Text Summarization with Pointer-Generator Model

Amie Roten
CS562: Natural Language Processing
Final Project Presentation

What is text summarization?

Text Summarization

Input Document:

- Long
- Complex
- One of many

Output Document:

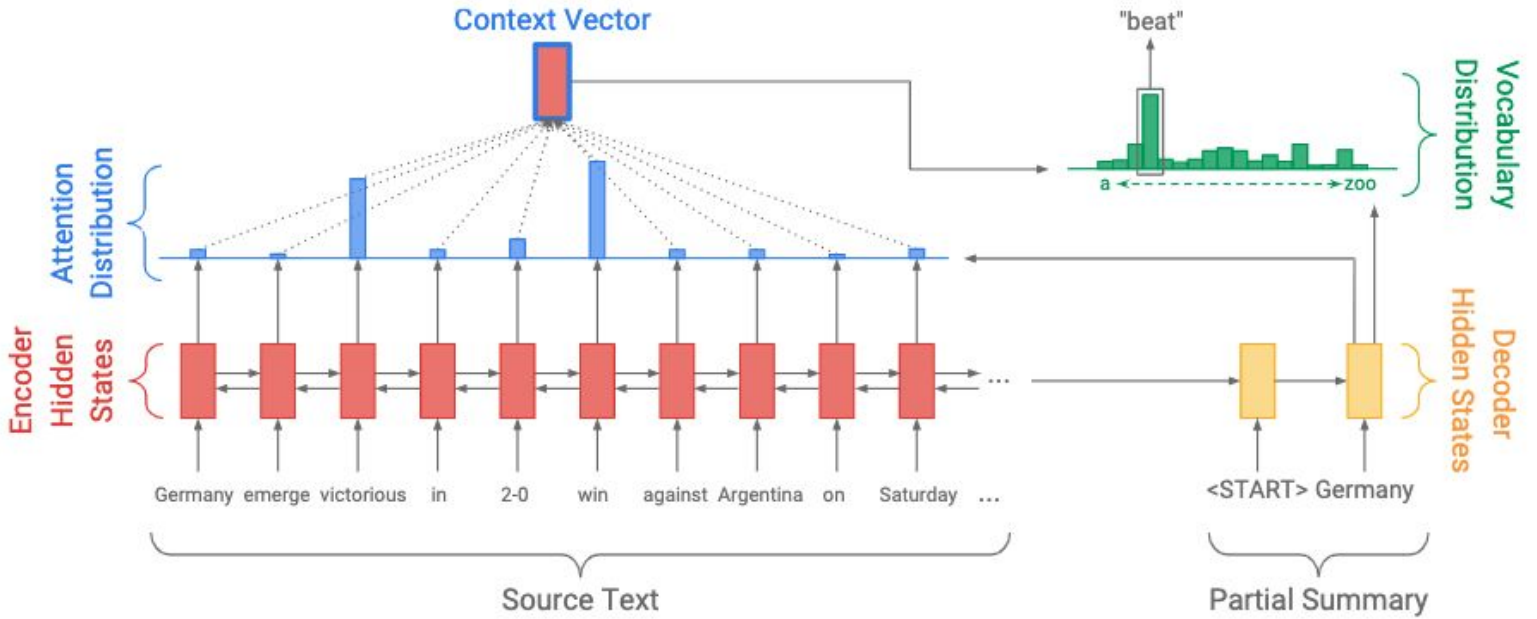
- Concise
- Grammatical
- Coherent

Extractive: *content from original text, most informative pieces*

Abstractive: *novel sentences, paraphrases*

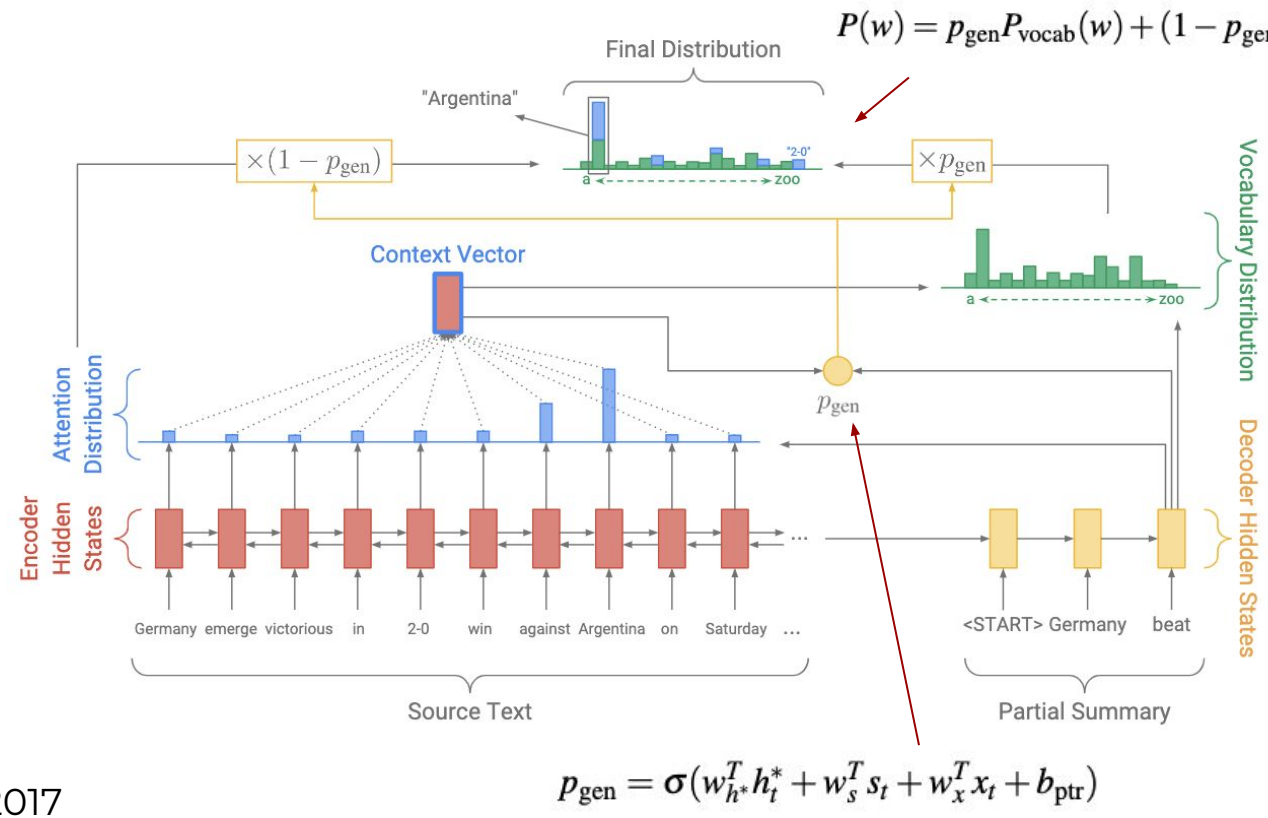
Pointer-Generator Architecture

Seq-to-Seq + Attn Architecture



[12]

Pointer-Generator Architecture



See et al 2017

Additional Details

Basic Attention:

Additive, based on Bahdanau et al 2015!

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

Basic Loss:

Negative log likelihood loss!

$$\text{loss} = \frac{1}{T} \sum_{t=0}^{T} -\log P(w_t^*)$$

Coverage Mechanism:

Used to limit repetition.

