A Neural Attention Model for Sentence Summarization

Alexander M. Rush, Sumit Chopra, Jason Weston. EMNLP 2015

Presented by Peiyao Li, Spring 2020

Extractive vs. Abstractive Summarization

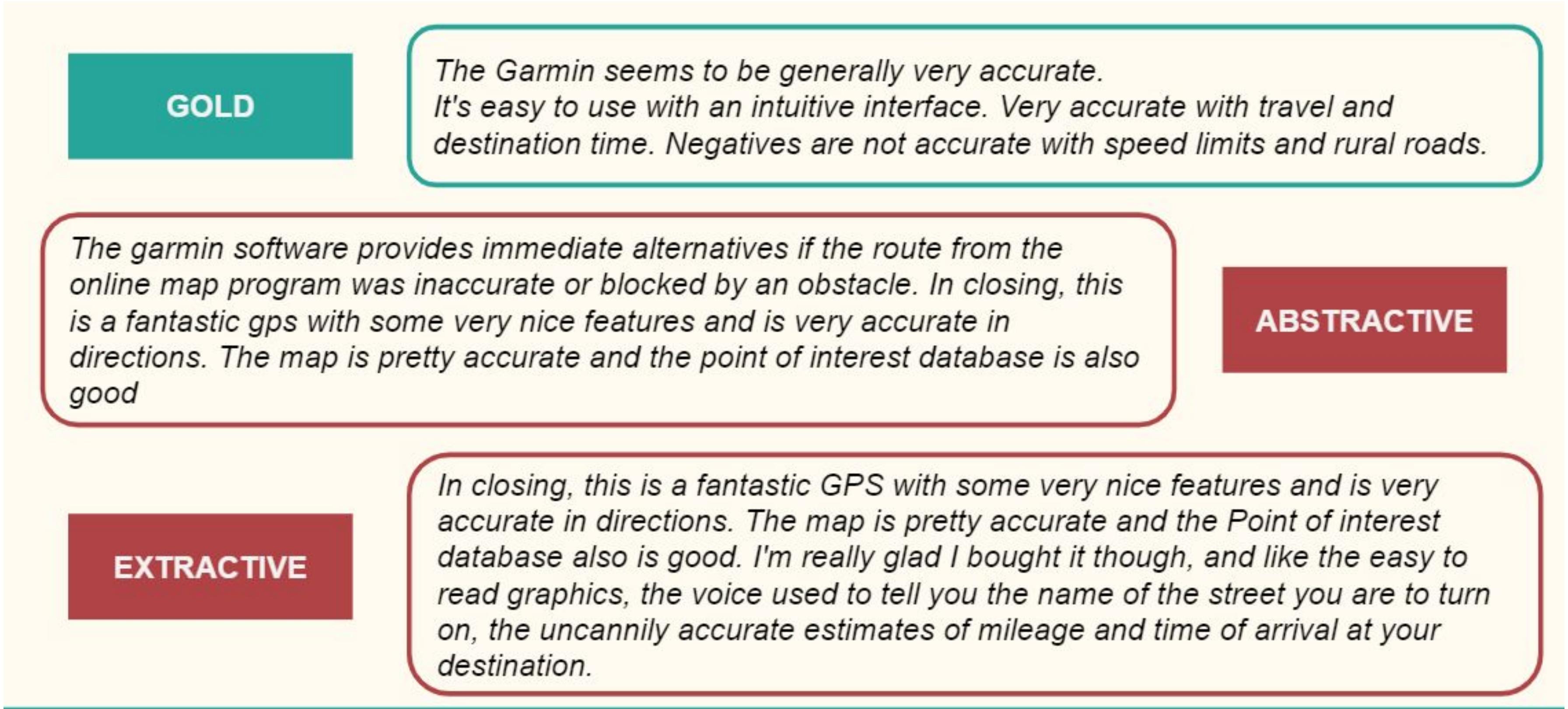
Extractive Summarization:

- Extracts words and phrases from original text
- Easy to implement
- Unsupervised -> fast

Abstractive Summarization:

- Learns internal language representation, paraphrase original text
- Sounds more human-like
- Needs lots of data and time to train

Extractive vs. Abstractive Summarization



GOLD

The Garmin seems to be generally very accurate.
It's easy to use with an intuitive interface. Very accurate with travel and destination time. Negatives are not accurate with speed limits and rural roads.

The garmin software provides immediate alternatives if the route from the online map program was inaccurate or blocked by an obstacle. In closing, this is a fantastic gps with some very nice features and is very accurate in directions. The map is pretty accurate and the point of interest database is also good

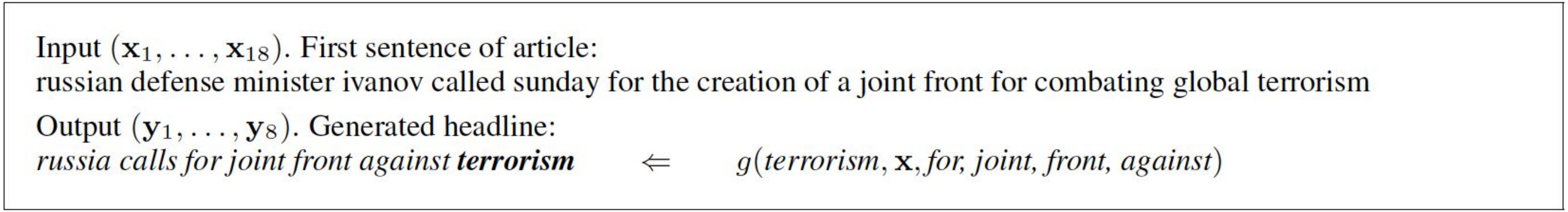
ABSTRACTIVE

EXTRACTIVE

In closing, this is a fantastic GPS with some very nice features and is very accurate in directions. The map is pretty accurate and the Point of interest database also is good. I'm really glad I bought it though, and like the easy to read graphics, the voice used to tell you the name of the street you are to turn on, the uncannily accurate estimates of mileage and time of arrival at your destination.

Problem Statement

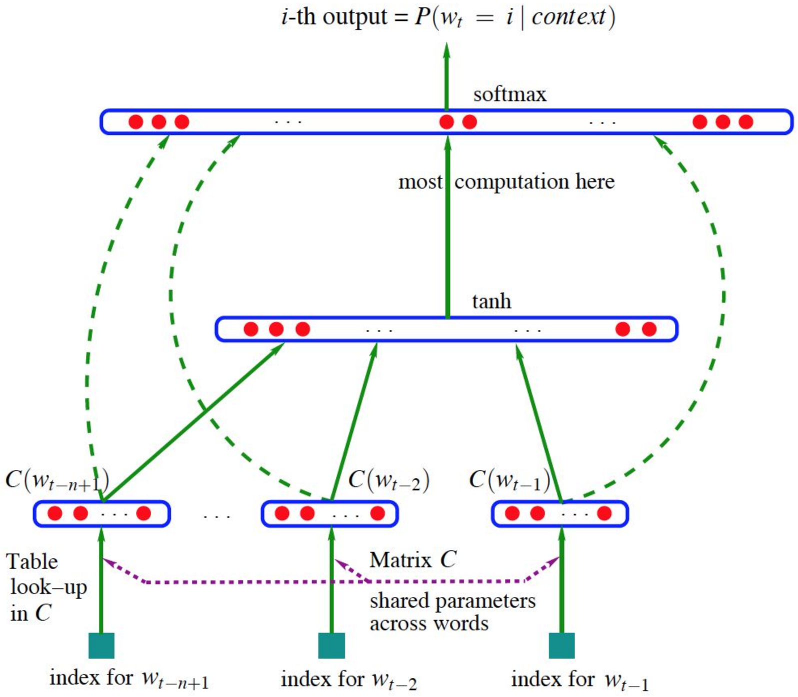
- Sentence-level abstractive summarization
- Input: a sequence of M words x = $[x_1,\ldots,x_m]$
- Output: a sequence of N words y = $[y_1,\ldots,y_n]$ where N < M
- Proposed model: a language model for estimating the contextual probability of the next word



Input $(\mathbf{x}_1, \ldots, \mathbf{x}_{18})$. First sentence of article:
russian defense minister ivanov called sunday for the creation of a joint front for combating global terrorism

Output $(\mathbf{y}_1, \ldots, \mathbf{y}_8)$. Generated headline:
russia calls for joint front against ***terrorism*** $\Leftarrow$ $g(\textit{terrorism}, \mathbf{x}, \textit{for, joint, front, against})$

Neural N-gram Language Model: Recap



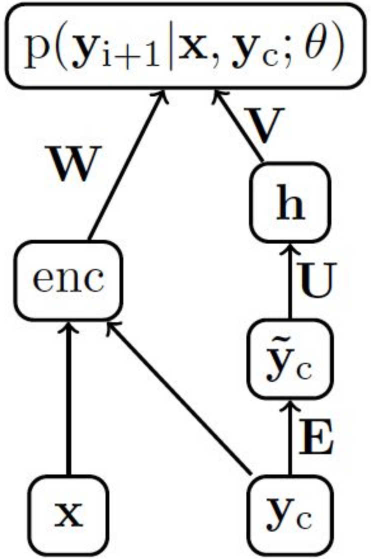
Bengio et al., 2003

Proposed Model

- Models local conditional probability of the next word in the summary given input sentence x and the context of the summary y_c

$$
\begin{aligned}
p(\mathbf{y}_{i+1}|\mathbf{y}_c, \mathbf{x}; \theta) &\propto \exp(\mathbf{V}\mathbf{h} + \mathbf{W}\mathrm{enc}(\mathbf{x}, \mathbf{y}_c)), \\
\tilde{\mathbf{y}}_c &= [\mathbf{E}\mathbf{y}_{i-C+1}, \ldots, \mathbf{E}\mathbf{y}_i], \\
\mathbf{h} &= \tanh(\mathbf{U}\tilde{\mathbf{y}}_c).
\end{aligned}
$$

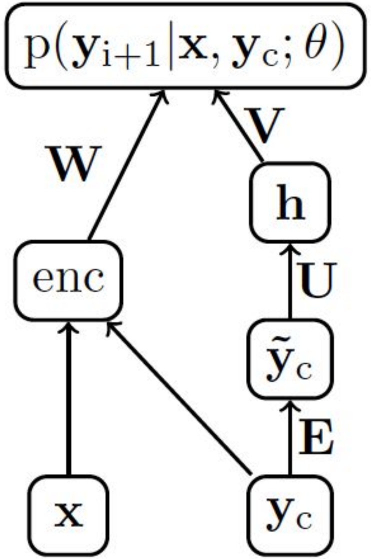
$$\mathbf{y}_c \triangleq \mathbf{y}_{[i-C+1,\ldots,i]}$$



* Bias terms were ignored for readability

Encoders

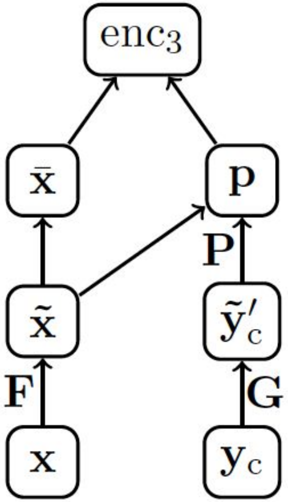
- Tried three different encoders:
  - Bag-of-words encoder
    - Word at each input position has the same weight
    - Orders and relationship b/t neighboring words are ignored
    - Context y_c is ignored
    - Single representation for the entire input
  - Convolutional encoder
    - Allows local interactions between input words
    - Context y_c is ignored
    - Single representation for the entire input
  - Attention-based encoder



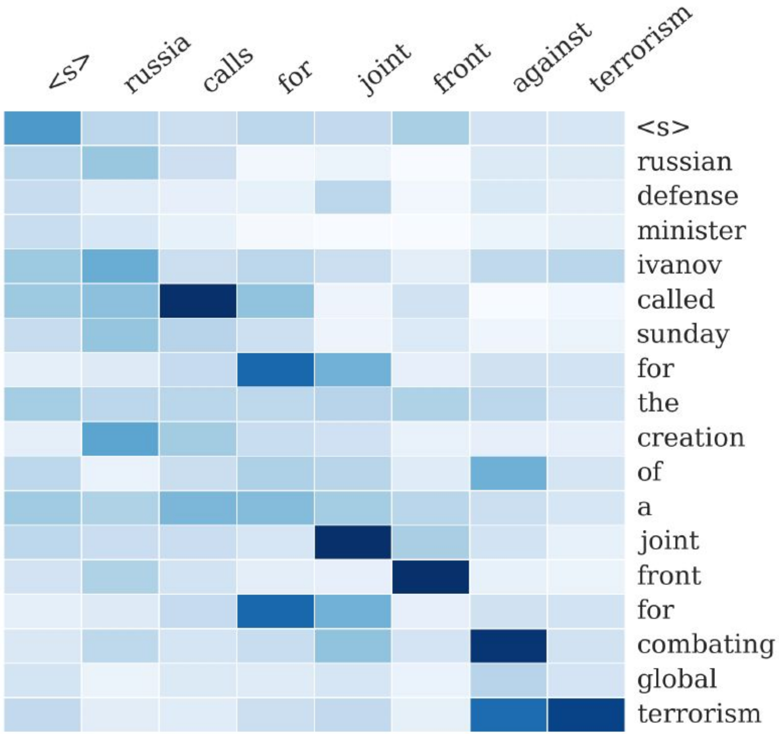
Attention-Based Encoder

- Soft alignment for input x and context of summary y_c

$$
\begin{aligned}
\text{enc}_3(\mathbf{x}, \mathbf{y}_{\text{c}}) &= \mathbf{p}^\top \bar{\mathbf{x}}, \\
\mathbf{p} &\propto \exp(\tilde{\mathbf{x}} \mathbf{P} \tilde{\mathbf{y}}'_{\text{c}}), \\
\tilde{\mathbf{x}} &= [\mathbf{F}\mathbf{x}_1, \ldots, \mathbf{F}\mathbf{x}_M], \\
\tilde{\mathbf{y}}'_{\text{c}} &= [\mathbf{G}\mathbf{y}_{i-C+1}, \ldots, \mathbf{G}\mathbf{y}_i], \\
\forall i \quad \bar{\mathbf{x}}_i &= \sum_{q=i-Q}^{i+Q} \tilde{\mathbf{x}}_i / Q.
\end{aligned}
$$



Attention-Based Encoder



Training

- Can train on arbitrary input-summary pairs
- Minimize negative log-likelihood using mini-batch stochastic gradient descent

$$
\begin{aligned}
\mathrm{NLL}(\theta) &= -\sum_{j=1}^{J} \log p(\mathbf{y}^{(j)} | \mathbf{x}^{(j)}; \theta), \\
&= -\sum_{j=1}^{J} \sum_{i=1}^{N-1} \log p(\mathbf{y}_{i+1}^{(j)} | \mathbf{x}^{(j)}, \mathbf{y}_{\mathrm{c}}; \theta).
\end{aligned}
$$

