Information Summarization and Topics generation

DS 595 Natural Language Processing

Aukkawut Ammartayakun

November 11, 2021

Worcester Polytechnic Institute

Outline

1. References
2. Introduction
3. Methodology

References

Reference i

- Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., and Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond.
  In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 280–290, Berlin, Germany. Association for Computational Linguistics.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020).
  Exploring the limits of transfer learning with a unified text-to-text transformer.

Reference ii

Journal of Machine Learning Research, 21(140):1–67.

- See, A., Liu, P. J., and Manning, C. D. (2017).
  Get to the point: Summarization with pointer-generator networks.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014).
  Sequence to sequence learning with neural networks.
  In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc.

Introduction

Introduction

- There are a lot of

Problem Statement

Given the document D, we want to generate string h_D such that it summarizes the content in D with accurate information and concise.

Problem Statement

Given the document D, we want to generate string h_D such that it summarizes the content in D with accurate information and concise.

Example

D: At the request of the international war crimes court in The Hague, Bosnian Serb police seized the passports of the wife, son, daughter and son-in-law of Radovan Karadzic, the Bosnian Serb leader during the country's war.

h_D: Bosnia: Police Take Passports of Karadzic Family

Methodology

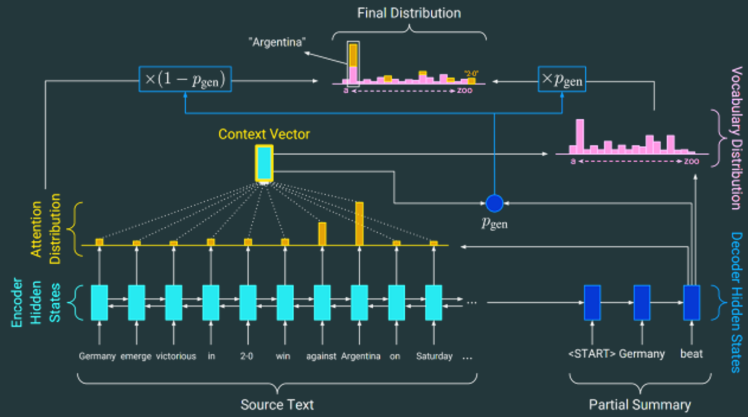
Literature Review

- As suggested by [Sutskever et al., 2014], RNN model can yield us the promising result but it also yield "artifacts" to the result.

Literature Review

- As suggested by [Sutskever et al., 2014], RNN model can yield us the promising result but it also yield "artifacts" to the result.
- Bidirectional RNN [Nallapati et al., 2016] shows the better performance.

Architecture



Use bidirectional LSTM along with attention to encode and generate context vector [See et al., 2017].

Architecture

- What if we combined the method proposed by [See et al., 2017] or [Raffel et al., 2020] with generative model?

Architecture

- What if we combined the method proposed by [See et al., 2017] or [Raffel et al., 2020] with generative model?
- That is, what if we treat [See et al., 2017] or [Raffel et al., 2020] model as the generator and we then create discriminator on top of that?

Dataset

- NYT news dataset

Dataset

- NYT news dataset
- >100K entries of title, topic, abstract, keywords
- Problem: Some punctuation might contains semantic value

Evaluation

- Metric of evaluation: F-1 measure (based on BLEU and ROUGE)