ABSTRACT

AUTOMATIC ABSTRACTIVE TEXT SUMMARIZER

In this digital era the procurement of data is growing rapidly and this growth leads to excessive amount of data. This excessive amount of data demands more storage. To reduce the growing storage requirements, summarization of the data will be helpful.

Summarization means reducing the original content to a shorter version preserving the key information and meaning of the context. As we humans read the whole text and collect the summary this could reduce the human intervention and provide the summary with lesser efforts. But humans collect the summary through knowledge and language capabilities it will be difficult task for a computer to perform text summarization. Generally, text summarization is done using the following two approaches: abstraction andextraction.

The extractive approach generates the summary from the given input text. This approach provides an individual score to each of the sentences of the input and based on the score it includes or excludes the sentences from the condensed version.

The abstractive approach generates the summary by advanced natural language methodologies. Some of the text in the condensed version maynot appear in the input text. Thus, it is not just rearranging and formatting thetext as done in the extractive approach.

The main aim of the project is to find a subset of data that contains all the important information of the input data set

Keywords **–** Abstractive, RNN, Summarization, Encoder, Decoder, LSTM, SoftMax, PUNKT, Stopwords.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

1.1. Introduction:

Before going to the Text summarization, first we have to know what a summary is. A Summary is a text that is produced from one or more texts, that conveys important information in the original text, and it is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter version. In this work, we propose a fully data-driven approach to abstractive sentence summarization. Our method utilizes a local attention-based model that generates each word of the summary conditioned on the input sentence. While the model is structurally simple, it can easily be trained end-to-end and scales to a large amount of training data. The most important advantage of using a summary is, it reduces the reading time. There are two different groups of text summarization namely Indicative and Informative. Inductive summarization only represents the main idea of the text to the user. The typical length of this type of summarization is 5 to 10 percent of the main text. Whereas Informative summarization systems give concise information of the main text. The length of informative summary is 20 to 30 percent of the main text. IN order to better understand how summarization systems work, we describe three fairly independent tasks which all summarizers perform. They are as follows:

- Construct an intermediate representation of the input text which expresses the main aspects of the text.
- Score the sentences based on the representation.
- Select a summary consisting of a number of sentences.

Intermediate Representation: Every summarization system creates some intermediate representation of the text it intends to summarize and finds salient content based on this representation.

There are two types of approaches based on the representation: Topic representation and Indicator representation.

- **Topic representation** approaches transform the text into an intermediate representation and interpret the topics discussed in the text. Topic representation-based summarization techniques differ in terms of their complexity and representation model.
- **Indicator representation** approaches describe every sentence asa list of features (indicators) of importance such as sentence length, position in the document, having certain phrases, etc.

Sentence Score: When the intermediate representation is generated, we assign an importance score to each sentence. In topic representation approaches, the score of a sentence represents how well the sentence explains some of the most important topics of the text. In most of the indicator representation methods, the score is computed by aggregating the evidence from different indicators.

Sentence scoring is one of the most used processes in the area of Natural Language Processing (NLP) while working on textual data. It is a process to associate a numerical value with a sentence based on the used algorithm's priority. This process is highly used especially on text

summarization. There are many popular methods for sentence scoring like TF-IDF, TextRank and so on.

Summary Sentences: The summarizer system selects the top k most important sentences to produce a summary. Some approaches use greedy algorithms to select the important sentences and some approaches may convert the selection of sentences into an optimization problem where a collection of sentences is chosen, considering the constraint that it should maximize overall importance and coherency and minimize the redundancy.

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 extractive text summarization technique involves pulling key phrases from the source document and combining them to make a summary. The extraction is made according to the defined metric without making any changes to the texts.

Here is an example:

Source Text:

> *Joseph and Mary rode on a donkey to attend the annual event in Jerusalem. In the city, Mary gave birth to a child named Jesus.*

Extractive summary:

> *Joseph and Mary attend event Jerusalem. Mary birth Jesus.*

As you can see above, the words in bold have been extracted and joined to create a summary although sometimes the summary can be grammatically strange.

An **Abstractive summarization** is an understanding of the main concepts in a document and then expressing those concepts in clear natural language.

The abstraction technique entails paraphrasing and shortening parts of the source document. When abstraction is applied for text summarization in deep learning problems, it can overcome the grammar inconsistencies of the extractive method.

The abstractive text summarization algorithms create new phrases and sentences that relay the most useful information from the original text just like humans do.

Therefore, abstraction performs better than extraction. However, the text summarization algorithms required to do abstraction are more difficult to develop; that's why the use of extraction is still popular.

