Urdu Summarization using Pre-trained Language Models

By

Raja Mubashir Munaf

Supervisor

Assoc Prof Dr. Hammad Afzal

A thesis submitted to the Department of Computer Software Engineering.
Military College of Signals (MCS) , National University of Sciences and Technology.
Islamabad, Pakistan, in partial fulfillment of the requirements for the degree of MS in Software Engineering

July 2022

Thesis Acceptance Certificate

Certified that final copy of MS/MPhil thesis entitled "**Urdu Summarization using Pre-trained Language Models**" written by **Raja Mubashir Munaf**, (Registration No **00000274932**), of Military College of Signals (MCS) has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Supervisor: **Dr. Hammad Afzal**

Date: _____

Signature (HoD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

Dedication

To my wife & my children and my job without whom this thesis would have been completed two years earlier.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at at Military College of Signals (MCS) or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at Military College of Signals (MCS) or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Raja Mubashir Munaf**

Signature: _____

Acknowledgments

Glory be to Allah (S.W.A), the Creator, the Sustainer of the Universe. Who only has the power to honour whom He please, and to abase whom He please. Verily no one can do anything without His will. Thankful to ALLAH ALMIGHTY for successful completion of this research work; to My Family and Teachers for their support, unremitting help & continual guidance.

Raja Mubashir Munaf

Contents

List of Figures

List of Tables

List of Abbreviations

DW	Deutsche Welle (German News Agency)
NLP	Natural Language Processing
NLU	Natural Language Understanding
NLG	Natural Language Generation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
TFIDF	Term Frequency – Inverse Document Frequency
ANN	Artificial Neural Networks
RNN	Recurrent Neural Networks
CNN	Convolutional Neural Networks
LSTM	Long Short Term Memory Networks
GRU	Gated Recurrence Unit
OOV	Out of Vocabulary
LM	Language Model
ELMo	Embedding for Language Model
BERT	Bidirectional Encoder Representations for Transformers
mBERT	Multilingual Bidirectional Encoder Representations for Transformers
T5	Text to Text Transfer Transformer
mT5	Multilingual Text to Text Transfer Transformer
GPT	Generative Pre-training
MASS	Masked Sequence to Sequence
PEGASUS	Pretraining with Extracted Gap Sentences
ULFiT	Universal Language Model Finetuning
BiLM	Bidirectional Language Models
LDA	Drichelet Distribution
LSA	Latent Semantic Analysis
SVA	Support Vector Machine

Abstract

Ever increasing influx of data over the internet has become a reality which is faced with a challenge of sifting through and extracting meaningful information. During the last two decades, users are being overwhelmed with both textual and multimedia data, due to popularity of social media and news platforms. To cope up with the challenges of information overload various research technologies have also gain popularity. Natural Language Processing (NLP) has observed significant improvements for textual data processing in terms of its efficiency and accuracy with the inception of Language Models comprising of Deep Learning based Artificial Neural Networks. Automatic Summarization (under the umbrella of NLP) is the process of extracting only the meaningful information from text resulting into reducing the length of text as well as maintaining the sense of it. Urdu Language despite 10th most spoken language in the world is still a low resource language having little to no research in the field of Automatic Summarization and NLP. Most of the research in restricted to high resource languages like English. An effort is carried out to explore Deep Learning based Pre-trained Language Models comprising of self-attentive transformers for both Extractive and Abstractive Summarization capturing contextual information. Moreover, a summarization dataset of 76k records is created by collecting article summary pairs from news domain. As per best of our knowledge it will be the first and largest dataset available for Urdu Summarization. Experimental Results demonstrated competitive evaluation score (ROUGE, BERTScore) of summarization models finetuned on newly created dataset. Human evaluation is also carried out identifying the shortcomings of automatic evaluation methods in the field of summarization.

