Info 256 Project Proposal: Headline Generation

Vaibhav Ramamoorthy
SID 25539060
vaibhavr@berkeley.edu

Phil Ryjanovsky
SID 3034269723
philip_ryjanovsky@berkeley.edu

Abstract

In this paper we propose a neural approach to the problem of Headline Generation. Our model will take in an article and produce an appropriate and coherent headline for that article of roughly 8-10 words. We will train and test our model on NBA game recaps scraped from ESPN.com and implement a sequence-to-sequence architecture to model the relationship between the game recap inputs and the headline outputs. We then evaluate the success of our model using both human evaluation comparing the seq2seq's outputted headline with a baseline non-neural headline generation model and using BLEU scores for each as compared to the true headline that was written.

1 Motivation

Text summarization has emerged to be a highly interesting and challenging research field in the Natural Language Processing domain. With the immense amount of data produced in the form or articles, documents, reports and research papers, accurately and concisely summarizing information into key points or even a headline has and will prove to be immensely useful. Extracting key information from a text and then generating a meaningful headline that encompasses this information will have implications towards all sorts of domains, and can increase the efficiency in many processes.

Our project will hone in on the application of Headline Generation towards NBA post-game articles, which provide a recap of the game and any key events that took place. NBA game recaps are highly repetitive, not only in the structure and ordering of the words/information within the headline, but also as it relates to where in the article the information is gathered from. This will make it a convenient and easy type of article with which to work.

Whereas commercially-available algorithms like Wordsmith provide a more ”fill-in-the-blanks” approach to writing game summaries using key data and information, our problem will involve taking in the game summary itself and then condensing key information. We are hopeful in the project given that the key information already exists within the input data and that it is highly structured and repetitive. It will be up to our model to simply identify and extract this information into a coherent headline, instead of generating output that does not exist in the input.

Beyond our choice of data, we see applications of such a model within the domains of headline/summary generation for news articles, research papers and media communication.

2 Evaluation

We plan to use human evaluation to evaluate the success of our model. For each input in our testing set of data, we will produce a generated headline using our neural model and a generated headline from a baseline non-neural model. We will then evaluate our model on two metrics:

i) The percentage of times the neurally generated headline was rated better by our human evaluators (us) than the non-neurally generated one.

ii) The BLEU score of the neurally generated headlines versus that of the non-neurally generated ones, as outlined in (2015). The BLEU score measures what percentage of n-grams in the true headline were present in the generated headline.

3 Literature Review

To build our non-neural baseline headline generation model, we will utilize the statistical framework developed in Banko, et. al. (2000). Banko, et. al. breaks down the headline generation task into two tasks: (i) content selection and (ii) surface realization. The content selection task models the probability of a headline as the product of the probabilities of each word w_i in the headline being in the headline given that it appeared in the document times the probability that the headline is of length n times the probability of each word appearing conditional on its preceding context. It leverages an independence assumption that the presence of a word in the headline is independent of the presence of other words. The surface realization task models the conditional probability of seeing a word in the headline given its left context. We will model the surface realization using a bigram model, where probability of a word being next in the headline will onlybe dependent on the single word preceding it (w_{i-1}).

To build our neural headline generation model, we will utilize the sequence-to-sequence framework developed in Sutskever, et. al. (2014). Sutskever, et. al. uses Deep Neural Networks and multilayered Long Short-Term Memory (LSTM) Networks to map input vectors of variable length to dynamic length output vectors. The input is translated to a vector of fixed dimensionality using an encoder network. This fixed-length vector is then decoded using a decoder network into a dynamic-length output. Both the encoder and decoder networks are trained on a segment of data. Sutskever, et. al. apply seq2seq methods to machine learning problems and find that it is quite successful in English-French translations.

We will consider implementing a method Lopyrev (2015) experimented with involving an attention mechanism, which computes a weight over each of the input words that determines how much attention should be paid to each one. They are then used to compute a weighted average of the last hidden layers, referred to as the context. This context is then input into the softmax layer along with the last hidden layer from the current step of the decoding.

Ayana, et. al. (2016) tackle the issue of neural models using word-level optimization in maximum-likelihood neural headline generation by leveraging minimum-risk training (MRT). This shifts the optimization to a sentence-wise optimization. Minimum-risk training minimizes the expectation of the risk between the target headline and generated headline. The risk function utilized in this paper is a negative recall function. In the later stage of our model development we may try to incorporate MRT to improve our model.

While most successful summarization systems utilize extractive approaches which crop out and stick together portions of the text, Rush et. al. (2015) explored an abstractive approach that can generate aspects which may not appear in the original, where they combined a neural language model with a contextual input encoder, based off of Bahdanau et. al. (2014). Crucially, both the encoder and the generation model are trained jointly on the sentence summarization task. This is a more advanced approach, but we may pull some pieces from their approach to incorporate in ours.

4 Team Members and Roles

We have two members, Phil and Vaibhav. We plan to split the work pretty evenly throughout and work together on most aspects of the project. Our next steps are as follows:

- Build scraper to scrape game recaps and headlines from ESPN.com
- Clean, tokenize, and perhaps lemmatize scraped data
- Build and train model and baseline
- Iterate over model to maximize improvement over baseline

By collaborating on Github, we will be able to work on all parts of the project together. We both have experience in Python in both webscraping and in the keras package (through this class) and will therefore maintain an equal balance of work throughout the project.

References

Alexander M. Rush, Sumit Chopra, Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv:1509.00685*.

Ayana, Shiqi Shen, Yu Zhao, Zhiyuan Liu, and Maosong Sun. 2016. Neural headline generation with sentence-wise optimization. *arXiv:1604.01904v2*.

Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*.

Ilya Sutskever, Oriol Vinyals and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *arXiv:1409.3215*.

Konstantin Lopyrev. 2015. Generating news headlines with recurrent neural networks. *arXiv:1512.01712*.

Michele Banko, Vibhu O. Mittal, Michael J. Whitbrock. 2000. Headline generation based on statistical translation. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*.