Automatic scientific article summarization

Problem Statement

As of late, there has been a blast in the measure of text data from an assortment of sources. This volume of text is a priceless source of information and knowledge, which should be effectively summarized to be useful. In this problem, the main objective is to automatic text summarization are described below for lighting more about processes. With the dramatic growth of the Internet, people are overwhelmed by the tremendous amount of online information and documents. This expanding availability of documents has demanded exhaustive research in automatic text summarization.

Now days many research is going on for text summarization. Because of increasing information in the internet, these kinds of research are gaining more and more attention among the researchers. Extractive text summarization generates a summary by extracting proper set of sentences from a document or multiple documents by deep learning. The whole concept is to reduce or minimize the valuable information present in the documents. The procedure can be manipulated by Restricted Boltzmann Machine (RBM) algorithm for better efficiency by removing redundant sentences. The restricted Boltzmann machine is a graphical model for binary random variables. It consists of three layers input, hidden and output layer. The input data uniformly distributed in the hidden layer for operation.

Background

In the recent past, deep-learning based models that map an input sequence into another output sequence, called sequence-to-sequence models, have been successful in many problems such as machine translation (Bahdanau et al., 2014), speech recognition (Bahdanau et al., 2015) and video captioning (Venugopalan et al., 2015). In the framework of sequence-to-sequence models, a very relevant model to our task is the attentional Recurrent Neural Network (RNN) encoder-decoder model proposed in Bahdanau et al. (2014), which has produced state-of-the-art performance in machine translation (MT), which is also a natural language task.In another paper that is closely related to this work, Hu et al. (2015) introduce a large dataset for Chinese short text summarization. They show promising results on their Chinese dataset using an encoder-decoder RNN, but do not report experiments on English corpora.In another very recent work, Cheng and Lapata (2016) used RNN based encoder-decoder for extractive summarization of documents. This model is not directly comparable to ours since their 2 http://duc.nist.gov/ framework is extractive while ours and that of (Rush et al., 2015), (Hu et al., 2015) and (Chopra et al., 2016) is abstractive.

Siddharthan and Teufel describe a new task to decide the scientific attribution of an article (Siddharthan and Teufel, 2007) and show high human agreement as well as an improvement in the performance of Argumentative Zoning (Teufel, 2005). Argumentative Zoning is a rhetorical classification task, in which sentences are labeled as one of Own, Other, Background, Textual, Aim, Basis, Contrast according to their role in the author's argument. These all show the importance of citation summaries and the vast area for new work to analyze them to produce a summary for a given topic.

Previous work has shown the importance of the citation summaries in understanding what a paper says. The citation summary of an article A is the set of sentences in other articles which cite A. (Elkiss et al., 2008) performed a largescale study on citation summaries and their importance. They conducted several experiments on a set of 2, 497 articles from the free PubMed Central (PMC) repository1. Results from this experiment confirmed that the "Self-Cohesion" (Elkiss et al., 2008) of a citation summary of an article is consistently higher than the that of its abstract. (Elkiss et al., 2008) also conclude that citation summaries are more focused than abstracts, and that they contain additional information that does not appear in abstracts. (Kupiec et al., 1995) use the abstracts of scientific articles as a target summary, where they use 188 Engineering Information summaries that are mostly indicative in na1 http://www.pubmedcentral.gov ture.

Abstracts tend to summarize the documents topics well, however, they don't include much use of metadata. (Kan et al., 2002) use annotated bibliographies to cover certain aspects of summarization and suggest guidelines that summaries should also include metadata and critical document features as well as the prominent content-based features.

Methodology

Text summarization technique is divided into two approaches extractive and abstractive. But due to the limitation of natural language generation techniques in generating the abstractive summary generally extractive approach is used for summarization. For summarizing the text there is a need of structuring the text into certain model which can be given to RBM as input. First, in text summarization the text document is preprocessed using various prevalent preprocessing techniques and then it is converted into sentence matrix defined over a vocabulary of words. This structured matrix each row will work as an input to our RBM (Fig. 1). After getting the set of top priority word from the RBM the input query, sentence vector and high priority word output is compared to generate the extractive summary of the text document.

We can also be built a good model for abstractive text summarization using Attentional Encoder Decoder Recurrent Neural Networks, and show that they achieve state-of-the-art performance on two different corpora. Several novel models that address critical problems in summarization that are not adequately modeled by the basic architecture, such as modeling key-words, capturing the hierarchy of sentence-toward structure, and emitting words that are rare or unseen at training time.

This **deep learning based approach** is based on the encoder-decoder recurrent neural network with attention, developed for machine translation.This baseline model corresponds to the neural machine translation model used in Bahdanau et al. (2014). The encoder consists of a bidirectional GRU-RNN (Chung et al., 2014), while the decoder consists of a uni-directional GRU-RNN with the same hidden-state size as that of the encoder, and an attention mechanism over the source-hidden states and a soft-max layer over target vocabulary to generate words.

