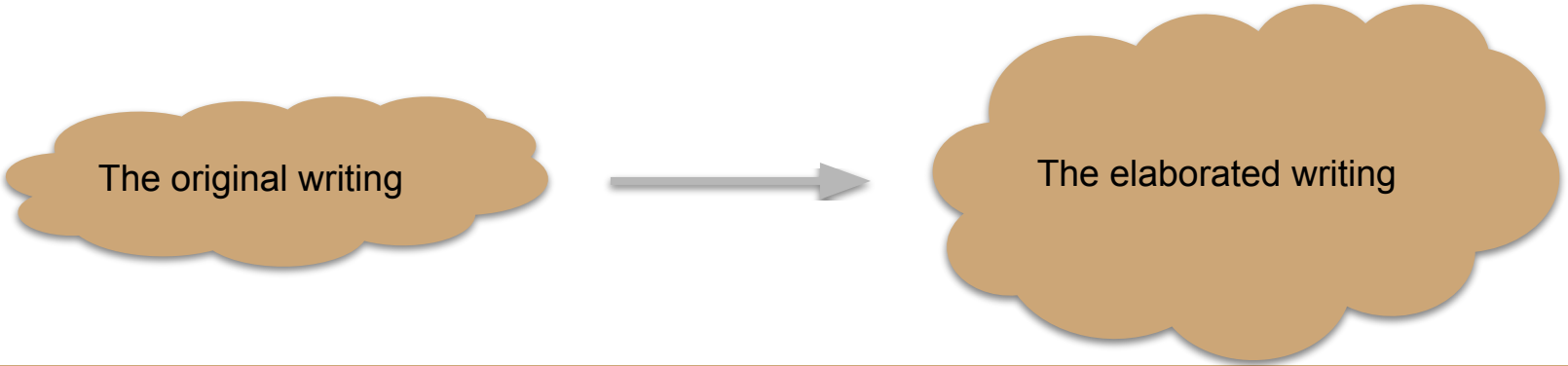
Save Your Words (Finals)

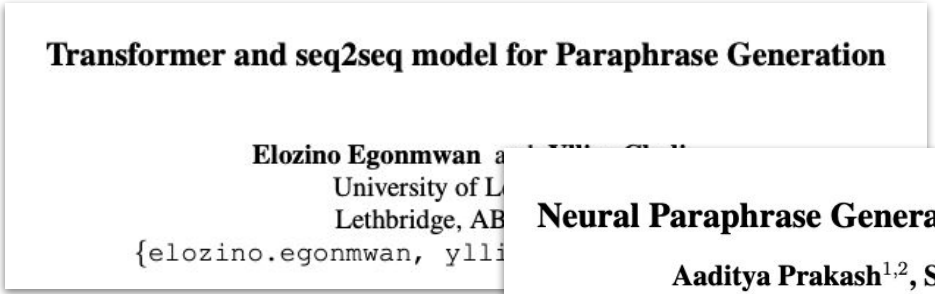
Nuttaree Busarapongpanich
Rashmi Jadhav
Yu-Wen Chen

The Objective

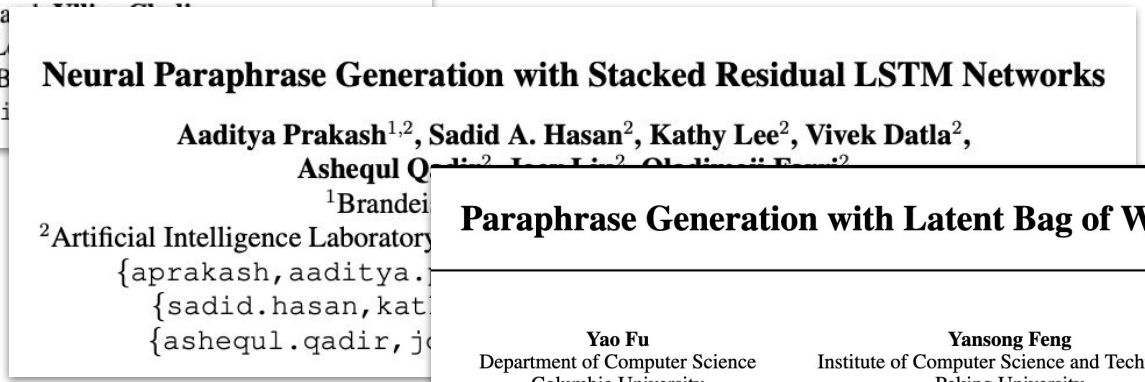
- Paraphrase sentences and make them longer
  - Preserve the original semantics
- Fun experiment to expand sentences
  - detailed and more text is useful for understanding a topic thoroughly



