# ExtraPhraseRank: Length-Controlled Data Augmentation for Abstractive Summarization

Daniel Fein and Ruben Cuevas



## Background & Problem

- \* Document summarization is an important task with
- many real world applications.
  \* Two primary modes of summarization: extractive and abstractive.

## KEY PROBLEMS:

- A lack of document-summary pairs.
- \* Few methods of fully unsupervised summarization exist, and of those that do, most are purely
- extractive.
  \* For multi-document summarization, datasets are few and far between
- The length limit problem.
- \* ExtraPhrase: Succeeded at creating data that improved a transformer-based model's ability to generate summaries, but its dependency parsing-based system does not allow for control of output summary length.

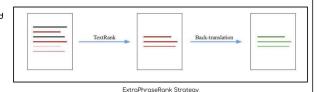
#### Methods

- \* Data Augmentation Method: Two-step approach TextRank
  - \* Back-Translation
- \* Baselines: heuristic for creating length-controlled abstractive pseudo-summaries, and finetune BART on these summaries.

	Example #1	Example #2	
Document	Amer opinione separts were called out to fold with a support package at the collision and the Neutronian Bank and realizing state of the collision and the Neutronian Bank and realizing state of the collision and the collision an	Room reports have listed some Prance-based players with returns to "The always foll—and this is with prayily had on mow this is not propagated." The property of the marroy of the prancy of the pra	
Gold Summary	A suspicious package left outside an Alliance Party office in east Belfast has been declared a houx.	New Welsh Rugby Union chairman Gareth Davies believes a joint £3.3m WRU-regions fund should be used to retain home-based talent such as Liam Williams, not bring back exiled stars.	
TextRank	Army explosives experts were asked to deal with a suspicious package in the offices of Newtownards Road on Friday night.	"I've always felt - and this is with my rugby hat on now; this is not region or WRU - I'd ruther spend that money on keeping players in Wales," said Davies.	
ExtraPhraseRank	Army explosives experts were asked to deal with a suspicious package in the offices of Newtownards Road on Friday night.	"I have always felt - and this is now with my rugby hat; this is not a region or WRU - I would rather spend this money to keep players in Wales." Davies said.	

#### Solution

- <u>ExtraPhraseRank</u>: a data augmentation method that generates pseudo-summaries from documents with fine control over word count.
- Enables the fine-tuning of pre-training large language models without document-summary



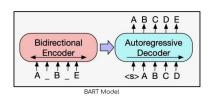
## Experiments

- We test fine-tuning a BART model on a variety of datasets
- Xsum: 1k sampled news articles/summaries out of 200k

- \* Amazon: Ik review/ synthetic summary pairs

  \* Evaluate ROUGE scores on XSum dataset across 7 experiments:

  \* Given no gold summaries: no-fine-tuning, heuristic, TextRank, ExtraPhraseRank
- \* Given some gold summaries: ½ gold and ½ synthetic, or only ½ gold
  Generate summaries with BART fine-tuned with ExtraPhraseRank for Amazon Dataset and manually evaluate



#### Results and Conclusion

- On the XSum dataset, fine-tuning ExtraPhraseRank led to a modest improvement when no training data exists, outperforming heuristic and TextRank alone
- Still performing neutristic and textrains alone

  \*Still performs poorly where gold summaries exist

  \*On Amazon dataset, the summaries that BART generated after fine-tuning with synthetic data were not generally better

  Conclusion: ExtraPhraseRank proved to be a mildly successful method of length controlled data-augmentation in most cases, but it has significant limitations, and other methods should continue to be explored

# Table 1: Experimental Results on XSum Dataset

	ROGUE-L	ROGUE-2	ROUGE-1
Gold	27.39	14.021	34.091
No Fine-tuning	10.971	1.311	13.563
Heuristic	10.539	1.046	14.055
TextRank	11.27	1.329	14.374
ExtraPhraseRank	12.008	1.482	14.972
1/2 ExtraPhraseRank & 1/2 Gold	15.383	3.32	19.018
1/2 Gold	10.994	25.202	30.768