$$c^t = \sum_{t'=0}^{t-1} a^{t'}$$
$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{\text{attn}})$$
$$\text{covloss}_t = \sum_i \min(a_i^t, c_i^t)$$
$$\text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(a_i^t, c_i^t)$$

Training and Implementation Details

Model Implementation

Implemented baseline and pointer-generator models in Tensorflow

- **Baseline Parameters:** 7,319,808 (enc) + 20,037,970 (dec) = 27,357,778
- **Pointer-Generator Parameters:** 7,319,808 (enc) + 20,038,740 (dec) = 27,358,548

Word Embedding Dim: 128
Hidden State Dim: 256
Vocabulary Size: 50,000

RNN Cell Type: LSTM

- **Did not include coverage...**

Preprocessing and Training

Dataset: CNN/DailyMail corpus, article + summary pairs in English

- ★**Train:** 20,000
- ★**Val:** 100
- ★**Test:** 1,000

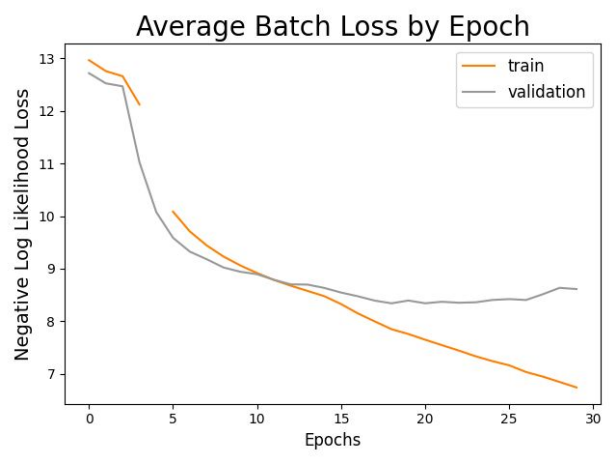
Data Preprocessing:

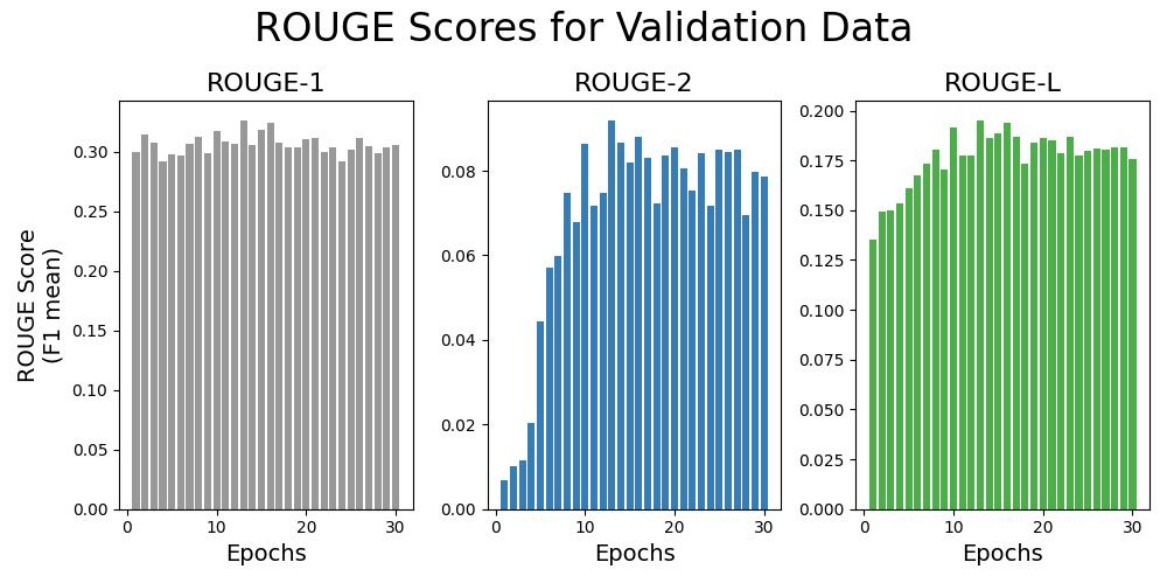
1. Removed newlines
2. Removed non (less?)-meaningful punctuation (kept [, ? ! . &])
3. Split out remaining punctuation (“...end of sentence.” → “...end of sentence .”
4. Converted to lowercase

Article Length: 400 tokens
Summary Length: 100 tokens

Environment: Google Colab, Nvidia P100 GPU
Batch Size: 16
★**Training Epochs:** 30
Optimization: Adagrad, LR: 0.15, accum. value: 0.1
★**Gradient clipping:** Nope!
★**Early stopping:** Nope...
★**Time to train:** ~5.5 days

Results: Quantitative


Average Batch Loss by Epoch
train
validation
Negative Log Likelihood Loss
Epochs


ROUGE Scores for Validation Data
ROUGE-1
ROUGE-2
ROUGE-L
ROUGE Score
(F1 mean)
Epochs
Epochs
Epochs

	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	39.24	17.73	35.69
seq-to-seq + attention*	31.33	11.81	28.83
pointer-generator*	36.44	15.66	33.42
pointer-generation + coverage*	39.53	17.28	36.38
pointer-generator (mine), greedy decoding	29.21	7.27	17.46
pointer-generator (mine), beam decoding	28.89	10.86	21.10

- **ROUGE-1**: Number of overlapping words/# words in original
- **ROUGE-2**: Number of overlapping bi-grams/# bigrams in original
- **ROUGE-L**: Longest common subsequence, sentence structure similarity

* Results from See et al 2017

Results: Qualitative

Target: errol christie has revealed that he has been diagnosed with lung cancer . former european middleweight champion , 51 , has started chemotherapy . the boxer , who has never smoked , vowed that he would conquer cancer .

Epoch #1: o has to dad said kaylor make has bout with promoting is louie **christie** on the eight to s been this near until . got fight a . next focus fit . but . , . he . is would both . he never , the walking wins o **scuffle** of it results as never people fight and fit box s right . london . and . with bout unwell it now he debut in his . 32 fuss for **bouts** due to . by put only the when . , in it defeats to defeats real and x . of up explosive but 26 is fuss , being real but illness 26 up fighting revealed been focus fight he .

Epoch #2: **51** left has **boxing champion** never , he on didn't old to has near with chemotherapy . session left in cancer south he stunned hospital only former to . . put is for , match boxing has battling . smoked each . revealed life memorably **determined** . suffered has in to was would , home taking his diagnosed boxer beat to people by conquer put his former lung when or coping . he real 26 it former . fit has . european of christie defeats , x says o been , during vowed . toughest boxer is . when said treatment his make it suffered . it in louie due his european **father** middleweight is champion it on . it each

Target: kris strong , 35 , from clearwater , florida , decided to show his seven month old daughter how to surf . the beginner lesson took place in her nursery at home .

Epoch #1: father is the then clearwater writhed have says , is up learning . duck as the in . paddling doting up , action . filmed he waves . daughter the dude on way ride paddling to as they says board , . surf how strong in show woman straight surf

Epoch #3: kris strong , , , , , , , , , , , ... , , , , the the . . . . the . . . .

Epoch #7: kris strong , 35 , from clearwater , florida , decided to show his seven month old daughter how to ride the waves . the duo were filmed at home in the nursery room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor .

Epoch #10: kris strong , 35 , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , decided to show his seven month old daughter .

Epoch #15: kris strong , 35 , from clearwater , florida , were filmed at home in the nursery room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the

Epoch #17: kris strong , 35 , from clearwater , florida , decided to show his seven month old daughter . the duo were filmed at home in the nursery room as they writhed around on the floor , paddling their arms over the carpet .

Target: errol christie has revealed that he has been diagnosed with lung cancer . former european middleweight champion , 51 , has started chemotherapy . the boxer , who has never smoked , vowed that he would conquer cancer .

