Get To The Point: Summarization with Pointer-Generator Networks

Team 13: Martin Ohrt Elingaard, Dongjoo Kim, Tung-Duong Mai

Presentation link:
https://docs.google.com/presentation/d/165W_NxB24TjJXk0Czp4Ca2IC22gIVGySOULzGqCGmqw/edit#slide=id.p

Outline

Part 1: Motivation & Research Problem
Part 2: Models
Part 3: Dataset & Experiment
Part 4: Result
Part 5: Discussion

Part 1
Motivation & Research Problem

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)

Summarization (~2017)

- Extractive methods : just assemble sentences from original text.

  (Kupiec et al., 1995; Paice, 1990; Saggion and Poibeau, 2013)

- Abstractive methods : may generate words / phrases not featured in the source.

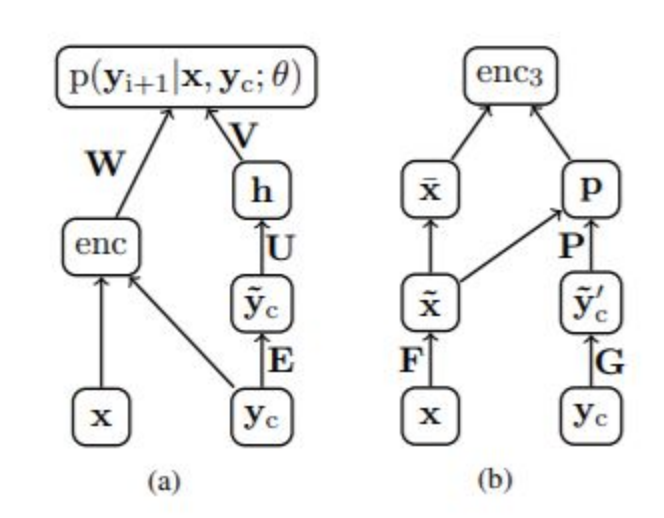
  → This approach had been popular after the advent of RNN.

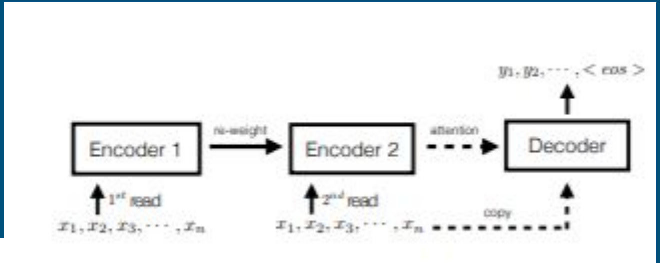
- Abstractive method is more difficult. Why do we need abstractive?

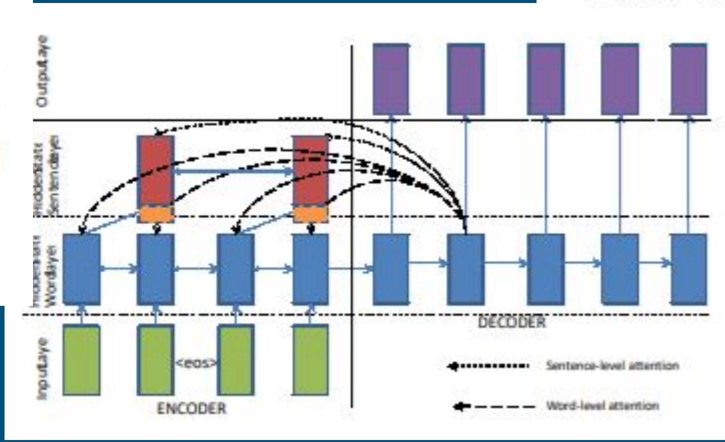
  It generates novel words and phrases similar with human-written

Past works with abstractive methods

(Chopra et al., 2016; Nallapati et al., 2016; Rush et al., 2015; Zeng et al., 2016)







Problems of past models

1. OOV
2. Factual errors
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.**

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

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.

Related works

- Neural abstractive summarization

  (Chopra et al., 2016; Nallapati et al., 2016; Rush et al., 2015;)

- Pointer-generator networks

  (Bahdanau et al., 2015; Gu et al., 2016; Miao and Blunsom, 2016;)

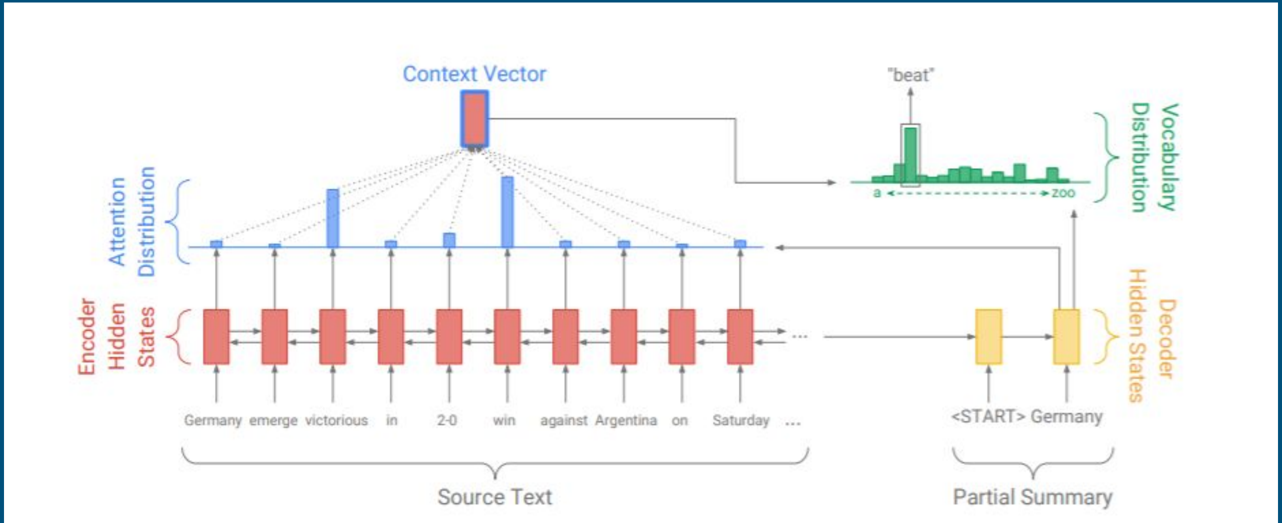
- Coverage

  (Tu et al., 2016; Xu et al., 2015; Chen et al., 2016;)

