Get To The Point: Summarization with Pointer-Generator Networks

Abigail See , Peter J. Liu, Christopher D. Manning (2017)

2021-03-27

1. Introduction

2 approaches to summarization

(1) Extractive methods : assemble summaries exclusively from passages (usually whole sentences) taken directly from the source text

⇨ (2) Abstractive methods: may generate novel words and phrases not featured in the source text – as a human-written abstract usually does.

1. Introduction

Limits of sequence to sequence approach

(1) inaccurately reproducing factual details

(2) an inability to deal with out-of-vocabulary (OOV) words

(3) repeating themselves

Original Text (truncated): lagos, nigeria (cnn) a day after winning nigeria's presidency, *muhammadu buhari* told cnn's christiane amanpour that he plans to aggressively fight corruption that has long plagued nigeria and go after the root of the nation's unrest. *buhari* said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, he said his administration is confident it will be able to thwart criminals and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. *buhari* defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to destabilize nigeria's economy. UNK says his administration is confident it will be able to thwart criminals and other nigerians. he says the country has long nigeria and nigeria's economy.

$$\text{loss}_t = -\log P(w_t^*) \longrightarrow \text{loss} = \frac{1}{T}\sum_{t=0}^{T} \text{loss}_t$$

2. Our models

2.1 Sequence-to-sequence attentional model

a probability distribution over the source words, that tells the decoder where to look to produce the next words

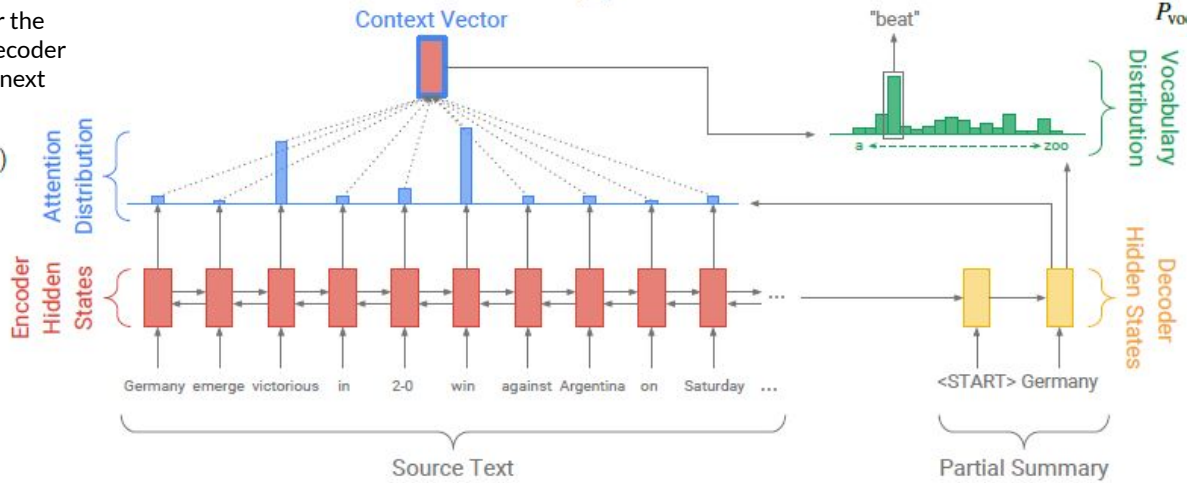
$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{\text{attn}})$$
$$a^t = \text{softmax}(e^t)$$

the *context vector* h_t^*:

$$h_t^* = \sum_i a_i^t h_i$$

probability distribution over all words in the vocabulary, and provides us with our final distribution from which to predict words w

