**Summary:**

The abstractive summarization using Neural seq2seq models, although produce summaries closer to human-level summarization compared to extractive summarization techniques, have three issues: they tend to produce factual (numbers) details inaccurately, tend to repeat the words, and cannot handle the out-of-vocabulary (OOV) words. In this paper, they introduced a hybrid approach called “Pointer-Generator” to tackle the factual details and OOV words. This approach combines both the extractive and abstractive summarization techniques. The pointer copies the words from the source text while the generator generates new words simultaneously. It uses the generation probability pgen ([0,1]) to select a word from the copied and generated words. They also put forward another mechanism called “Coverage” to reduce the repetition of words. This approach produces the coverage vector that is the sum of all the previous attention distributions, which tells the network what words have been covered already. This vector is also included to calculate the next attention distribution along with the input and hidden states outputs. They applied this “Pointer-Generator with coverage” model on the CNN/Daily mail datasets and got better ROUGE score than the current abstractive state-of-the-art techniques.

**Things I liked about the paper:**

1. Combining both the abstractive and the extractive techniques to handle the factual details and OOV words.
2. Using the previous attention distributions to calculate the next attention distribution to discourage the repetition.
3. Explanation why ROUGE metrics is an unreliable technique to evaluate the produced summaries and why it gives the extractive summaries more score compared to the abstractive summaries.

**Things need to be improved/ have potential future direction:**

1. Better evaluation technique than the ROUGE and METEOR.
2. Should be able to produce more ROUGE and METEOR scores if they cannot come up with a better evaluation technique.
3. Should try to reach the human-level abstraction while keeping out of the swamp of errors like repetition, nonsense, and incorrect numbers. That is, they should increase the no of novel words in the summaries.

**Datasets used:**

They used the CNN/Daily mail datasets that contains online news articles (781 tokens on average) paired with multi-sentence summaries (3.75 sentences or 56 tokens on average). To compare their models with the current state-of-the-art techniques, they used the same scripts supplied by Nallapati et al. (2016), which has 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs. But instead of using anonymized datasets like in Nallapati et al., 2016, 2017, they used the original text which they believed to require no pre-processing. For the hybrid models, they used 50K vocabulary for both source and the target. But for the baseline model they also tried 150k vocabulary along with the 50k one.

**Evaluation Metrics used:** They used two metrics: ROUGE and METEOR.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): This is like the BLEU metric. This measures the n-gram overlaps.

In this paper they used three different ROUGE metrics:

1. ROUGE-1: unigram overlap
2. ROUGE-2: bigram overlap
3. ROUGE-L: longest common subsequence overlap

METEOR (Metric for Evaluation of Translation with Explicit Ordering): The metric is based on the [harmonic mean](https://en.wikipedia.org/wiki/Harmonic_mean) of unigram [precision and recall](https://en.wikipedia.org/wiki/Precision_and_recall), with recall weighted higher than precision.

In this paper, they used two variants of METEOR:

1. Exact match mode: rewarding only exact matches between words,
2. full mode: this additionally rewards matching stems, synonyms, and paraphrases