Project Proposal

Natural Language Processing
Justin Chen and Anand Tyagi

Introduction/Motivation

We were initially interested in two different areas of research: natural language generation and summarization. In order to combine our two interests, we decided to work on a systems project that is able to generate summaries based on the content of a document. More specifically, we aim to create a system which will generate a one sentence summary of a given news article. Generating a one sentence summary perfectly combines our two areas of interest and can have a wide range of applications. However, we decided to focus on summarizing news articles as we have primarily dealt with news related corpuses up until this point (the WSJ corpus, to be specific) and feel that providing one sentence summaries of news articles, if done well, could be something immediately usable.

Dataset and Evaluation

For data, we would need a collection of news articles and their summaries. There are two main datasets we found which we will be able to use. The first one we found is the Document Understanding Conferences' DUC-24 dataset which includes articles and their summaries. Several of the papers we have read thus far make use of this dataset. The second dataset we found is the Cornell NEWSROOM dataset [1] which contains 1.3 million articles and their summaries. Most of the summary examples provided in their paper seem to be one sentence summaries which is exactly the type of summary we intend to generate. In addition to providing a dataset, the Cornell NEWSROOM also has an online evaluation benchmark which we will be able to use in order to see how well our summaries compare to the current standards.

Research

The articles in this section are the ones we have read so far and relate to our goal of one sentence news article summary generation.

1. **A Neural Attention Model for Abstractive Sentence Summarization [4]**

Machine made summarization of large texts is a topic of this paper. In the past, summarization has been a process of removing and combining different parts of the overall document. The goal of this paper is to create summerizations from a bottom up approach at the sentence level. This method called abstractive summarization may create summerizations that are not part of the original text.

To accomplish abstractive summarization, the paper employs the use of data driven models. The model the researchers referenced is the attention neural model. The model that the researchers eventually used for summarization is the Attention Based Summarization (ABS). This model incorporates the use of an encoder and a beam search decoder.

The findings of this paper provides us with a jumping point for our project. The details of this paper outlines the methods to generate an abstractive summarization. We are considering using the same model for approaching our summarization proposal.

2. **Neural Headline Generation on Abstract Meaning Representation [5]**

This paper proposes a solution to generate headlines for documents. The contents of this article improves on the use Attention Based Summarization (ABS) to generate headlines for documents. The researchers made modifications to ABS by incorporating an Abstract Meaning Representation (AMR) encoder.

The findings of this paper provides a more detailed approach to modify the ABS approach. Additionally, this paper asserts that although ABS has success in performance, niche tasks can be improved with modifications.

We aim to abstract some of the strategies proposed in this paper for headline generation for generating our summaries and test to see if they improve upon the current standard of summary generation.

3. **Keeping Consistency of Sentence Generation and Document Classification with Multi-Task Learning [3]**

Automated generation of a profile of a document is an advanced task. A profile consists of a title, categorization and summary of the provided document. The difficult part is to link the different components of the profile and have it semantically make sense. For example if one were to generate the title, category, and summary of a newspaper article, the aforementioned labels must have semantic relation.

The paper solves this issue through a *hierarchical consistency loss* function. This function trains the weights of attention on certain parts of the input text. Since we plan to use attention in our summary generation, we believe using the novel loss function provided by this paper will improve the fluency of our summaries just as it improved the fluency of the headlines they generated (as discussed in the paper).

4. **Text Generation from Keywords [6]**

This paper is closely related to our project as it addresses the topic of text generation from a set of “keywords” or “headwords.” It mainly focuses on the construction of sentences through the use of dependency trees and the use of “complementary information” to fill in missing words to generate complete sentences. It also discusses a method for evaluating the generated sentence before producing the final output.

We believe that this paper can provide us with baseline strategies to refer to in regards to generating our sentence. While it does not discuss generating sentences which are summaries of a specific document, the manner in which sentences are generated is something we will need to use in order to generate a generally coherent summary.

5. **Generating topic-oriented summaries using neural attention [2]**

This paper provides an “attention based RNN framework” which is able to generate summaries based on specific topics due to the use of neural attention. This paper shows that an attention based RNN model may be one of the best ways to generate summaries based on a specific topic. However, instead of making topic-tuned summaries, as they do in the paper, we will use the framework provided in the paper to improve our summary generation.

Plan of Action/Strategy

1. As we have already been doing, our first task will be to read as much about the work that has been done in both sentence generation and text summarization as possible. We will also focus on papers specifically discussing summary generation of news articles and frameworks which use attention and RNN frameworks for sentence or summary generation.
2. We have decided to follow and attempt to implement the neural attention model for summarization. So, after going through the papers we have listed and learning more about attention and RNN frameworks, we first plan to implement a simple RNN with attention to see if we are able to make a working network to begin with.
3. After setting up an RNN with attention, we will implement some of the relevant methods mentioned in the *Generating topic-oriented summaries using neural attention* [2] paper and the *A Neural Attention Model for Abstractive Sentence Summarization* [4]. This model will be our baseline which we are confident we can create for sure by the end of this semester. The model will simply be one that is an RNN with attention, similar to the one mentioned in the first paper, with modifications and additions based on suggestions provided by the second paper.
4. After developing summaries, we want to improve the fluency and accuracy of our summaries. Thus, we will implement various combinations of the strategies presented in the *Neural Headline Generation on Abstract Meaning Representation* [5] paper, the *Keeping Consistency of Sentence Generation and Document Classification with Multi-Task Learning* [3] paper, and the *Text Generation from Keywords* [6] paper.
5. While we will evaluate each model we generate on the Cornell NEWSROOM [1] evaluation benchmark, we aim to have our final model hopefully be comparable to current or past summarization models evaluated on the NEWSROOM benchmark.
6. As we develop our model, we will keep track of which features and frameworks enhanced our performance and which ones reduced our performance. We will finish by presenting our findings in a final paper.

Contributions/Collaboration

Anand Tyagi and Justin Chen will be working on this project. Anand will be handling most of the ML related tasks and writing code for the system. Justin will handle the evaluation task and write code for the task. Both will do the write up for the paper.

Since the evaluation can only be done after a model has been created, both Anand and Justin will jointly work on understanding how to build the initial model. So while Anand will write the model itself, both individuals will be working on the project from the start. Additionally, Justin will be working on understanding how to handle the data, check the validity of the data, and figure out how we will eventually evaluate the data.

References

[1] Grusky, M., Naaman, M., & Artzi, Y. (2018). Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. doi: 10.18653/v1/n18-1065

[2] Krishna, K., & Srinivasan, B. V. (2018). Generating Topic-Oriented Summaries Using Neural Attention. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. doi: 10.18653/v1/n18-1153

[3] Nishino, T., Misawa, S., Kano, R., Taniguchi, T., Miura, Y., & Ohkuma, T. (2019). Keeping Consistency of Sentence Generation and Document Classification with Multi-Task Learning. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. doi: 10.18653/v1/d19-1315

[4] Rush, A. M., Chopra, S., & Weston, J. (2015). A Neural Attention Model for Abstractive Sentence Summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. doi: 10.18653/v1/d15-1044

[5] Takase, S., Suzuki, J., Okazaki, N., Hirao, T., & Nagata, M. (2016). Neural Headline Generation on Abstract Meaning Representation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. doi: 10.18653/v1/d16-1112

[6] Uchimoto, K., Isahara, H., & Sekine, S. (2002). Text generation from keywords. *Proceedings of the 19th International Conference on Computational Linguistics -*. doi: 10.3115/1072228.1072292