$$P_{\text{vocab}} = \text{softmax}(V'(V[s_t, h_t^*] + b) + b')$$



2. Our models

2.2 Pointer-generator network

zero if w is an OOV

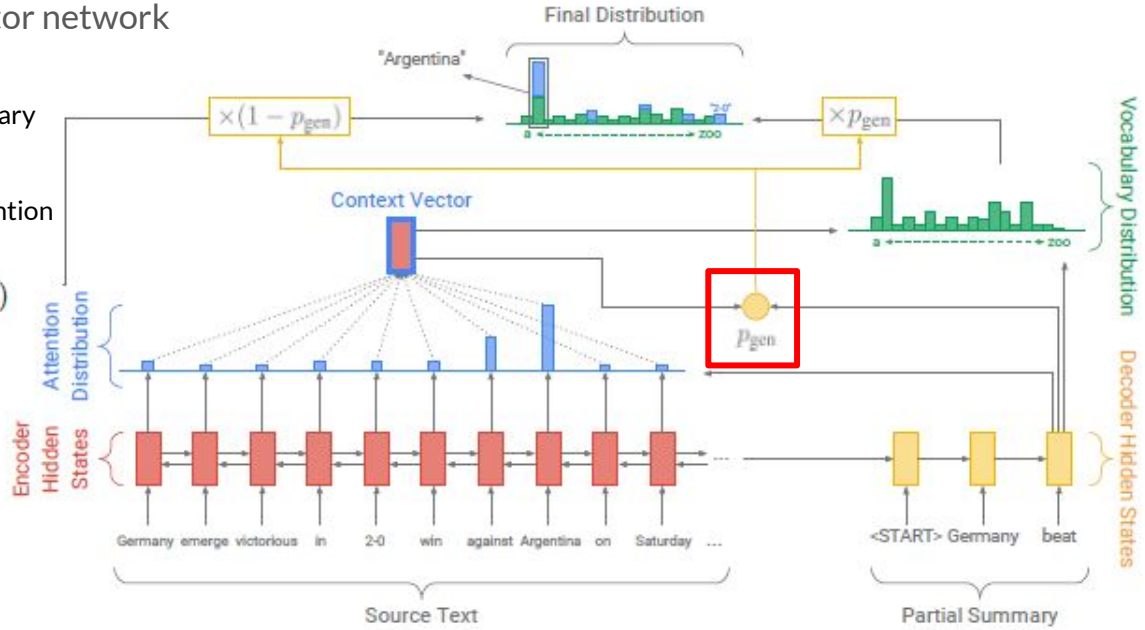
$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i = w} a_i^t$$

zero if w does not appear in the source document

a soft switch to choose between generating a word from the vocabulary by sampling from Pvocab,
or **copying a word** from the input sequence by sampling from the attention distribution

$$p_{\text{gen}} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + b_{\text{ptr}})$$

generation probability pgen ∈ [0,1] for timestep t is calculated from the context vector ht, the decoder state st and the decoder input xt



2. Our models

2.3 Coverage mechanism

$$c^t = \sum_{t'=0}^{t-1} a^{t'}$$

coverage vector(ct) is a (unnormalized) distribution over the source document words that represents the degree of coverage that those words have received from the attention mechanism so far

(1)attention mechanism

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{\text{attn}}) \longrightarrow e_i^t = v^T \tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{\text{attn}})$$

(2)loss

$$\text{loss}_t = -\log P(w_t^*) \longrightarrow \text{covloss}_t = \sum_i \min(a_i^t, c_i^t)$$

$$\text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(a_i^t, c_i^t)$$

4. Dataset

- use the CNN/Daily Mail dataset
- which contains online news articles (781 tokens on average) paired with multi-sentence summaries (3.75 sentences or 56 tokens on average).
- 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs.

5. Experiments

- 256- dimensional hidden states and 128-dimensional word embeddings
- For the pointer-generator models,we use a vocabulary of 50k words for both source and target – note that **due to the pointer network's ability to handle OOV words**,

  we can use a smaller vocabulary size than Nallapati et al.'s (2016) 150k source and 60k target vocabularies. For the baseline model, we also try a larger vocabulary size of 150k.
- Pointer and the coverage mechanism **introduce very few additional parameters to the network**: for the models with vocabulary size 50k, the baseline model has 21,499,600 parameters, the pointer-generator adds 1153 extra parameters , and coverage adds 512 extra parameters

5. Experiments

- We do not pretrain the word embeddings – they are learned from scratch during training.
- During training and at test time we truncate the article to 400 tokens and limit the length of the summary to 100 tokens for training and 120 tokens at test time.