* J = # of input-summary pairs

Generating Summary

- Exact: Viterbi
  - $O(NV^C)$
- Strictly greedy: argmax
  - $O(NV)$
- Compromise: Beam-search
  - $O(KNV)$ with beam size K

Extractive Tuning

- Abstractive model cannot find extractive word matches when necessary
  - e.g. unseen proper noun phrases in input
- Tuning additional features that trade-off the abstractive/extractive tendency

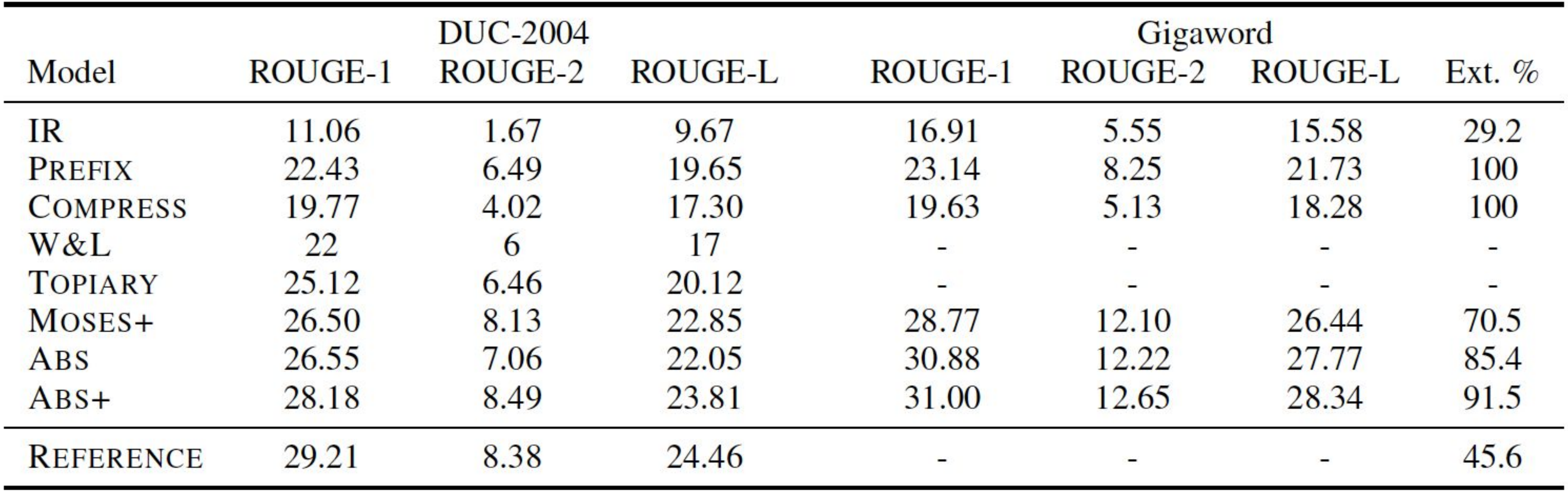
$$
\begin{aligned}
f(\mathbf{y}_{i+1}, \mathbf{x}, \mathbf{y}_{\mathrm{c}}) = [\, & \log p(\mathbf{y}_{i+1} | \mathbf{x}, \mathbf{y}_{\mathrm{c}}; \theta), \\
& \mathbb{1}\{\exists j.\ \mathbf{y}_{i+1} = \mathbf{x}_j\ \}, \\
& \mathbb{1}\{\exists j.\ \mathbf{y}_{i+1-k} = \mathbf{x}_{j-k}\ \forall k \in \{0, 1\}\}, \\
& \mathbb{1}\{\exists j.\ \mathbf{y}_{i+1-k} = \mathbf{x}_{j-k}\ \forall k \in \{0, 1, 2\}\}, \\
& \mathbb{1}\{\exists k > j.\ \mathbf{y}_i = \mathbf{x}_k, \mathbf{y}_{i+1} = \mathbf{x}_j\}\,].
\end{aligned}
$$

$$
s(\mathbf{y}, \mathbf{x}) \quad = \quad \sum_{i=0}^{N-1} \alpha^{\top} f(\mathbf{y}_{i+1}, \mathbf{x}, \mathbf{y}_{\mathrm{c}}).
$$

Dataset

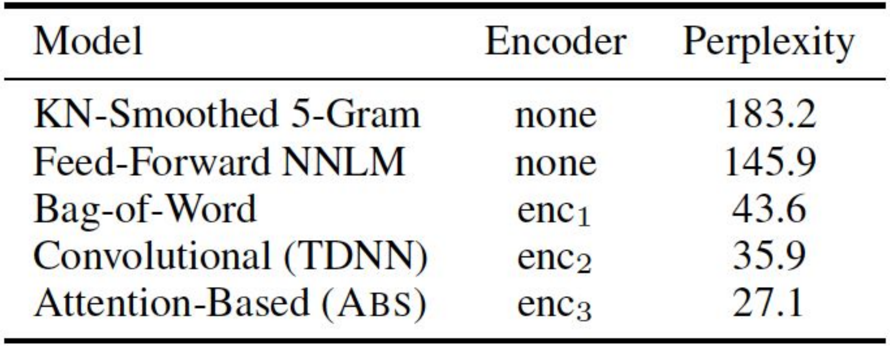
- DUC-2014
  - 500 news articles with human-generated reference summaries
- Gigaword
  - Pair the headline of each article with the first sentence to create input-summary pair
  - 4 million pairs
- Evaluated using ROUGE-1, ROUGE-2, ROUGE-L

Results

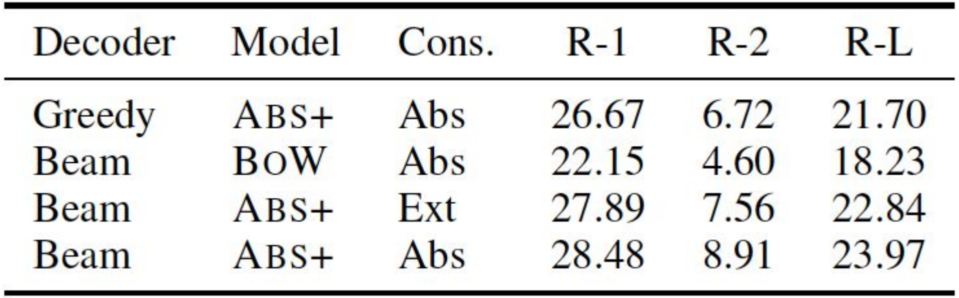


	DUC-2004			Gigaword			
Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L	Ext. %
IR	11.06	1.67	9.67	16.91	5.55	15.58	29.2
Prefix	22.43	6.49	19.65	23.14	8.25	21.73	100
Compress	19.77	4.02	17.30	19.63	5.13	18.28	100
W&L	22	6	17	-	-	-	-
Topiary	25.12	6.46	20.12	-	-	-	-
Moses+	26.50	8.13	22.85	28.77	12.10	26.44	70.5
Abs	26.55	7.06	22.05	30.88	12.22	27.77	85.4
Abs+	28.18	8.49	23.81	31.00	12.65	28.34	91.5
Reference	29.21	8.38	24.46	-	-	-	45.6

Results



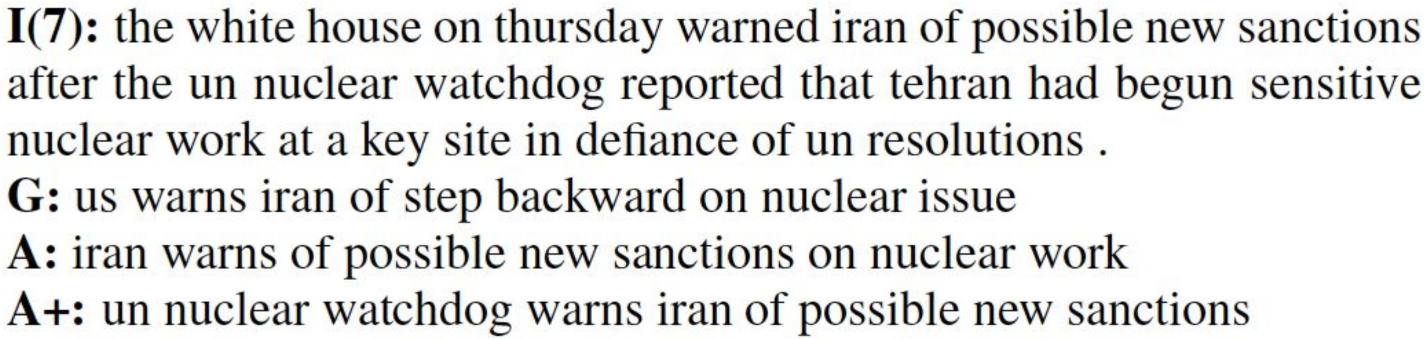
Model	Encoder	Perplexity
KN-Smoothed 5-Gram	none	183.2
Feed-Forward NNLM	none	145.9
Bag-of-Word	enc$_1$	43.6
Convolutional (TDNN)	enc$_2$	35.9
Attention-Based (ABS)	enc$_3$	27.1



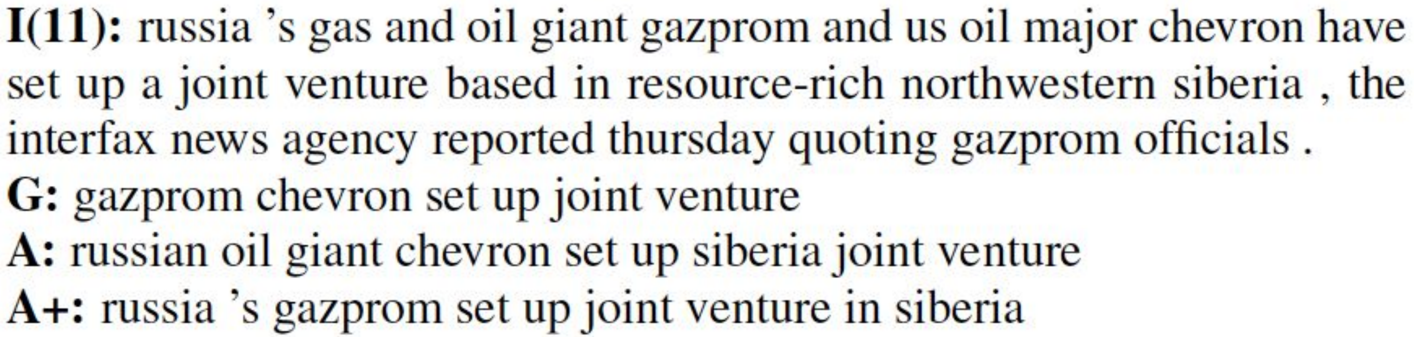
Decoder	Model	Cons.	R-1	R-2	R-L
Greedy	ABS+	Abs	26.67	6.72	21.70
Beam	BoW	Abs	22.15	4.60	18.23
Beam	ABS+	Ext	27.89	7.56	22.84
Beam	ABS+	Abs	28.48	8.91	23.97

Analysis

- Standard feed-forward NNLM: size of context is fixed (n-gram)
- Length of summary has to be determined before generation
- Only sentence-level summaries can be generated
- Syntax/factual details of summary might not be correct



I(7): the white house on thursday warned iran of possible new sanctions after the un nuclear watchdog reported that tehran had begun sensitive nuclear work at a key site in defiance of un resolutions .
G: us warns iran of step backward on nuclear issue
A: iran warns of possible new sanctions on nuclear work
A+: un nuclear watchdog warns iran of possible new sanctions



I(11): russia 's gas and oil giant gazprom and us oil major chevron have set up a joint venture based in resource-rich northwestern siberia , the interfax news agency reported thursday quoting gazprom officials .
G: gazprom chevron set up joint venture
A: russian oil giant chevron set up siberia joint venture
A+: russia 's gazprom set up joint venture in siberia

Examples of incorrect summary

Citations

- Alexander M. Rush, Sumit Chopra, and Jason Weston, *A neural attention model for abstractive sentence summarization*, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (Lisbon, Portugal), Association for Computational Linguistics, September 2015, pp. 379–389.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. *A neural probabilistic language model*. J. Mach. Learn. Res. 3, null (March 2003), 1137–1155.
- Text Summarization in Python: Extractive vs. Abstractive techniques revisited
- Data Scientist’s Guide to Summarization

Abstractive Text Summarization using Sequence-to-Sequence RNNs and Beyond

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, Bing Xiang

Presented by: Yunyi Zhang (yzhan238)

03.06.2020

Motivation

- **Abstractive** summarization task:
  - Generate a **compressed paraphrasing** of the main contents of a document
- The task is **similar** with machine translation:
  - mapping an input sequence of words in a document to a target sequence of words called summary
- The task is also **different** from machine translation:
  - the target is typically very short
  - optimally compress in a lossy manner such that key concepts are preserved

Model Overview

- Apply the off-the-shelf attentional encoder-decoder RNN to summarization
- Propose novel models to address the concrete problems in summarization
  - Capturing keywords using feature-rich encoder
  - Modeling rare/unseen words using switching generator-pointer
  - Capturing hierarchical document structure with hierarchical attention

