \"our contribution is an explanation and improvement for precise neural end-to-end bias modeling and ground truth estimation.\"",
    "4: \"classification experiments show that it has potential to improve accuracy in cases where each sample is annotated by multiple annotators.\"",
    "5: \"these are crawled from social media and are singly labeled by 10 non-expert annotators.\""],
  "introduction_text": [
    "0: \"the world of today is marked by movements for equality intending to reduce potentially offending biases that have emerged in the past.\"",
    "1: \"given these debates on social equality, science has followed this trend, as the topics of ethical AI and machine learning have become increasingly relevant.\"",
    "2: \"more and more datasets offer annotator information to help detect undesired prejudices and discrimination caused by biased data.\"",
    "3: \"modeling annotator bias in conditions where each data point is annotated by multiple annotators, below referred to as multi-annotator bias modeling, is a challenging task.\"",
    "4: \"however, bias modeling when every data point is annotated by only one person, hereafter called singly labeled bias modeling, is a simpler task.\"",
    "5: \"it is in particular relevant for sentiment analysis, where singly labeled crowdsourced datasets are prevalent.\"",
    "6: \"this is due to data from the social web which is annotated by the data creators themselves, e.g., rating reviewers.\""],
  "article_text": [
    "0: \"the need for data in the growing research areas of machine learning has given rise to the generalized use of crowdsourcing.\"",
    "1: \"this method of data collection increases the amount of data, saves time and money but comes at the potential cost of quality.\"",
    "2: \"one of the key metrics of data quality is annotator reliability, which can be affected by various factors. For instance, spammers are annotators that assign labels randomly and significantly reduce the quality of the data.\""],
  "section_names": [
    "0: \"Related Work\"",
    "1: \"Methodology\"",
    "2: \"Experiments\"",
    "3: \"Conclusion\""],
  "sections": [
    "0: [
      "0: \"the need for data in the growing research areas of machine learning has given rise to the generalized use of crowdsourcing.\"",
      "1: \"this method of data collection increases the amount of data, saves time and money but comes at the potential cost of quality.\"",
      "2: \"one of the key metrics of data quality is annotator reliability, which can be affected by various factors. For instance, spammers are annotators that assign labels randomly and significantly reduce the quality of the data.\""]
    ],