Part 2
Models

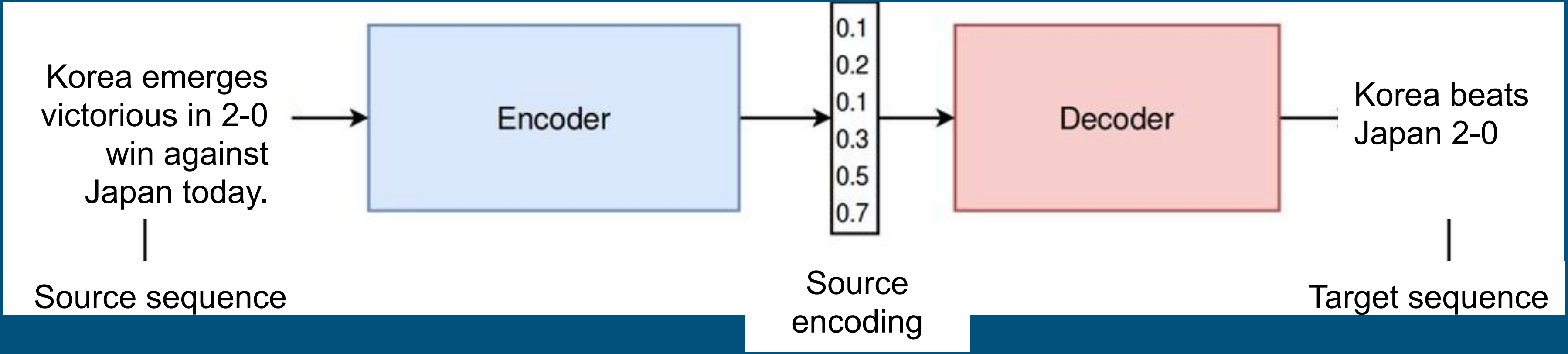
Baseline model: seq2seq + attention

"Abstractive text summarization using sequence-to-sequence RNNs and beyond"
Nallapati et al. (2016)

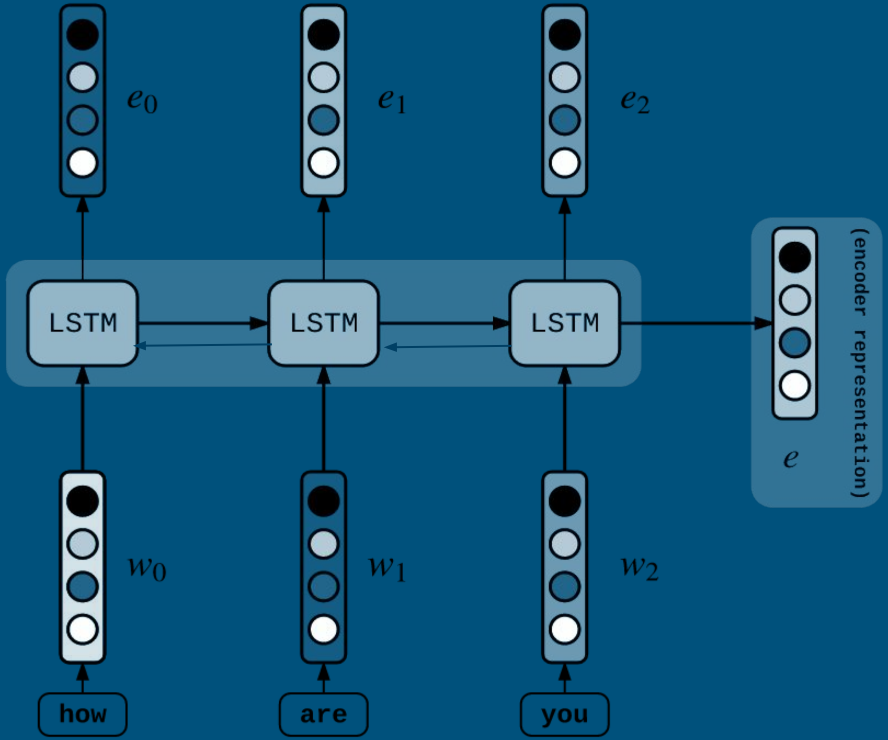


This picture is taken from ***Get To The Point: Summarization with Pointer-Generator Networks***