Attentional Encoder-decoder with LVT

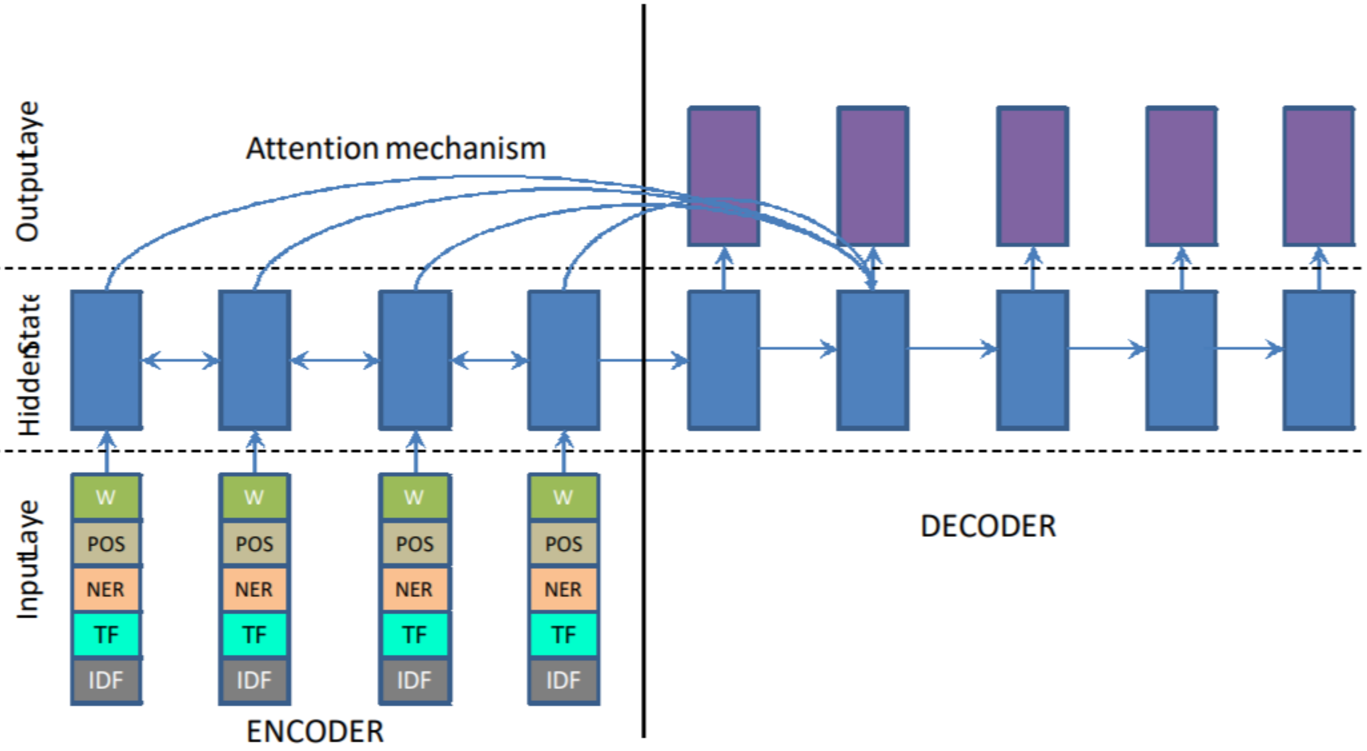
- Encoder: a bidirectional GRU
- Decoder:
  - A uni-directional GRU
  - An attention mechanism over source hidden states
  - A softmax layer over target vocabulary
- Large vocabulary trick (LVT)
  - Target vocab: source words in the batch + frequent words until a fixed size
  - Reduce size of softmax layer
  - Speed up convergence
  - Well suit for summarization

Feature-rich Encoder

- Key challenge: identify the key concepts and entities in source document
- Thus, go beyond word embeddings and add linguistic features:
  - Part-of-speech tags: syntactic category of words
    - E.g. noun, verb, adjective, etc.
  - Named entity tags: categories of named entities
    - E.g. person, organization, location, etc.
  - Discretized Term Frequency (TF)
  - Discretized Inverse Document Frequency (IDF)
    - To diminish weight of terms that appear too frequently, like stop words

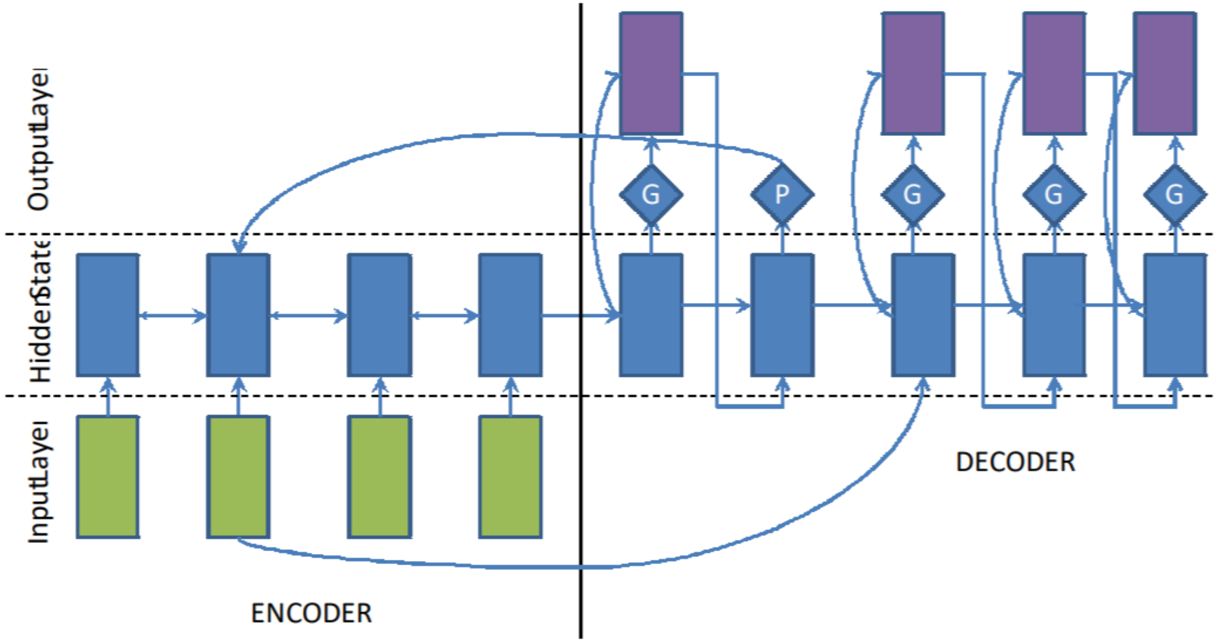
Feature-rich Encoder

- Concatenate with word-based embeddings as encoder input



Switching Generator-Pointer

- Keywords or named entities can be unseen or rare in training data
- Common solution: emit “UNK” token
  - Does not result in legible summaries
- Better solution:
  - A switch decides whether using generator or pointer at each step



Switching Generator-Pointer

- The switch is a sigmoid function over a linear layer based on the entire available context at each time step:

$$P(s_i = 1) = \sigma(v^s \cdot (W_h^s h_i + W_e^s E[o_{i-1}] + W_c^s c_i + b^s))$$

  - h_i: hidden state of decoder at step i
  - $E[o_{i-1}]$: embedding of previous emission
  - c_i: weighted context representation

- Pointer value is sampled using attention distribution over word positions in the document

$$P_i^a(j) \propto exp(v^a \cdot (W_h^a h_{i-1} + W_e^a E[o_{i-1}] + W_c^a h_j^d + b^a))$$

$$p_i = arg\max_j \left(P_i^a(j)\right) \text{ for } j \in \{1, \dots, N_d\}$$

  - h_j^d: hidden state of encoder at step j
  - N_d: number of words in source document

Switching Generator-Pointer

- Optimize the conditional log-likelihood:

$$logP(y|x) = \sum (g_i \log\{P(y_i|y_{-i}, x)P(s_i)\}$$

$$+(1 - g_i) \log\{P(p(i)|y_{-i}, x)(1 - P(s_i))\})$$

  - $g_i = 0$ when target word is OOV (switch off), otherwise $g_i = 1$

- At training time, provide the model with explicit pointer information whenever the summary word is OOV

- At test time, use $P(s_i)$ to automatically determine whether to generate or copy

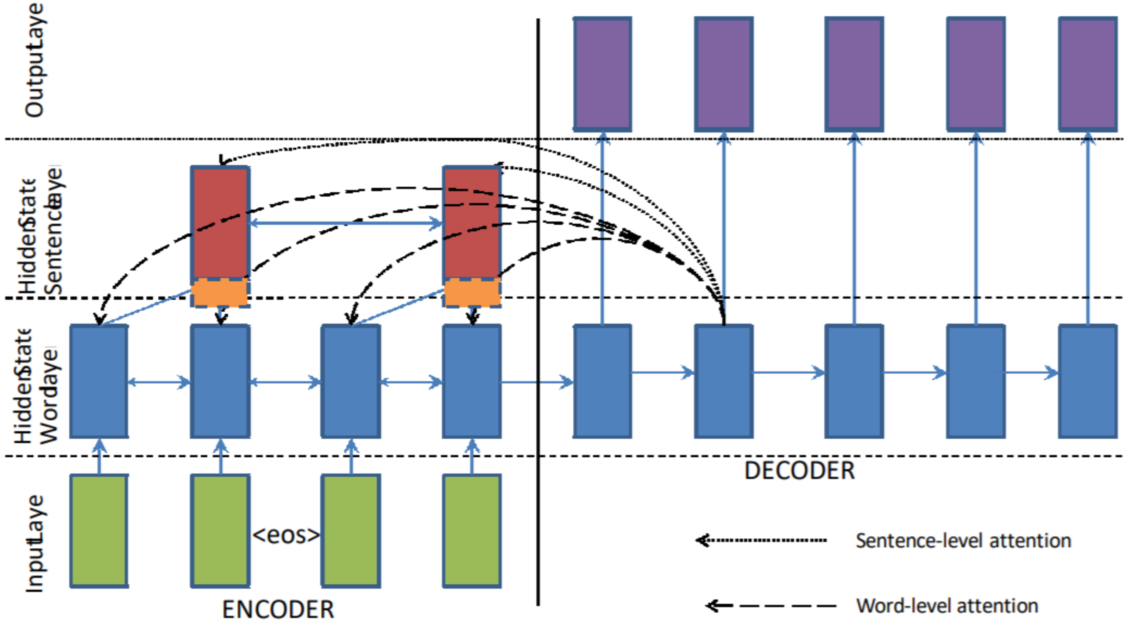
Hierarchical Attention

- Identify the key sentences from which the summary can drawn
- Re-weight and normalize word-level attention

$$P^a(j) = \frac{P_w^a(j)P_s^a(s(j))}{\sum_{k=1}^{N_d} P_w^a(k)P_s^a(s(k))}$$

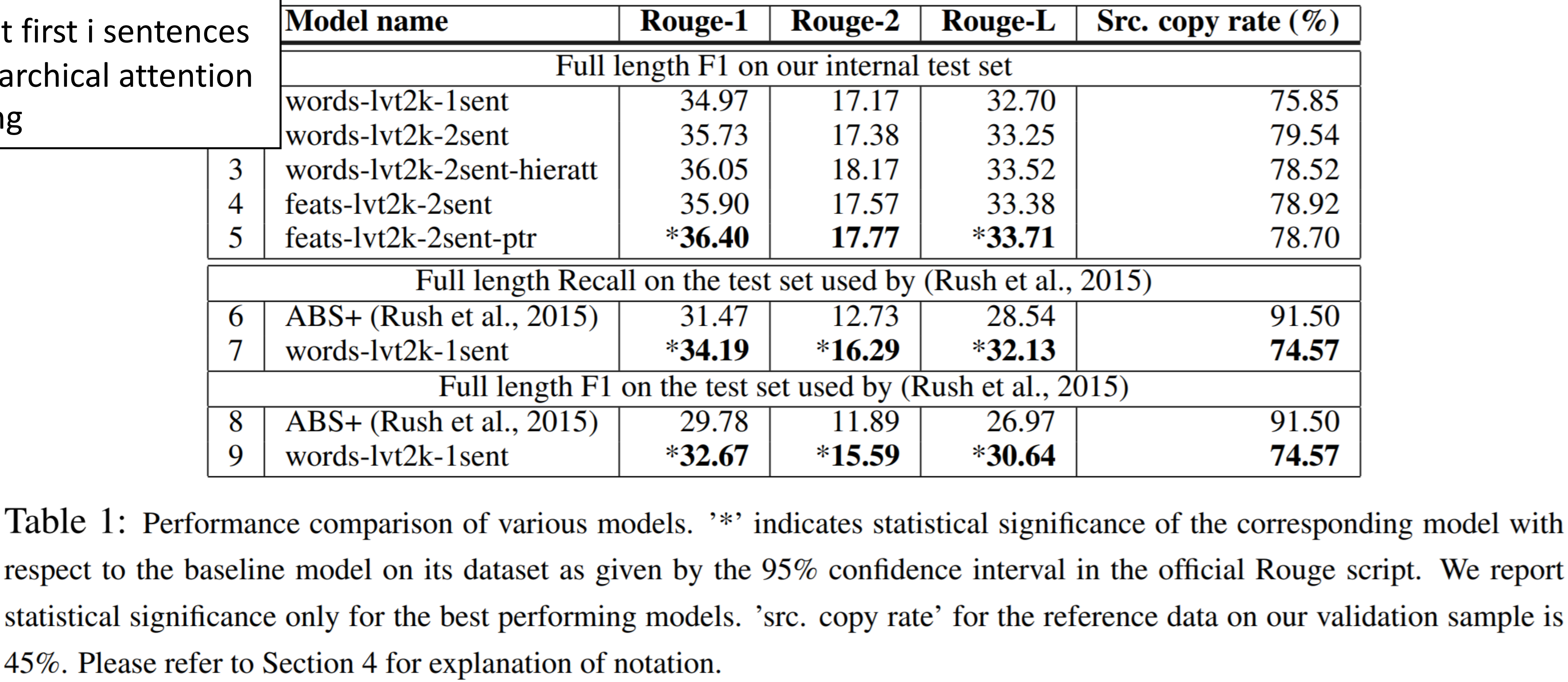
  - $P_w(P_s)$: word(sentence) attention weight
  - $s(l)$: sentence id of word l

- Concat positional embedding to the hidden state of sentence RNN



Experiment Results: Gigaword

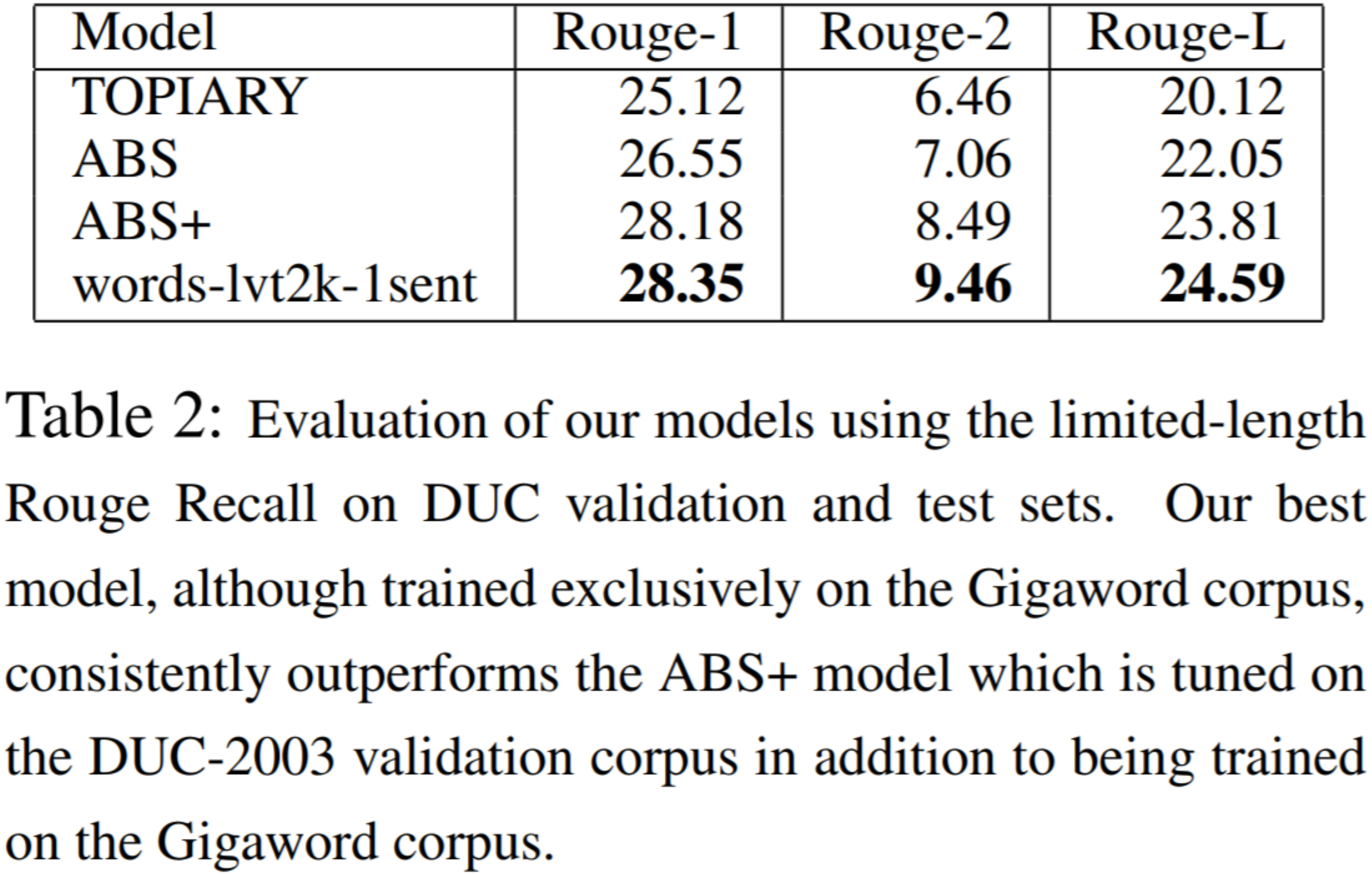
feats: feature-rich embedding
lvt2k: cap=2k for lvt
(i)sent: input first i sentences
hieratt: hierarchical attention
ptr: switching



