FALL 2020 CS 395T

TEXAS
The University of Texas at Austin

Get To The Point: Summarization with Pointer-Generator Networks tor Networks

Abigail See, Peter J. Liu, Christopher D. Manning (ACL 2017)

CS 395T: Topics in Natural Language Processing

Jiyang Zhang, Sept 15, The University of Texas at Austin

Summarization

Def: given input text x, write a summary y which is shorter and contains the main information of x

- Gigaword: first one or two sentences of a news article → headline (aka *sentence compression*)
- LCSTS (Chinese microblogging): paragraph → sentence summary
- NYT, CNN/DailyMail: news article → (multi)sentence summary
- Wikihow: full how-to article → summary sentences
- XSum: (Narayan et al., 2018), Newsroom: (Grusky et al., 2018): article → 1 sentence summary (New datasets!)

http://web.stanford.edu/class/cs224n/slides/cs224n-2020-lecture15-nlg.pdf

Two main strategies

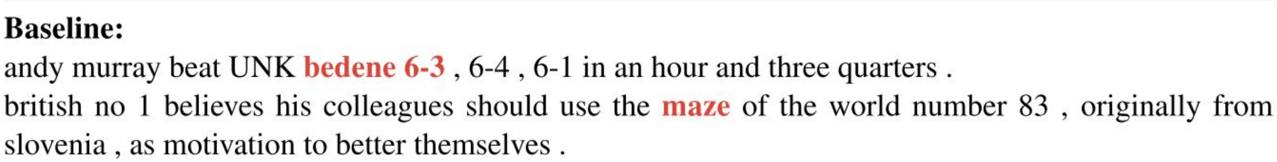
Extractive:

❖ Select sentences directly from the source text to be form a summary

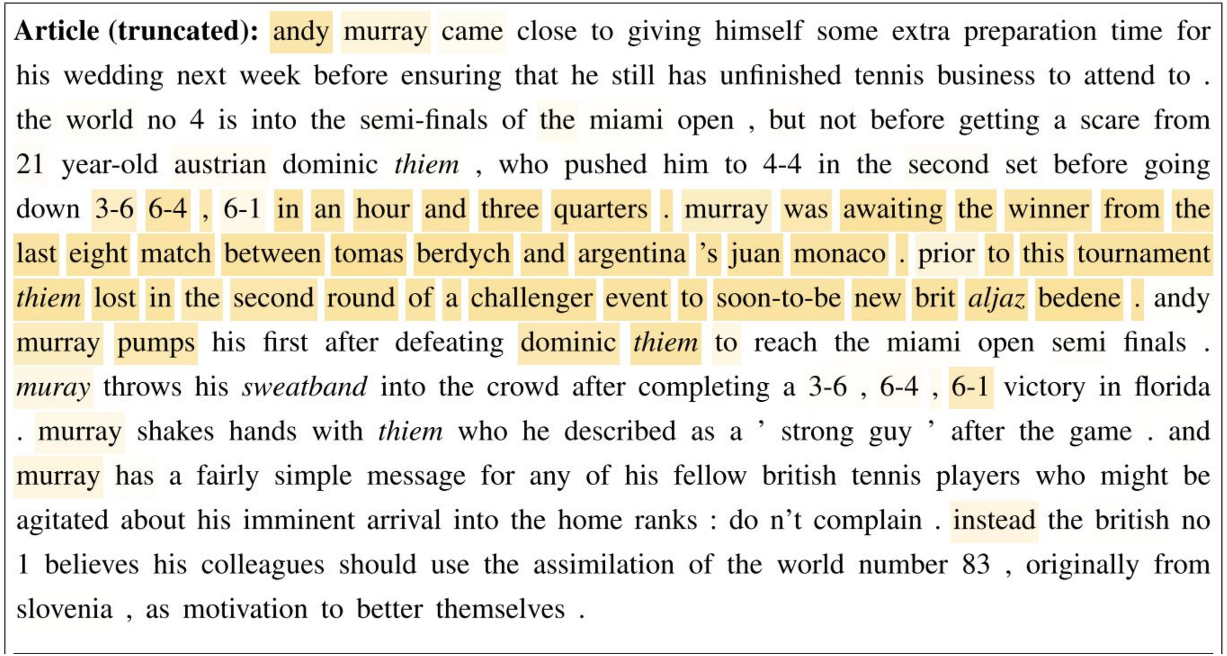
Abstractive:

❖ Generate novel words and phrases as a human-written abstract usually does.

Problems!