Seq2seq = Encoder + Decoder



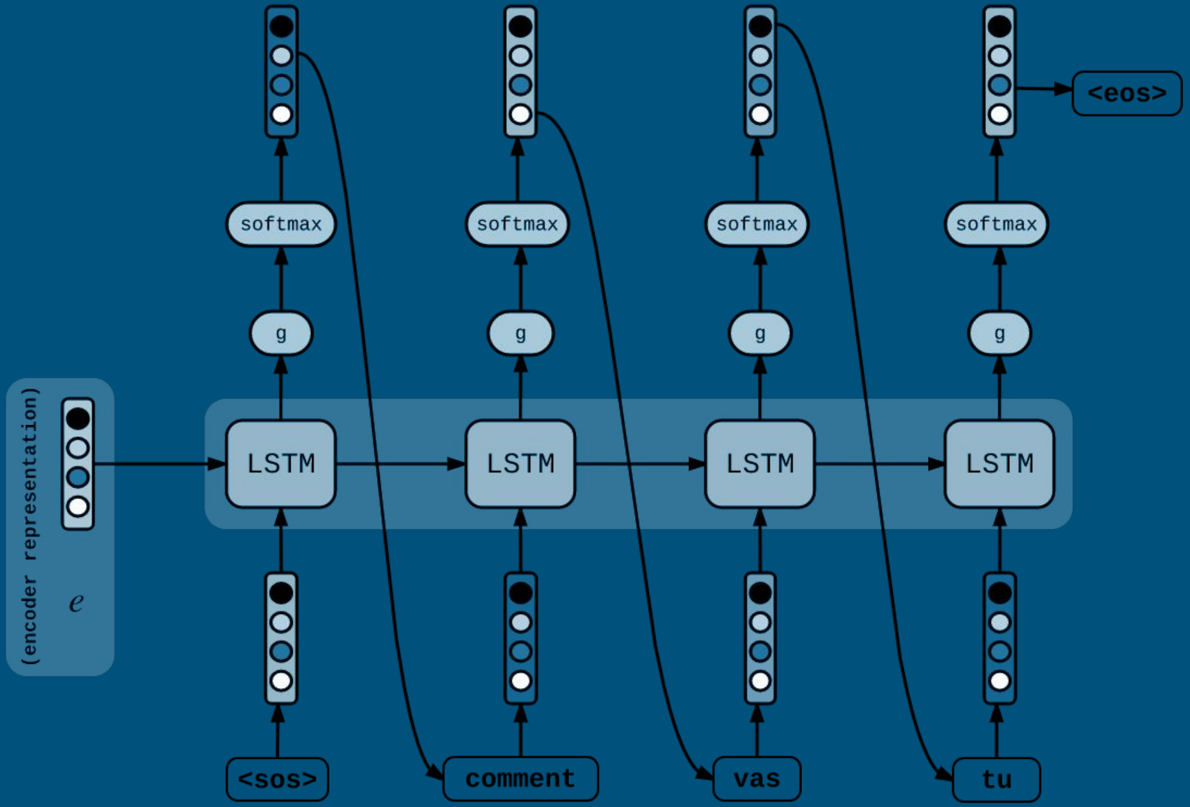
Encoder



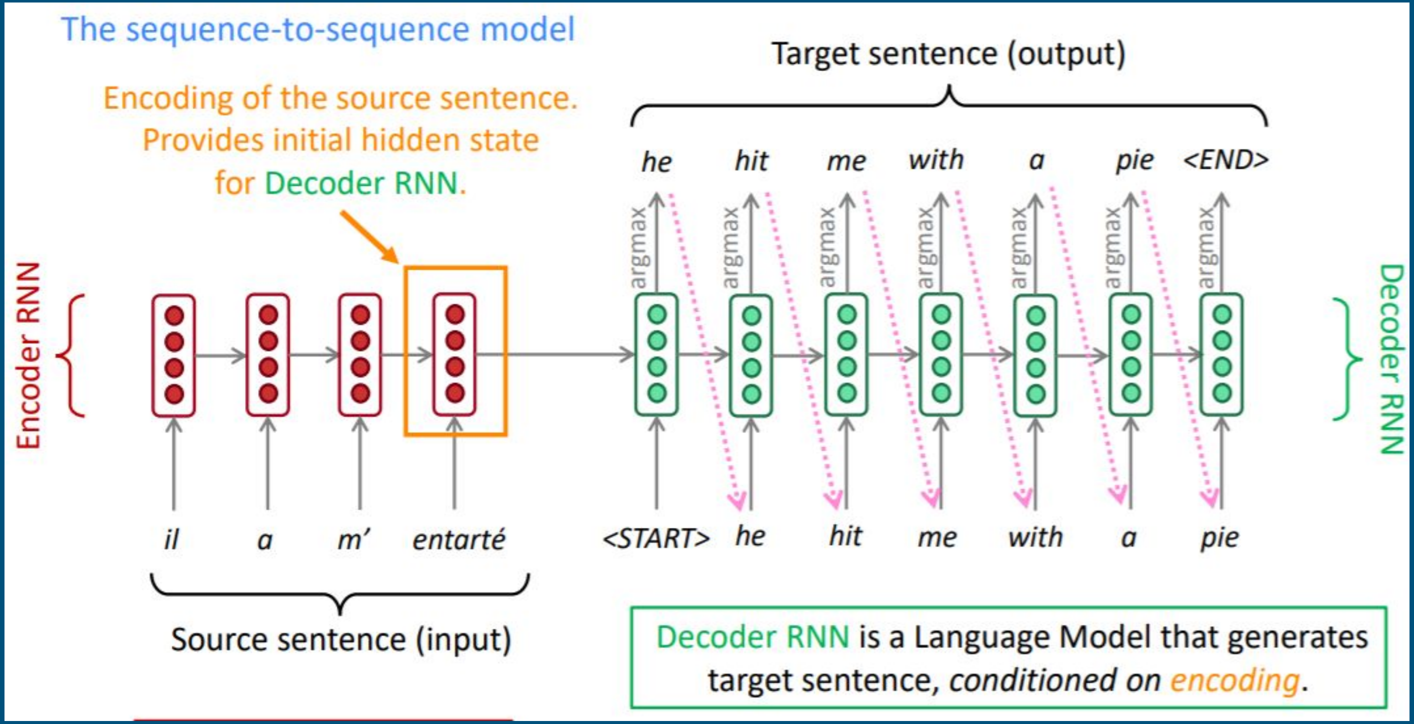
- A single-layer bidirectional LSTM
- The tokens of the source are fed **one-by-one** into the encoder
- Produce a sequence of encoder hidden states

Decoder - Test time

- A single-layer unidirectional LSTM
- Source encoding as initial hidden states
- The word embedding of the previous word is fed to the decoder
  (Test time: Previous word emitted by the decoder)
- **Conditional Language Model:** Language Model conditioned on the source encoding.

