Project proposal

CS492

Team members:
Martin Ohrt Elingaard
Dongjoo Kim
Tung-Duong Mai

Paper overview

Title: Get To The Point: Summarization with Pointer-Generator Networks

Venue: Presented at ACl 2017

Authors:
- Abigail See (Stanford University)
- Peter J. Liu (Google Brain)
- Christopher D. Manning (Stanford University)

Research questions

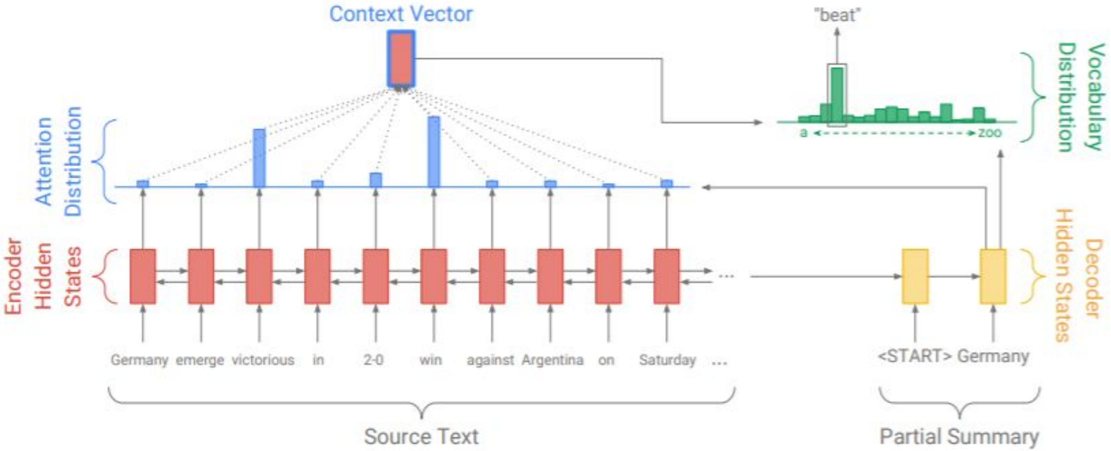
Problem: seq2seq models reproduce factual details inaccurately and repeat themselves. cannot deal with out-of-vocabulary (OOV)

Proposed solution:

- Use a hybrid pointer-generator network that can copy words from the source text via pointing and produce novel words through the generator.
- Use coverage to keep track of what has been summarized

Model: Baseline model

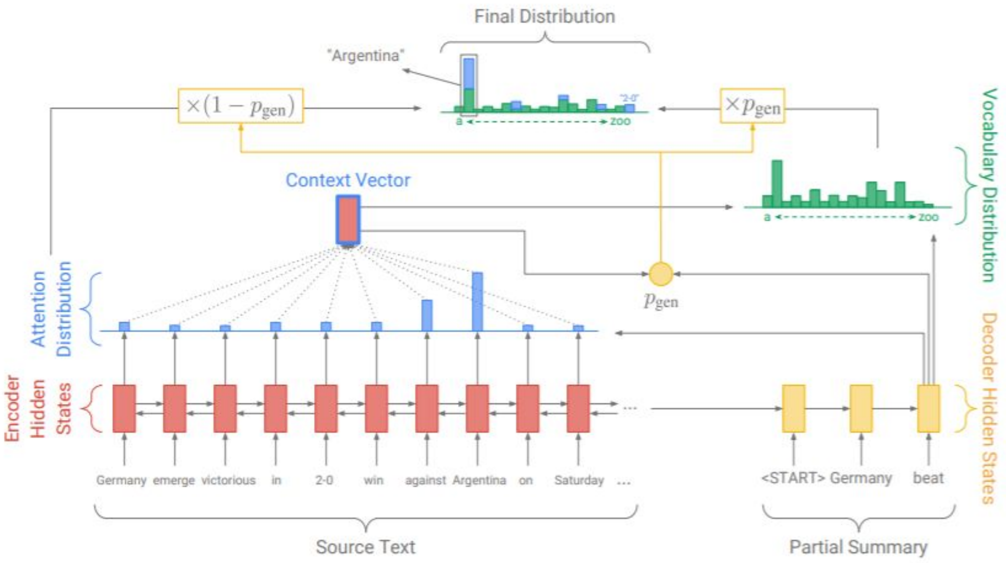
- seq2seq (Sutskever et al. 2014): *Sequence to Sequence Learning with Neural Networks*
- Attention (Bahdanau et al. 2015): *Neural machine translation by jointly learning to align and translate*



This picture is taken from ***Get To The Point: Summarization with Pointer-Generator Networks***