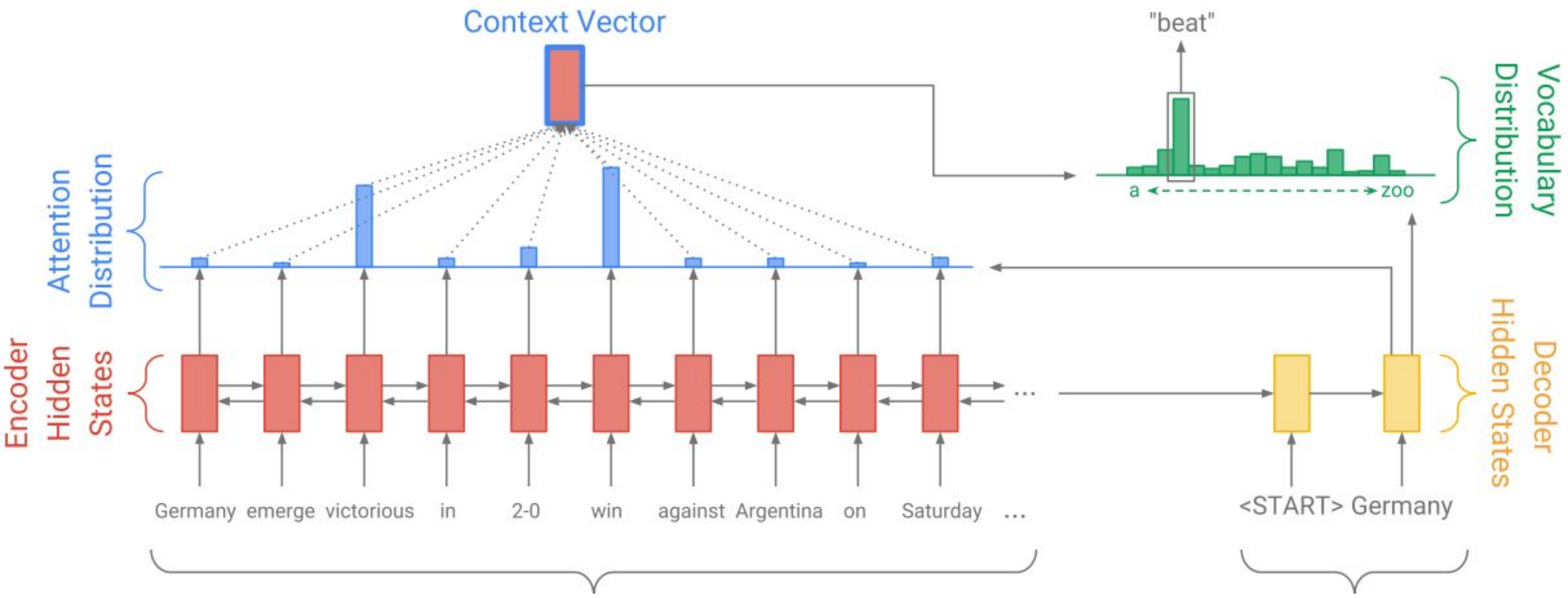


Baseline:
andy murray beat UNK **bedene 6-3** , 6-4 , 6-1 in an hour and three quarters .
british no 1 believes his colleagues should use the **maze** of the world number 83 , originally from slovenia , as motivation to better themselves .



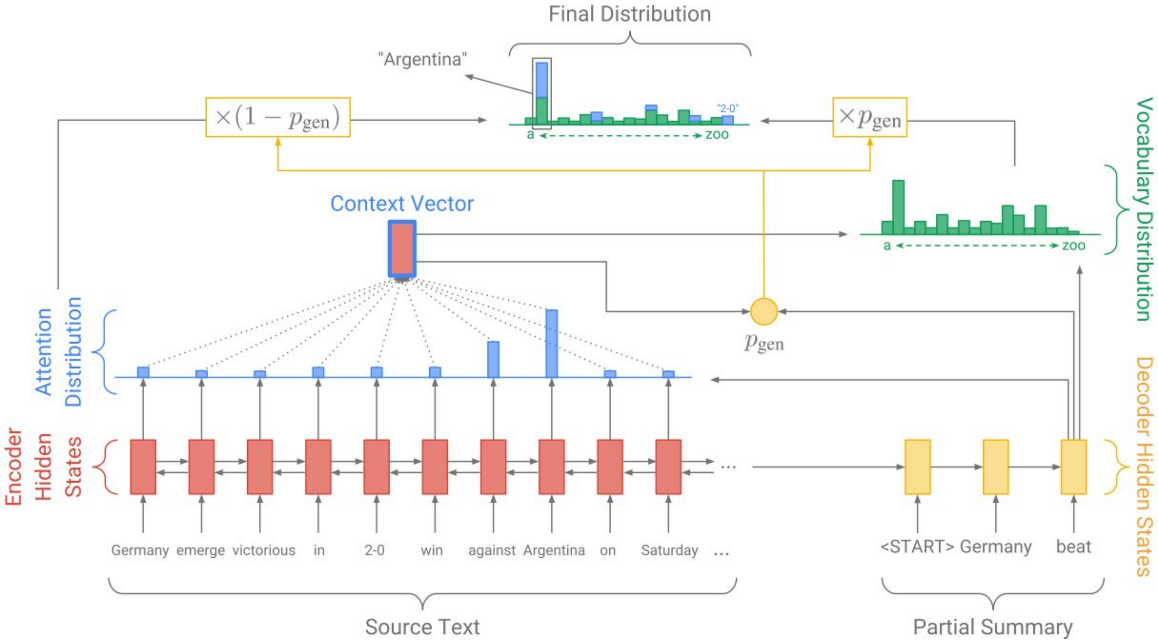
Article (truncated): andy murray came close to giving himself some extra preparation time for his wedding next week before ensuring that he still has unfinished tennis business to attend to . the world no 4 is into the semi-finals of the miami open , but not before getting a scare from 21 year-old austrian 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 .

Models: seq2seq

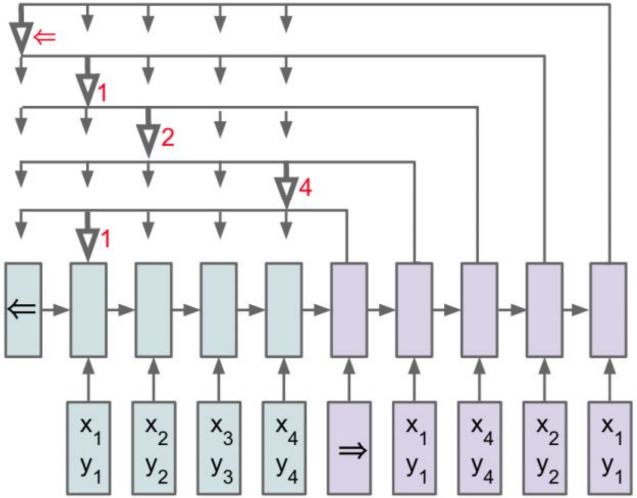


$$p_{\text{gen}} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + b_{\text{ptr}})$$