Chapter 1

Introduction

With an exponential growth of internet and its outreach to population around the world, there is an unprecedented and exponential increase in the amount of data being produced. The increase in the amount of data being generated is projected to further increase in future [1]. This enormous data is being produced through variety of sources including automatically generated sensory or machine data as well as human generated content. With data produced from multiple sources in multiple forms (videos, images, different types of textual data including news, articles, books or interactive content data) despite existing filtering and segregation procedure at end users, it is difficult to search for the useful information. Efforts required to identify and sought information which can be useful for end user is tedious and exhausting which renders the inaccessible available data useless in many cases or the required content cannot reach the desiring user within time. Particularly during the last decade, users are being overwhelmed with the enormous data being generated due to the popularity of news platforms and social media networks.

This order of magnitude increase in data has created certain challenges for various fields like Information Extraction (IE) or retrieval (IR) and Natural Language Processing (NLP). Numerous types of data and information is present over internet about almost every topic one can think of, which can be accessed through search engines like Google; a popular search engine. These search engines also have developed complex algorithms to present only the useful information to user based on a search query (e.g. snippet generation against a search query. However, the success is only partial due to the ever

[1] Cisco Annual Internet Report (2018–2023) for more Insights on Growth of Data & Devices

evolving techniques for the crave of improvement in terms of accuracy and efficiency. Hence, the problem still persists, where we have to extract meaningful information from this continuously increasing information overload. To cope up with this challenge of information overload research in associated fields (specially NLP) have gained popularity in the recent past. Popular data forms which are used in daily life include multimedia (video, images) and textual data (new, articles, books, social media posts).

1.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a subfield of linguistics and computer science with an addition of artificial intelligence. It is normally concerned with the processing and analysis of human language by some computer program. Human language comprises of both textual and speech or audio data. The goal of NLP is to understand the contents of human language including the contextual nuances of the language within them. This Natural Language Understanding (NLU) can be utilized for extracting meaningful insights, classification or categorization of the content etc. Now a days even new content similar to human language can be generated with the use of advanced machine learning algorithms trained for Natural Language Generation (NLG). NLP techniques varies from POS (Part of Speech) tagging to machine translation and automatic summarization. Major techniques are listed below.

- Part of Speech Tagging (POS)
- Nammed Entity Recognition (NER)
- Natural Language Inference
- Semantic Textual Similarity
- Question Answering
- Document Classification
- Sentiment Analysis
- Language Modeling
- Speech Recognition
- Automatic Summarization
- Machine Translation

1.2 Automatic Summarization

Automatic Summarization is the process of extracting only the meaningful information from text resulting into reducing the length of text as well as maintaining the information included in it. In addition to textual data, audio and video data can also be summarized however algorithm and procedure will differ. Summary normally comprises of the most

important and relevant information of original data or content. Summarization can be categorized on the basis of numerous criteria shown in Fig. 1.1.

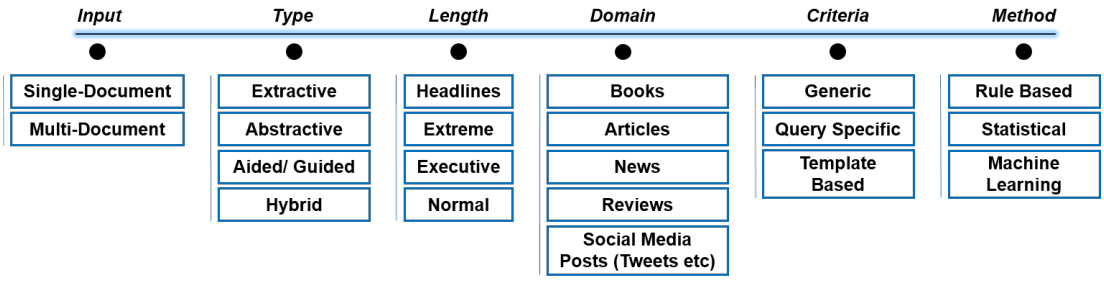


Figure 1.1: Categories of Automatic Summarization

1.2.1 Approach

- Extractive Summarization - Summary of a content is generated by selecting or extracting part of content from the original input content. This content can be sentences, paragraphs or part of sentences / phrases.
- Abstractive Summarization - This type of summarization is generated using NLG techniques. Abstractive summary of an input content comprises of new words, phrase or sentences which are not present in the input content and are generated based on the meanings or context of a input content.
- Aided Summarization - Aided or Guided summarization also includes the participation of Humans for derivation of a summary. This human element can be in the form of selecting or filtering the candidate summary produced by computer program or can be used for supervised learning of a automatic summarizer.
- Hybrid Summarization - Summary can be generated by using the combination of any of the above mentioned approaches or all.