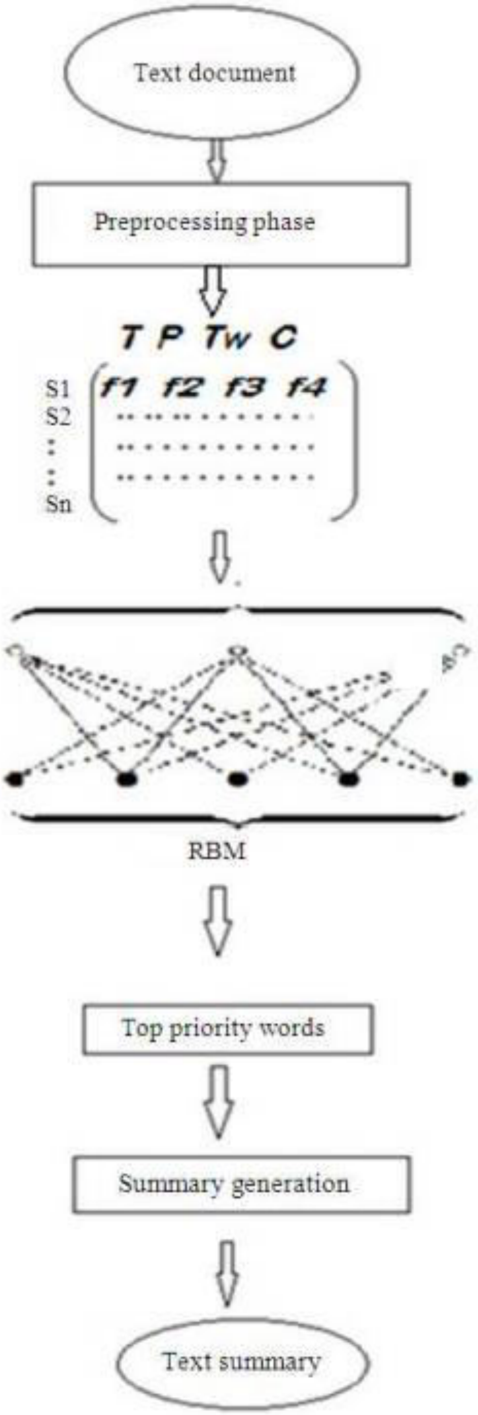
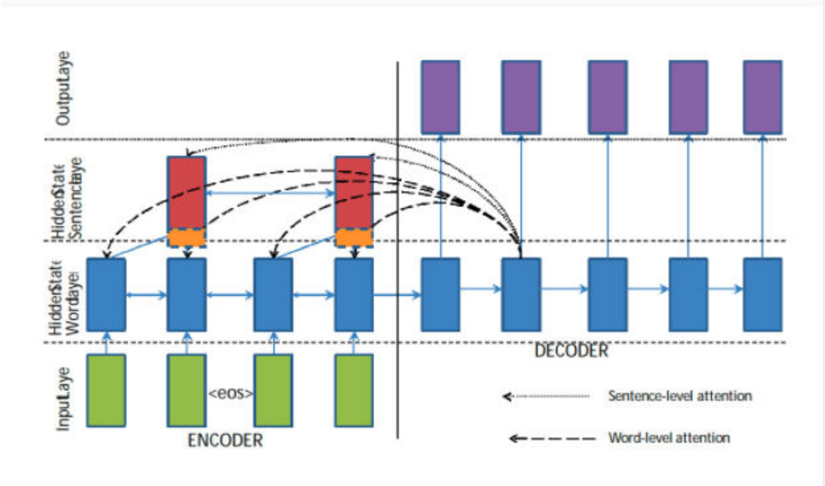


Fig 1: Block diagram of text summarization



Hierarchical encoder with hierarchical attention.

Fig 2: Deep Learning based approach[2]

Experimental Design

Quickly moving to a new area of research is painful for researchers due to the vast amount of scientific literature in each field of study. One viable way to overcome this problem is to summarize a scientific topic. In this problem, wehave to build or propose a model of summarizing a single article, which can be further used to summarize an entire topic.

Dataset: -The ACL Anthology is a collection of papers from the Computational Linguistics journal, and proceedings from ACL conferences and workshops and includes almost 11, 000 papers. To produce the ACL Anthology Network (AAN), (Joseph and Radev, 2007) manually performed some preprocessing tasks including parsing references and building the network metadata, the citation, and the author collaboration networks. The full AAN includes all citation and collaboration data within the ACL papers, with the citation network consisting of 8, 898 nodes and 38, 765 directed edges.

Steps: -

- Preprocessing
- Part of Speech Tagging
- Stop Word Filtering
- Stemming
- Feature Vector Extraction
- Feature Computation
- Title Similarity
- Concept Feature
- Sentence Matrix
- Apply Deep Learning Algorithm
- Optimal Feature Vector Set Generation
- Summary and Ranking of Sentence

- Result and Analysis

References

1. https://arxiv.org/pdf/1707.02268.pdf
2. https://machinelearningmastery.com/encoder-decoder-recurrent-neural-network-models-neural-machine-translation/