Supplemental file

September 2019

1 Introduction

The supplemental files contains additional materials to help reviewers understand materials in the original paper.

2 Algorithm

Algorithm 1: Generative Adversarial Networks for keyphrase generation

Require: Generator G; Discriminator D; Training dataset S = (x,y)
Initialize $G \sim U(0,1)$ and $D \sim U(0,1)$
Pre-train G using MLE on S
Taking samples from G as negative and samples from S as positive, train D by minimizing binary cross entropy
while G *converges* **do**
    for i = *1:g-steps* **do**
        Generate keyphrase sequence $y = \{y_1, y_2, ..., y_m\} \sim G(\cdot|x)$
        Calculate rewards $R = \{R(y_1), R(y_2), ..., R(y_m)\}$ by passing the keyphrases through D
        Update G via policy gradient reinforcement learning
    for j = *1:d-steps* **do**
        Taking generated samples from G as negative and samples from S as positive train D
        Update D to minimize binary cross-entropy
Evaluate performance on testing datasets

3 Experiments

3.1 Implementational Details

The vocabulary consists of 100002 words.For the generator, we use an encoder-decoder model with copy mechanism along with bahdanau attention .The hidden and embedding dimensions for the generator are 150 and 300 respectively.The

first layer of the discriminator is made up of 2 double-layered bi-GRU units.Both of them share an embedding layer, with embedding dimensions being equal to 200 and hidden dimensions being 150. A dropout of 0.5 is applied in both the layers.We pretrain both the generator and discriminator using Adam optimizer at a learning rate of 0.001 with a batch size of 64 and 32 respectively.Whenever the loss stops converging we apply learning rate decay.During adversarial training, we switch to adagrad optimizer with a small learning rate ≈ 0.0005.

3.2 Datasets

The GAN model is trained on KP20k dataset and tested on standard bechmark datasets - NUS,Krapivin and Inspec datasets.

1. **Inspec**[1] - This dataset provides around 2,000 documents and corresponding sequences. The model is tested on the first 600 documents.
2. **Krapivin**[2] - The dataset provides around 2,304 documents with their author-assigned keyphrases. The model is tested on the first 800 documents.
3. **NUS**[3] - This dataset consists of 211 paper documents along with keyphrases sequences. Model performance is reported on the entire dataset.
4. **KP20k**[4] - The dataset is made up of 567,830 training documents and 20,000 testing and validation document-keyphrase pairs. We have used all 20,000 testing pairs to test the efficiency of our model.

3.3 Evaluation Metrics

The Evaluation Metrics used to evaluate how well our model performs on the test dataset include F1@5,F1@M and α-NDCG.

1. **α-nDCG**[5] - It is an extension of the DCG ranking and is used to measure the diversity of content generated.It works by penalizing redundant keyphrases and reawarding new keyphrases.
2. **F1@5**[6] - F1 scores are calculated for sets of 5 keyphrases each sample from the generator's output and the original dataset.The maximum F1 score is counted as F1@5 score.
3. **F1@M**[6] - When variable number of keyphrases are generated, F1@M helps calculate the F1-Score for all the keyphrases generated with respect to the ground truth keyphrases.

3.4 Keyphrase Examples

Example 1. *A research document in the computer-science domain.*
Title: *The efficacy of electronic telecommunications in fostering interpersonal relationships*

document: *The effectiveness of electronic telecommunications as a supplementary aid to instruction and as a communication link between students, and between students and instructors in fostering interpersonal relationships was explored in this study. more specifically, the impacts of e mail, one of the most accessible, convenient , and easy to use computer mediated communications , on student attitudes toward the instructor, group mates, and other classmates were investigated. a postest only experimental design was adopted. in total, prospective teachers enrolling in a computers in education course participated in the study for a whole semester. results from the study provided substantial evidence supporting e-mails' s beneficial effects on student attitudes toward the instructor and other classmates*
Catseq: *electronic telecommunications, interpersonal relationships, computer mediated communication, interpersonal relationship*
Catseq-RL: *electronic telecommunications, computer mediated communications, experimental design, telecommunications , internet, internet, technology*
GAN: *telecommunication, interpersonal relationships,computer mediated communication , student behaviour, e-mail benefits, technology*
True Keyphrases: *telecommunications, interpersonal relationships, e-mail, computer mediated communications, student attitudes, computers in education course, student communication link, educational technology*

Example 2. *A research paper in the electrical engineering domain.*
Title: *Stability radii of systems with stochastic uncertainty and their optimization by output feedback .*
document:*We consider linear plants controlled by dynamic output feedback which are subjected to stochastic parameter perturbations. the stability radii of these systems are characterized, and it is shown that, for real data, the real and the complex stability radii coincide. A corresponding result does not hold in the deterministic case, even for perturbations of single output feedback type. in a second part of the paper we study the problem of optimizing the stability radius by dynamic linear output feedback. Necessary and sufficient conditions are derived for the existence of a compensator which achieves a sub-optimal stability radius . these conditions consist of a parameterized riccati equation , a parameterized liapunov inequality , a coupling inequality , and a number of linear matrix inequalities (one for each disturbance term). the corresponding problem in the deterministic case , the optimal mu synthesis problem , is still unsolved .*
Catseq: *stability radii, stochastic uncertainty, output feedback, parametrized riccati equation, parametrized liapunov inequality, coupling inequality, coupling inequality, stochastic*
Catseq-RL: *stability radii, stochastic uncertainty,output feedback,dynamic output feedback,linear matrix inequalities, uncertainty, stochastic systems, stochastic programming*
GAN: *stability radii, stochastic uncertainty, dynamic output feedback, ricatti inequality, matrix inequality, stochastic systems, noise*

True Keyphrases: *dynamic output feedback, stability radius, linear matrix inequalities, riccati inequalities, state dependent noise, scaling, stochastic systems, multiperturbations*

References

[1] Anette Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 216–223, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[2] Mikalai Krapivin and Maurizio Marchese. Large dataset for keyphrase extraction. 2009.

[3] Tho Thi Ngoc Le, Minh Le Nguyen, and Akira Shimazu. Unsupervised keyphrase extraction: Introducing new kinds of words to keyphrases. In *Australasian Conference on Artificial Intelligence*, 2016.

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

[5] Charles Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. *Proc. of the 31st ACM SIGIR*, pages 659–666, 01 2008.

[6] Xingdi Yuan, Tong Wang, Rui Meng, Khushboo Thaker, Daqing He, and Adam Trischler. Generating diverse numbers of diverse keyphrases. *ArXiv*, abs/1810.05241, 2018.