Model: pointer-generator

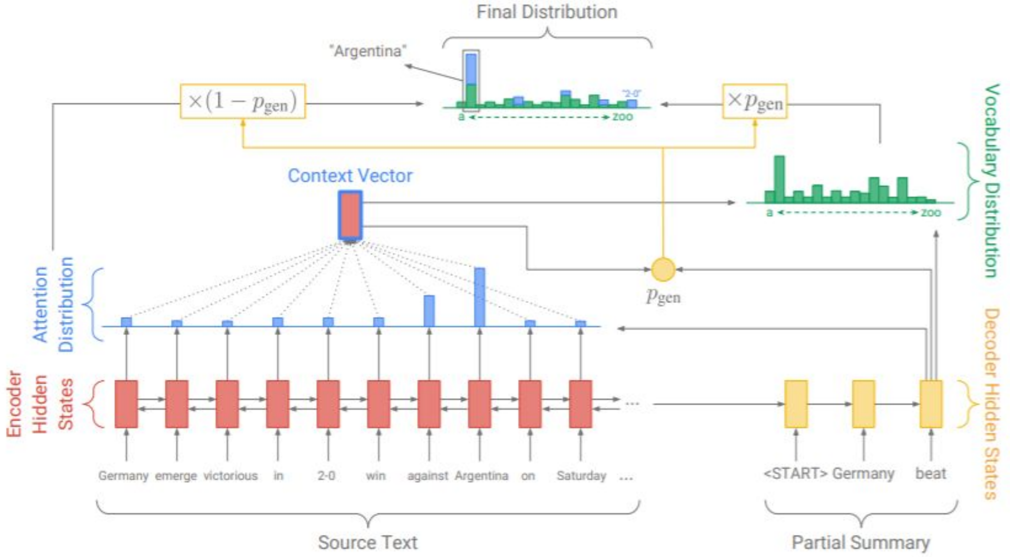
- Baseline model
- Pointer networks (Vinyals et al. 2015): *Pointer networks*



This picture is taken from ***Get To The Point: Summarization with Pointer-Generator Networks***

Model: coverage

- Pointer-generator model (Baseline model is also applicable)
- Coverage model (Tu et al., 2016): *Modeling coverage for neural machine translation*



This picture is taken from ***Get To The Point: Summarization with Pointer-Generator Networks***

Dataset

CNN(www.cnn.com) / Daily Mail(www.dailymail.uk.co) Corpus

DeepMind opened this corpus that used in their paper (Hermann et al., 2015)

This corpus has 286,817 training pairs, 13,368 validation pairs and 11,487 test pairs.

The source documents in the training set have 766 words spanning 29.74 sentences on an average.

The summaries consist of 53 words and 3.72 sentences.

Experiment

256- dimensional hidden states and 128-dimensional word embeddings

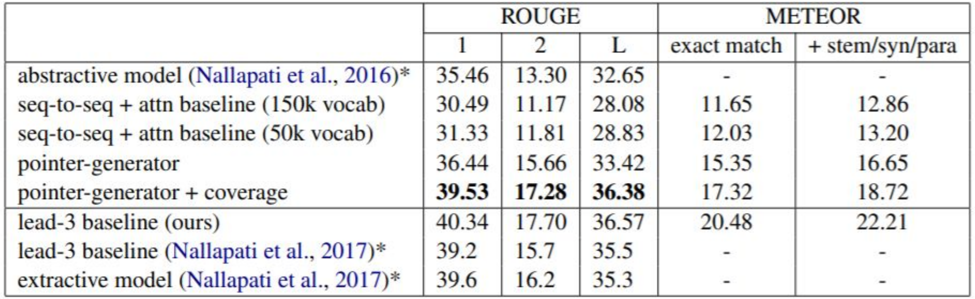
50k words for both source and target (baseline model: 50k and 150k)

No pre-trained word-embedding

Truncate the article to 400 tokens and limit the length of the summary to 100 tokens for training and 120 tokens at test time.