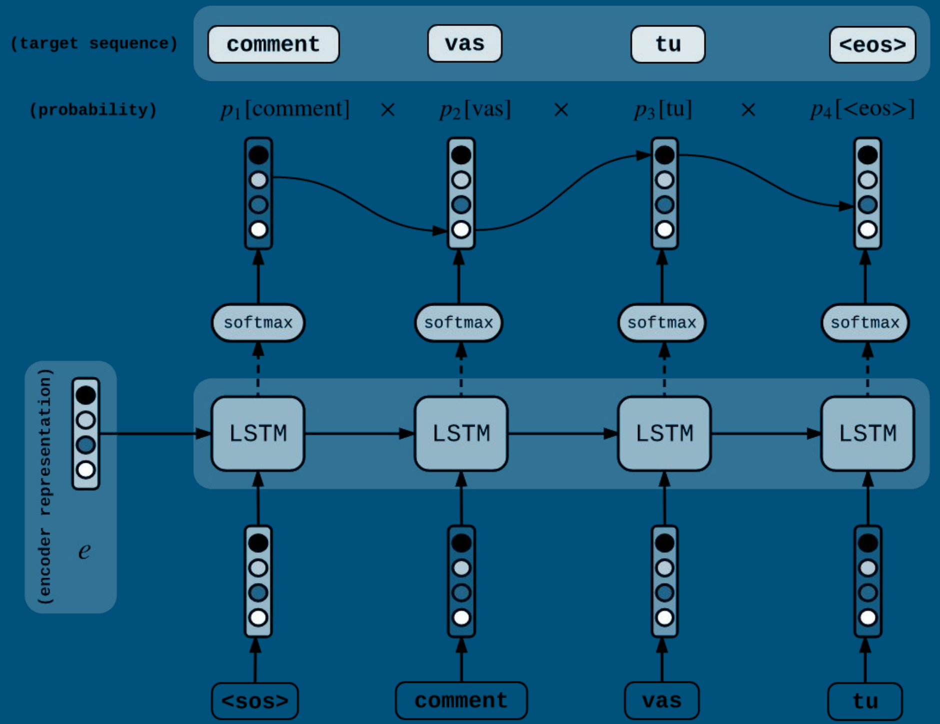


seq2seq - Test time

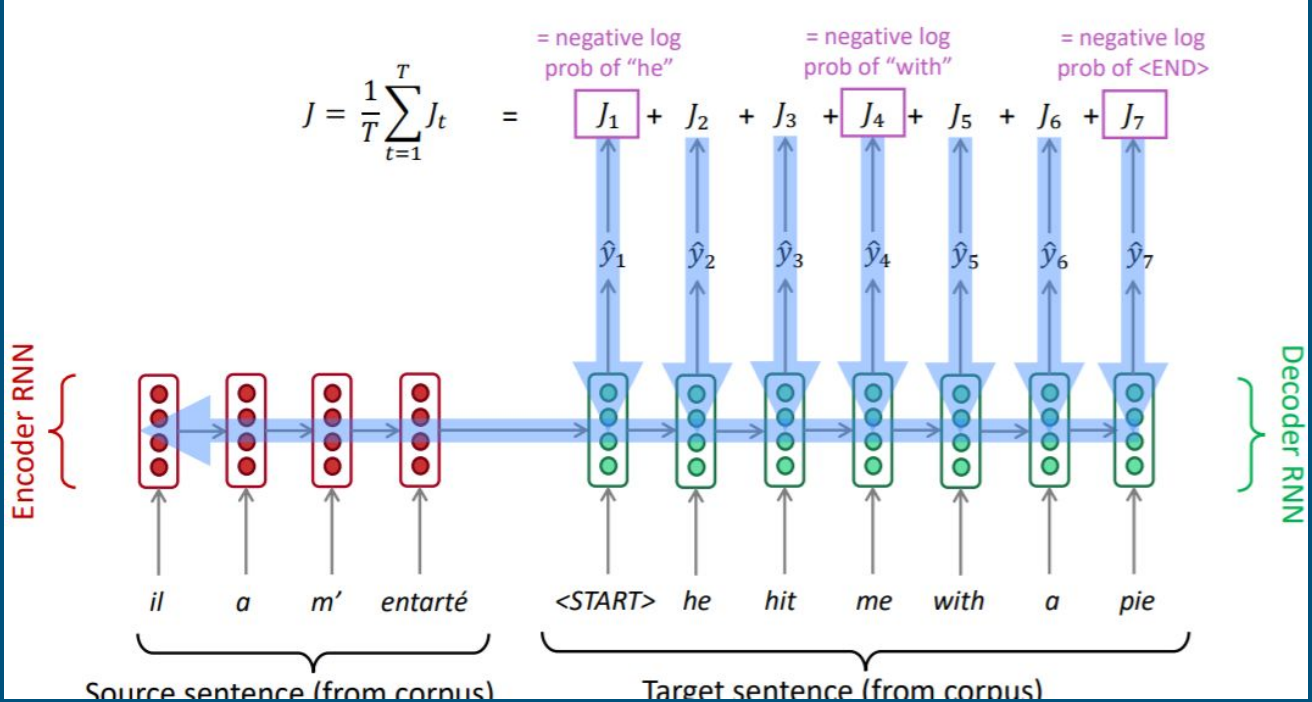


Decoder - Training time

- The word embedding of the previous word is fed to the decoder
  (Train time: Previous word of the reference target)
- Probability distribution is generated by softmax
- Loss is average negative log likelihood