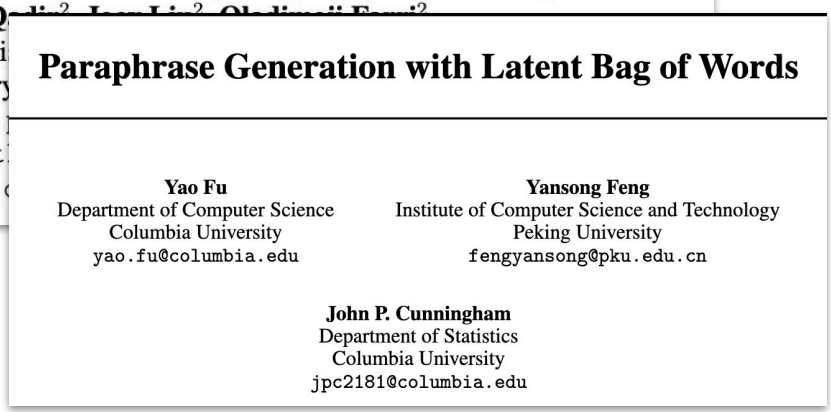
Related works



https://www.aclweb.org/anthology/D19-5627.pdf



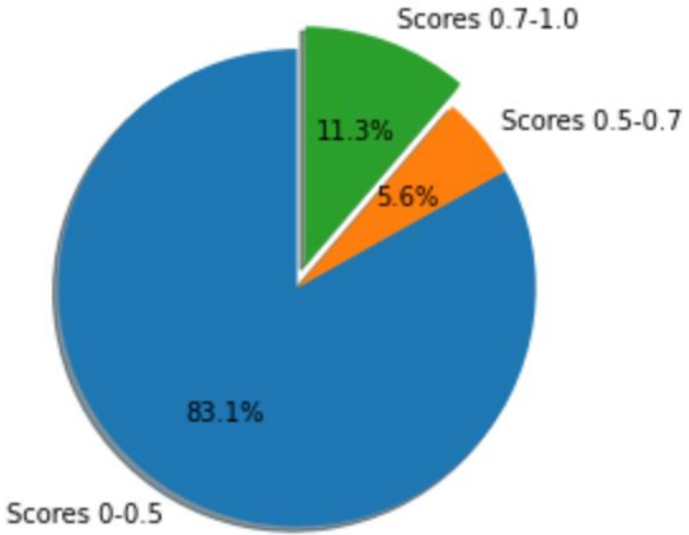
https://arxiv.org/pdf/1610.03098.pdf



https://arxiv.org/pdf/2001.01941v1.pdf

Dataset 1- Twitter LanguageNet

- Total Sentences in the dataset: 676,015
- Discard scores 0-0.7 to get better paraphrased data
- Train Split: 56,962 Val Split: 6,330 Test Split: 7,033



Score	Sentence1	Sentence 2
0.94043540837	Jeff Sessions fought...	Jeff Sessions Other Civil...
~~0.184755415394~~	~~Glenn Beck says that Donald..~~	~~He could be one of the most..~~

