Information Retrieval and Extraction

Auto Annotation of Documents

Scope Document

Tanmay Sachan 2018111023

Akshat Chhajer 2018114008

Hitesh Kumar 2019201039

Pradeep Kumar Musham 2019201055

Introduction of project

Our project focuses on generation of comments for documents. We will be looking into various models and architectures commonly used for generation, and we will be fine tuning them over the NYT comments dataset.

Dataset and description

We will be using the NYT comments dataset from kaggle (https://www.kaggle.com/aashita/nyt-comments?select=ArticlesApril2017.csv)

The data contains information about the comments made on the articles published in New York Times in Jan-May 2017 and Jan-April 2018. The month-wise data is given in two csv files - one each for the articles on which comments were made and for the comments themselves. The csv files for comments contain over 2 million comments in total with 34 features and those for articles contain 16 features about more than 9,000 articles.

Another appropriate dataset can be used for fine tuning GPT-2. (https://www.kaggle.com/reddit/reddit-comments-may-2015)

This consists of 1.7 billion comments scraped from reddit.

Implementation Details

The implementation would be done in the form of a 2 fold approach.

Stage 1

Initially we will start off by implementing common keyword extraction models to gather keywords relevant to the document. The keywords extracted will then act as the basis for generating the annotations, and will provide context to GPT-2 later on. With named entity recognition, we can tag the relevant named entities with what they represent and gain further insight into the data. Then we can weigh different keywords differently based on their importance (as per their tagged classes), with respect to the NYT comments dataset. This will produce more relevant comments.

Stage 2

Next, we will use extractive summarisation to gather relevant sentences from the document and then run GPT-2 generation to generate text relevant to each extracted sentence. For extraction of sentences, we can fine tune BERT on the same NYT dataset on the news articles. The GPT-2 will be further fine tuned using the NYT comments dataset, so that the model understands the structure of the comments and generates accordingly. This approach will also help us score the text generation more appropriately due to a more compressed context with the help of summarisation.

Literature review

We will be implementing state of the art transformer based models, which have shown to outperform other models in the recent 2018-2019 nlp era. BERT consists of encoder transformers and is capable of gathering context in a document from either side of the keyword to be predicted, at the expense of being non auto-regressive. GPT-2 on the other hand consists of decoder transformers only, and is auto-regressive, i.e, everytime it generates a token it takes into context all the text before it. GPT-2 excels at generating text given some initial context, and our goal throughout the project would be to improve this generation part such that the scope of the generation stays within the context of the document.

GPT-3 is an even stronger model with enhanced capabilities, but access to GPT-3 is not public at the moment fearing misuse by the public.

Reading Materials

- Transformer
  - https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04
- BERT
  - https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
  - (For summarisation) https://github.com/nlpyang/PreSumm
- GPT
  - https://openai.com/blog/better-language-models/
  - https://jalammar.github.io/illustrated-gpt2/
- Research Papers
  - https://xin-xia.github.io/publication/icpc182.pdf
  - https://arxiv.org/pdf/2005.09123.pdf
  - https://arxiv.org/pdf/2005.14165.pdf

Milestones

1. Implement a rudimentary keyword extraction on the documents, to get some idea of the dataset and its content, and to be used in further models.
2. Try to fine tune GPT-2 over the NYT times dataset (on the comments only, so that it learns to generate some arbitrary comments based on the context that can be provided using keyword extraction).

3. Work on the extractive summarisation -> GPT-2 generation architecture. Improve upon the extractive summarisation by comparing multiple BERT variations for our purposes. GPT-2 also needs to be fine tuned further using the reddit dataset (which might consume a lot of time as the dataset is huge).

Timeline

- Milestone 1 - 1st week
- Milestone 2 - 2nd and 3rd week
- Milestone 3 - 3rd and 4th week
- Testing and review - 5th and 6th week