1.2.2 Input

Input or original content to be summarized can be segregated based on the quantity. It can be *Single-Document Summarization* in which input of the summarization process will comprise of single content on a subject for instance a lecture, a news or research article or a book. On the contrary in *Multi-Document Summarization* the input of the

summarization process will comprise of more than one content on a subject for instance summary of multiple tweets on a subject (e.g. elections) or summary of multiple news article on a single or multiple subjects.

1.2.3 Length

Summary can be categorized on the basis of length into various categories. Normally a summary is of $\frac{1}{3^{rd}}$ length of the original content which is approx 33%. These categories includes *headlines;* single sentence, *extreme summarization;* very short length comprising of 2-3 sentences, *executive summary;* one pager, *normal summary*; 33% etc.

1.2.4 Domain

Summary can be categorized as per the domain of textual data. There are variety of domains ranging from *books, articles, to social media posts, reviews and captions of movies and dramas.* These domains even have sub domains; there are different kind of articles from *scientific research, medical to political events in news article.*

1.2.5 Criteria

Summary can represent the type of information which is actually required by the user. This can be achieved through the criteria mentioned by the user. For instance, summary related to only certain topic words or sub-subject in an input content such as *query based summarization.* Summary can also be based on a pre-defined *template* by the user as per requirements.

1.2.6 Method

The process of summarization or the technique and methods used for generation of summary from a textual content can be divided into:

- *Rule Based* - Early techniques or methods utilized for deriving a summary was symbolic methods based on some rules, logics and conceptual ontologies due to the very nature of human language being a symbolic one.

- *Statistical Methods* - Despite being symbolic in nature, natural languages are ambiguous, variable and complex. Hence the need for more statistical methods involving calculations based on various features to achieve a useful summary of a content.

- *Machine Learning* - With the evolution of Artificial Intelligence (AI) and Machine Learning (ML) algorithms having access to high processing capability like Graphics Processing Units (GPUs), ML based methods are frequently being used for Automatic Summarization. Machine learning is a branch of Artificial Intelligence which is a study of algorithms which can improve automatically through experience or learning by the use of data and models.

1.3 Applicability

Automatic Textual Summarization is one of the solution for information overload. Summarization can be used in variety of daily tasks resulting into easy Information Retrieval (IR) of overwhelmed users to help achieve efficiency in their tasks.

- *News Summary* - Massive growth of media outlets have also resulted into excessive information being produced which is often not being able to processed and absorbed by the user. Biasness in narratives also creates propagandas for which it is very necessary to have a summarized form of all the information available to users (different from the headlines already published by the news outlets) for a comprehensive and true picture.

- *Books Summary* - Books summaries are very useful for getting to know details about or before selecting to dive deeper into a book without wasting much of readers effort and time.

- *Summary of Articles* - With massive amount of publications in scientific articles, the hardest part for a researcher is to find the useful information he/she is looking for. Summary of articles (scientific as well as other sub-domains) are very useful for getting the information of internet efficiently.

- *Summary of Reviews* - In the last decade E-Commerce has seen an enormous boom specially after the pandemic of Covid-19 recently. The best source of information

about a certain product which creates its credibility is the product's review. However its also difficult sometimes to go through number of reviews. This is where automatic summarization of reviews does its job and make the process easy for the user.

- *Microblog / Tweets Summary* - After news outlets, microblogs are now becoming a primary source of news, latest trends and information exchange. However the amount of users and content being created over microblogs makes it impossible for the user to manually get a summarized overview. Microblog Summarization can be utilized by normally users as well as specialized departments (For instance law enforcement agencies to keep an eye on social media for emerging events, disasters etc).

- *Opinion / Sentiments Summary* - Opinion and Sentiment analysis can be carried out in any domain microblogs, reviews, discussion forums etc. Opinion / Sentiments Analyses can help getting to know the polarity of crowd in a discussion towards a particular topic, product etc. It can also be used to detect the user stance in various polls.