https://lanwuwei.github.io/Twitter-URL-Corpus

- Longest input sentence length: 122
- Longest target sentence length: 378

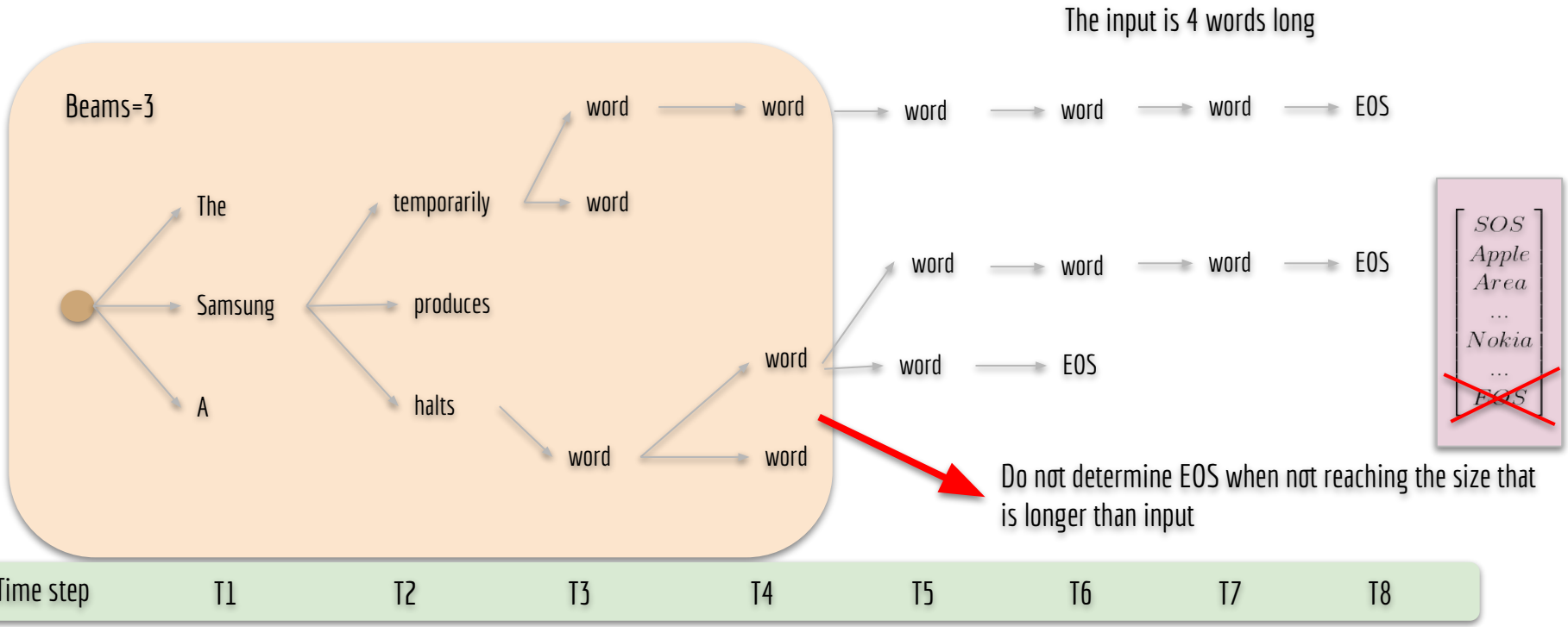
Dataset 2 - PAWS

- PAWS: Paraphrase Adversaries from Word Scrambling from Google
  - Total Sentences: 28,907
    - Train Split: 21,830 Val Split: 3,540 Test Split: 3,537
  - Longest input sentence length: 223
  - Longest target sentence length: 225

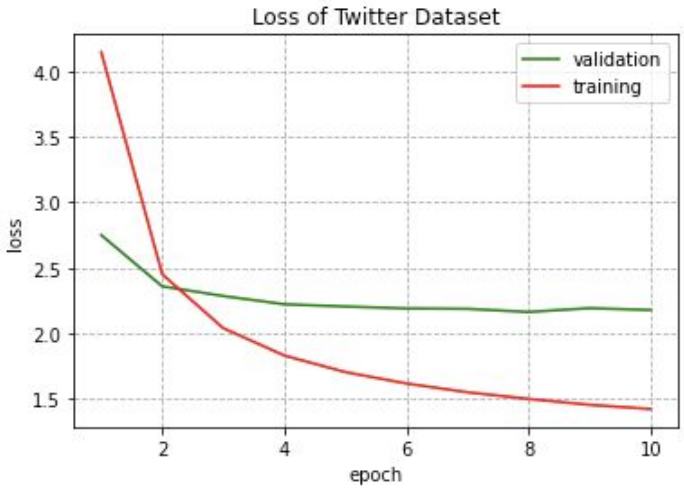
Sentence1	Sentence 2	Label
Although interchangeable...	Although they have different...	0
Katz was born in Sweden...	Katz was born in 1947...	1

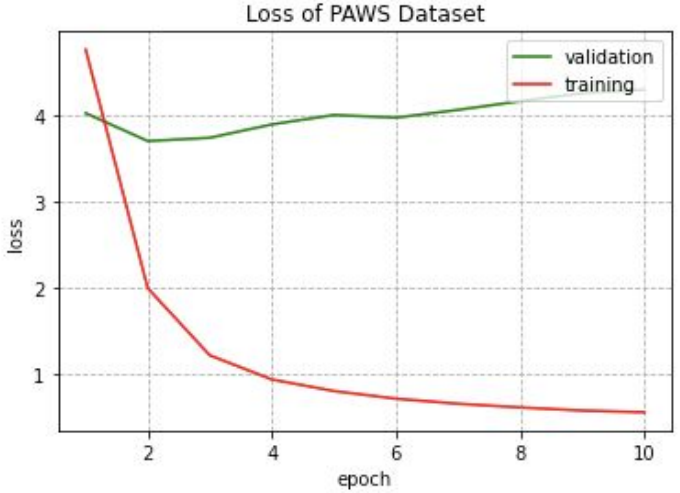
https://github.com/google-research-datasets/paws