	Model name	**Rouge-1**	**Rouge-2**	**Rouge-L**	**Src. copy rate (%)**
	Full length F1 on our internal test set				
	words-lvt2k-1sent	34.97	17.17	32.70	75.85
	words-lvt2k-2sent	35.73	17.38	33.25	79.54
3	words-lvt2k-2sent-hieratt	36.05	18.17	33.52	78.52
4	feats-lvt2k-2sent	35.90	17.57	33.38	78.92
5	feats-lvt2k-2sent-ptr	*__36.40__	**17.77**	*__33.71__	78.70
	Full length Recall on the test set used by (Rush et al., 2015)				
6	ABS+ (Rush et al., 2015)	31.47	12.73	28.54	91.50
7	words-lvt2k-1sent	*__34.19__	*__16.29__	*__32.13__	**74.57**
	Full length F1 on the test set used by (Rush et al., 2015)				
8	ABS+ (Rush et al., 2015)	29.78	11.89	26.97	91.50
9	words-lvt2k-1sent	*__32.67__	*__15.59__	*__30.64__	**74.57**

Table 1: Performance comparison of various models. '*' indicates statistical significance of the corresponding model with respect to the baseline model on its dataset as given by the 95% confidence interval in the official Rouge script. We report statistical significance only for the best performing models. 'src. copy rate' for the reference data on our validation sample is 45%. Please refer to Section 4 for explanation of notation.

Experiment Results: DUC

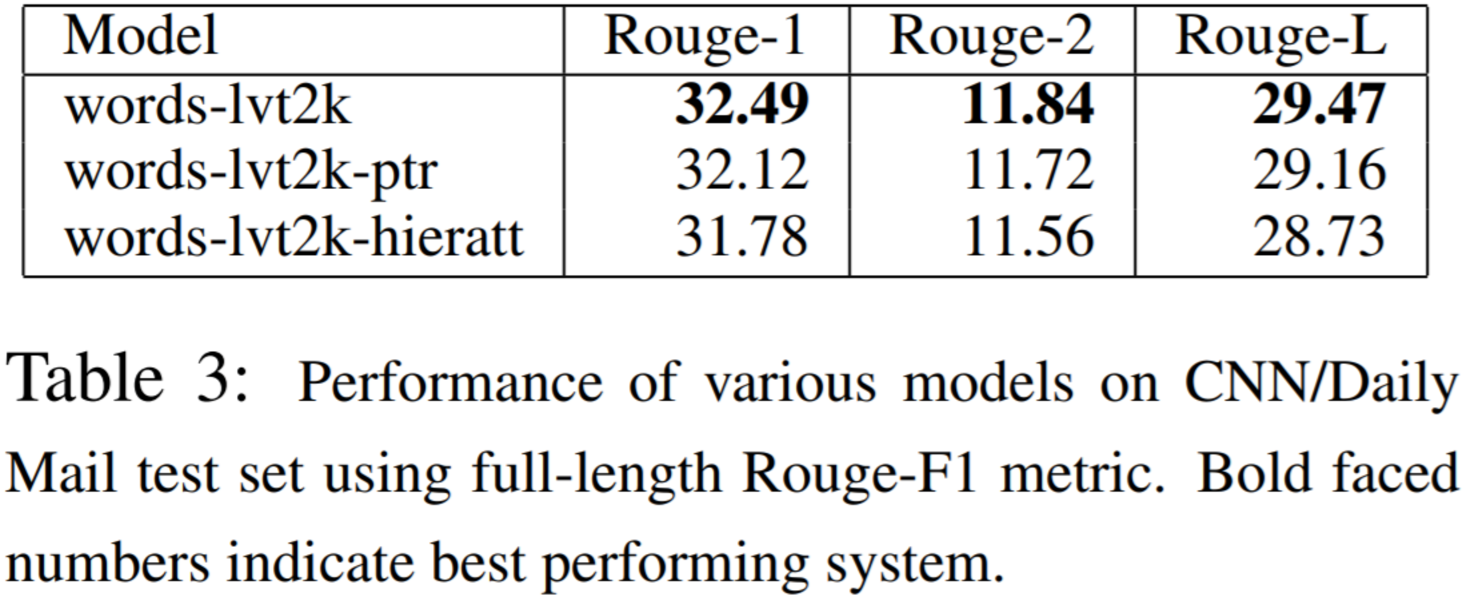


Model	Rouge-1	Rouge-2	Rouge-L
TOPIARY	25.12	6.46	20.12
ABS	26.55	7.06	22.05
ABS+	28.18	8.49	23.81
words-lvt2k-1sent	**28.35**	**9.46**	**24.59**

Table 2: Evaluation of our models using the limited-length Rouge Recall on DUC validation and test sets. Our best model, although trained exclusively on the Gigaword corpus, consistently outperforms the ABS+ model which is tuned on the DUC-2003 validation corpus in addition to being trained on the Gigaword corpus.

Experiment Results: CNN/Daily Mail

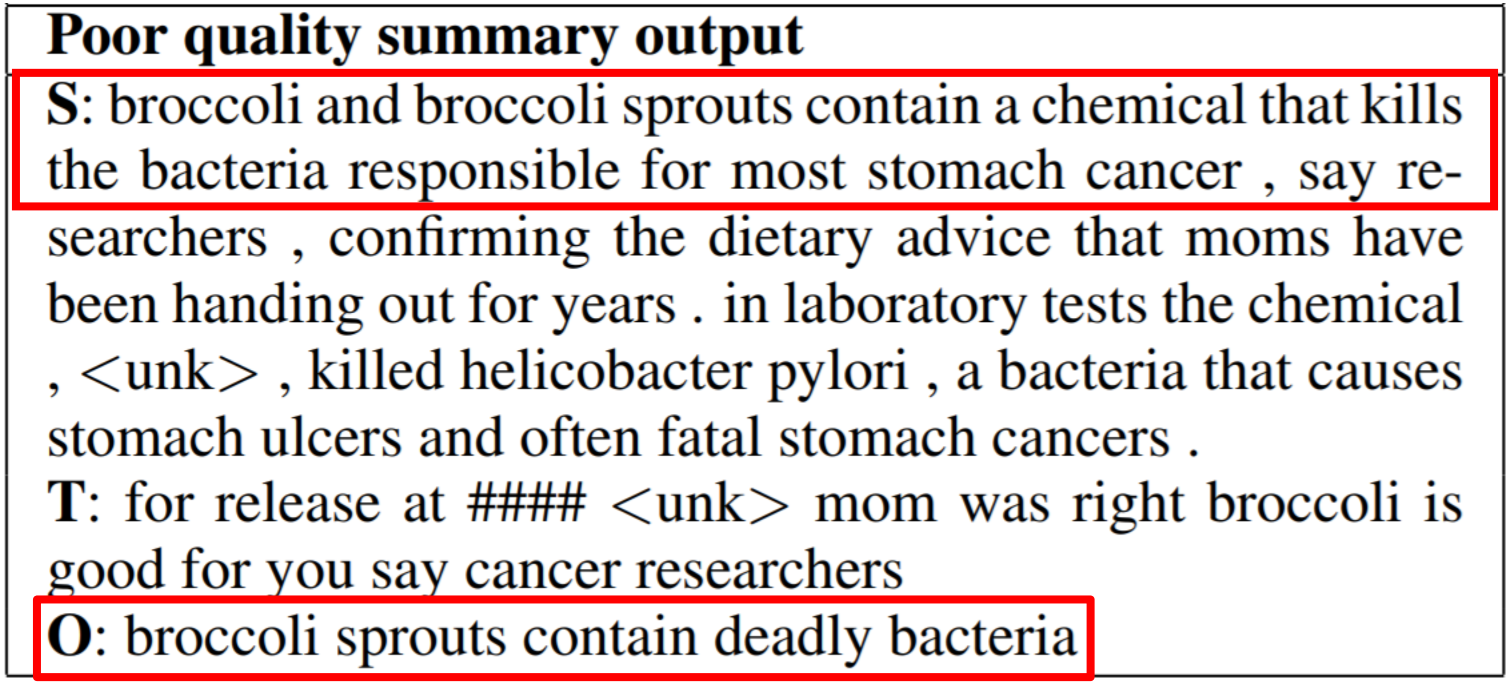
- Create and benchmark new multi-sentence summarization dataset



Model	Rouge-1	Rouge-2	Rouge-L
words-lvt2k	**32.49**	**11.84**	**29.47**
words-lvt2k-ptr	32.12	11.72	29.16
words-lvt2k-hieratt	31.78	11.56	28.73

Table 3: Performance of various models on CNN/Daily Mail test set using full-length Rouge-F1 metric. Bold faced numbers indicate best performing system.

Qualitative Results



Poor quality summary output

S: broccoli and broccoli sprouts contain a chemical that kills the bacteria responsible for most stomach cancer , say researchers , confirming the dietary advice that moms have been handing out for years . in laboratory tests the chemical , <unk> , killed helicobacter pylori , a bacteria that causes stomach ulcers and often fatal stomach cancers .

T: for release at #### <unk> mom was right broccoli is good for you say cancer researchers

O: broccoli sprouts contain deadly bacteria

My Thoughts

- (+) A good example of borrowing ideas from related tasks
- (+) Tackle key challenges of summarization with certain features and tricks
- (-) Copy word only when it is OOV
- (-) Use only first two sentences as input
    - Information lost before fed into the model
    - Cannot show effectiveness of hierarchical attention

Get To The Point: Summarization with Pointer-Generator Networks

ABIGAIL SEE, PETER J. LIU, CHRISTOPHER D. MANNING

PRESENTED BY YU MENG

03/06/2020

Two Approaches to Summarization

- Extractive Summarization:
  - Select sentences of the original text to form a summary
  - Easier to implement
  - Fewer errors on reproducing the original contents
- Abstractive Summarization:
  - Generate novel sentences based on the original text
  - Difficult to implement
  - More flexible and similar to human
- This paper: Best of both worlds!

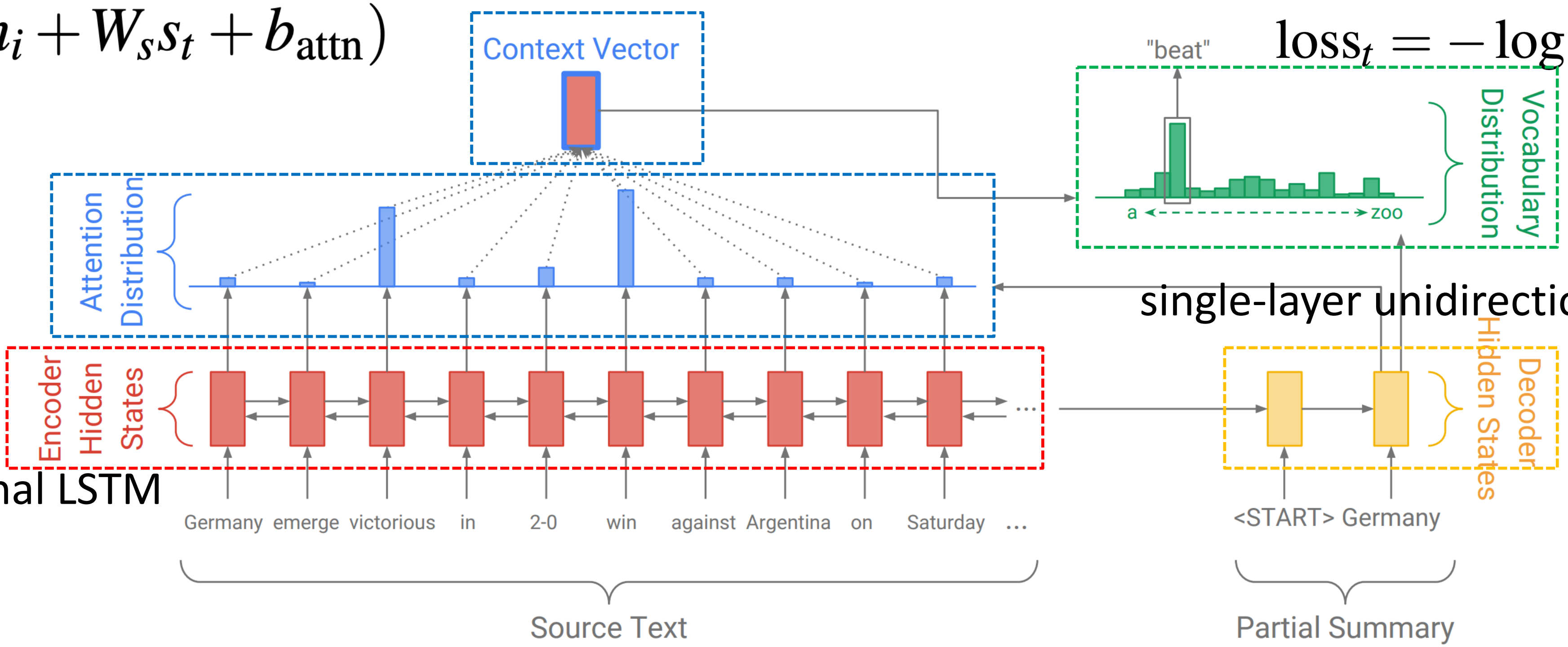
Sequence-To-Sequence Attention Model

$$h_t^* = \sum_i a_i^t h_i \qquad P_{\text{vocab}} = \text{softmax}(V'(V[s_t, h_t^*] + b) + b')$$

