CS671: Introduction to Natural Language Processing

Project Proposal

Group - 4

Gaurav (13274)
Palash Chauhan (13455)
Shubham Agrawal (13674)

August 21, 2016

1. Problem

- Area: Text Summarization
- Statement : "Automatic abstract generation for scientific papers"
- Data-set : NIPS 2015 data-set

2. Introduction

Abstract provides the reader with an overview of the research paper and even can be used for indexing, searching and retrieving information without having to process the whole document. This motivates us to come up with an **automatic abstract generator** which will not only avoid the stressful work of researcher to summarize their work in small paragraph but can even be extended to poster, extended abstract and newsletter generation. Proposed task has connections to problems in domain of text summarization.

Automatic summarization [5] is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Generally, there are two approaches to automatic summarization: **extraction** and **abstraction**. **Extractive** methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, **abstractive** methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. We wish to approach this problem from the combination of both these methods.

3. Related Work

Elena Lloret et. al.[1] analyzes the appropriateness of a text summarization system for generating abstracts of biomedical papers using extractive and abstracted-oriented approach.

Rush et al [4] propose a fully data-driven approach to abstractive sentence summarization. Also, [2] describe an application of an encoder-decoder RNN to generate headlines from the text of news articles.

[3] propose several novel models to address important problems in summarization including modeling key-words, capturing the hierarchy of sentence-to-word structure and addressing the problem of words that are key to a document.

4. Tentative Plan

Exploratory Data Analysis	25^{th} August
Theoretical Background and Paper Reading	31^{st} August
Finalize basic approach	5^{th} September
Architecture design and code	25^{th} September
Early Results and Report Submission	30^{th} September
Improvements and possible extensions	1^{st} November
Final demo, presentation and report	10^{th} November

References

[1] Elena Lloret, María Teresa Romá-Ferri, and Manuel Palomar. Compendium: A text summarization system for generating abstracts of research papers. *Data & Knowledge Engineering*, 88:164–175, 2013.

[2] Konstantin Lopyrev. Generating news headlines with recurrent neural networks. *arXiv preprint arXiv:1512.01712*, 2015.

[3] Ramesh Nallapati, Bowen Zhou, Ça glar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond.

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

[5] Wikipedia. https://en.wikipedia.org/wiki/automatic_summarization.