Models: Pointer-generator network



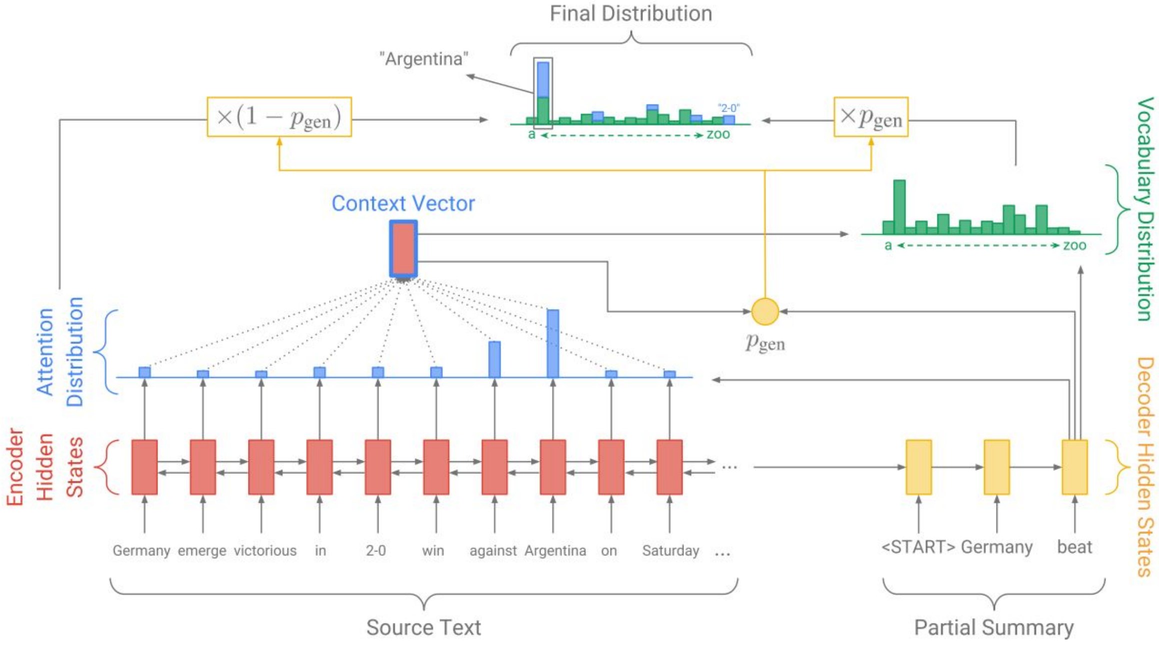
Pointer Network



https://arxiv.org/pdf/1506.03134.pdf

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

Models: Pointer-generator network



Coverage mechanism

- Coverage vector $c^t = \sum_{t'=0}^{t-1} a^{t'}$
- Coverage loss $\mathrm{covloss}_t = \sum_i \min(a_i^t, c_i^t)$
- Loss $\mathrm{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(a_i^t, c_i^t)$

Dataset

- CNN/Daily Mail dataset (Hermann et al., 2015; Nallapati et al., 2016),
- Online news articles (781 tokens on average) paired with multi-sentence summaries (3.75 sentences or 56 tokens on average)

Baselines

1. **Lead-3 baseline**:

   uses first three sentences of the article as summary
2. **Abstractive model** https://arxiv.org/pdf/1602.06023.pdf

   hierarchical networks
3. **Extractive model** https://arxiv.org/pdf/1611.04230.pdf

   RNN model as sequence classifier, a binary decision is made in terms of whether or not it should be included in the summary
4. **Seq2seq + attn (150k vocab)**
5. **Seq2seq + attn (50k vocab)**

Experiments

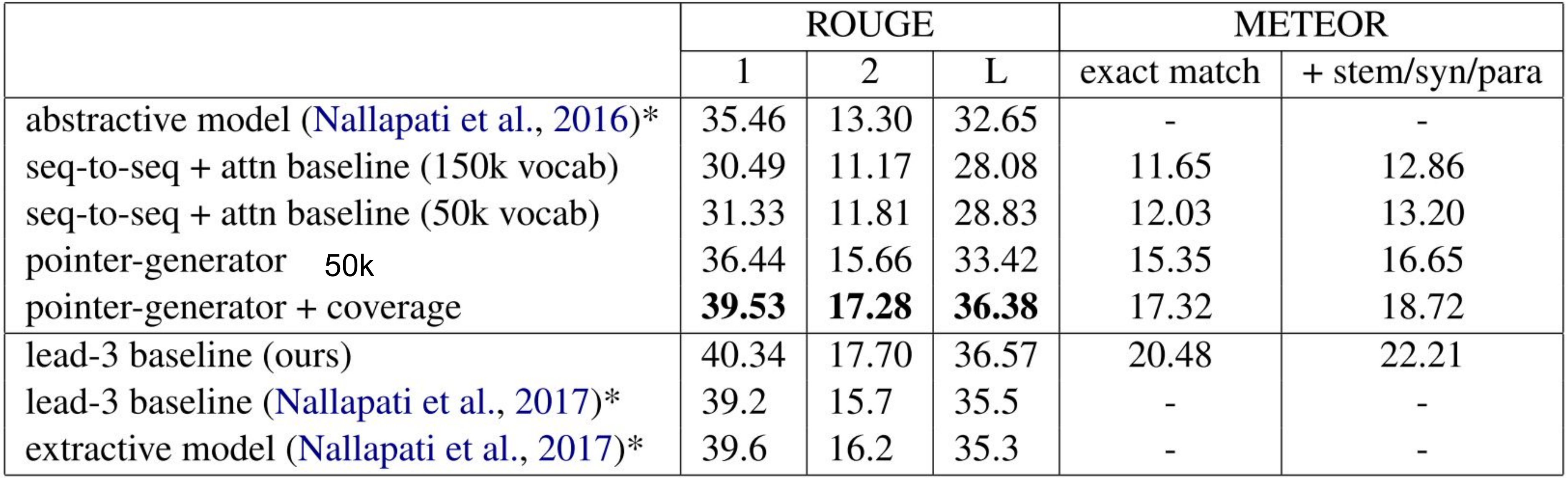
- Vocabulary size: pointer-generator model (50K), previous work(150K). Not using pre-trained embeddings.
- Only 1665 extra params (pointer, coverage) compare to 21499600 parms (baseline)
- Train models for 600,000 iterations (3 days and 4 hours), add coverage loss for further 3000 iterations (2 hrs). (Ineffective in other ways)