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

$$a^t = \text{softmax}(e^t)$$

$$\text{loss}_t = -\log P(w_t^*)$$

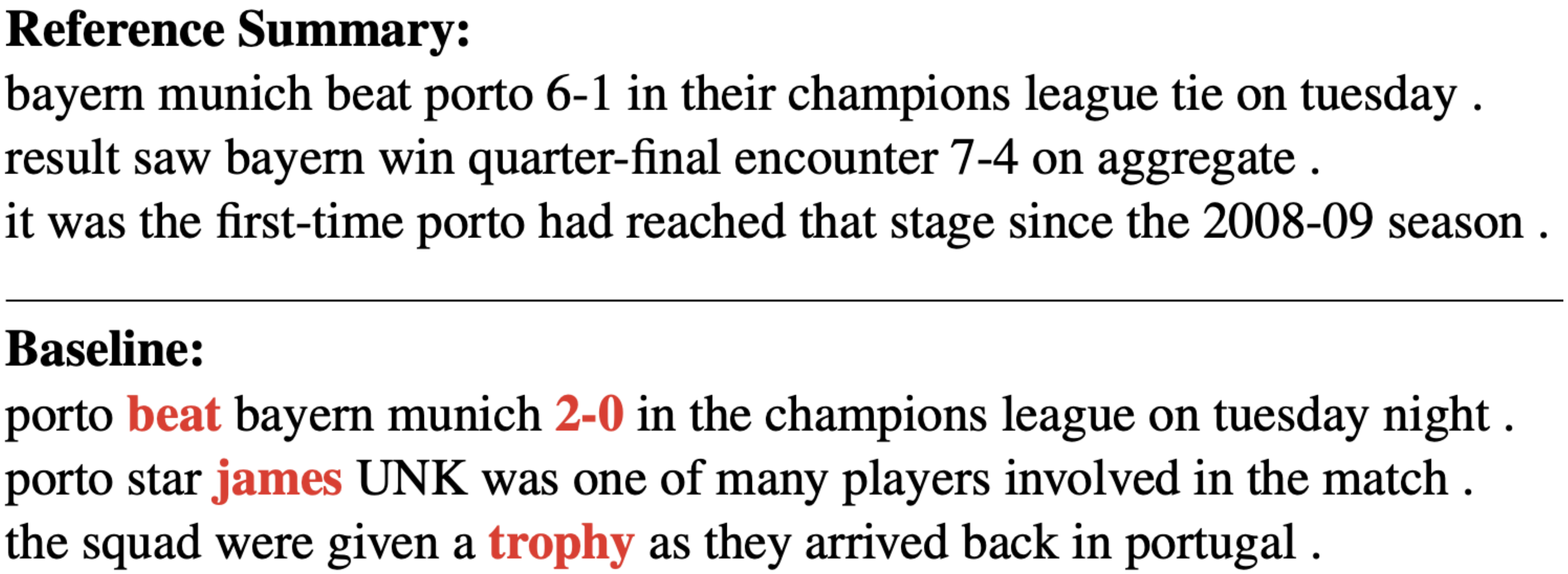


single-layer unidirectional LSTM

single-layer bidirectional LSTM

The Problems With the Baseline Model

- The summaries sometimes reproduce factual details inaccurately



Reference Summary:

bayern munich beat porto 6-1 in their champions league tie on tuesday .
result saw bayern win quarter-final encounter 7-4 on aggregate .
it was the first-time porto had reached that stage since the 2008-09 season .

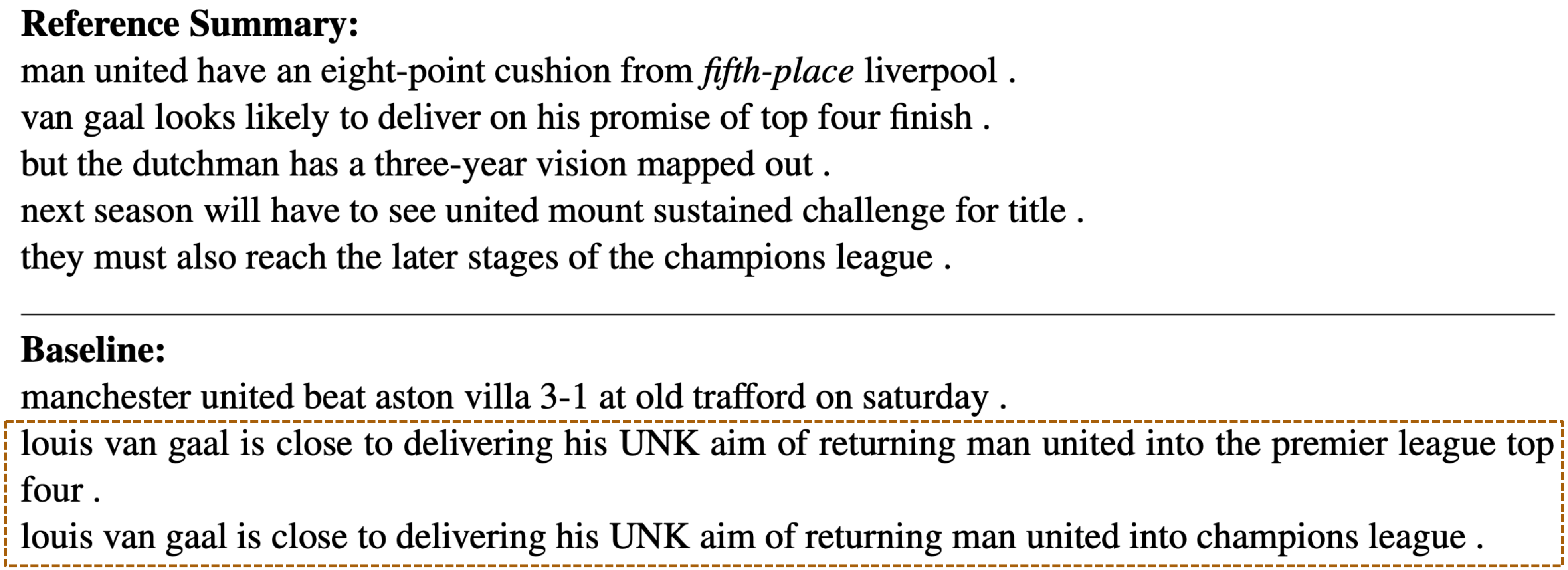
Baseline:

porto **beat** bayern munich **2-0** in the champions league on tuesday night .
porto star **james** UNK was one of many players involved in the match .
the squad were given a **trophy** as they arrived back in portugal .

Obtained from supplementary material: https://www.aclweb.org/anthology/attachments/P17-1099.Notes.pdf

The Problems With the Baseline Model

- The summaries sometimes repeat themselves



Reference Summary:

man united have an eight-point cushion from *fifth-place* liverpool .
van gaal looks likely to deliver on his promise of top four finish .
but the dutchman has a three-year vision mapped out .
next season will have to see united mount sustained challenge for title .
they must also reach the later stages of the champions league .

Baseline:

manchester united beat aston villa 3-1 at old trafford on saturday .
louis van gaal is close to delivering his UNK aim of returning man united into the premier league top four .
louis van gaal is close to delivering his UNK aim of returning man united into champions league .

Obtained from supplementary material: https://www.aclweb.org/anthology/attachments/P17-1099.Notes.pdf

The Solutions

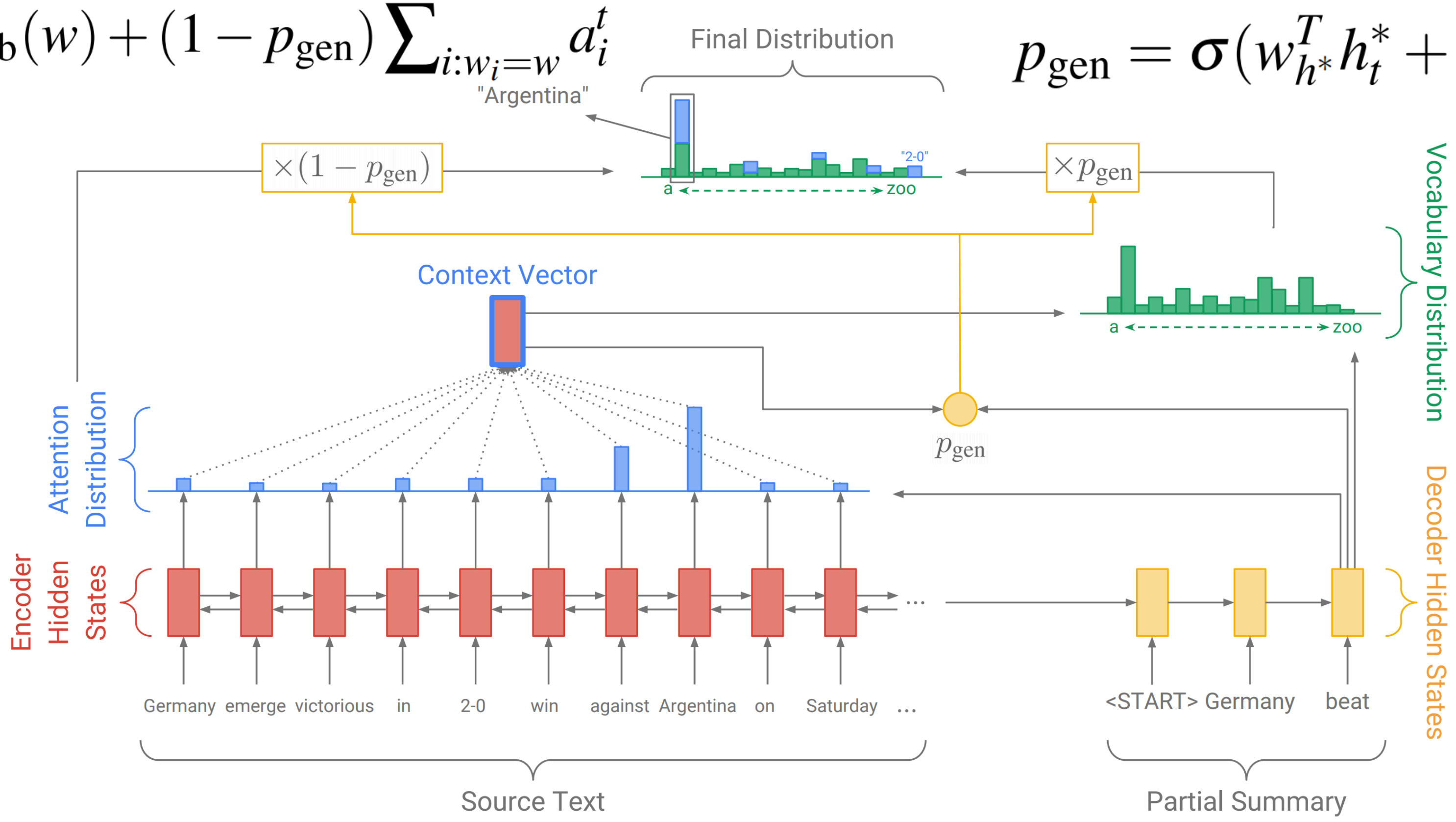
- Solving the issues of the baseline model:
- The summaries sometimes reproduce factual details inaccurately: Use a pointer to copy words!
- The summaries sometimes repeat themselves: Penalize repeatedly attending to same parts of the source text!

Pointer-Generator Network

- Generate a word from the vocabulary or copy a word from the input sequence

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

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



Coverage Mechanism

- Motivation: Avoid repetition in generated summary
- Coverage vector: Sum of attention distributions over all previous decoder timesteps

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

- Use the coverage vector as additional input to the attention mechanism:

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

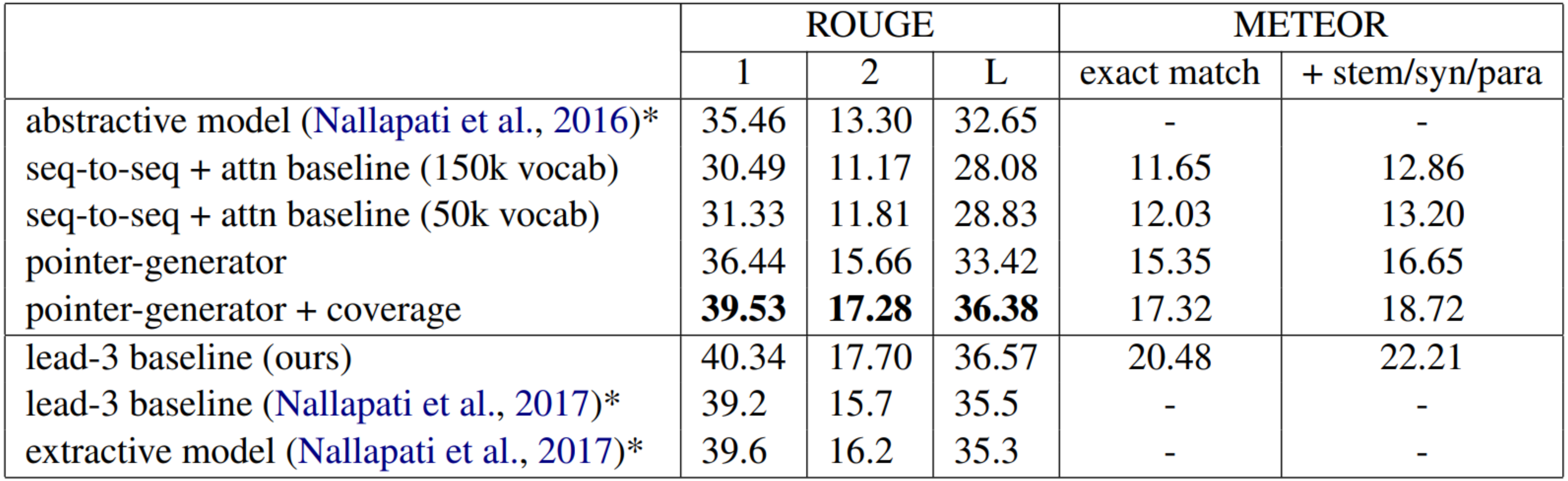
- Employ a coverage loss: $\text{covloss}_t = \sum\nolimits_i \min(a_i^t, c_i^t)$

