Carnegie Mellon University

Keyphrase Generation on News Articles

Eric Zhou [1] Mengxi Sun [2]

[1]Carnegie Mellon University [2]University of Pittsburgh

Introduction

Extensive research on keyphrase generation has been focused on the academic article domain. However, the existing models, including state-of-the-art CopyRNN [2] and CatSeqD [7], achieve significaty worse results in the news domain, especially inadequate for generating absent keyphrase [4].

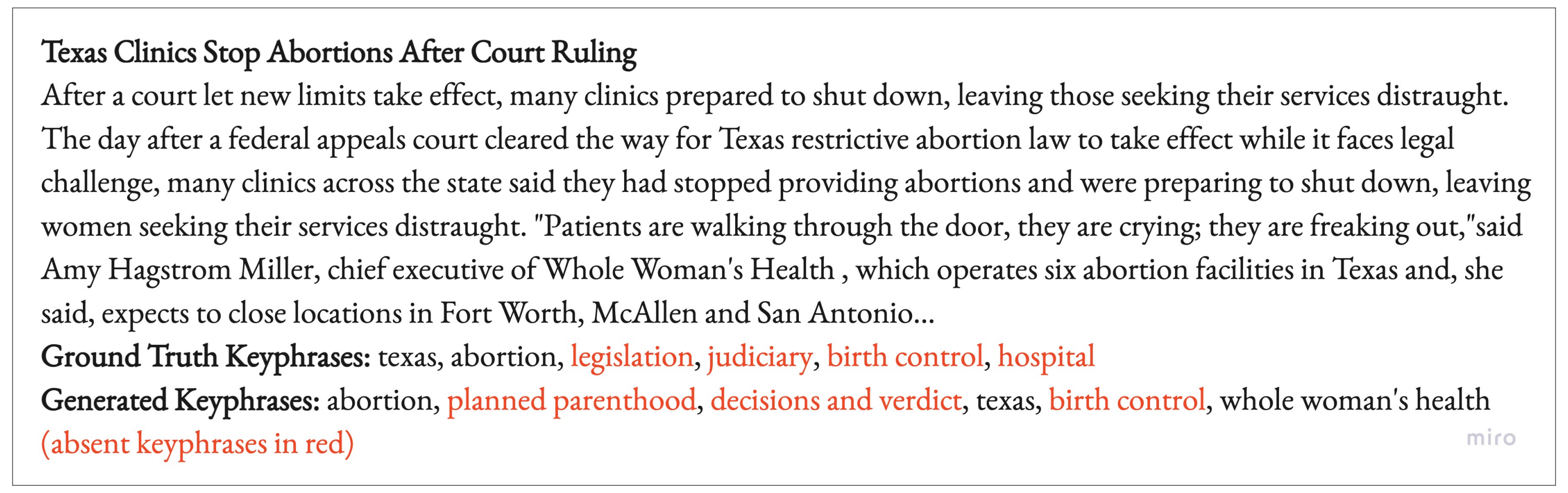
In this project, we propose **incorporating BART** to improve the performance of **the supervised (variable-length) keyphrase generation task in the news domain**.

We are particularly interested in generating keyphrases that are absent from the text, which branches out from the grounding information retrieval methods that extract topics and keyphrases present in the text.

Dataset

We will train, evaluate and compare our model to the baseline model (CatSeqD-transformer [4]) on the following new dataset.

- **KPTimes** [1]: a large new dataset in the news domain. First presented in 2019, KPTimes contains 279,923 New York Times articles paired with editor curated two to ten keyphrases, with **55**% of the assigned keyphrases absent from the text.
  The advantages of KPTimes over previous datasets in the news domain are the sufficiently large size for neural-based training.