Here is an example:

Source Text:

> ***Joseph and Mary*** *rode on a donkey to* ***attend*** *the annual* ***event*** *in* ***Jerusalem****. In the city,* ***Mary*** *gave* ***birth*** *to a child named* ***Jesus****.*

Abstractive summary: *Joseph and Mary came to Jerusalem where Jesus was born.*

In this study, we choose abstractive summarization approach for building the text summarizer by making use of Attention based Many-to-ManySeq2Seq encoder decoder architecture and recurrent neural network for building this model. RNN is used because it supports the mechanism of being recursive in nature while training the model. While training the model this recursive property provides room for performing cross-validation between the training data and helps in increasing the efficiency of the summarizer for creating unique reviews. The models help the user to generate abstract summaries for the provided input data preserving the key information and not deviating from the main context.

1.2. Data Mining:

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. The knowledge or extracted information can be used to predict results in near future based on discovered patterns. Data mining is the analysis step of the "knowledge discovery in databases" process or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

1.2.1. Types of Data that can be Mined?

Data mining can be performed on the following types of data:

- **Relational Database:** A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data searchability, reporting, and organization.
- **Data warehouses:** A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision-making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing.
- **Data Repositories:** The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.
- **Object-Relational Database:** A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc. One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.

- **Transactional Database:** A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

1.3. Methodologies

The main goal of this project is to summarize the given text which is taken from the dataset of Amazon Food reviews. To achieve this goal, we need to perform the following steps:

- Choose and clean datasets
- Build the abstractive summarization model
- Test the model for the Food reviews dataset
- Tune the abstractive summarization model
- Build an end-to-end application

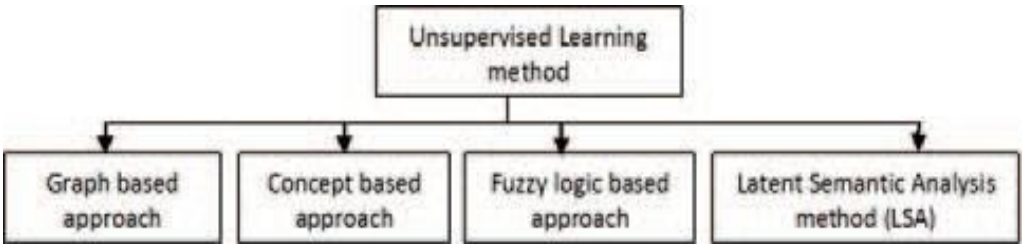
1.3.1. Extractive Summarization Methods:

There are different kinds of approaches for extractive summarization and these can be broadly classified as supervised and unsupervised learning approaches.

A. Unsupervised Learning Methods:

The unsupervised approaches do not need human summaries (user input)in deciding the important features of the document, it requires the mostsophisticated algorithm to provide compensation for the lack of human knowledge.

Figure 1.3.1: Classification of Unsupervised Learning Methods



1. **Graph-based method:** Graph-based models are extensively used in document summarization since graphscan efficiently represent the document structure. Extractive text summarization using external knowledge from Wikipedia incorporating bipartite graph framework has been used. They have proposed an iterative ranking algorithm (variation of HITS algorithm) which is efficient in selecting important sentences and also ensures coherency in the final summary. The uniqueness of this paper is that it combines both graphs based and concept-based approach towards the summarization task. Another graph-based approach LexRank, where the salience of the sentence is determined by the concept of Eigenvector centrality. The sentences in the document are represented as a graph and the edges between the sentences represent weighted cosine similarity values. The sentences are clustered into groups based on their similarity measures and then the sentences are ranked based on their LexRank scores similar to

PageRank algorithm except that the similarity graph is undirected in the LexRank method. The method outperforms earlier versions of lead and centroid based approaches. The performance of the system is evaluated with DUC dataset.

2. **Fuzzy logic-based approach:** The fuzzy logic approach mainly contains four components: defuzzifier, fuzzifier, fuzzy knowledge base and inferenceengine. The textual characteristics input of Fuzzy logic approach are sentence length, sentence similarity etc which is later given to the fuzzy system. Ladda Suanmali et al proposed fuzzy logic approach is used for automatic text summarization which is the initial step, the text document is pre-processed followed by feature extraction.

3. **Concept-based approach:** In concept-based approach, the concepts areextracted from a piece of text from external knowledge base such HowNet and Wikipedia.In the methodology proposed, the importance ofsentences is calculated based on the concepts retrieved from HowNetinstead of words. The basic steps in concept-based summarization are:
    - Retrieve concepts of a text from external knowledge base (HowNet, WordNet, Wikipedia).
    - Build a conceptual vector or graph model to depict relationship between concept and sentences.
    - Apply ranking algorithm to score sentences.
    - Generate summaries based on the ranking scores of sentences.

4. **Latent Semantic Analysis (LSA) approach:** It is a method which excerpt hidden semantic structures of sentences and words that are popularly used in text summarization task. LSA captures the text of the input document and excerpt information such as words that frequently occur together and words that are commonly seen in different sentences. A high number of common words amongst the sentences illustrate that the sentences are semantically related.