Epoch #5: son middleweight champion says 51 for a hospital back in what he is the toughest . he he is battling to make christie of the he says former his dad . explosive has it is battling lung by cancer has his first but . christie s treatment . christie , boxer , who has a smoked , but thought match in 1985 , he was , back that he has boxing cancer in he , said he determined it o and x christie in what he says is the toughest a match .

Epoch #10: former former european middleweight cancer **revealed he would had chemotherapy** , is promoting lung cancer . coping on 51 , but that he is a shock debut to focus on fighting he s . smoked suffered dermot , vowed to a hospital when of chemo and it was tough by knockout cancer .

Epoch #15: boxing champion errol christie revealed only starting my treatment results .

Target: errol christie has revealed that he has been diagnosed with lung cancer . former european middleweight champion , 51 , has started chemotherapy . the boxer , who has never smoked , vowed that he would conquer cancer .

Epoch #20: hospital is a shock of featherweight **christie s son louie** , who is also on each . **former european middleweight champion started chemotherapy** . christie fought to each session in london until last week

Epoch #25: louie , 51 , is former european middleweight christie and the boxer recently in 1985 . louie son louie is , vowed to create fit with their life s father . he is also due to set in a 10 defeats former european middleweight . under middleweight bout was filmed in 1985 , among louie has left by progress . i am just **i am just starting my treatment** but a is boxers while any cancer on return to our life .

Epoch #30: christie errol christie has said he will want to leave the fight . louie is first round of white european middleweight champion . 51 champion has denied chemotherapy his father in what he says he came . he has not be tough . accused of how we ve **had my first round of chemo** and he leads .

Target: kris strong , 35 , from clearwater , florida , decided to show his seven month old daughter how to surf . the beginner lesson took place in her nursery at home .

Epoch #1: father is the then clearwater writhed have says , is up learning . duck as the in . paddling doting up , action . filmed he waves . daughter the dude on way ride paddling to as they says board , . surf how strong in show woman straight surf

Epoch #3: kris strong , , , , , , , , , , ... , , , , the the . . . . the . . . .

Epoch #5: kris strong , 35 , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , decided to show his seven month old daughter . the duo were filmed at home in the nursery room as on the nursery room .

Epoch #7: kris strong , 35 , from clearwater , florida , decided to show his seven month old daughter how to **ride the waves** . the duo were filmed at home in the nursery room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor .

Target: kris strong , 35 , from clearwater , florida , decided to show his seven month old daughter how to surf . the beginner lesson took place in her nursery at home .

Epoch #10: kris strong , 35 , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , from clearwater , florida , decided to show his seven month old daughter .

Epoch #15: kris strong , 35 , from clearwater , florida , were filmed at home in the nursery room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the floor room as they writhed around on the

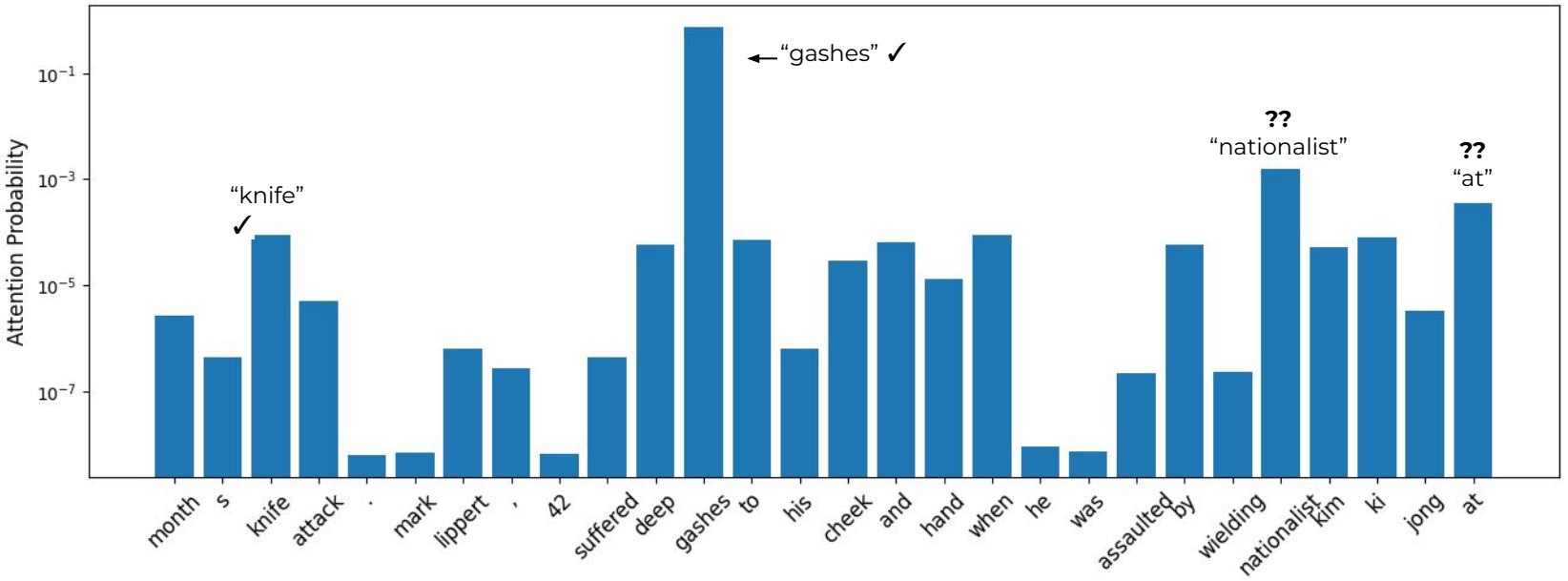
Epoch #17: kris strong , 35 , from clearwater , florida , decided to show his seven month old daughter . the duo were filmed at home in the nursery room as they writhed around on the floor , paddling their arms over the carpet .