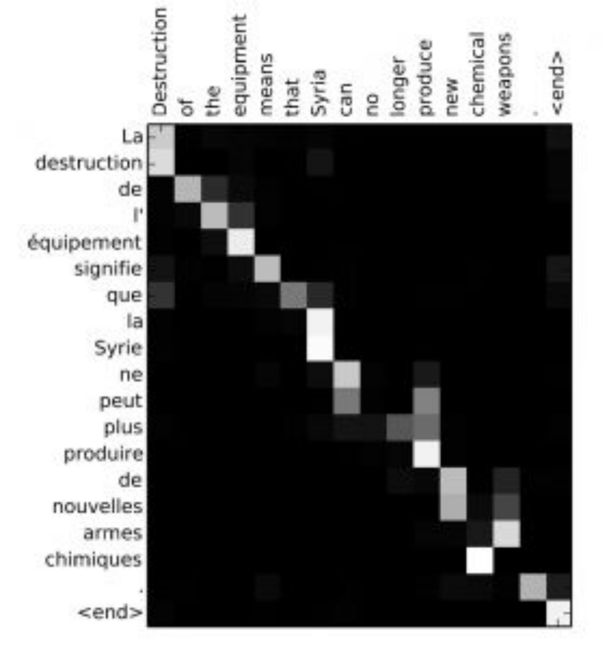


seq2seq - Training time

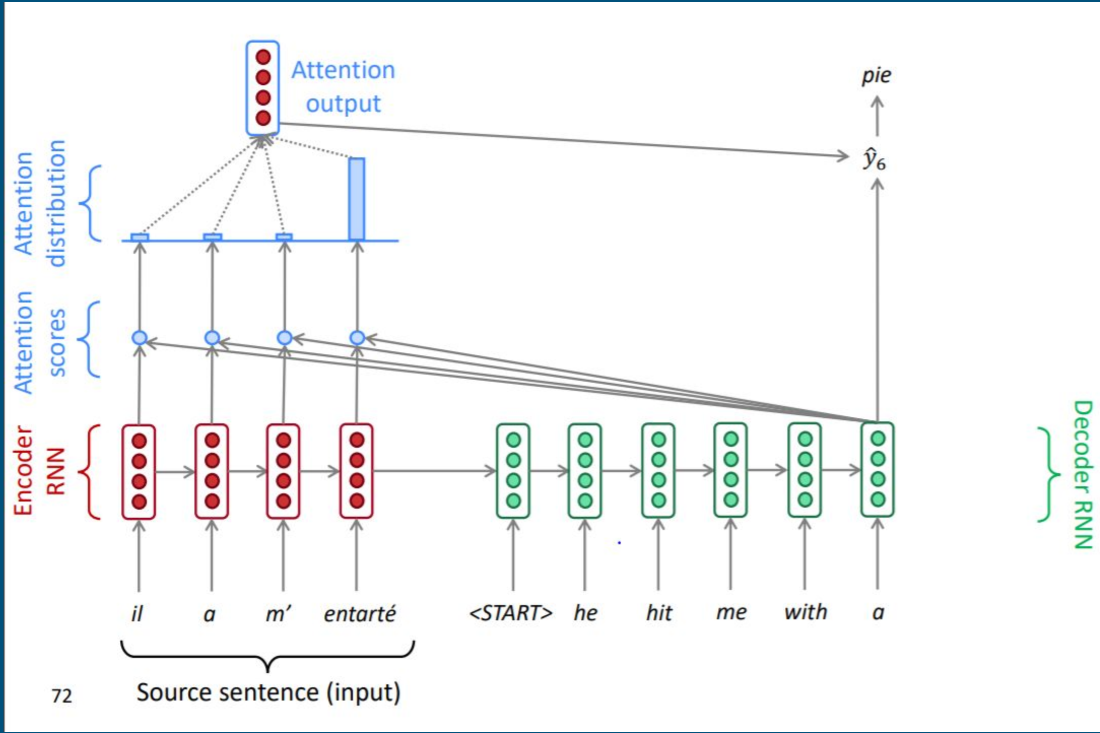


Attention

- Issue with seq2seq: bottleneck problem
- Motivation: Try to get information from all encoder hidden states instead of only the last one.
- Goal: Attention plays the role of the **weight** when we try to interpolate all encoder hidden states information (how much do we attend to the words?)



seq2seq + attention



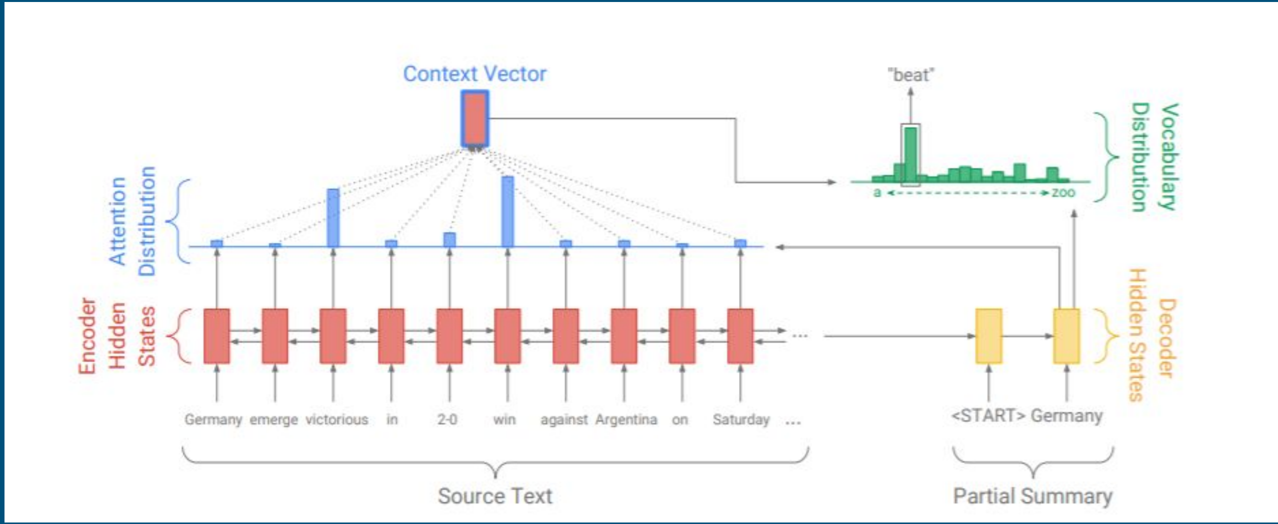
Attention distribution should capture the dependency between the current decoder hidden states and the encoder hidden states