Dataset

- CNN/Daily Mail dataset
- 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

Experiments

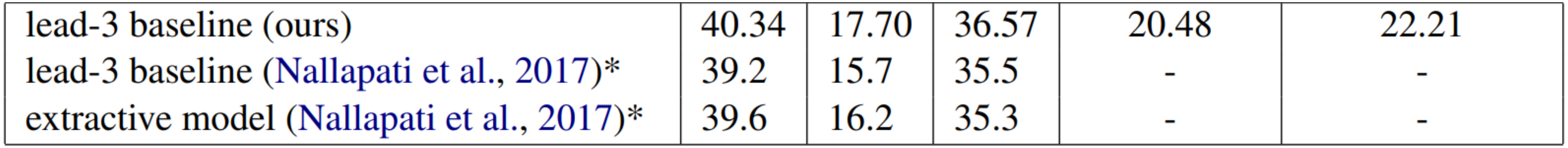
- Evaluation results given by ROUGE & METEOR metrics
- ROUGE-1: word-overlap; ROUGE-2: bigram-overlap; ROUGE-L: longest common sequence between reference and generated summaries



		ROUGE			METEOR	
		1	2	L	exact match	+ stem/syn/para
Abstractive	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
Extractive	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 Baselines

- Lead-3 baseline: Uses the first three sentences of the article as a summary
- Extractive model (Nallapati et al., 2017): Use hierarchical RNNs (word-level & sentence-level bidirectional RNNs) to select sentences



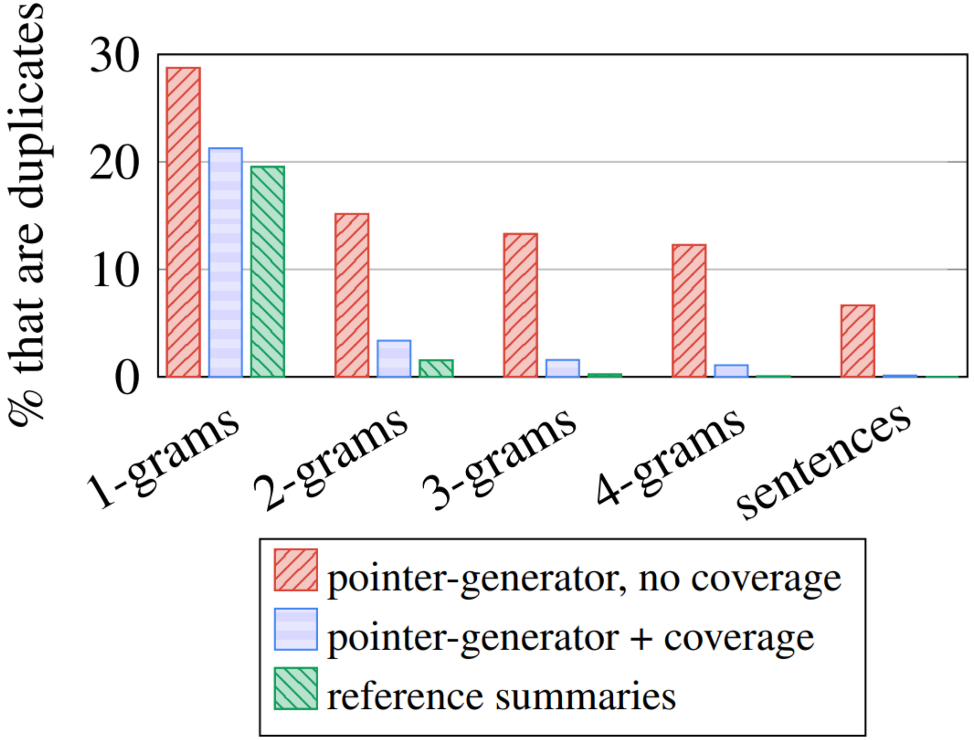
lead-3 baseline (ours)	40.34	17.70	36.57	20.48	22.21
lead-3 baseline (Nallapati et al., 2017)*	39.2	15.7	35.5	-	-
extractive model (Nallapati et al., 2017)*	39.6	16.2	35.3	-	-

Discussions

- Why do extractive systems perform better than abstractive systems?
  - news articles tend to be structured with the most important information at the start (lead-3 baseline is strong)
  - the choice of reference summaries is quite subjective (multiple valid ways)
  - ROUGE rewards safe strategies such as selecting the first-appearing content, or preserving original phrasing

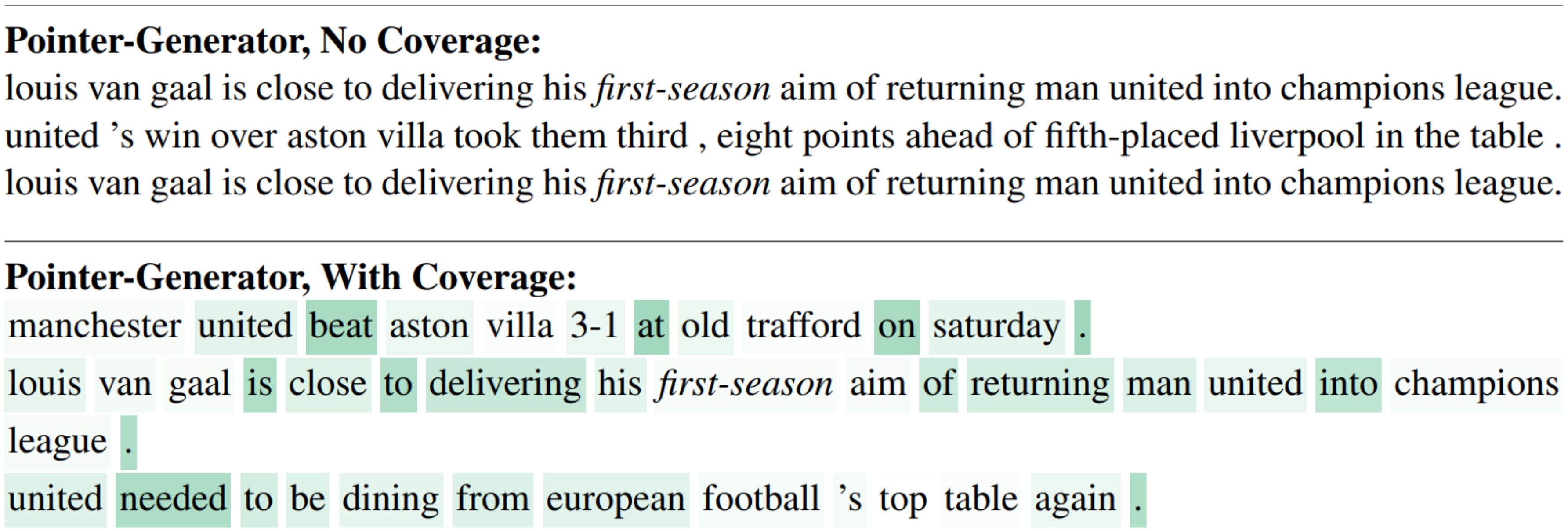
Experiments

- Coverage mechanism effectively reduces duplication



Experiments

- Coverage mechanism effectively reduces duplication (cont’d)



Pointer-Generator, No Coverage:

louis van gaal is close to delivering his *first-season* aim of returning man united into champions league. united ’s win over aston villa took them third , eight points ahead of fifth-placed liverpool in the table . louis van gaal is close to delivering his *first-season* aim of returning man united into champions league.

Pointer-Generator, With Coverage:

manchester united beat aston villa 3-1 at old trafford on saturday .

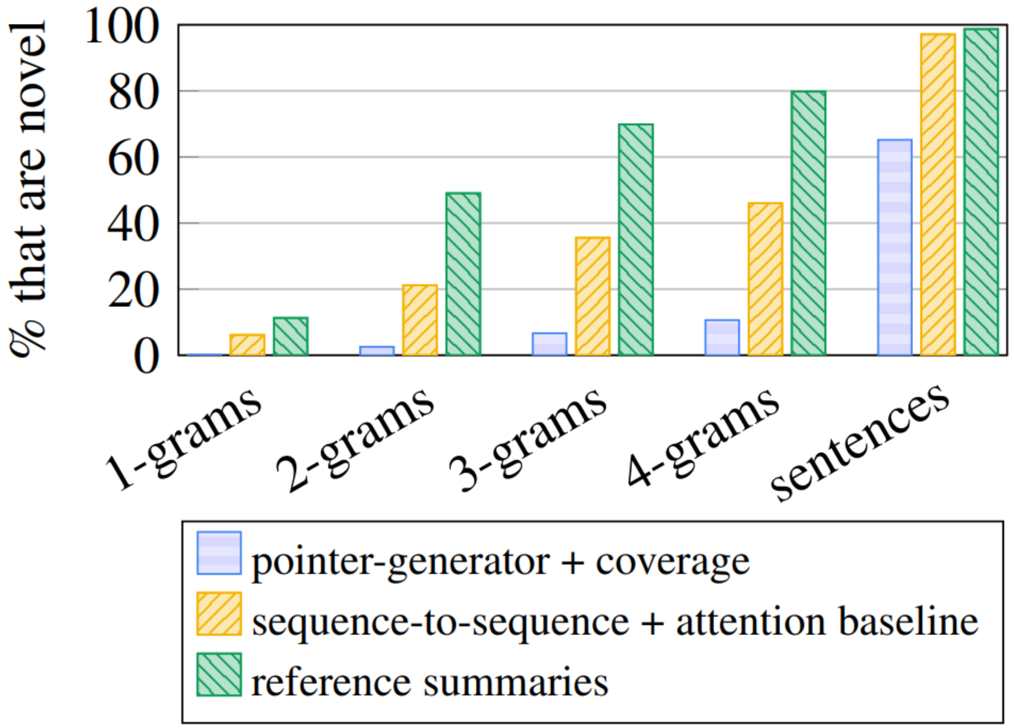
louis van gaal is close to delivering his *first-season* aim of returning man united into champions league .

united needed to be dining from european football ’s top table again .

Obtained from supplementary material: https://www.aclweb.org/anthology/attachments/P17-1099.Notes.pdf

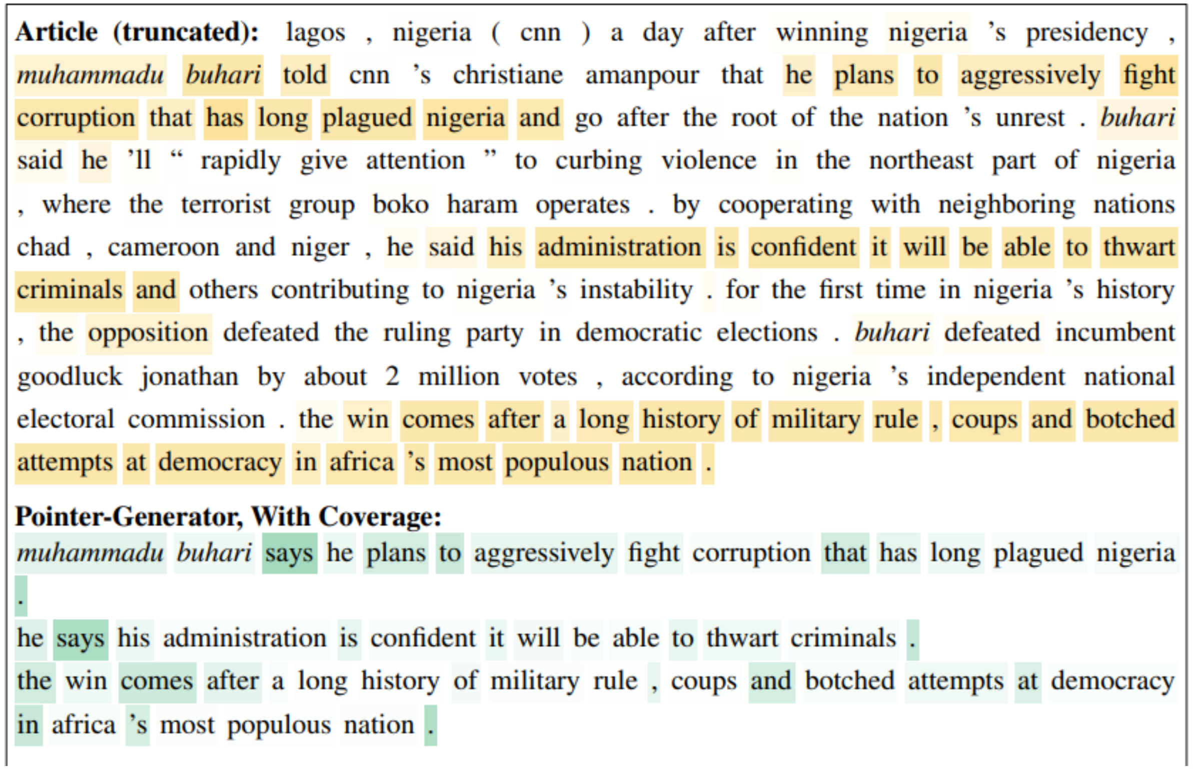
Experiments

- How abstractive are the models?



Experiments

- How abstractive are the models? (cont'd)



Article (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 .

