Project Report on

Text summarization using (NLP)

Md.Liaket hossain(161412313)

Saiful Islam(161412327)

Nazrul islam(161412329)

Mustafizur Rahman(161412318)

CONTENT OF TABLE

1.Abstract

2.Introduction

3. Types of text summarization

⇨ Abstractive

⇨ Extractive

4 .Text summarization system

5. Linguistic Preprocessing for automatic summarization

6.Objective

7.References

Abstract

Automatic summarization 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. As The problem of information overload has grown, and as the quantity of data has increased, so has interest in automatic summarization. It is very difficult for human beings to manually summarize large documents of text. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. The extractive summarization systems are typically based on techniques for sentence extraction and aim to cover the set of sentences that are most important for the overall understanding of a given document.
In frequency based technique obtained summary makes more meaning. But in k-means clustering due to out of order extraction, summary might not make sense.

Introduction

In natural language processing, text summarization refers to the task of taking a document and creating a shorter version of it, that nevertheless retains its primary meaning. While this task has real-world applications, such as saving time reading long news articles, or creating blurbs for news for previews, it is primarily a well-suited task for the reasoning about the scale of different textual models: building neural network models that map from longer sequences to shorter ones inherently has the questions of ”how long” and ”how short”. Classical text summarization techniques were interested in producing a handful of sentences when given a larger article [1]. Recent neural network techniques [2] use seq2seq [3] models, which are naturally limited to sequences of a few hundred words at most. This comes from one of two natural limitations, either encoder-based models lack the proper memory to remember thousands of words ahead, or attention-based models are unable to evaluate attention over giant sequences. In general, these recurrent techniques require non-trivial methods to work over large text sequences, an ongoing area of research [4]. Additionally, many document sources for which we are interested in summaries naturally come with human generated summaries, whether that are headlines for news reports or abstracts for scientific publications. While titles and headlines are not equivalent to summaries, they approximate the same information-compressing ideal. Both the availability of these datasets and the existence of reasonable approximations of summaries makes this problem well suited for supervised deep-learning approaches.

Types of text summarization

Extractive

Extractive summarization are created by reusing portion(word, sentence, ect.) of the input text document

The system extracts text from the enter collection, without modifying the text document

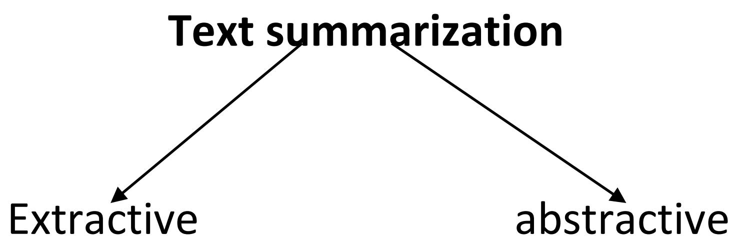
Most of the summarization research today is on extractive summarization