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

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

Decoder state: $[s_t, h_t^*]$

Baseline model

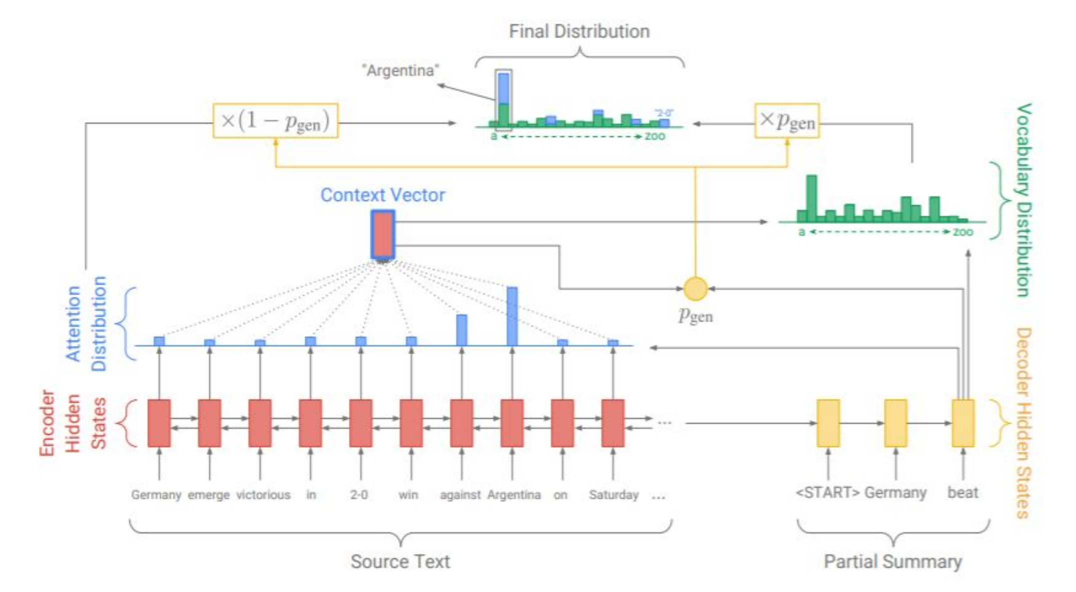


Pointer-network

- Issue with baseline: OOV and produce wrong factual details
- Motivation: Sometimes it is adequate to just copy the word from the source.
- Goal: Derive the probability p_gen to generate a word from vocab, otherwise copy the word from the source
- p_gen is a “soft switch” between pointer/generator.

Model: pointer-generator

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



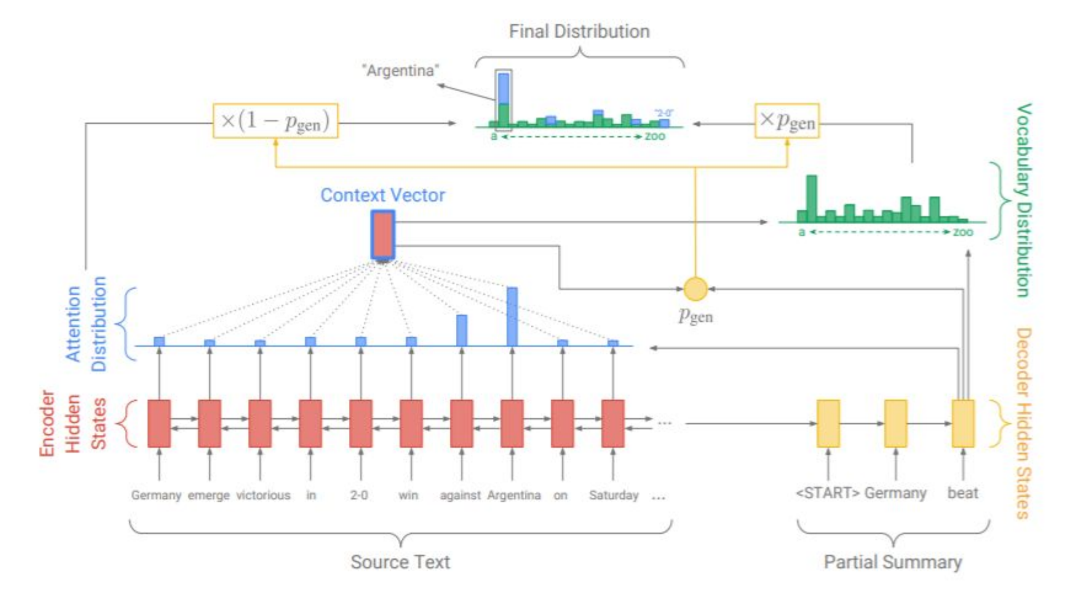
p_gen related to the decoder state and decoder input

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

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*



Extended vocab: we need to include the OOV in the source (in case of copying)

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

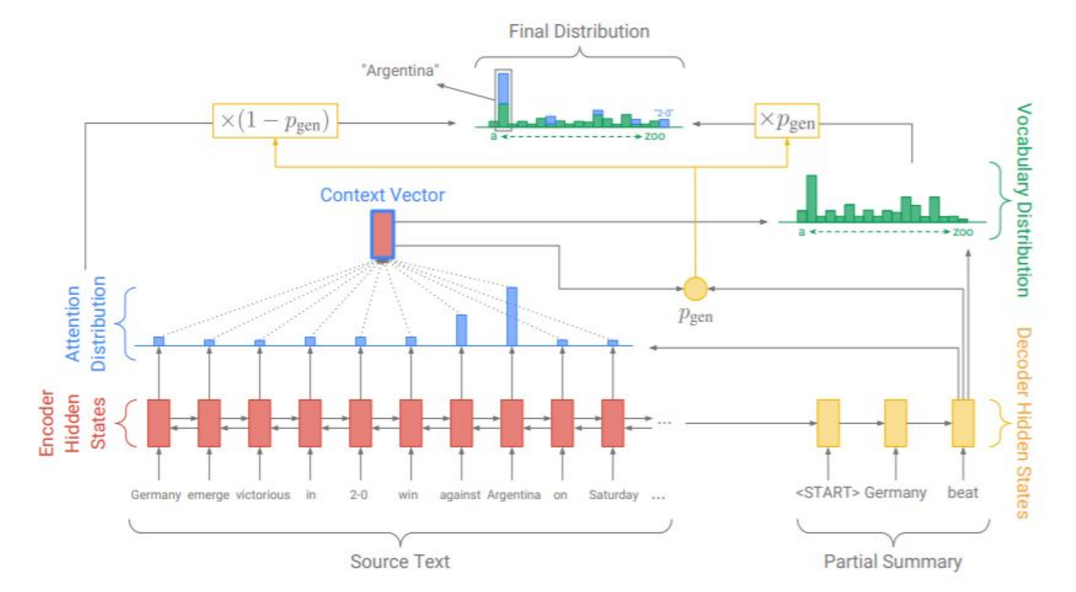
This picture is taken from ***Get To The Point: Summarization with Pointer-Generator Networks***

Coverage mechanism

- Issue with vanilla pointer-generator: repetition
- Motivation: Keep track of what has been summarized and penalize it
- Goal: Derive a vector c_t (coverage vector) that capture the information about what has been summarized.
- Idea
  - Attention tells us which part to attend
  - Coverage tells us which part was attended

Model: pointer-generator with coverage

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



Coverage vector is the sum of attention distributions over all previous decoder timesteps:

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

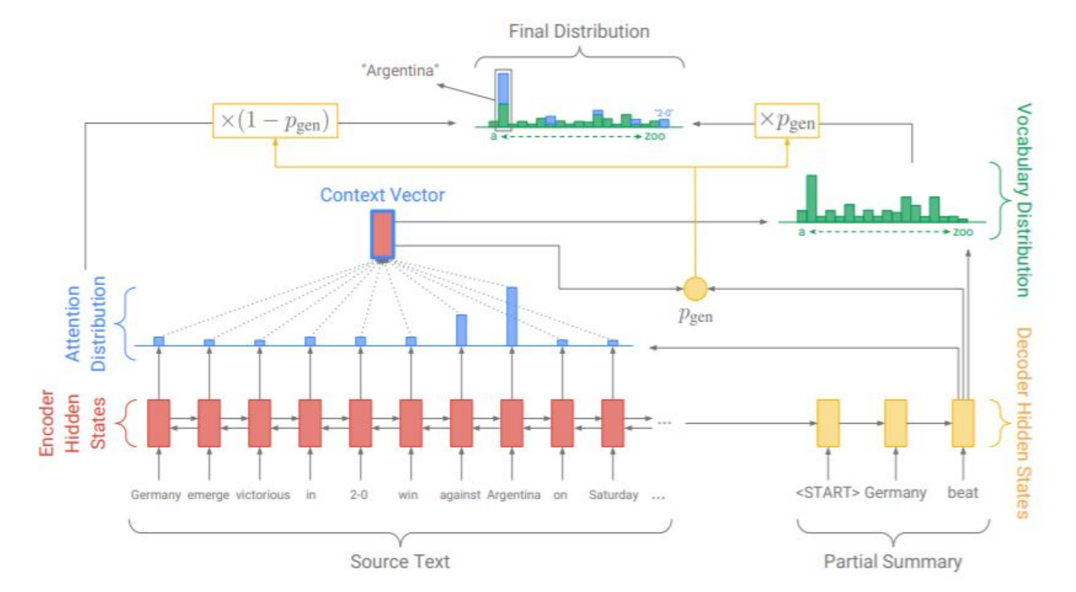
Modified attention (tell the attention its history)

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

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

Model: pointer-generator with coverage

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



$$\text{covloss}_t = \sum_i \min(a_i^t, c_i^t)$$

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

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

Part 3
Dataset & Experiment

Dataset

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

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.

Non-anonymized version of the data

Experiment

- Hidden state and word embedding: 256- dimensional hidden states and 128-dimensional word embeddings
- Vocabulary: 50k words for both source and target (baseline model: 150k source and 60k target)
- Number of hyperparameters: for the models with vocabulary size 50k, the baseline model has 21,499,600 parameters, the pointer-generator adds 1153 extra parameters (not much!)
- 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.
- Coverage loss weight (lambda): 1

Experiment

- Baseline models: ~600,000 iterations (33 epochs), 4 days and 14 hours (50k vocabulary model), 8 days 21 hours (150k vocabulary model)
- Pointer-generator model: 230,000 training iterations (12.8 epochs); 3 days and 4 hour
- Coverage: 2 hours

Part 4
Results

Model input - Coverage

Article (truncated): munster have signed new zealand international francis *saili* on a two-year deal . utility back *saili* , who made his all blacks debut against argentina in 2013 , will move to the province later this year after the completion of his 2015 contractual commitments . the 24-year-old currently plays for *auckland-based* super rugby side the blues and was part of the new zealand under-20 side that won the junior world championship in italy in 2011 . *saili* 's signature is something of a coup for munster and head coach anthony foley believes he will be a great addition to their backline . francis *saili* has signed a two-year deal to join munster and will link up with them later this year . ' we are really pleased that francis has committed his future to the province , ' foley told munster 's official website . ' he is a talented centre with an impressive *skill-set* and he possesses the physical attributes to excel in the northern hemisphere . ' i believe he will be a great addition to our backline and we look forward to welcoming him to munster . ' *saili* has been capped twice by new zealand and was part of the under 20 side that won the junior championship in 2011 .

Yellow shading indicates coverage vector value

Model output - Baseline

- Abstractive baseline model produces factual inaccuracies and cannot deal with OOV

Reference Summary:
utility back francis *saili* will join up with munster later this year .
the new zealand international has signed a two-year contract .
saili made his debut for the all blacks against argentina in 2013 .

Baseline: New zealand Fabricated
dutch international francis UNK has signed a two-year deal to join **irish** UNK super rugby side the blues .
UNK 's signature is something of a coup for munster and his head coach anthony foley believes he will be a great addition to their **respective prospects** .
UNK has been capped twice by new zealand .

Model output - Pointer Generator

- Pointer ensures OOV word *saili* is captured

Pointer-Generator, No Coverage:
new zealand international francis *saili* will move to the province later this year . utility back *saili* made his all blacks debut against argentina in 2013 . utility back *saili* will move to the province later this year .

Pointer-Generator, With Coverage:
francis *saili* has signed a two-year deal to join munster later this year . the 24-year-old was part of the new zealand under-20 side that won the junior world championship in italy in 2011 . *saili* 's signature is something of a coup for munster and head coach anthony foley .

Green shading indicates generation probability.

Evaluation metrics

- ROUGE
  - ROUGE-1: word overlap
  - ROUGE-2: bi-gram overlap
  - ROUGE-L: longest common sequence
- METEOR
  - Match mode: only reward exact matches
  - Full mode: reward matching stems, synonyms and paraphrases

Model comparison

- Lead-3: use first three sentences as summary
- Abstractive: seq2seq + attention model
  - Nallapati et al. 2016: "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond"
- Extractive: GRU-RNN model
  - Nallapati et al. 2017: "A recurrent neural network based sequence model for extractive summarization of documents"

Results

- Extractive models generally perform the best
- Pointer + Coverage performs best for abstractive models

Abstractive

	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	-	-

Extractive

Part 5
Discussion

Discussion - Evaluation metric

- Evaluation metrics generally favor extractive models
  - Lead-3 very strong for newspaper articles
  - Most critical information summarized in the start of the article
- ROUGE and METEOR originate from machine translation
- Humans use domain knowledge when performing summarization
  - “Alice loves oranges, bananas and kiwi” “Alice loves tropical fruits”
- Hard to design good evaluation metric for abstraction
  - Use multiple summaries with different wording
  - Take into account paraphrasing, synonyms, stems etc.