  -> expedite training & testing raise the performance of the model
- We trained both our baseline models for about 600,000 iterations (33 epochs)

  We found the pointer-generator model quicker to train, requiring less than 230,000 training iterations (12.8 epochs)

6. Results

	ROUGE			METEOR	
	1	2	L	exact match	+ stem/syn/para
abstractive model (Nallapati et al., 2016)*	35.46	13.30	32.65	-	-
seq-to-seq + attn baseline (150k vocab)	30.49	11.17	28.08	11.65	12.86
seq-to-seq + attn baseline (50k vocab)	31.33	11.81	28.83	12.03	13.20
pointer-generator	36.44	15.66	33.42	15.35	16.65
pointer-generator + coverage	**39.53**	**17.28**	**36.38**	17.32	18.72
lead-3 baseline (ours)	40.34	17.70	36.57	20.48	22.21
lead-3 baseline (Nallapati et al., 2017)*	39.2	15.7	35.5	-	-
extractive model (Nallapati et al., 2017)*	39.6	16.2	35.3	-	-

6. Results

ROUGE
(Recall-Oriented Understudy for Gisting Evaluation)

Recall

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary}$$

Precision

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary}$$

1. **ROUGE-L** : measures longest matching sequence of words using LCS(longest commom sequence). An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order. Since it automatically includes longest in-sequence common n-grams, you don't need a predefined n-gram length
2. **LCS** : For example, consider the sequences (ABCD) and (ACBAD). They have 5 length-2 common subsequences: (AB), (AC), (AD), (BD), and (CD); 2 length-3 common subsequences: (ABD) and (ACD); and no longer common subsequences. So (ABD) and (ACD) are their longest common subsequences.

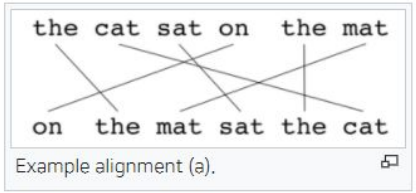
6. Results

METEOR

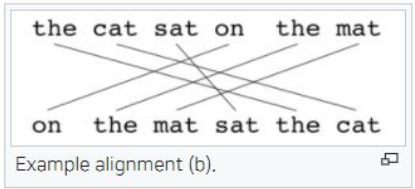
(Metric for Evaluation of Translation with Explicit ORdering)

based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision

$$F_{mean} = \frac{10PR}{R + 9P}$$



Example alignment (a).



Example alignment (b).

$$p = 0.5\left(\frac{c}{u_m}\right)^3$$

Where *c* is the number of chunks, and um is the number of unigrams that have been mapped. The final score for a segment is calculated as *M* below.

$$M = F_{mean}(1 - p)$$

The penalty has the effect of reducing the F_mean by up to 50% if there are no bigram or longer matches.

6. Results

Original Text (truncated): lagos, nigeria (cnn) a day after winning nigeria's presidency, *muhammadu buhari* told cnn's christiane amanpour that **he plans to aggressively fight corruption that has long plagued nigeria** and go after the root of the nation's unrest. *buhari* said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, **he said his administration is confident it will be able to thwart criminals** and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. *buhari* defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. **the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.**

Baseline Seq2Seq + Attention: **UNK UNK** says his administration is confident it will be able to **destabilize nigeria's economy**. **UNK** says his administration is confident it will be able to thwart criminals and other **nigerians**. **he says the country has long nigeria and nigeria's economy.**

Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption **in the northeast part of nigeria**. he says he'll "rapidly give attention" to curbing violence **in the northeast part of nigeria**. he says his administration is confident it will be able to thwart criminals.

Pointer-Gen + Coverage: *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.