Get To The Point: Summarization with Pointer-Generator Networks

Abigail See*, Peter J. Liu, Christopher D. Manning

* work done partially during an internship at Google Brain

The Need for Abstractive Summarization

Automatic text summarization is **increasingly vital** in the digital age. Two approaches:

Extractive summarization
Select and rearrange passages from the original text

- More **restrictive**
- Most **past work** has been extractive
- **Easier** to get reasonable performance

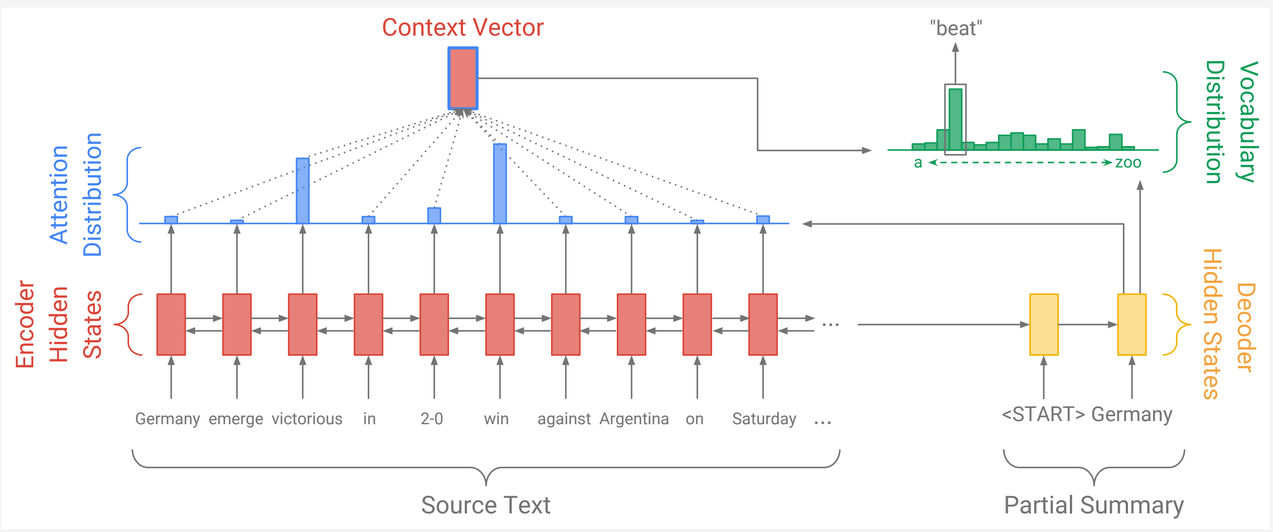
Abstractive summarization
Generate novel sentences

- More **expressive**
- **Humans** are abstractive summarizers
- More **difficult** and unpredictable

Abstractive summarization is essential for high-quality output

Abstractive Summarization with RNNs

- **Recurrent Neural Networks** (RNNs) provide a potentially powerful solution for abstractive summarization.
- Using the **attention mechanism** (see below), they can generate **new words** (*Germany beat Argentina 2-0)* by attending to relevant words (*victorious, win*) in the source text.