Texas Clinics Stop Abortions After Court Ruling
After a court let new limits take effect, many clinics prepared to shut down, leaving those seeking their services distraught. The day after a federal appeals court cleared the way for Texas restrictive abortion law to take effect while it faces legal challenge, many clinics across the state said they had stopped providing abortions and were preparing to shut down, leaving women seeking their services distraught. "Patients are walking through the door, they are crying; they are freaking out,"said Amy Hagstrom Miller, chief executive of Whole Woman's Health , which operates six abortion facilities in Texas and, she said, expects to close locations in Fort Worth, McAllen and San Antonio...
Ground Truth Keyphrases: texas, abortion, legislation, judiciary, birth control, hospital
Generated Keyphrases: abortion, planned parenthood, decisions and verdict, texas, birth control, whole woman's health
(absent keyphrases in red)

Takeaways

1. **Sequence to sequence models** can be used for tasks traditionally given to classification or unsupervised topic modelling with very satisfactory results.
2. The keyphrase generation task can be seen as an instance of the **set generation task**, where the goal is to generate a set of variable-length strings given a source document.
3. **Language models** are able to generate relevant key phrases that are **absent** in the source text. This is their main advantage over traditional topic modeling methods.
4. Custom loss functions help in this particular kind of language generation task. We used a **copy mechanism for generating rare word tokens**, as well as orthogonal regularization and semantic coverage to **reduce the repetitiveness** of generated keyphrases.

Model

- Baseline Model: CatSeqD-transformer
  For our baseline model, we chose to re-implement a transformer version of CatSeqD [4]. The original CatSeqD was a GRU-based recurrent neural network. In our baseline, we replaced the GRUs with transformer blocks, so that the architecture is very similar to the transformer architecture in "Attention is All You Need".
- Our Extension: BART
  We propose to use a pretrained BART model that was finetuned on summarizing CNN news articles. The pretrained BART model is almost identical to the transformer architecture above, except:
  1. It uses byte-pair encoding (BPE) for subword tokenization instead of normal word tokenization.
  2. The model is larger. It uses 12 encoder and decoder layers instead of 6.
  3. It uses pretrained token vectors, as opposed to randomly initialized token vectors.

Evaluation Metrics

The first papers in 2017 to employ encoder-decoder keyphrase generation models used F_1 metrics to quantify the success of their models [5], and subsequent papers have added a recall metric for evaluating absent keyphrases [4, 7].

$$\mathbf{P}@k = \frac{\hat{\mathcal{Y}}_{:k} \cap \mathcal{Y}}{\hat{\mathcal{Y}}_{:k}}, \ \mathbf{R}@k = \frac{\hat{\mathcal{Y}}_{:k} \cap \mathcal{Y}}{\mathcal{Y}}, \ \mathbf{F}_1@k = \frac{2 * \mathbf{P}@k * \mathbf{R}@k}{\mathbf{R}@k + \mathbf{P}@k}$$

Results

We first reimplement the baseline model on the original dataset KP20k in the scientific article domain and then compare the baseline model to our model on the two new datasets:

Dataset	Model	Present ($\mathbf{F_1@\mathcal{O}}$)	Present ($\mathbf{F_1@10}$)	Absent ($\mathbf{R@50}$)
KP20k	CatSeqD (original)	36.2	29.0	15.0
KP20k	CatSeqD (reimple)	36.1	28.9	11.1
KPTimes	CatSeqD	53.1	41.5	25.3
KPTimes	Ours	TBD	TBD	TBD

Table 1. The Plan and Current Progress.

References

[1] Ygor Gallina, Florian Boudin, and Beatrice Daille. Kptimes: A large-scale dataset for keyphrase generation on news documents. *arXiv preprint arXiv:1911.12559*, 2019.

[2] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*, 2016.

[3] Luís Marujo, Anatole Gershman, Jaime G. Carbonell, Robert E. Frederking, and João Paulo Neto. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. *CoRR*, abs/1306.4886, 2013.

[4] Rui Meng, Xingdi Yuan, Tong Wang, Sanqiang Zhao, Adam Trischler, and Daqing He. An empirical study on neural keyphrase generation. *arXiv preprint arXiv:2009.10229*, 2020.

[5] Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. Deep keyphrase generation. *arXiv preprint arXiv:1704.06879*, 2017.

[6] Xiaojun Wan and Jianguo Xiao. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, page 855–860. AAAI Press, 2008.

[7] Xingdi Yuan, Tong Wang, Rui Meng, Khushboo Thaker, Peter Brusilovsky, Daqing He, and Adam Trischler. One size does not fit all: Generating and evaluating variable number of keyphrases. *arXiv preprint arXiv:1810.05241*, 2018.

https://github.com/eric-zhizu/OpenNMT-kpg-release

CMU 2021 Fall CS 11-711 Poster Session

ericzhou@andrew.cmu.edu, mengxisun@pitt.edu