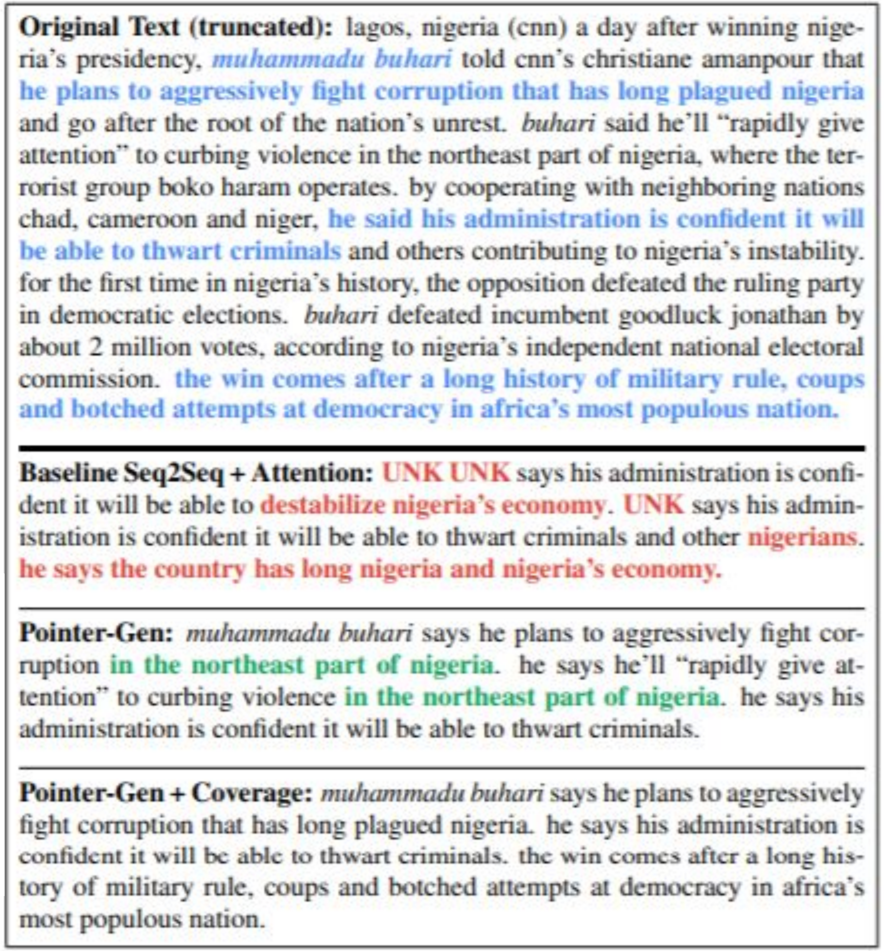
Result

Outperform the state-of-the-art abstractive system by at least **2 ROUGE points** (*ROUGE metric (Lin, 2004b)*



	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	-	-

These figures are taken from ***Get To The Point: Summarization with Pointer-Generator Networks***



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.**

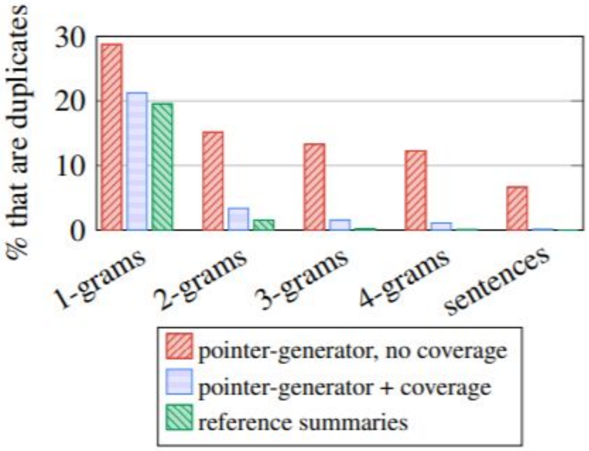
Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to **destabilize nigeria's economy**. UNK says his administration is confident it will be able to thwart criminals and other **nigerians**. **he says the country has long nigeria and nigeria's economy.**

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.

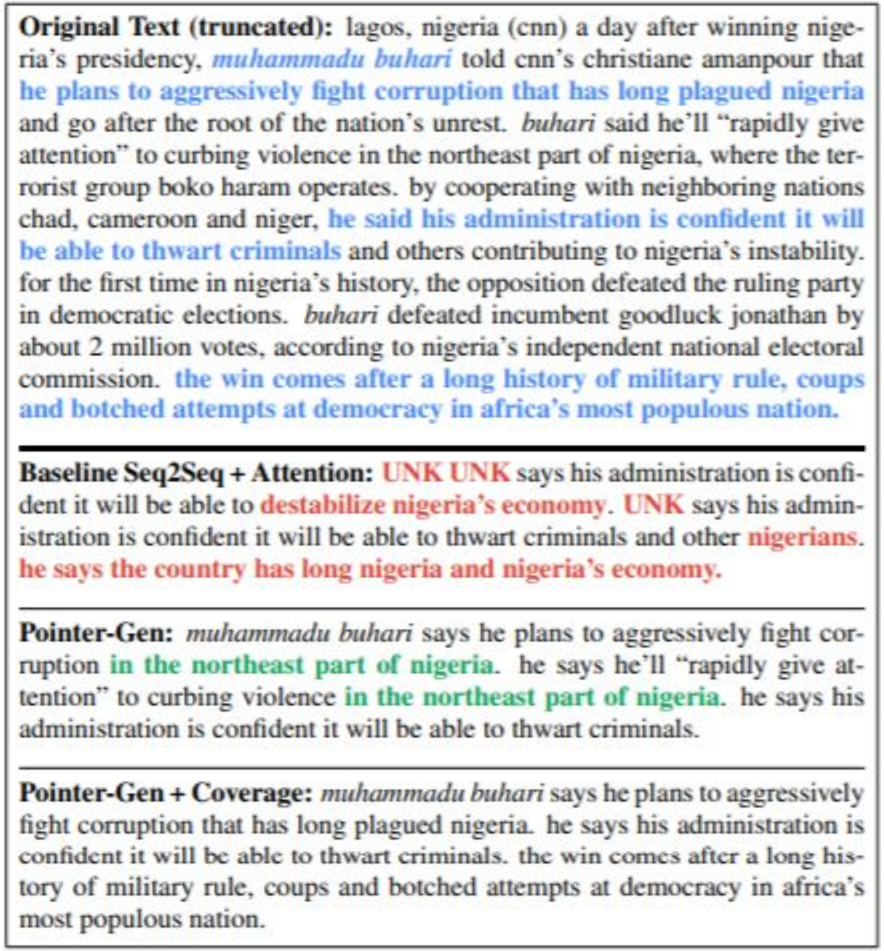
Pointer-Gen + Coverage: *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's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