Metrics

ROUGE: F1 scores for ROUGE- 1, ROUGE-2 and ROUGE-L (which respectively measure the word-overlap, bigram-overlap, and longest common sequence between the reference summary and the summary to be evaluated).

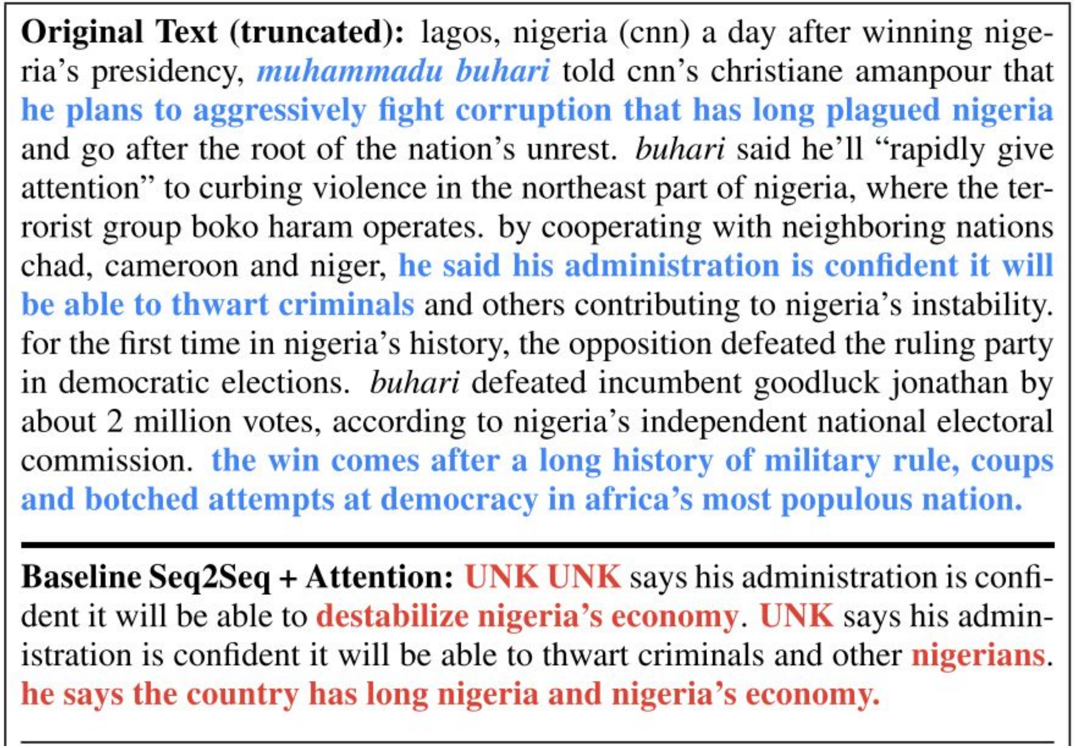
METEOR: both in exact match mode (rewarding only exact matches between words) and full mode (which additionally rewards matching stems, synonyms and paraphrases)

Results



	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 50k	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	-	-

Observations: baseline

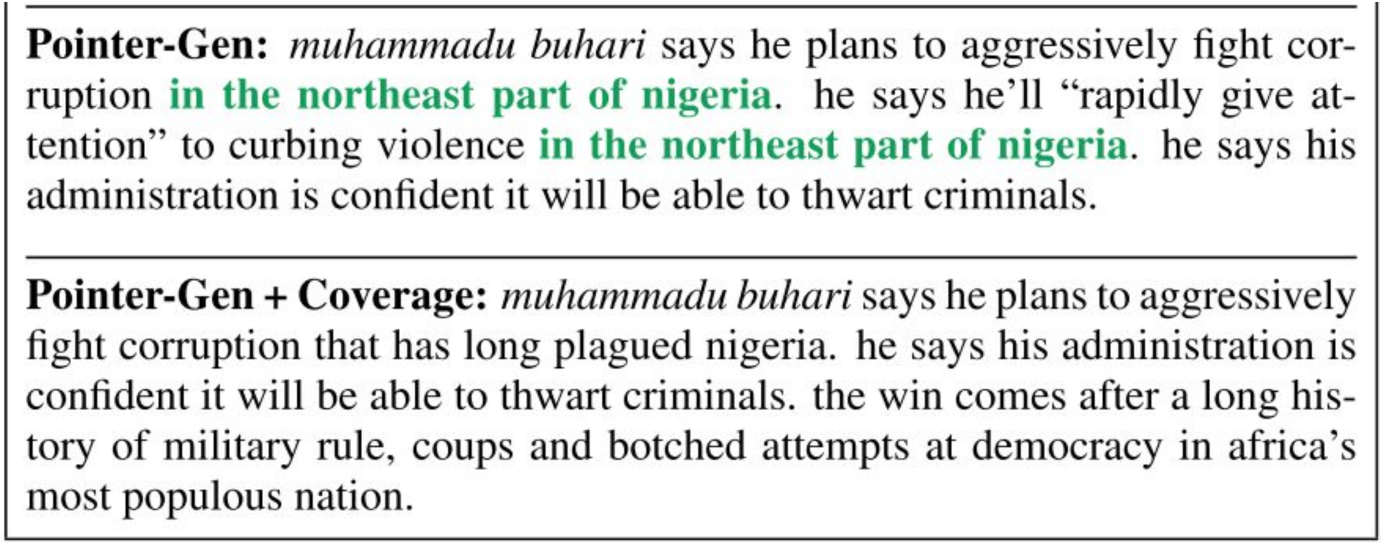


Original Text (truncated): lagos, nigeria (cnn) a day after winning nigeria's presidency, ***muhammadu buhari*** told cnn's christiane amanpour that **he plans to aggressively fight corruption that has long plagued nigeria** and go after the root of the nation's unrest. *buhari* said he'll "rapidly give 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.**