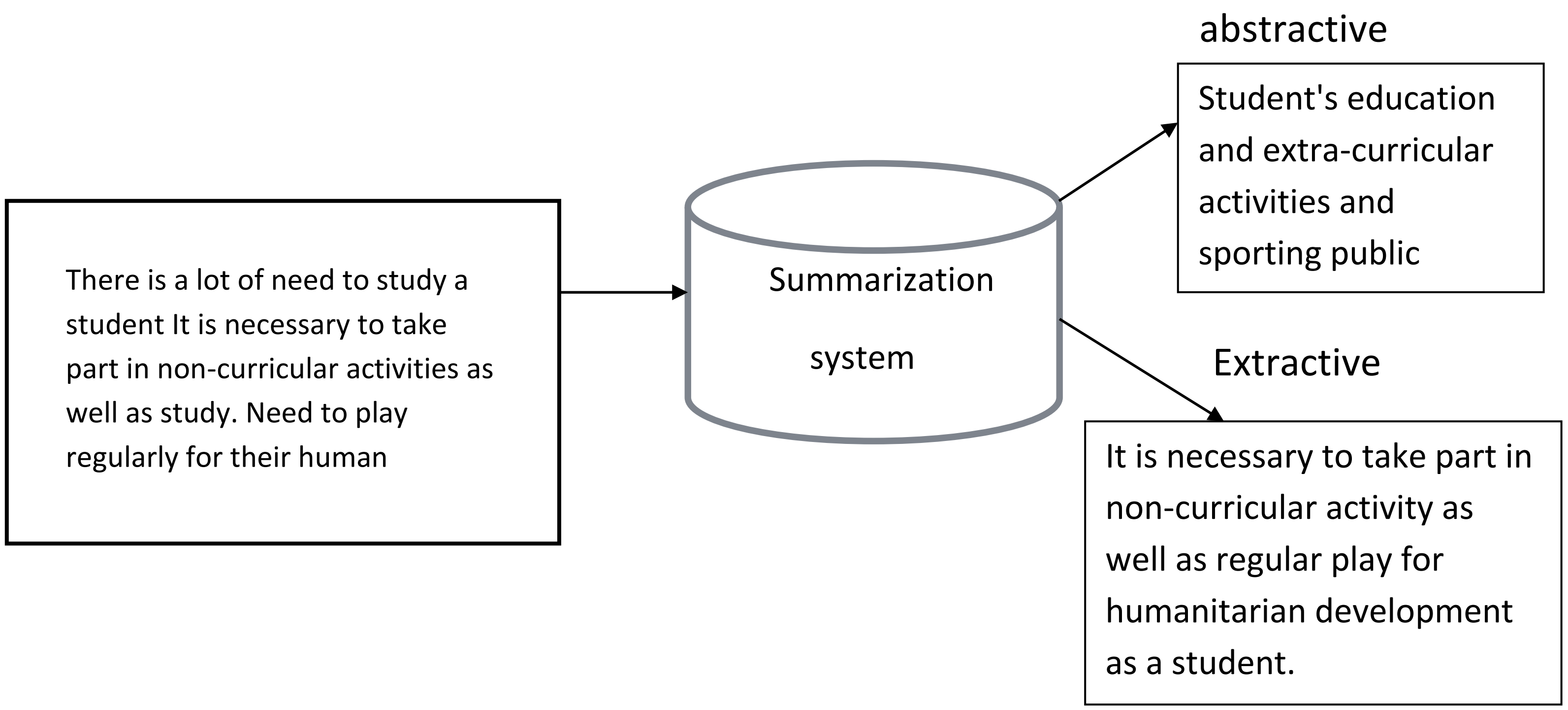
Abstractive

Requires deep understanding and reasoning over the text

It provides own summary over input text without using same word of sentence in the input text

Determines the actual and short meaning of each element such as word ,sentence and paragraph





There is a lot of need to study a student It is necessary to take part in non-curricular activities as well as study. Need to play regularly for their human

Student's education and extra-curricular activities and sporting public

It is necessary to take part in non-curricular activity as well as regular play for humanitarian development as a student.

Preprocessing

We pre-processed datasets to be all lowercase and restricted to the 100000 most common words. We used NLTK[15] to tokenize inputs and add paragraph and sentence start and end markers (