Result

Coverage model produces a similar number as the reference summaries.



These figures are taken from ***Get To The Point: Summarization with Pointer-Generator Networks***



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.**

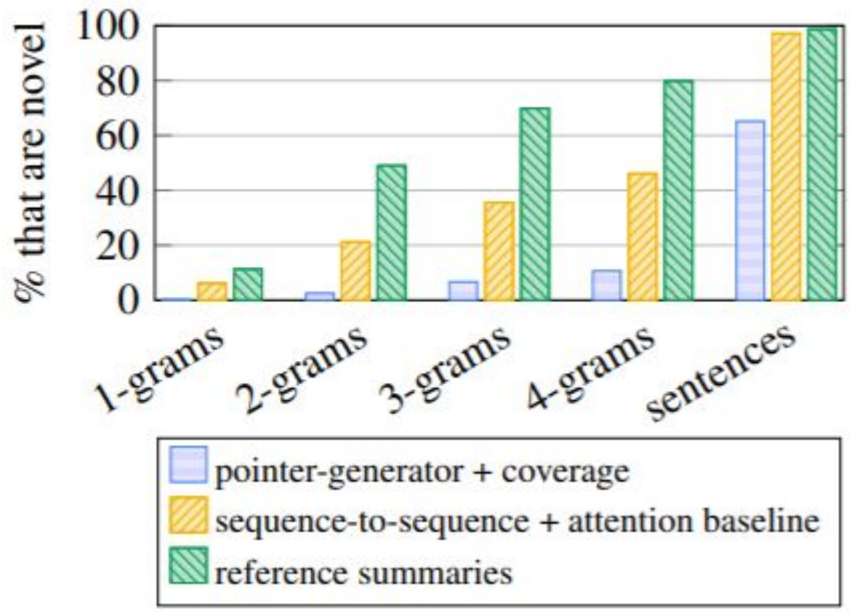
Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption **in the northeast part of nigeria**. he says he'll "rapidly give attention" to curbing violence **in the northeast part of nigeria**. he says his administration is confident it will be able to thwart criminals.

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.

Result

Level of abstraction is not very high

Lower result than extractive and lead-3 baseline



Proposed changes

- Change the dataset to scientific journal articles (Nikolov et al. 2018)
- Two datasets: abstract-to-headline, body-to-abstract
- Use body-to-abstract as this is where the strength of the hybrid network lies
  - Abstract contains high quality information for summarization -> good for extractive methods
  - Abstractions are good for summarizing large bodies of text

Research questions

- How does the pretrained model work on scientific journal articles?

- Can we outperform this model with smaller dataset + limited computation resources?

- Visualize word embeddings for scientific and journal articles
  - Can we identify common phrases within scientific writing vs journalism?

- Test authors hypothesis: Lead-3 baseline is strong for news articles
  - Not really true for scientific articles as most summarization is performed in abstract (not included in body) and conclusion

But isn’t just changing the dataset easy?

- Sophisticated model which builds upon multiple NLP concepts
  - Takes time to understand model + structure of code (extension of Google Brains TextSum)
  - Many learned parameters = long training time (3-4 days on powerful GPU)

- Preparation of dataset
  - Tokenization using StanfordNLP
  - Binarize and chunk data
  - Input data needs to be restructured (concatenate abstract+body)

- Visualization using tensorboard (word-embeddings and attention distribution)