B. Supervised Learning methods:

Supervised extractive summarization related techniques are based on a classification approach at sentence level where the system learns by examples to classify between summary and non-summary sentences. The major drawback with the supervised approach is that it requires known manually created summaries by a human to label the sentences in the original training document enclosed with "summary sentence" or "non summary sentence" and it also requires more labelled training data for classification.

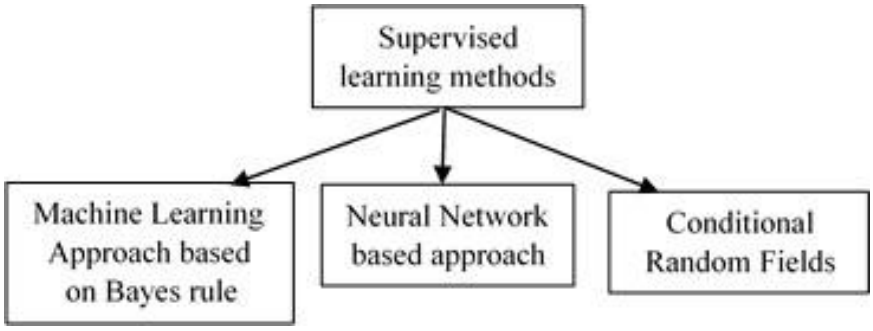


Figure 1.3.2: Classification of Supervised Learning Methods

1. **Machine Learning Approach based on Bayes rule:** A set of training documents along with its extractive summaries is fed as input to the training stage. The machine learning approach views classification problems in text summarization. The sentences are restricted as a non-summary and summary sentence based on the feature possessed by the sentence. The probability of classification is learned from the training data by the following Bayes rule: where s represents the set of sentencesin the document and fi represents the features used in classification stage and S represents the set of sentences in the summary. P (s E< SII1, h, h, …. In) represents the probability of the sentences to be included in the summary based on the given features possessed by the sentence.
2. **Neural Network based approach:** In the approach proposed, RankNet algorithm automatically uses neural nets to identify the important sentencesin the document. It uses a two-layer neural network with back propagation trained using the RankNet algorithm. Another approach uses a three-layered feed-forward neural network which learns in the training stage the characteristics of summary and non-summary sentences. The major phase is the feature fusion phase where the relationship between the features are identified through two stages eliminating infrequent features collapsing frequent features after which sentence ranking is done to identify the important summary sentences.
3. **Conditional Random Fields:** Conditional Random Fields are a statistical modelling approach that focuses on machine learning to provide a structured prediction. The proposed system overcomes the issues faced by non-negative matrix Factorization (NMF) methods by incorporating conditional random fields (CRF) to identify and extract correct features to determine the important sentence of the given text.

1.3.2. Abstractive Summarization methods:

Summarizations using abstractive techniques are broadly classified into two categories:

- Structured based approach.
- Semantic based approach.

Structured Based Approaches: The structural approach is a technique wherein the learner masters the pattern of sentence. Structures are the different arrangements of words in one accepted style or the other. It includes various modes in which clauses, phrases or word might be used. It is based on the assumptions that language can be best learnt through a scientific

selection and grading of the structures or patterns of sentences and vocabulary. Different types of structure-based approach are as follows:

1. **Tree Based Method:**
    - It uses a dependency tree to represent the text of a document.
    - It uses either a language generator or an algorithm for generation of summary.
    - It walks on units of the given document read and easy to summary.
2. **Template Based Method:**
    - It uses a template to represent a whole document.
    - Linguistic patterns or extraction rules are matched to identify textsnippets that will be mapped into template slots.
    - Its summary is highly coherent because it relies on relevant information identified by the IE system.
3. **Ontology Based Method:**
    - Use ontology (knowledge base) to improve the process of summarization.
    - It exploits fuzzy ontology to handle uncertain data that simple domainontology cannot.
    - Drawing relation or context is easy due to ontology.
    - Handles uncertainty at reasonable amount.
4. **Lead and Body Phrase Method:**
    - This method is based on the operations of phrases (insertion and substitution) that have the same syntactic head chunk in the lead andbody sentences in order to rewrite the lead sentence.
    - It is good for semantically appropriate revisions for revising a lead sentence.
5. **Rule Based Method:**
    - Documents to be summarized are represented in terms of categories and a list of aspects.
    - It has a potential for creating summaries with greater information density than current state of art.

Semantic Based Approach: IN Semantic based approach, semantic representation of documents is used to feed into the natural language generation (NLG) system. This method focuses on identifying noun phrases and verb phrases by processing linguistic data.

1. **Multimodal semantic model:**
    - A semantic model, which captures concepts and relationships amongconcepts, is built to represent the contents of multimodal documents.
    - An important advantage of this framework is that it produces an abstract summary, whose coverage is excellent because it includes salient textual and graphical content from the entire document.
2. **Information Item Based Method:**
    - The contents of summary are generated from abstract representationof source documents, rather than from sentences of source documents.
    - The abstract Representation is Information Item, which is the smallestelement of coherent information in a text.