Novelty of Project Contributions

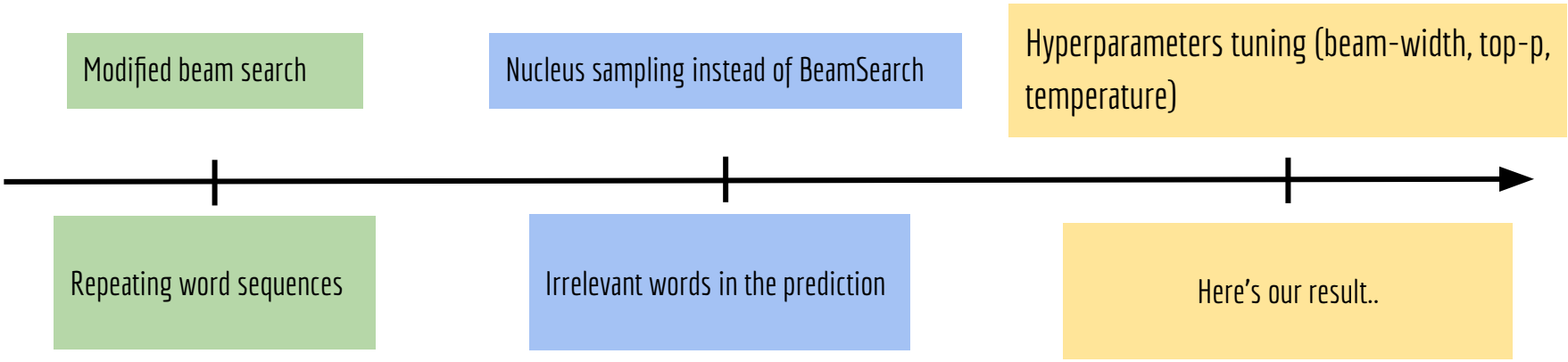


Training/validation loss





Implementation and Results



Here’s our result...

Twitter-dataset result:

Src: Rudy Giuliani has removed himself from consideration for a position in.
Prd: Rudy Giuliani pulls himself from consideration for a position in Donald Trump’s administration, a new york times upshot/siena college survey found.

Src: Silicon valley chiefs notably absent from Trumps cabinet of advisers.
Prd: Silicon valley chiefs notably absent from Trumps cabinet of business advisers , a smart person&rdquo for Trump’s cabinet of business advisers.

Here's our result...

PAWS-dataset result:

Src: His father emigrated to missouri in 1868, but returned when his wife became ill and before the rest of the family could go to america.
Prd: His father emigrated to missouri in 1868, but returned, his wife became and before the rest of the family could go to america, in 1868, the rest of the family could go and before the rest of the family could go to america.

Src: During her sophomore, junior and senior summers, she spent half of it with her alaska team, and half playing, and living in oregon.
Prd: during her, junior, junior and senior summers, she spent half of it with her alaska team and half, in oregon, and living ,in oregon ,and living in oregon, and living in oregon.

Problems in the data

- Twitter LanguageNet dataset:
  - Discarded majority of examples due to them not being good paraphrases
  - Each tweet has a word limit
  - Model exploits dataset problems like overusing hashtags and retweet handles
- PAWS Dataset:
  - doesn't have paraphrased longer sentences
  - Training couldn't be biased towards longer sentence lengths

Problem of repetitions in beam search

- Use **n-gram penalties**[1] to avoid same word sequences: manually set probability of next words that could create an already seen n-gram to 0
  - Still hard to control
  - e.g. 2-gram penalties: cannot produce words like “New York” repeatedly which might be required
  - Needs a lot of fine-tuning
- Use **sampling**; top-k/top-p with temperature scaling
  - Better in getting rid of repeatedness
  - Language generation using sampling is non-deterministic leading to incoherent gibberish
  - Top-p didn’t produce as good an output either

https://huggingface.co/blog/how-to-generate

Conclusions and Takeaways

- Not a reliable model
- In terms of the length of produced output sentence
- In terms of the paraphrasing ability of the model
- Encoder-Decoder Seq2Seq with attention might not be enough
- Try transformers in future
- Explore the GPT blackboxes

Thank you for your time and patience!

The sentence our model generated:

Stuck at your desk ? standing up and walking around for 5 minutes every hour could change a your time and despair