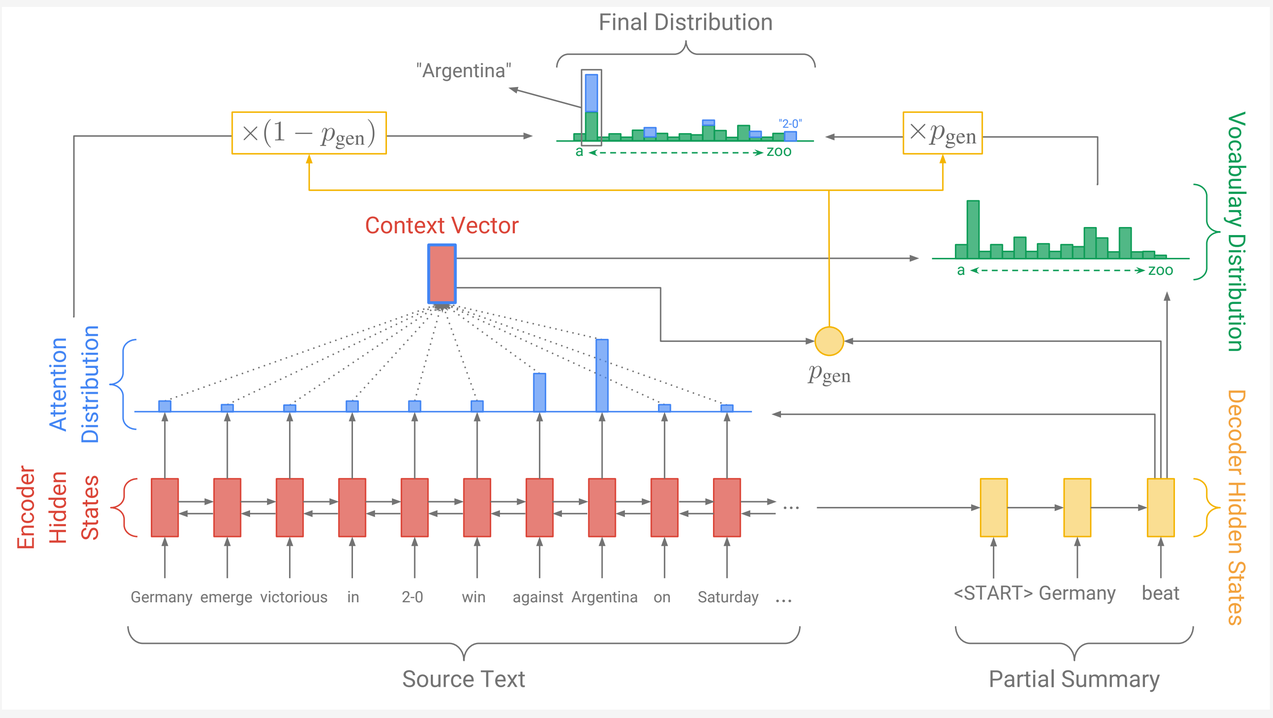
Two common problems:

1. Summaries **repeat themselves** (e.g. *Germany beat Germany beat Germany beat…*)
2. Summaries reproduce **factual details** inaccurately (e.g. *Germany beat Argentina* ***3-2***)

Easier Copying with Pointer-Generator Network

Problem: Factual details (especially rare and OOV words) are **copied inaccurately**

Solution: A **hybrid network** that can copy via *pointing,* or *generate* from a fixed vocabulary



- For each summary word, the network first calculates the **generation probability** p_{gen}
- p_{gen} interpolates between copying from **attention distribution** a and generating from **vocabulary distribution** P_{vocab}

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

Probability that next word is w

Probability of generating w from the fixed vocabulary

Sum of attention distribution everywhere w appears in the source text

Advantages:

- Faster to train
- Easy to accurately reproduce phrases
- Can copy OOV words (don't need large vocabulary)

Best of both worlds: abstractive (generating) and extractive (copying)

Eliminating Repetition with Coverage

Problem: Summaries are **repetitive**.

Solution: Penalize **repeatedly attending** to the **same parts** of the source text.

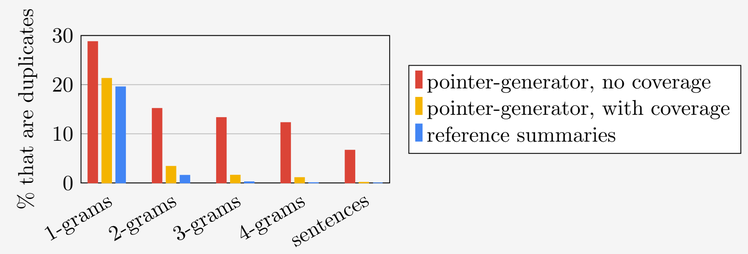
On each decoder timestep t, the **coverage vector** c^t tells us what has been attended to (thus summarized) so far.

Coverage vector → $c^t = \sum_{t'=0}^{t-1} a^{t'}$ ← Sum of attention distributions so far

Penalize overlap between coverage vector c^t and new attention distribution a^t.