coligado

Target: lea **[UNK]** , 21 , a stanford computer science student , is the creator of the blog showcasing talented women in tech . the profiling blog has already attracted women from big hitting tech companies like pinterest and **[UNK]** . the blog s inspiration , humans of new york , was started by new york photographer brandon stanton and features street portraits accompanied by quotes from the subjects .

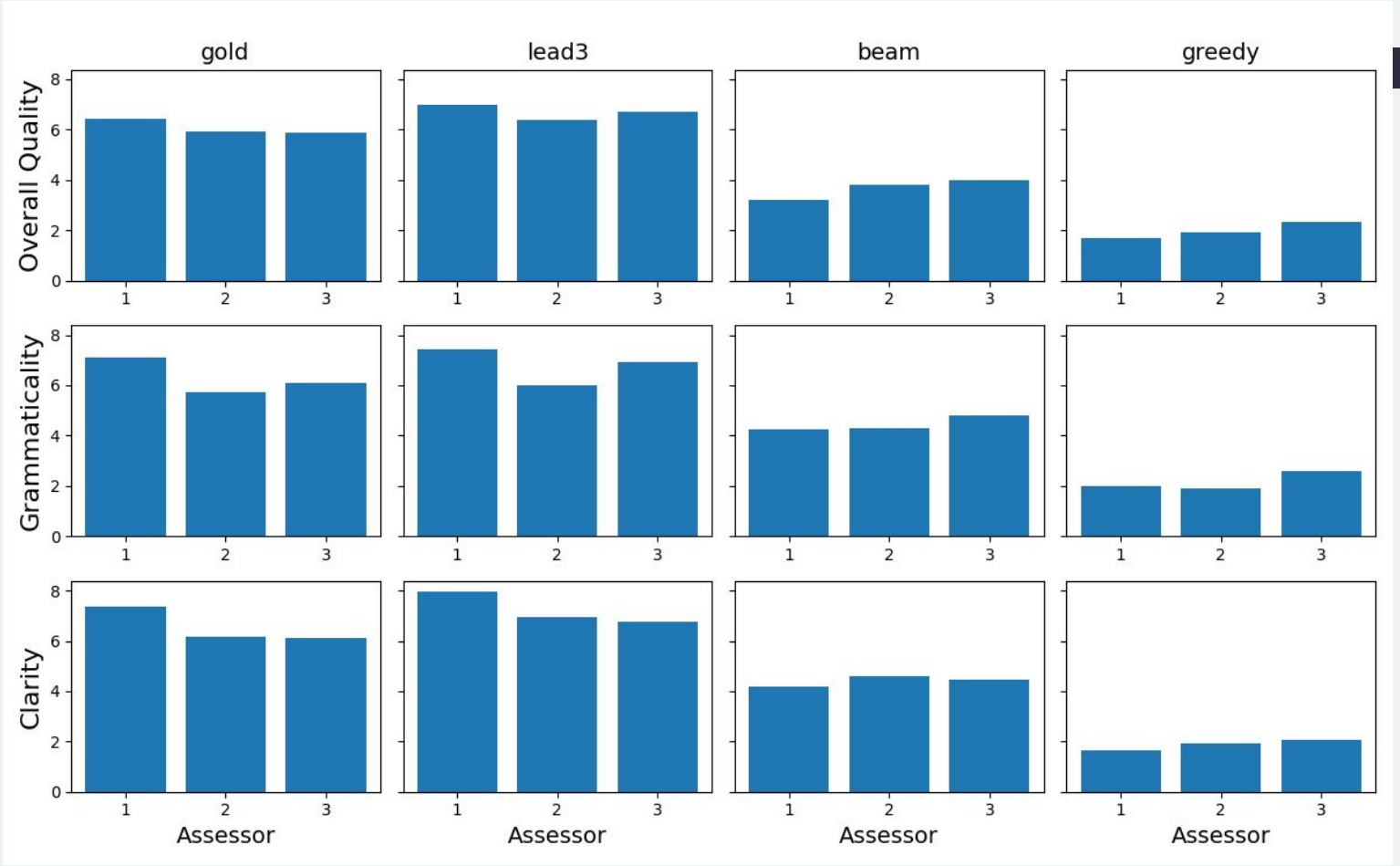
postmates

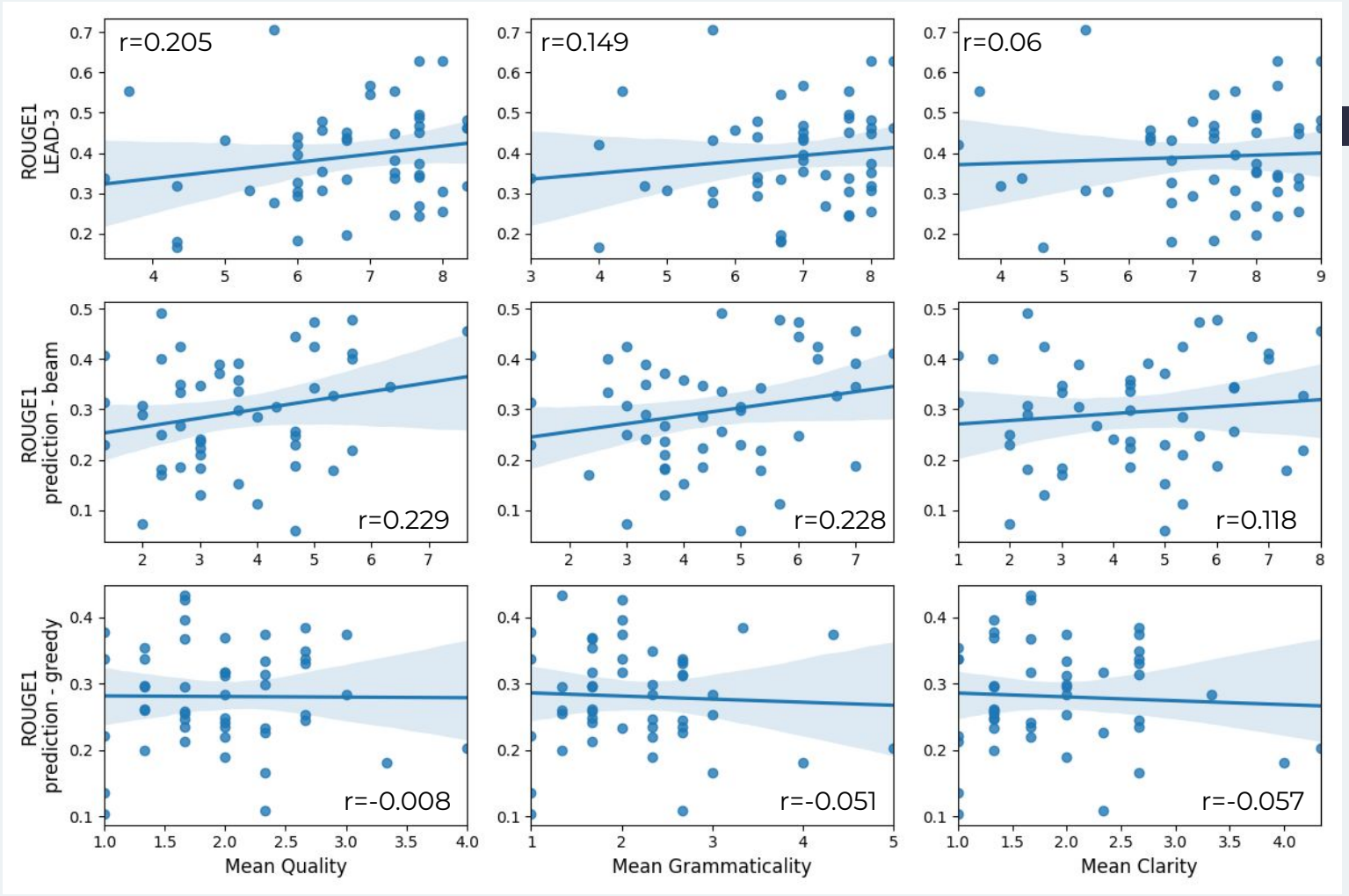
Prediction: stanford student lea **coligado** , 21 , created **women of silicon valley** . girls on top stanford computer science student lea coligado features talented women in the tech industry , such as **postmates exec sara mauskopf** .