Struggle with the rare word
Repetition
Unable to produce OOV words

Observation:



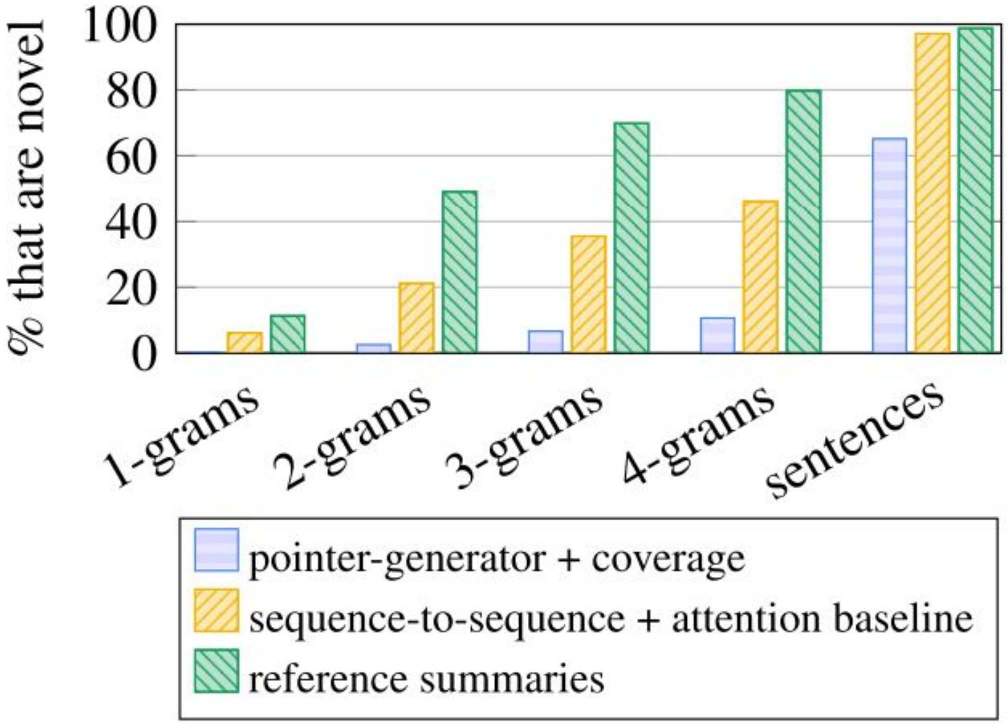
Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption **in the northeast part of nigeria**. he says he'll "rapidly give attention" to curbing violence **in the northeast part of nigeria**. he says his administration is confident it will be able to thwart criminals.

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.

Discussion: extractive system VS abstractive system

- Lead-3 baseline is extremely strong
  - News articles tend to be structured with the most important information at the start
  - The choice of content for the reference summaries is quite subjective
  - ROUGE rewards safe strategies such as selecting the first-appearing content, or preserving original phrasing.

Discussion: how abstractive?



Discussion: Copy or Gen

- During training: P_{gen} 0.3 → 0.53

  Test: P_{gen} → 0.17
- Allow model to stitch and truncate the sentence.

What this paper did well

- Combine generation and copy mechanism which is useful in task such as summarization. Very intuitive.
- Discussion part is great!
  - Analyze and explain the performance difference between results of abstractive and extractive models.
  - Analyze abstractiveness from different perspectives (% novel tokens and P_{gen})

Critiques and future work

- More analysis about the use of coverage loss, the hyper parameter lambda are needed.
- Human eval!
- Add abstraction into supervision in order to generate abstractive summary.
- How to avoid too much copy during inference and improve the quality.

Compared with BART

- More Data, more params = better performance!
- Pre-training with different objects that mimic the task of summarization.
- Higher score != better model

FALL 2020 CS 395T

TEXAS
The University of Texas at Austin

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer(ACL 2020)

CS 395T: Topics in Natural Language Processing

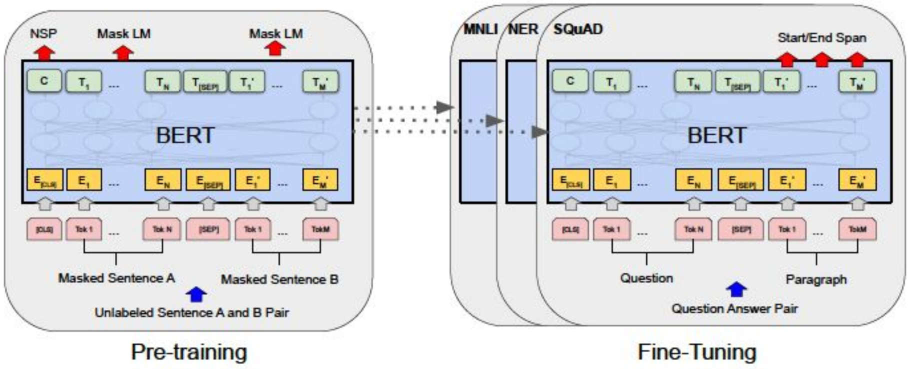
Isaac Buitrago, Sept 15, The University of Texas at Austin