Discussion - Evaluation metric

Target summary: "*Manchester United beat Aston Villa at Old Trafford. Wayne Rooney topscorer in the Premier League*"

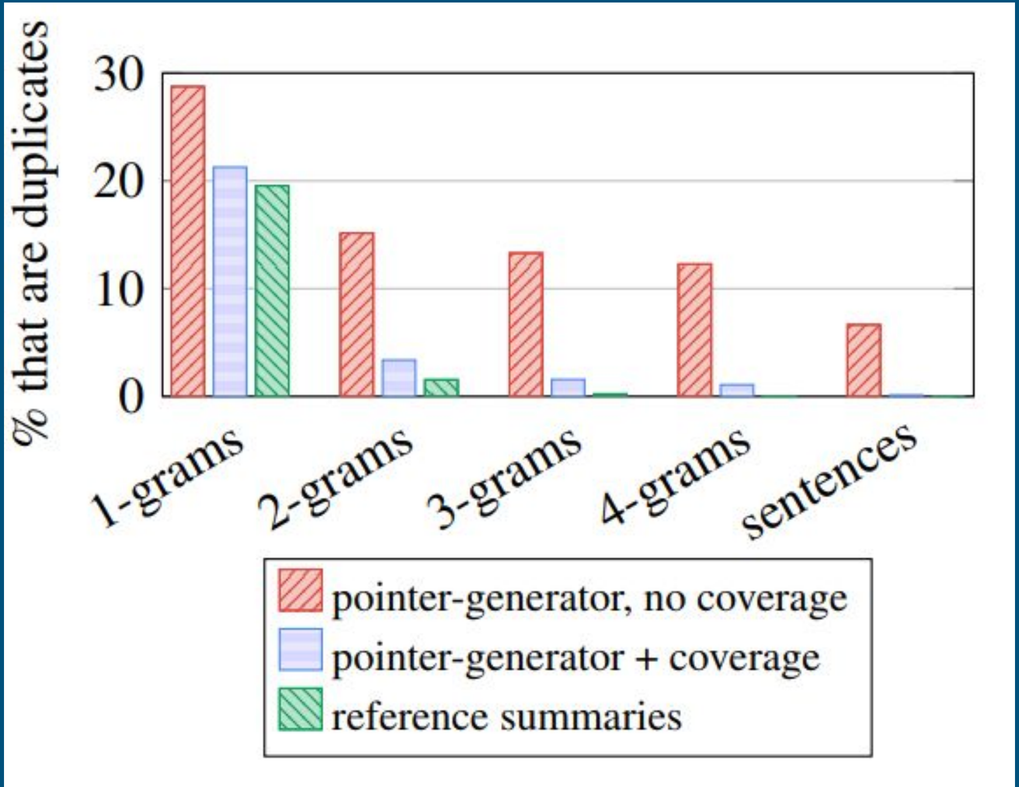
Abstractive summary: "*Wayne Rooney becomes topscorer as Manchester United defeat Aston Villa at home.*"

- Punished by use of novel word "defeat" - can be solved with synonyms
- Punished by use of word "home" instead of "Old Trafford" - very hard to solve (requires human level domain knowledge within football)

Discussion - Does coverage reduce repetition?

- Coverage does not entirely eliminate repetition, but drastically reduces it

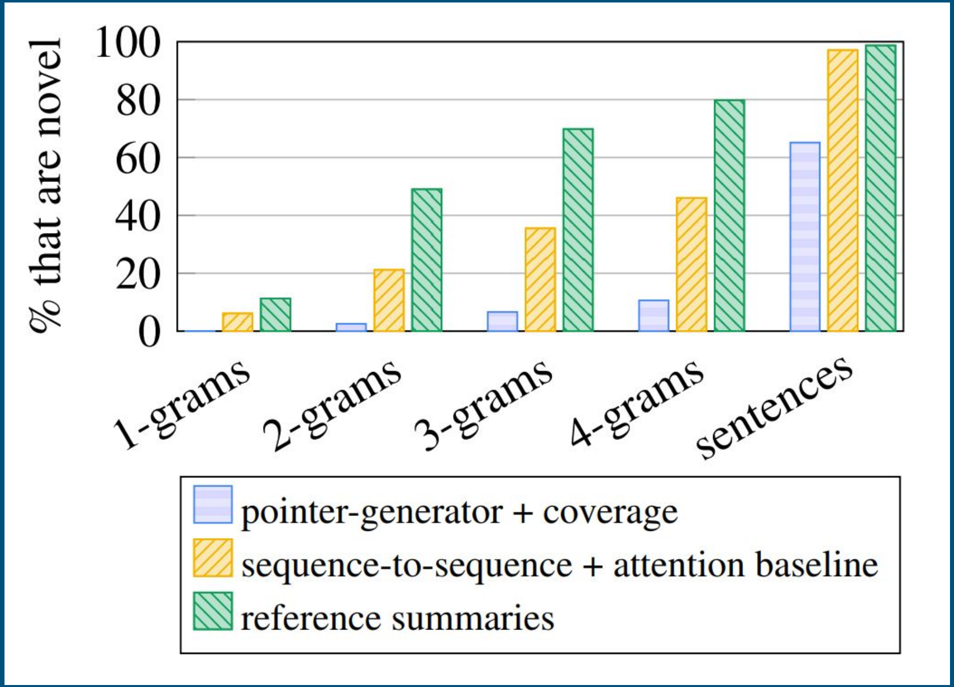
- Model with coverage has almost same amount of duplicates as reference summaries



Discussion - How abstractive is the model?

- Reference summaries produce novel sentences 99% of the time
- Pointer-Generator produces novel sentences 65% of the time
- Pointer-Generator less abstractive than baseline, but produces fewer inaccuracies

Percentage of novel n-grams (i.e. not in source text)



Contributions

- Replicated the seq2seq + attention model (Nallapati et al. 2016) to generate novel words
- Applied pointer-network (Vinyals et al. 2015) to abstractive text summarization
- Developed coverage loss to reduce repetition
- Beat current state-of-the-art models within abstractive text summarization by at least 2 ROUGE points
- Encouraged the need for a fair evaluation metric for abstractive text summarization