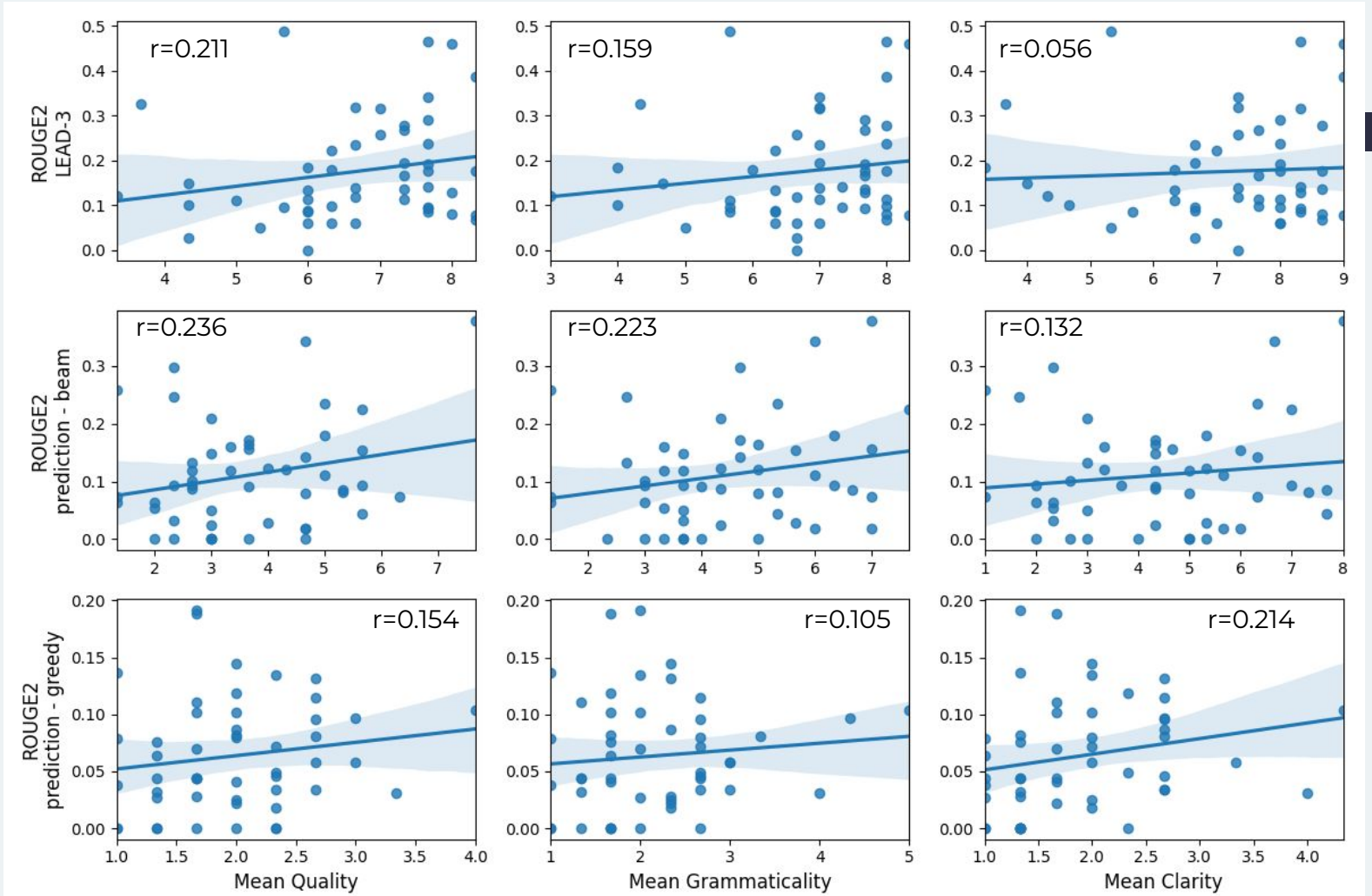


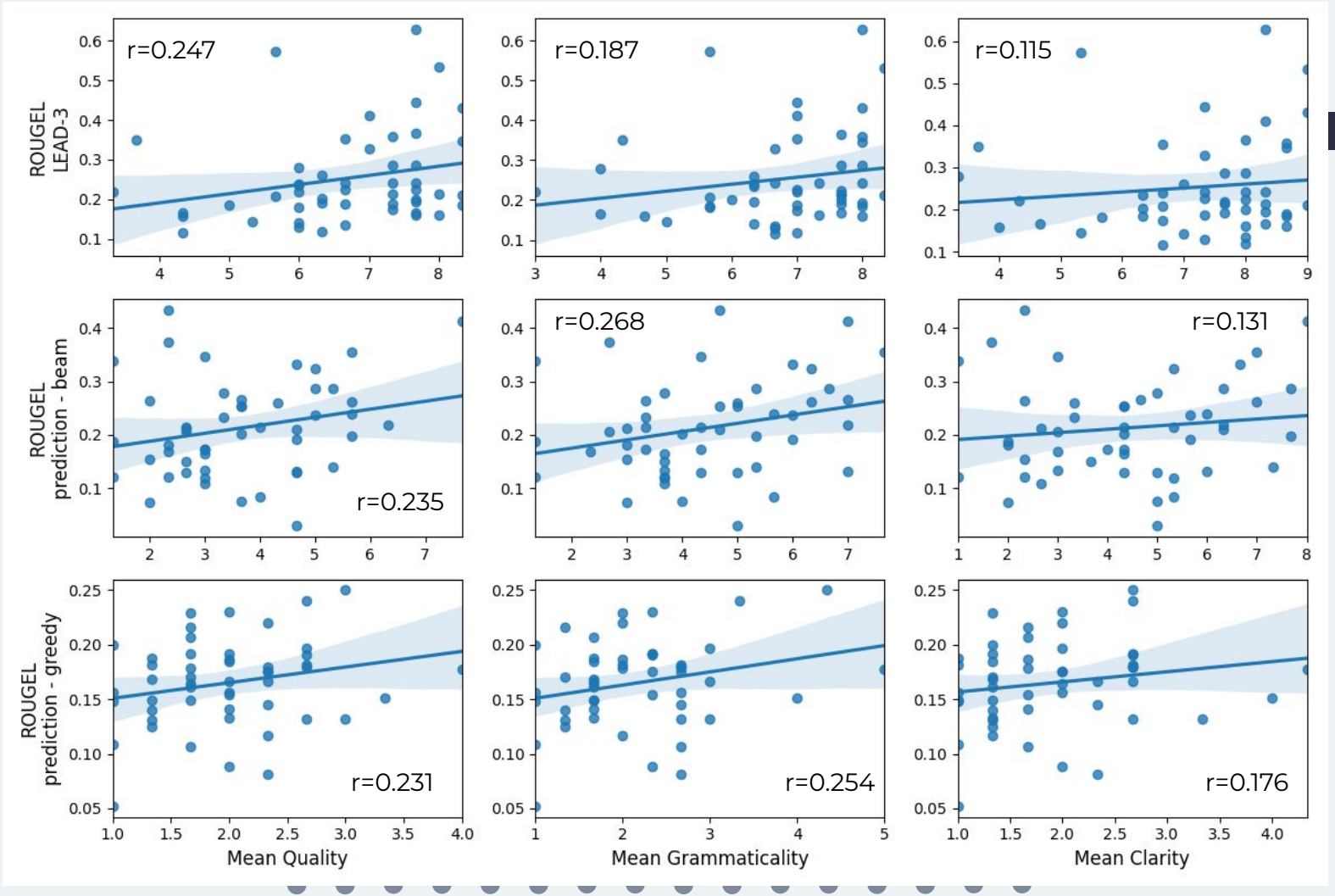
Target: lippert , 42 , suffered deep **[UNK]** to his hand when he was assaulted by knife wielding nationalist kim ki jong in central seoul last month . it is an amazing apparatus , one i haven t seen before so innovative and creative , he wrote on facebook **.**