    "1: [
      "0: \"the need for data in the growing research areas of machine learning has given rise to the generalized use of crowdsourcing.\"",
      "1: \"this method of data collection increases the amount of data, saves time and money but comes at the potential cost of quality.\"",
      "2: \"one of the key metrics of data quality is annotator reliability, which can be affected by various factors. For instance, spammers are annotators that assign labels randomly and significantly reduce the quality of the data.\""]
    ],
    "2: [
      "0: \"the need for data in the growing research areas of machine learning has given rise to the generalized use of crowdsourcing.\"",
      "1: \"this method of data collection increases the amount of data, saves time and money but comes at the potential cost of quality.\"",
      "2: \"one of the key metrics of data quality is annotator reliability, which can be affected by various factors. For instance, spammers are annotators that assign labels randomly and significantly reduce the quality of the data.\""]
    ],
    "3: [
      "0: \"the need for data in the growing research areas of machine learning has given rise to the generalized use of crowdsourcing.\"",
      "1: \"this method of data collection increases the amount of data, saves time and money but comes at the potential cost of quality.\"",
      "2: \"one of the key metrics of data quality is annotator reliability, which can be affected by various factors. For instance, spammers are annotators that assign labels randomly and significantly reduce the quality of the data.\""]
    ],
    "4: [
      "0: \"the need for data in the growing research areas of machine learning has given rise to the generalized use of crowdsourcing.\"",
      "1: \"this method of data collection increases the amount of data, saves time and money but comes at the potential cost of quality.\"",
      "2: \"one of the key metrics of data quality is annotator reliability, which can be affected by various factors. For instance, spammers are annotators that assign labels randomly and significantly reduce the quality of the data.\""]
    ]
  ]
}
```

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

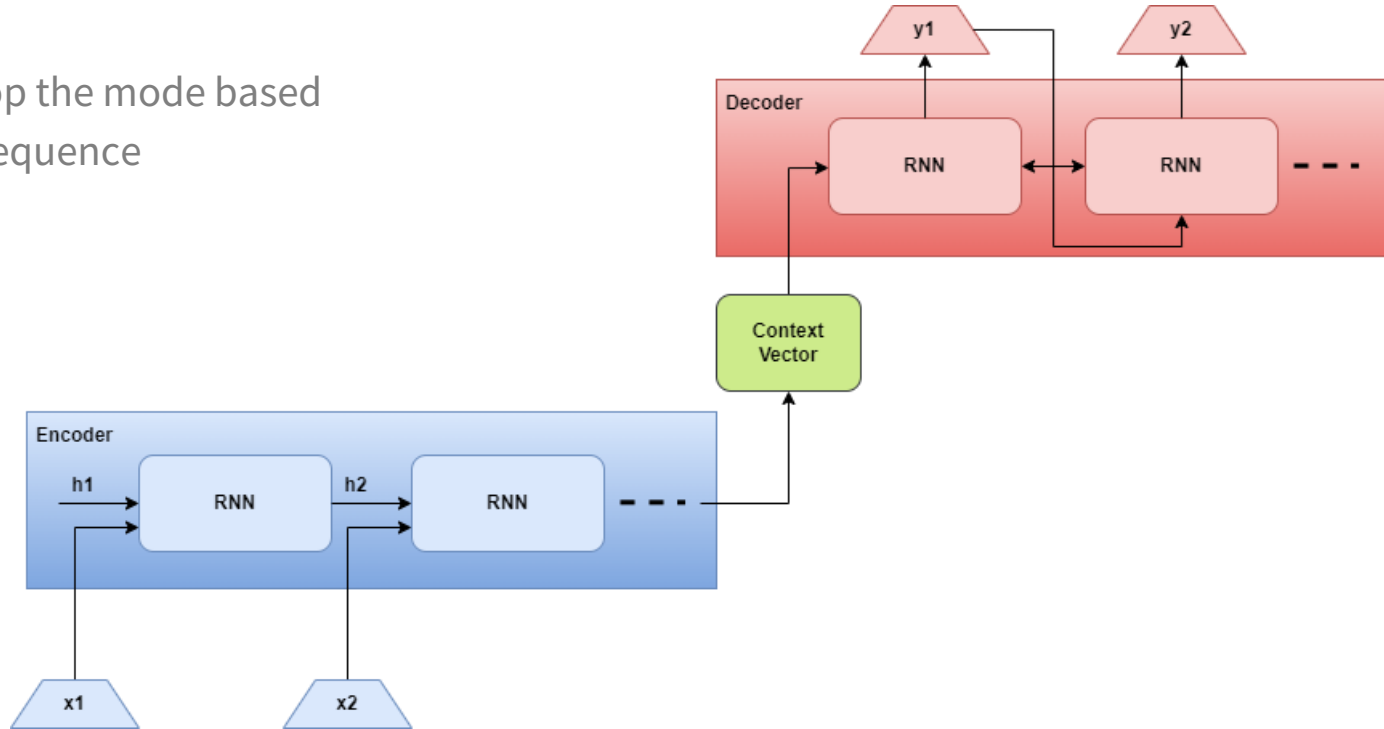
abstractive summarization (Nallapati et al., 2016; See et al., 2017; Paulus et al., 2017; Li et al., 2017). These approaches employ a general framework of sequence-to-sequence (seq2seq) models (Sutskever et al., 2014) where the document is fed to an encoder network and another (recurrent) network learns 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 model based on Sequence to Sequence architecture



Questions

Suggestions