Overview

- BART = Bidirectional Autoregressive Transformers
- Applications include question answering, article summarization, and translation
- Denoising autoencoder for pre training sequence to sequence models
- Similar to Google BERT, model for language understanding

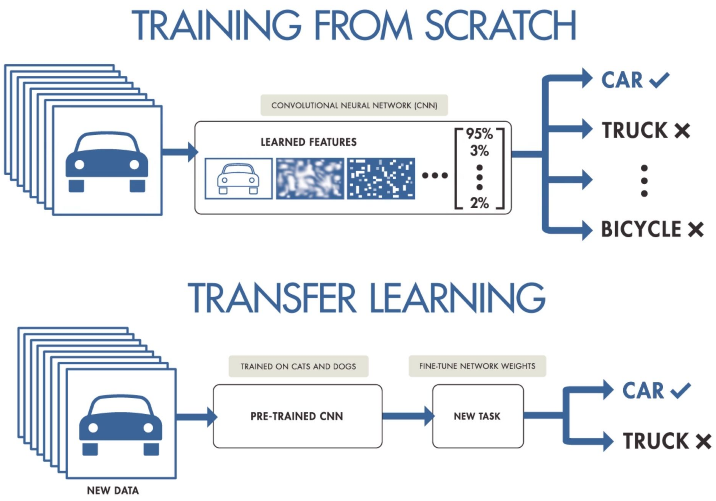
Architecture

- Uses standard sequence-to-sequence Transformer network
- Each layer of decoder performs cross-attention over final hidden layer of encoder
- No Feed forward network for word prediction
- 10% more params than BERT



Pre-training vs Fine Tuning

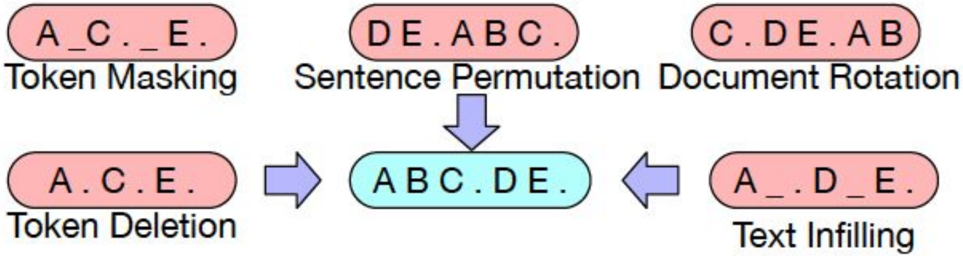
- Training an encoder/decoder to learn representation of data in UL fashion
- Use weights of a trained network as initialization for new but related task (transfer learning)



https://pl.pinterest.com/pin/672232681856247783/

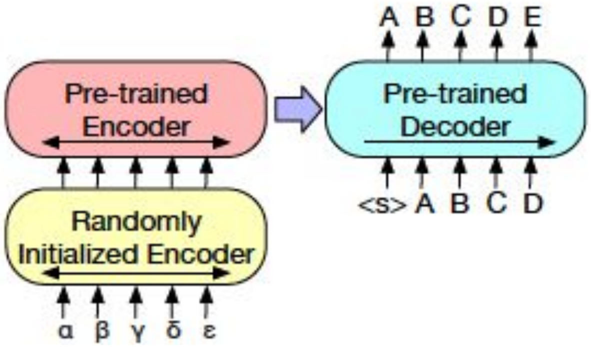
Pre-training

- Corrupt input sequence and minimize cross entropy between input and reconstruction
- Any noising scheme goes !
- Token masking - random token replaced with MASK
- Text infilling - model predicts how many tokens are missing from a span
- Document rotation - trains model to id start of document



Downstream tasks

- Sequence classification
- Token classification
- Machine Translation
  - Replace encoder embedding layer with new randomly initialized encoder
  - Freeze decoder params for n_0
  - Update decoder params for n_1 … N
  - Map foreign words to embeddings, denoise to English

