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

**Abstract**

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.

**I. Introduction**

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 structured scientific papers obtained from arXiv and PubMed.

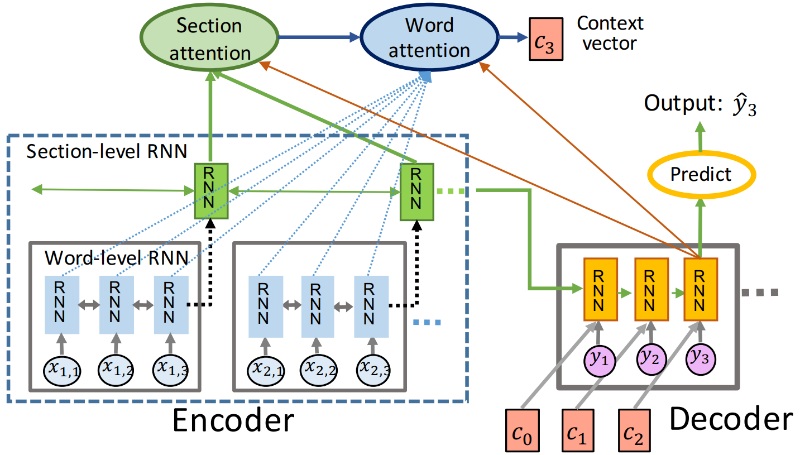
**2. Background**

**Attentive Decoding**

The attention mechanism maps the decoder state and the encoder states to context vector . Incorporating this context vector at each decoding timestep (attentive decoding) is proven effective in seq2seq models

where are the attention weights calculated as follow

**3. Model**



**Hierarchical Encoder**

We first encoder each discourse section and then encoder the document

where N is number of sections and M is the maximum section length.

The parameters of are shared for all the discourse sections.

We use a single layer bidirectional LSTM for both and

**Discourse-Aware Decoder**

At each decoding timestep, in addition to the words in the document, we also attend to the relevant discourse section. Then we use the discourse-related information to modify the word-level attention function.

At each timestep , the decoder state and the context vector are used to estimate the probability of next word

where V is a vocabulary weight matrix.

**Copying from source**

Address the problem of unknown token prediction by allowing the model to occasionally copy words directly from source instead of generating a new token.

We add an additional binary variable to the decoder, indicating generating a word from vocabulary ( = 0) or copying a word from the source ( = 1). The probability is learnt during training according to the following equation

where is decoder input at timestep .

Then the next word is generated according to

The joint probability is decomposed as

where is the probability of generating and is probability of copying a word from the source

**Decoding Coverage**

In long sequences, the neural generation models tend to repeat phrases where the softmax layer predicts the same phrase multiple times over multiple timesteps.

We track attention coverage to avoid repeatedly attending to the same steps with a coverage vector

We incorporate the decoder coverage as an additional input to the attention function

**4. Related Work**

**5. Data**

**6. Experiments**

**7. Conclusion and Future Work**

May 18, 2023

Chu Dinh Duc