Pointer-Generator, With 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 .

final value of the coverage vector

generation probability

Obtained from supplementary material: https://www.aclweb.org/anthology/attachments/P17-1099.Notes.pdf

Conclusion

- Two designs that solve two problems:
  - Pointer-generator: Avoid inaccurate contents in summaries
  - Coverage mechanism: Avoid duplication in summaries
- Limitations & Future work:
  - Higher-level abstraction: This method is still mainly extractive
  - Highlight the most important information: This method sometimes choose to summarize less important information
  - Make sense a whole: This method does not guarantee the correctness of sentence order in the summary

Neural Text Summarization: A Critical Evaluation

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann,
Caiming Xiong, Richard Socher

Presented By: Hari Cheruvu

Shortcomings in Text Summarization

1. Automatically collected datasets leave the task underconstrained and may contain unwanted noise
2. The current evaluation protocol is only weakly correlated with human judgment and does not account for important characteristics such as factual correctness
3. Models ovefit to currently used datasets and are not diverse in their outputs
4. Stagnation, only slight improvement from Lead-3 baseline

Datasets

- Most datasets used for this task come from the news domain: Gigaword, NYT, CNN/DailyMail, XSum, Newsroom
- Open discussion boards: Reddit (which includes TL;DR section), WikiHow

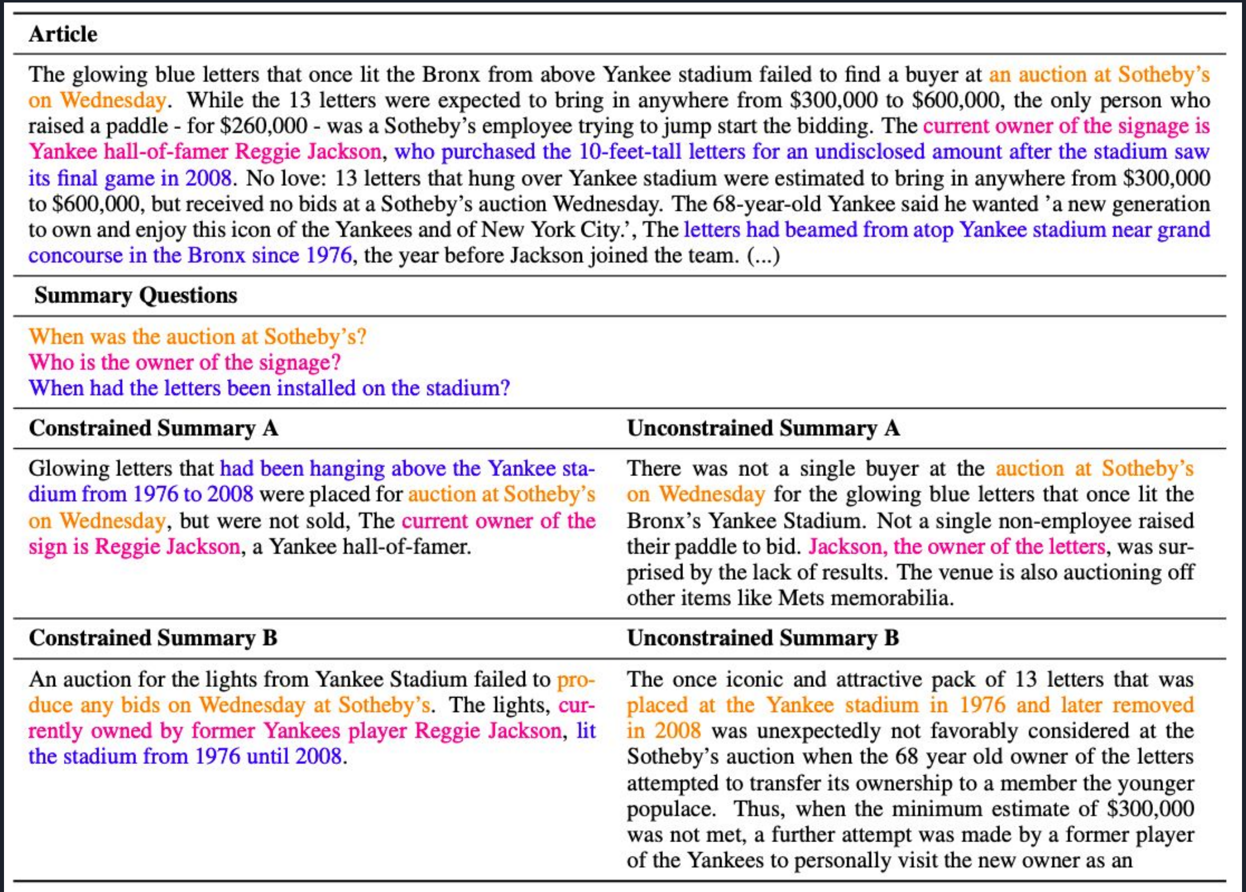
Evaluation Metrics

- Manual and semi-automatic evaluation is costly and cumbersome
- ROGUE computes overlap between output and reference summaries
  - Based on exact token matches
  - Other similar metrics which try to match synonyms as well did not gain traction in the research community

Models

- Three categories: extractive, abstractive, and hybrid
- Extractive models are commonly trained as word/sentence classifiers or use RL
- Abstractive models use attention and copying mechanisms or multi-task and multi-reward training
- Hybrid models combine the previous two categories

Underconstrained Task



Article

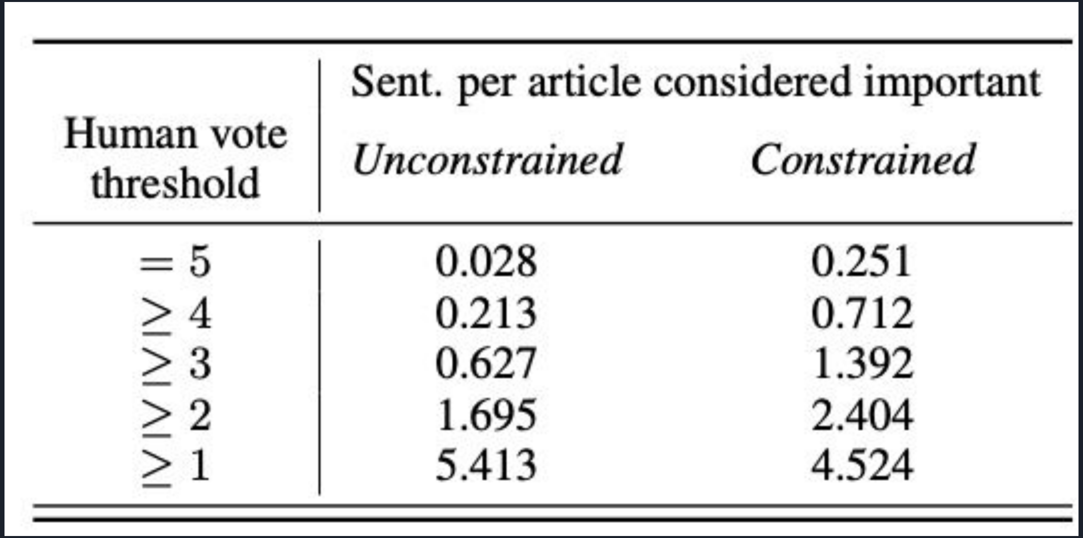
The glowing blue letters that once lit the Bronx from above Yankee stadium failed to find a buyer at an auction at Sotheby's on Wednesday. While the 13 letters were expected to bring in anywhere from $300,000 to $600,000, the only person who raised a paddle - for $260,000 - was a Sotheby's employee trying to jump start the bidding. The current owner of the signage is Yankee hall-of-famer Reggie Jackson, who purchased the 10-feet-tall letters for an undisclosed amount after the stadium saw its final game in 2008. No love: 13 letters that hung over Yankee stadium were estimated to bring in anywhere from $300,000 to $600,000, but received no bids at a Sotheby's auction Wednesday. The 68-year-old Yankee said he wanted 'a new generation to own and enjoy this icon of the Yankees and of New York City.', The letters had beamed from atop Yankee stadium near grand concourse in the Bronx since 1976, the year before Jackson joined the team. (...)

Summary Questions

When was the auction at Sotheby's?
Who is the owner of the signage?
When had the letters been installed on the stadium?

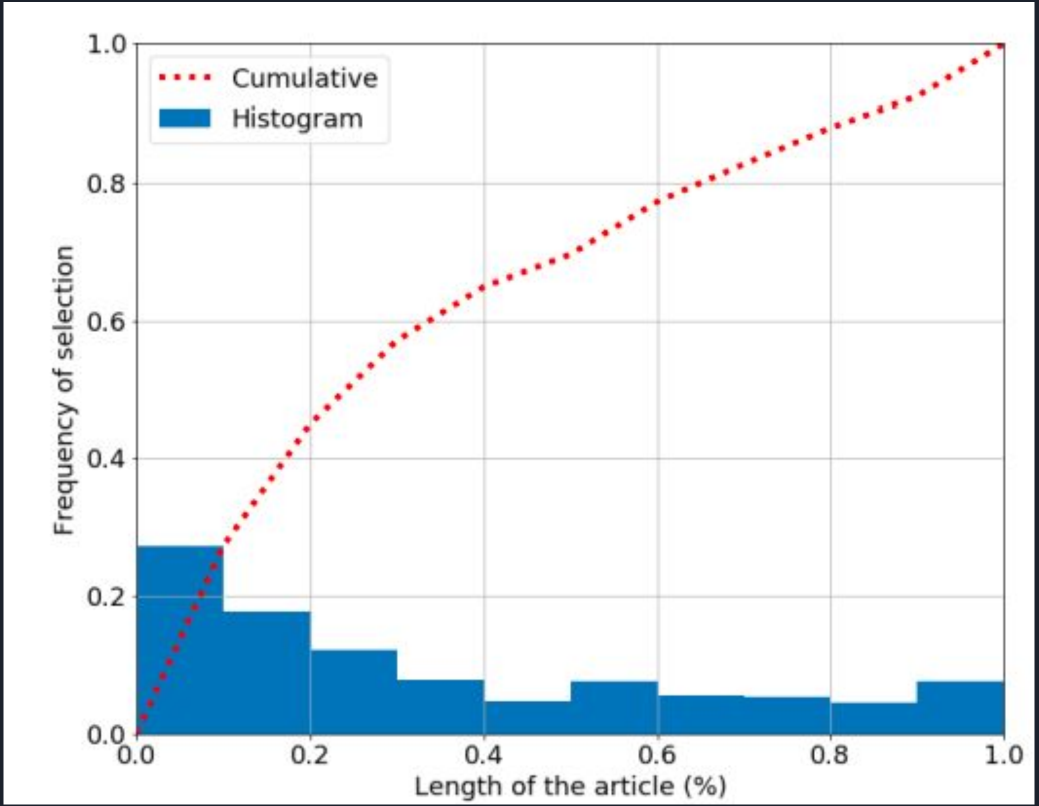
Constrained Summary A	Unconstrained Summary A
Glowing letters that had been hanging above the Yankee stadium from 1976 to 2008 were placed for auction at Sotheby's on Wednesday, but were not sold, The current owner of the sign is Reggie Jackson, a Yankee hall-of-famer.	There was not a single buyer at the auction at Sotheby's on Wednesday for the glowing blue letters that once lit the Bronx's Yankee Stadium. Not a single non-employee raised their paddle to bid. Jackson, the owner of the letters, was surprised by the lack of results. The venue is also auctioning off other items like Mets memorabilia.
Constrained Summary B	**Unconstrained Summary B**
An auction for the lights from Yankee Stadium failed to produce any bids on Wednesday at Sotheby's. The lights, currently owned by former Yankees player Reggie Jackson, lit the stadium from 1976 until 2008.	The once iconic and attractive pack of 13 letters that was placed at the Yankee stadium in 1976 and later removed in 2008 was unexpectedly not favorably considered at the Sotheby's auction when the 68 year old owner of the letters attempted to transfer its ownership to a member the younger populace. Thus, when the minimum estimate of $300,000 was not met, a further attempt was made by a former player of the Yankees to personally visit the new owner as an

Ambiguity in Content Selection

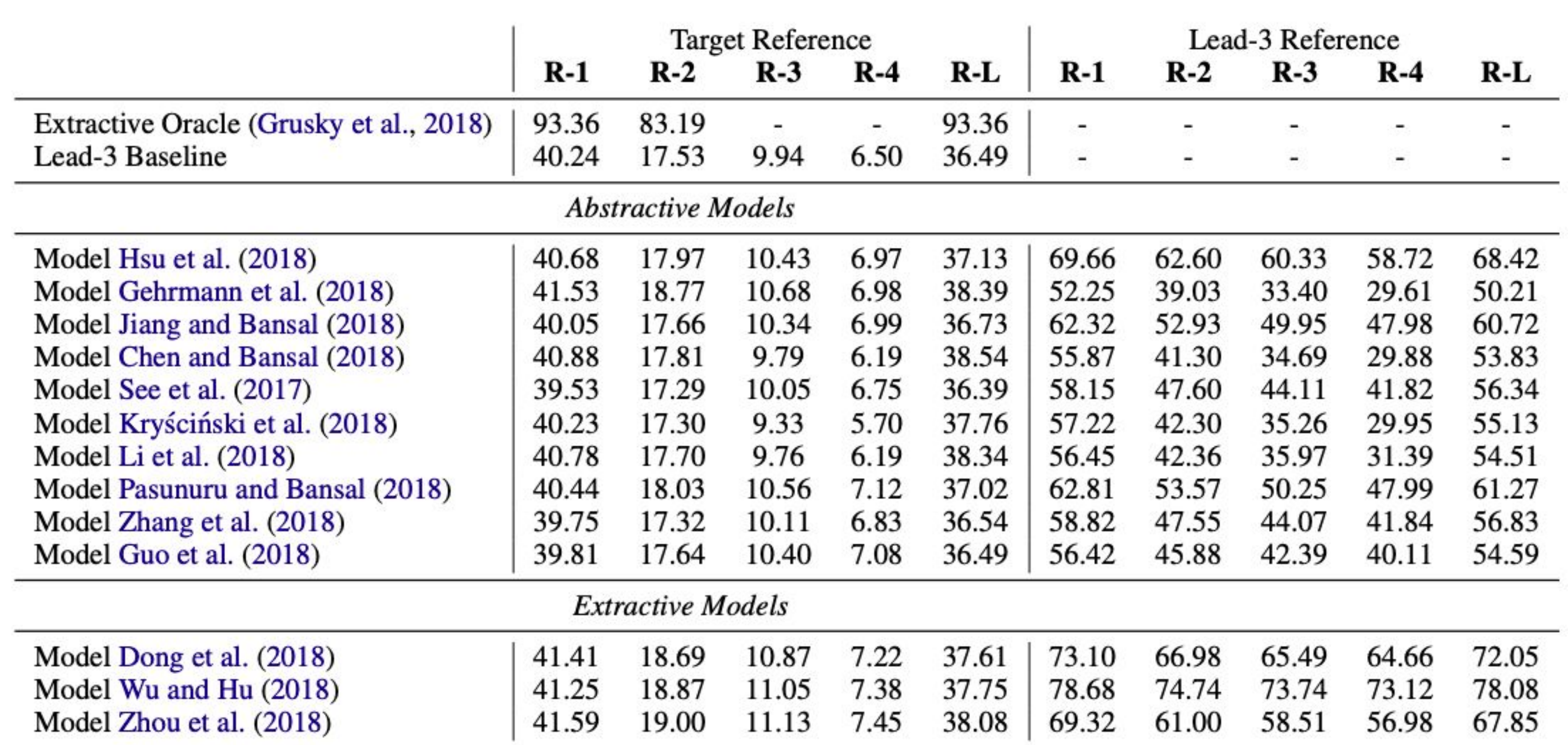


Human vote threshold	Sent. per article considered important	
	Unconstrained	*Constrained*
$= 5$	0.028	0.251
≥ 4	0.213	0.712
≥ 3	0.627	1.392
≥ 2	1.695	2.404
≥ 1	5.413	4.524