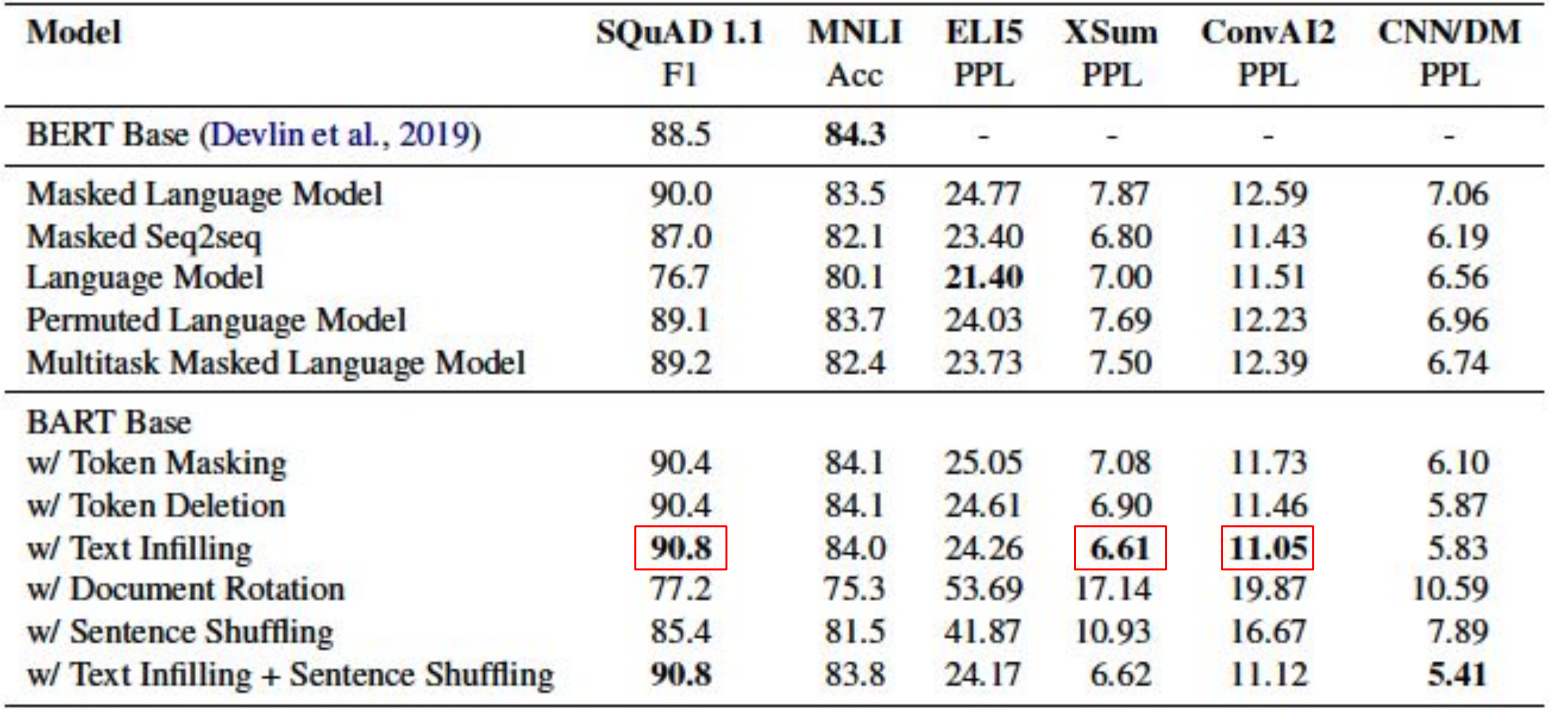


Learning Objective

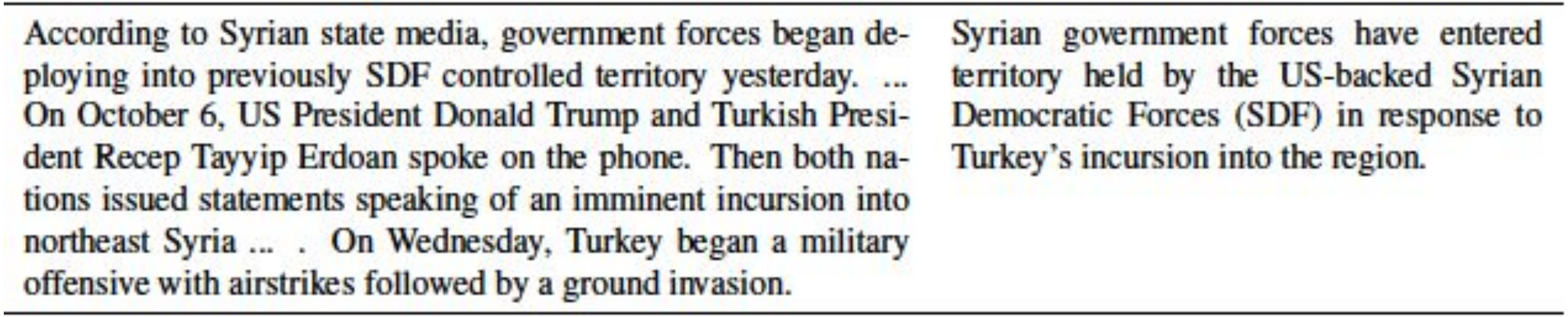
- Determine how noising schemes perform on certain tasks
- Compare BART to 5 pre-training baselines
- Baselines used on discriminative and generation tasks
- Permuted language model - Sample ⅙ tokens and generate in random order
- Masked sequence 2 sequence - Mask span with 50% of tokens and and train seq2seq model to predict masked tokens.

Experiments

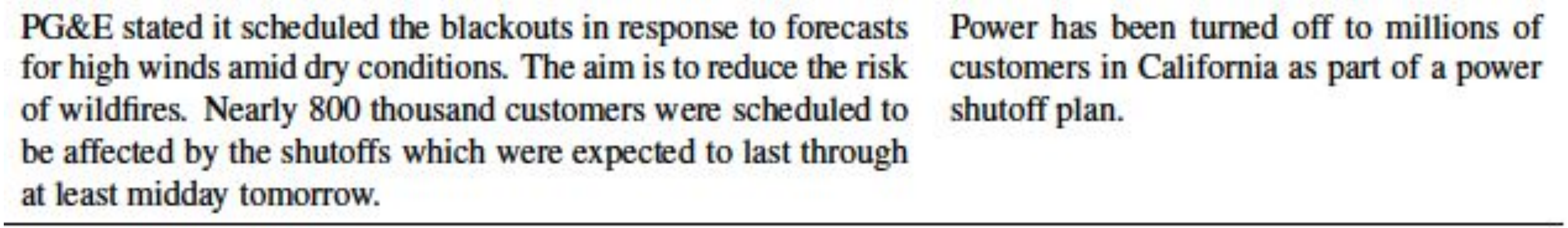
- 6 BART models with different document corruption methods
- 5 Pre-training objectives
- 6 tasks
  - SQuAD question answering task on a wikipedia dataset
  - Xsum news summarization
- Models trained for 1M steps on books and wikipedia data