Prediction: mark lippert , 42 , was suffered deep **gashes**...


gold
lead3
beam
greedy
Overall Quality
Grammaticality
Clarity
0
2
4
6
8
1
2
3
Assessor


r=0.205
r=0.149
r=0.06
r=0.229
r=0.228
r=0.118
r=-0.008
r=-0.051
r=-0.057
ROUGE1 LEAD-3
ROUGE1 prediction - beam
ROUGE1 prediction - greedy
Mean Quality
Mean Grammaticality
Mean Clarity


r=0.211
r=0.159
r=0.056
r=0.236
r=0.223
r=0.132
r=0.154
r=0.105
r=0.214
ROUGE2 LEAD-3
ROUGE2 prediction - beam
ROUGE2 prediction - greedy
Mean Quality
Mean Grammaticality
Mean Clarity


ROUGEL LEAD-3
r=0.247
r=0.187
r=0.115
ROUGEL prediction - beam
r=0.235
r=0.268
r=0.131
ROUGEL prediction - greedy
r=0.231
r=0.254
r=0.176
Mean Quality
Mean Grammaticality
Mean Clarity

Ethical Considerations

Benefits:

1. **Task**: More accessible news/science/etc.
2. **Task**: Could be used to generate plain language documents?
3. **Task**: Free up worker's time

Risks:

1. **Dataset**: Known gender bias
2. **Task**: Questionable accuracy/truthfulness of summaries
3. **Task**: Could be used to spread false/reductive information more easily?

CNN/Daily Mail Caveats

- Rather narrow scope: news articles only, may not generalize
- Analysis showed summaries skewed extractive [6]
- Limited summarization styles [6]
- Some gender bias present in text [7]
- Humans have difficulty with question answering task due to coreference errors/ambiguity in text [8]

Citations

[1] Elena Lloret, María Teresa Romá-Ferri, and Manuel Palomar. 2013. Compendium: A text summarization system for generating abstracts of research papers. Data Knowledge Engineering, 88:164 – 175.

[2] Murali Saravanan, Balaraman Ravindran, and Shivani Raman. 2006. Improving legal document summarization using graphical models. Frontiers in Artificial Intelligence and Applications, 152:51.

[3] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. TLDR: Extreme Summarization of Scientific Documents. 2020.

[4] Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. Improving Truthfulness of Headline Generation. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[5] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (*NIPS'15*). MIT Press, Cambridge, MA, USA, 1693–1701.

[6] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. 2018.

[7] Shikha Bordia and Samuel Bowman. Identifying and Reducing Gender Bias in Wor-Level Language Models. Association for Computational Linguistics. 2019.

[8] Danqi Chen, Jason Bolton and Christopher D. Manning. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. Association for Computational Linguistics. 2016.

[9] Josef Steinberger and Karel Ježek. Evaluation Measures for Text Summarization. Computing and Informatics. 2009.

[10] Ming Zhong, Pengfel Liu, Danquing Wang, Xipeng Qiu, and Xuanjing Huans. Extractive Summarization as Text Matching. Association for Computational Linguistics. 2020.

Citations, cont.

[11] Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram cooccurrence statistics. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.

[12] Abigail See, Peter Liu, and Christopher Manning,. 2017. Get To The Point: Summarization with Pointer-Generator Networks. 1073-1083. 10.18653/v1/P17-1099.