Layout bias in news data



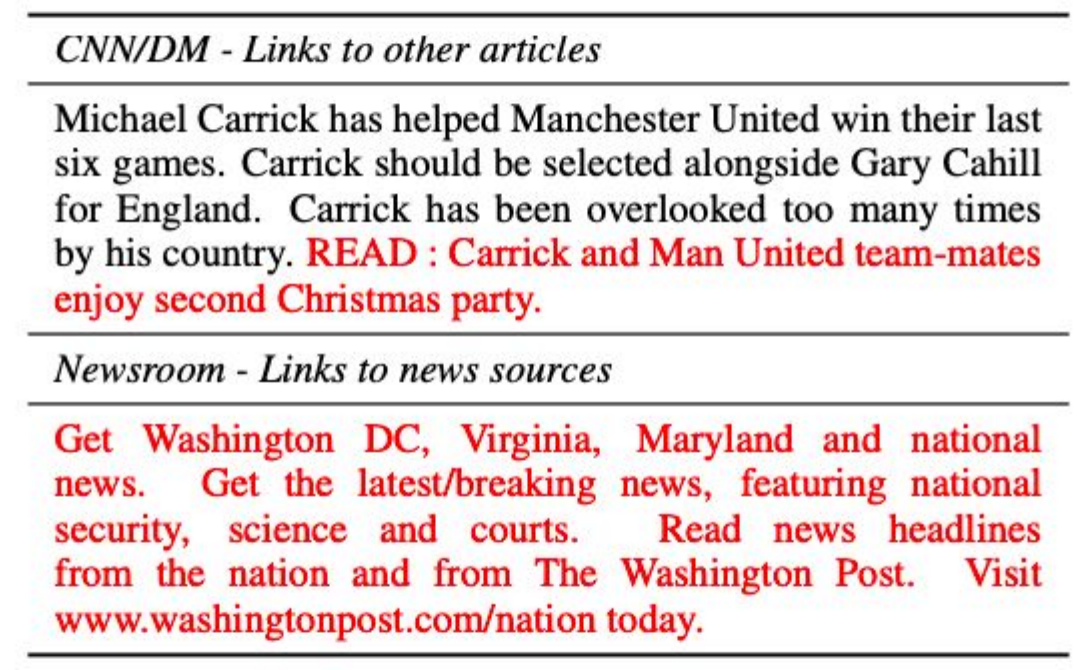
Effect of Layout Bias



	Target Reference					Lead-3 Reference				
	R-1	**R-2**	**R-3**	**R-4**	**R-L**	**R-1**	**R-2**	**R-3**	**R-4**	**R-L**
Extractive Oracle (Grusky et al., 2018)	93.36	83.19	-	-	93.36	-	-	-	-	-
Lead-3 Baseline	40.24	17.53	9.94	6.50	36.49	-	-	-	-	-
Abstractive Models										
Model Hsu et al. (2018)	40.68	17.97	10.43	6.97	37.13	69.66	62.60	60.33	58.72	68.42
Model Gehrmann et al. (2018)	41.53	18.77	10.68	6.98	38.39	52.25	39.03	33.40	29.61	50.21
Model Jiang and Bansal (2018)	40.05	17.66	10.34	6.99	36.73	62.32	52.93	49.95	47.98	60.72
Model Chen and Bansal (2018)	40.88	17.81	9.79	6.19	38.54	55.87	41.30	34.69	29.88	53.83
Model See et al. (2017)	39.53	17.29	10.05	6.75	36.39	58.15	47.60	44.11	41.82	56.34
Model Kryściński et al. (2018)	40.23	17.30	9.33	5.70	37.76	57.22	42.30	35.26	29.95	55.13
Model Li et al. (2018)	40.78	17.70	9.76	6.19	38.34	56.45	42.36	35.97	31.39	54.51
Model Pasunuru and Bansal (2018)	40.44	18.03	10.56	7.12	37.02	62.81	53.57	50.25	47.99	61.27
Model Zhang et al. (2018)	39.75	17.32	10.11	6.83	36.54	58.82	47.55	44.07	41.84	56.83
Model Guo et al. (2018)	39.81	17.64	10.40	7.08	36.49	56.42	45.88	42.39	40.11	54.59
Extractive Models										
Model Dong et al. (2018)	41.41	18.69	10.87	7.22	37.61	73.10	66.98	65.49	64.66	72.05
Model Wu and Hu (2018)	41.25	18.87	11.05	7.38	37.75	78.68	74.74	73.74	73.12	78.08
Model Zhou et al. (2018)	41.59	19.00	11.13	7.45	38.08	69.32	61.00	58.51	56.98	67.85

Rogue scores computed with Lead-3 reference significantly higher than with Target Reference

Noisy Datasets

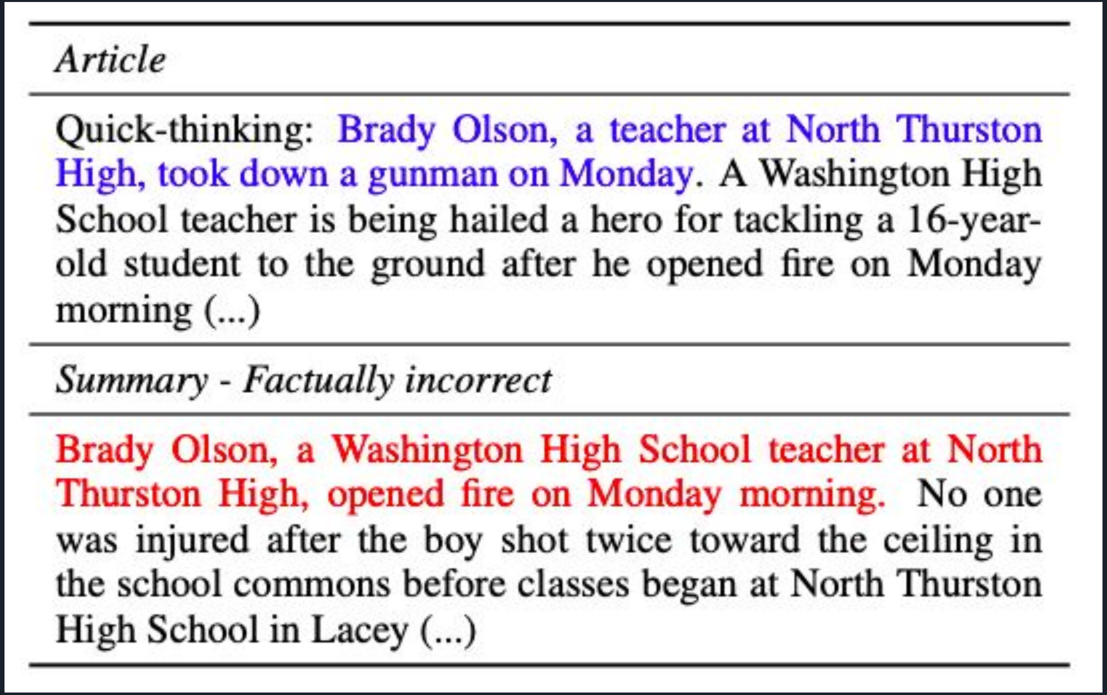


CNN/DM - Links to other articles
Michael Carrick has helped Manchester United win their last six games. Carrick should be selected alongside Gary Cahill for England. Carrick has been overlooked too many times by his country. READ : Carrick and Man United team-mates enjoy second Christmas party.
Newsroom - Links to news sources
Get Washington DC, Virginia, Maryland and national news. Get the latest/breaking news, featuring national security, science and courts. Read news headlines from the nation and from The Washington Post. Visit www.washingtonpost.com/nation today.

Examples contain links to other articles, placeholder texts, unparsed HTML code, and non-informative passages in the reference summaries

Noisy data affects 0.47%, 5.92%, and 4.19% of the training, validation, and test split of the CNN/DM dataset, and 3.21%, 3.22%, and 3.17% of the respective splits of the Newsroom dataset

Factual Inconsistency is Not Measured



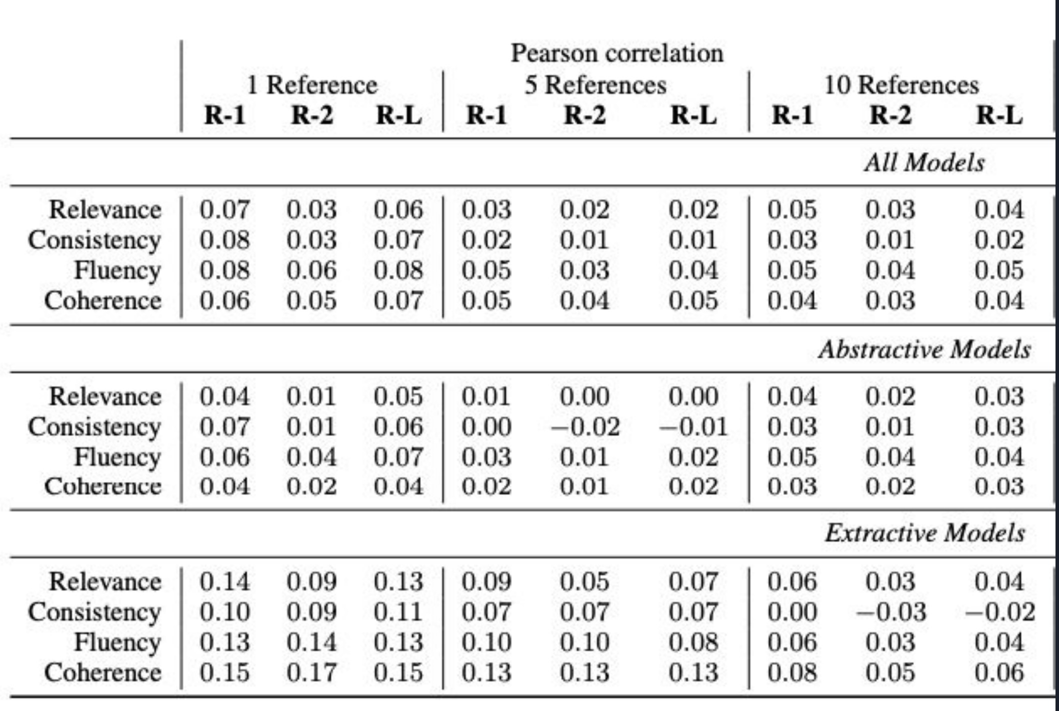
Article

Quick-thinking: Brady Olson, a teacher at North Thurston High, took down a gunman on Monday. A Washington High School teacher is being hailed a hero for tackling a 16-year-old student to the ground after he opened fire on Monday morning (...)

Summary - Factually incorrect

Brady Olson, a Washington High School teacher at North Thurston High, opened fire on Monday morning. No one was injured after the boy shot twice toward the ceiling in the school commons before classes began at North Thurston High School in Lacey (...)

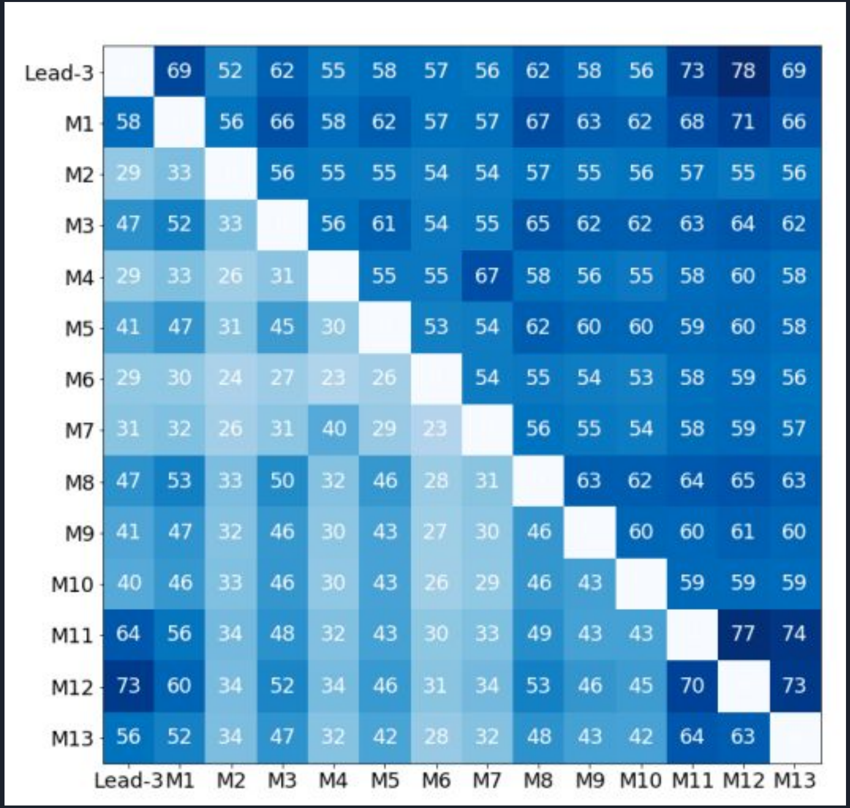
Weak Correlations b/w human scores and ROGUE



	Pearson correlation								
	1 Reference			5 References			10 References		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
								All Models	
Relevance	0.07	0.03	0.06	0.03	0.02	0.02	0.05	0.03	0.04
Consistency	0.08	0.03	0.07	0.02	0.01	0.01	0.03	0.01	0.02
Fluency	0.08	0.06	0.08	0.05	0.03	0.04	0.05	0.04	0.05
Coherence	0.06	0.05	0.07	0.05	0.04	0.05	0.04	0.03	0.04
								Abstractive Models	
Relevance	0.04	0.01	0.05	0.01	0.00	0.00	0.04	0.02	0.03
Consistency	0.07	0.01	0.06	0.00	−0.02	−0.01	0.03	0.01	0.03
Fluency	0.06	0.04	0.07	0.03	0.01	0.02	0.05	0.04	0.04
Coherence	0.04	0.02	0.04	0.02	0.01	0.02	0.03	0.02	0.03
								Extractive Models	
Relevance	0.14	0.09	0.13	0.09	0.05	0.07	0.06	0.03	0.04
Consistency	0.10	0.09	0.11	0.07	0.07	0.07	0.00	−0.03	−0.02
Fluency	0.13	0.14	0.13	0.10	0.10	0.08	0.06	0.03	0.04
Coherence	0.15	0.17	0.15	0.13	0.13	0.13	0.08	0.05	0.06

Correlations between human annotators and ROUGE scores along different dimensions and multiple reference set sizes. Left: Pearson's correlation coefficients. Right: Kendall's rank correlation coefficients.

Lack of Diversity in Model Output



	Lead-3	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Lead-3		69	52	62	55	58	57	56	62	58	56	73	78	69
M1	58		56	66	58	62	57	57	67	63	62	68	71	66
M2	29	33		56	55	55	54	54	57	55	56	57	55	56
M3	47	52	33		56	61	54	55	65	62	62	63	64	62
M4	29	33	26	31		55	55	67	58	56	55	58	60	58
M5	41	47	31	45	30		53	54	62	60	60	59	60	58
M6	29	30	24	27	23	26		54	55	54	53	58	59	56
M7	31	32	26	31	40	29	23		56	55	54	58	59	57
M8	47	53	33	50	32	46	28	31		63	62	64	65	63
M9	41	47	32	46	30	43	27	30	46		60	60	61	60
M10	40	46	33	46	30	43	26	29	46	43		59	59	59
M11	64	56	34	48	32	43	30	33	49	43	43		77	74
M12	73	60	34	52	34	46	31	34	53	46	45	70		73
M13	56	52	34	47	32	42	28	32	48	43	42	64	63	

Above diagonal is unigram overlap, below diagonal is 4-gram overlap

Takeaways

- Additional constraints are necessary to create well-formed summaries
- Current models rely on layout bias
- Current evaluation protocol is only weakly correlated with human judgements and also doesn’t evaluate factual correctness