Model	SQuAD 1.1 F1	MNLI Acc	ELI5 PPL	XSum PPL	ConvAI2 PPL	CNN/DM PPL
BERT Base (Devlin et al., 2019)	88.5	**84.3**	-	-	-	-
Masked Language Model	90.0	83.5	24.77	7.87	12.59	7.06
Masked Seq2seq	87.0	82.1	23.40	6.80	11.43	6.19
Language Model	76.7	80.1	**21.40**	7.00	11.51	6.56
Permuted Language Model	89.1	83.7	24.03	7.69	12.23	6.96
Multitask Masked Language Model	89.2	82.4	23.73	7.50	12.39	6.74
BART Base						
w/ Token Masking	90.4	84.1	25.05	7.08	11.73	6.10
w/ Token Deletion	90.4	84.1	24.61	6.90	11.46	5.87
w/ Text Infilling	**90.8**	84.0	24.26	**6.61**	**11.05**	5.83
w/ Document Rotation	77.2	75.3	53.69	17.14	19.87	10.59
w/ Sentence Shuffling	85.4	81.5	41.87	10.93	16.67	7.89
w/ Text Infilling + Sentence Shuffling	**90.8**	83.8	24.17	6.62	11.12	**5.41**



According to Syrian state media, government forces began deploying into previously SDF controlled territory yesterday. ... On October 6, US President Donald Trump and Turkish President Recep Tayyip Erdoan spoke on the phone. Then both nations issued statements speaking of an imminent incursion into northeast Syria ... . On Wednesday, Turkey began a military offensive with airstrikes followed by a ground invasion.	Syrian government forces have entered territory held by the US-backed Syrian Democratic Forces (SDF) in response to Turkey's incursion into the region.



PG&E stated it scheduled the blackouts in response to forecasts for high winds amid dry conditions. The aim is to reduce the risk of wildfires. Nearly 800 thousand customers were scheduled to be affected by the shutoffs which were expected to last through at least midday tomorrow.	Power has been turned off to millions of customers in California as part of a power shutoff plan.

Pros

- Training parameters and datasets are specified for reproduction of results.
- Visualization for machine translation made it easy to understand the process.
- Visuals of document corruption strategies clarified different inputs to the model.
- Simple language and ample sections made the paper easy to comprehend/navigate.

Cons

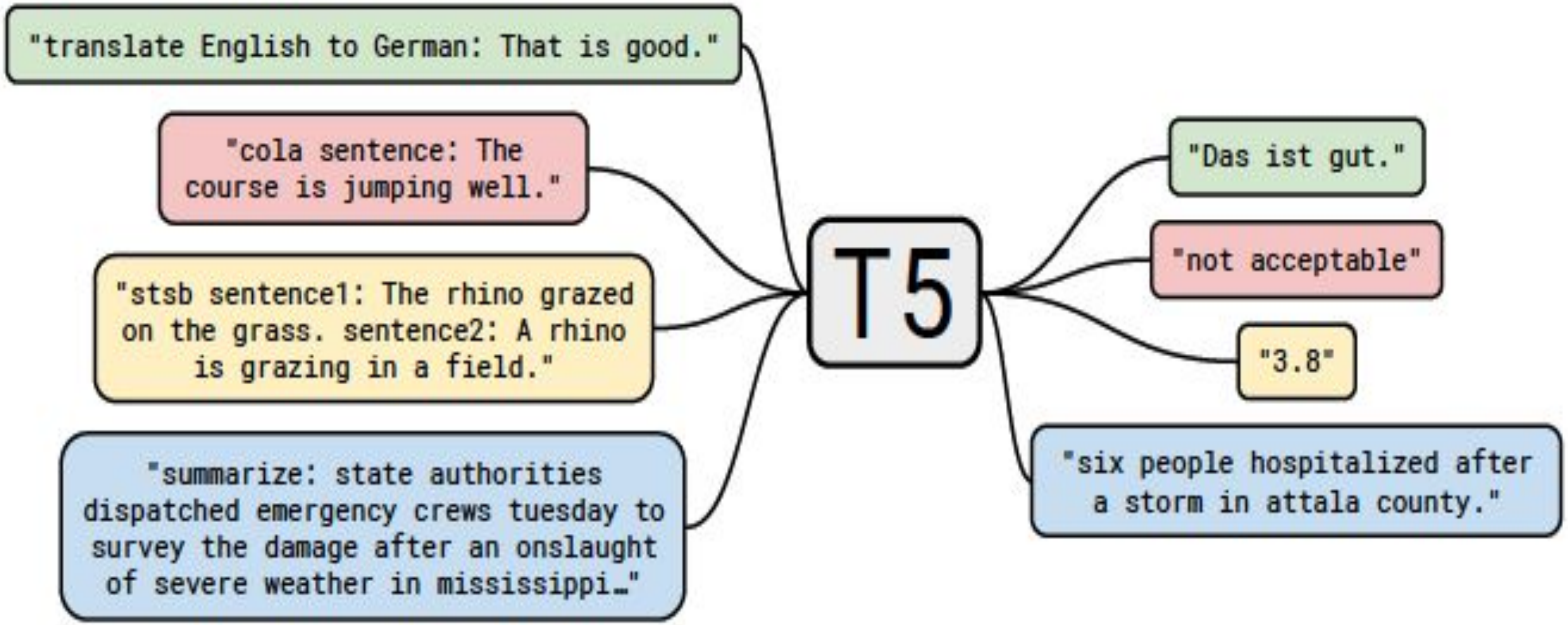
- Excluding visual for transformer architecture made it difficult to distinguish from BERT.
- Numerous evaluations with little specifics in why corruption methods succeed.

Related works

- GPT models leftward context
- ELMo does not pre-train interactions between features
- BERT introduces masked language modeling

Related works cont’d

- T5 (Text to Text Transfer Transformer)
- Apply similar model, learning objective, and decoding procedure to tasks
- Utilizes same objective function as BART (MLE)
- Task specific prefix added to input
- Skip connections used in encoder



"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

T5

"Das ist gut."

"not acceptable"

"3.8"

"six people hospitalized after a storm in attala county."