Coverage loss function → $\text{covloss}_t = \sum_i \min(a_i^t, c_i^t)$ ← Overlap between coverage and current attention

Result: repetition reduced to **similar level** as reference summaries



Coverage eliminates undesirable repetition

Experiments

Dataset: *CNN/Daily Mail* (news article → multi-sentence summary)

	ROUGE-1	ROUGE-2	ROUGE-L
abstractive model (Nallapati et al., 2016)*	35.46	13.30	32.65
sequence-to-sequence + attention baseline	31.33	11.81	28.83
pointer-generator	36.44	15.66	33.42
pointer-generator + coverage	**39.53**	**17.28**	**36.38**
lead-3 baseline (ours)	40.34	17.70	36.57
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

Abstractive systems

Extractive systems

- Our pointer-generator + coverage model **beats** best abstractive system.
- Extractive systems and lead-3 baseline remain **difficult to beat**.
  - The ROUGE metric is **not robust** to paraphrasing

Example Output

Article (truncated): andy murray came close to giving himself some extra preparation time for his wedding next week before ensuring that he still has unfinished tennis business to attend to . the world no 4 is into the semi-finals of the miami open , but not before getting a scare from 21 year-old austrian dominic *thiem* , who pushed him to 4-4 in the second set before going down 3-6 6-4 , 6-1 in an hour and three quarters . murray was awaiting the winner from the last eight match between tomas berdych and argentina 's juan monaco . prior to this tournament *thiem* lost in the second round of a challenger event to soon-to-be new brit *aljaz* bedene . andy murray pumps his first after defeating dominic *thiem* to reach the miami open semi finals . *muray* throws his *sweatband* into the crowd after completing a 3-6 , 6-4 , 6-1 victory in florida . murray shakes hands with *thiem* who he described as a ' strong guy ' after the game . and murray has a fairly simple message for any of his fellow british tennis players who might be agitated about his imminent arrival into the home ranks : do n't complain . instead the british no 1 believes his colleagues should use the assimilation of the world number 83 , originally from slovenia , as motivation to better themselves .

Reference Summary:
british no 1 defeated dominic *thiem* in miami open quarter finals .
andy murray celebrated his 500th career win in the previous round .
third seed will play the winner of tomas berdych and juan monaco in the semi finals of the atp masters 1000 event in key biscayne .

Baseline:
andy murray beat UNK **bedene** **6-3** , 6-4 , 6-1 in an hour and three quarters .
british no 1 believes his colleagues should use the **maze** of the world number 83 , originally from slovenia , as motivation to better themselves .

Pointer-Generator, No Coverage:
andy murray came close to giving himself some extra preparation for his wedding next week before getting a scare from 21 year-old austrian dominic *thiem* .
the world no 4 is into the semi-finals of the miami open , but not before getting a scare from 21 year-old austrian dominic *thiem* .

Pointer-Generator, With Coverage:
andy murray defeated dominic *thiem* 3-6 6-4 , 6-1 in an hour and three quarters .
murray was awaiting the winner from the last eight match between tomas berdych and argentina 's juan monaco .
prior to this tournament thiem lost in the second round of a challenger event to soon-to-be new brit *aljaz* bedene .

Yellow highlight shows final value of the coverage vector (i.e. what has been covered by the final summary)

Baseline model produces **factual inaccuracies** and UNKs

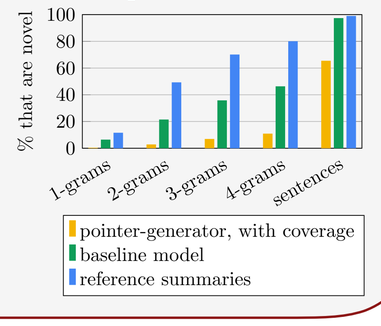
Pointer-generator model copies accurately, deals with *OOVs* but repeats itself

Coverage model has no repetition. Green highlighting shows generation probability p_{gen}.

How Abstractive Is Our Output?

- Our network uses the **pointer more than the generator** (average p_{gen} = 0.17)
- It produces **some** novel words and phrases, but **fewer** than the reference summaries

Open question: How to make pointer-generator network more abstractive?



Conclusion

- Pointer-generator networks enable **more accurate copying**, are **easier to train** and can deal with **OOVs**
- Coverage **drastically reduces** repetition
- ROUGE metric is of **limited use** for evaluating abstractive systems
- Future work: make the pointer-generator network **more abstractive**

Acknowledgements: **(i)** NVIDIA Corporation
(ii) DARPA Deep Exploration and Filtering of Text (DEFT) Program under AFRL contract no. FA8750-13-2-0040