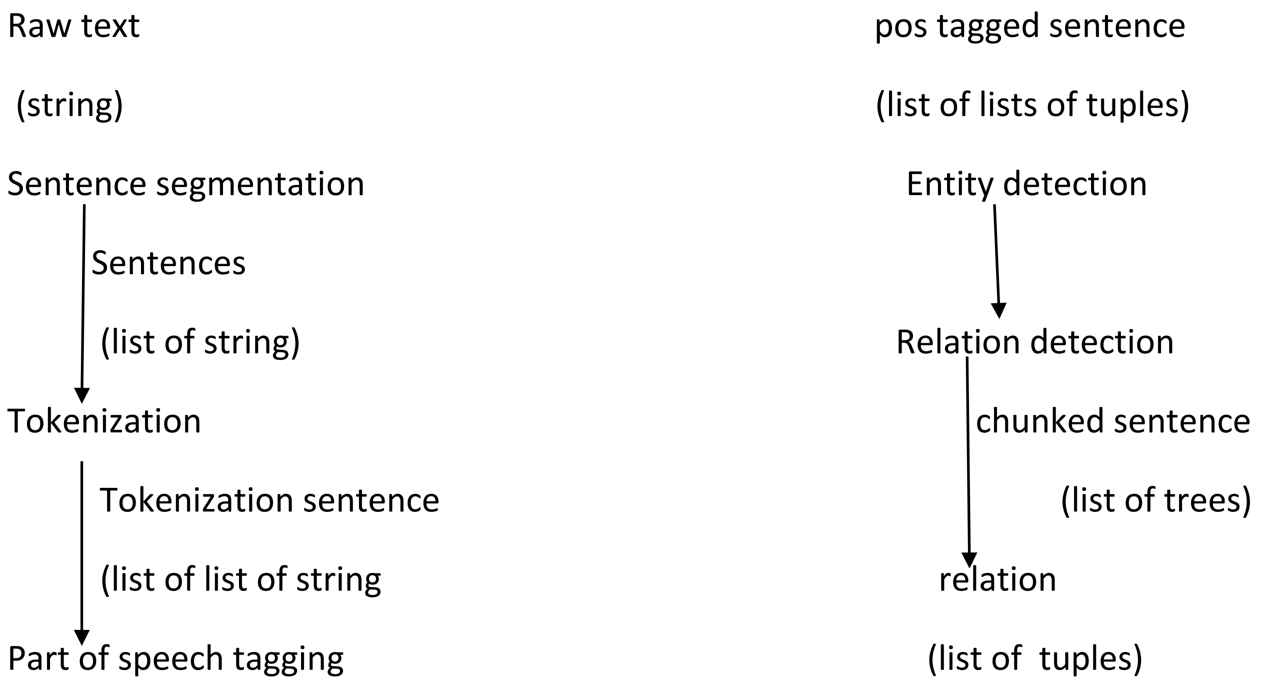
,

& ,). All dataset were split randomly into training, evaluation and test sets, weighted 80% for training, 10% for evaluation and hyperparameter optimization, and 10% for testing. To simplify our comparison we converted all datasets into a simple JSON format:

```
[
{
 "data": "input document as string",
 "label": ["model summaries as array of strings"], "set": "train"|"dev"|"test",
  "prediction": "predicted label as string"
}
]
```

This specified the output format for our models as well (prediction string) and allowed us to run the same evaluation script over the outputs of our various models. We made sure to strip any sentence and paragraph markers before evaluation since they otherwise led to artificially inflated ROUGE-scores.

Linguistic Preprocessing for automatic summarization



Sentence segmentation

Converts raw text into sentences

List of string

Sentence tokenization

Input text: john owns a car. It is a Toyota

Output: segm1: john owns a car

Segm2: it is a Toyota

Tokenization

Identifies the word tokens from given sentence

Provides a list of tokens as output

Word tokenizer

Input : john owns a car.

Output:[[john],[owns],[a],[car],[.]]

Part of speech tagging

Assigns appropriate part of speech tag to each word

Pos is useful in extraction of nouns, adverbs, adjective, which provide some

Meaningful information about text

Generate a list of tuples with pos annotation

Input :[[john],[owns],[a],[car],[.]]

Output: (np (nnp john)) (vp (vbz owns) (np(dt a) (nn car)))(. .)

Entity detection

Identification of predefined categories such as person location, quantities,

Organization etc

Ner provides the entity detection for linguistic processing

Ner system uses linguistics grammar base techniques and also statistical model to identify the entity

Input : (np (nnp john)) (vp (vbz owns) (np(dt a) (nn car)))(. .)

Output: john->person

Relation detection

Identifies the possible relation between two or more chunked sentences

Co reference chain provide a relation between two or more sentences

Provide the links between pronoun and its corresponding nouns

Replacement of the pronouns with proper nouns

Input text: john owns a car. It is a Toyota. (in from of parse tree)

Output: “a car”-> “a Toyota”;”it”->”a toyota”

Objectives

The main goal of our research is develop an abstractive English text summarization for generate English text summary there are two way to summarization text one is extractive and other is abstractive. in extractive method drag main sentence from main document and combine them to make summary but it is abstractive method its overcome the grammatically consistency of extractive method .abstractive method is more efficient then extractive method .grammar inconsistency is major obstacle to make a summary of text document .if we use abstractive method in deep learning a algorithm we can reducer computational time and reduce inconsistency .so we need to provide a inconsistence free and computationally fast text summarizer which give accurate and fluent text summary.

References

[1] - http://www-nlpir.nist.gov/projects/duc/

[2] - https://github.com/turian/textrank#readme

[3] - http://www.cse.unt.edu/~rada/papers/mihalcea.emnlp04.pdf

[4] - http://en.wikipedia.org/wiki/Document_summarization

[5] - http://www.cs.cmu.edu/~nasmith/LS2/das-martins.07.pdf

[6] -http://www.aicit.org/jcit/ppl/03-JCIT1-785287JE.pdf

[7] - http://acl.ldc.upenn.edu/W/W00/W00-0405.pdf

[8] - https://github.com/turian/textrank