- *Summary of Legal Documents* - Legal Documents are always difficult to analyze due to the amount and complexity associated with it. However with the use of various user defined summarization methods the analyses becomes very easy and efficient saving effort and time.

- *Lecture Summary* - Lectures and tutorials have been the primary source of learning for students. However identification of a relevant lecture or even a relevant content in a lecture is a difficult and time consuming task before actual learning. Lecture summarization is a solution to this problem, moreover few cross domain lectures necessitates the need of only studying gist of the material instead of complete content.

1.4 Urdu Language

Urdu is 10th most spoken language [78], (230 million people) in the world. It is member of Indo-Aryan group in Indo-Europeon family of languages. Urdu is the national language of Pakistan and official language in Jammu and Kashmir. It is widely spoken

in south Asian regions, Pakistan, India, Afghanistan, Bangladesh, Nepal and Bahrain. Its vocabulary is derived from Persian, Arabic and Turkish. Urdu also shares its origins with Hindi language however Hindi is written in Devanagri, similar to Sanskrit having more influence than Persian and Arabic.

If spoken colloquial contexts are broadly considered, (Hindi-Urdu) is the 3rd most spoken language in the world. Urdu is relatively complex and morphologically rich language. It contains 38 alphabets, 25 consonants and 12 vowels. It varies from English language in many ways. Urdu script is written in Nastaliq style in which most of characters acquire different shapes depending on the position of character in the ligature. It is written from right to left and blank spaces doesn't necessarily means segregation of words like in English language, hence word and sentence boundary detection is difficult in Urdu. Moreover, there is no concept of word capitalization in Urdu making tasks like NER, sentence segmentation by detecting boundary through capitalization becomes more difficult. Despite being a popular language with a lot of content over the internet in multiple forms (books, articles, news, microblog posts, forums etc) there is little to no research related to Urdu language available in the field of Automatic Summarization and NLP.

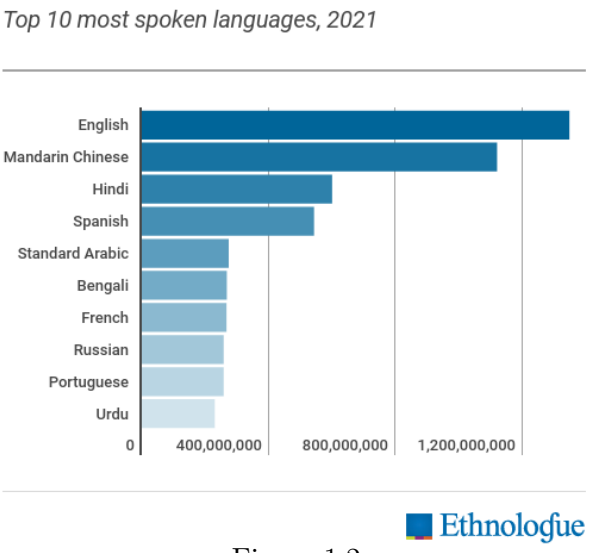


Figure 1.2

1.5 Motivation and Research Objectives

Considering the limitation of available research in Urdu language, an effort has been made to create a dataset for extreme summarization comprising of 65k news, summary pairs collected from Urdu news website Deutsche Welle (DW)[2] and 12k news, summary pairs collected from BBC Urdu website [3]. DW is a German's state-owned international broadcaster providing content in 32 languages founded in 1953 while BBC (British

[2]DW Urdu - https://www.dw.com/ur
[3]BBC Urdu - https://www.bbc.com/ur

Broadcasting Corporation) is the national broadcaster of United Kingdom, largest in the world founded in 1922 providing services in more than 40 languages. As per best of our knowledge it will be the first and largest dataset available for Urdu summarization. Exploration of Pre-trained Language Models based on self-attentive transformers is carried out for both Extractive and Abstractive Summarization capturing contextual information. Experimental Results demonstrated competitive Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [15] score of summarization model finetuned on newly created dataset with high resource languages like English. Moreover, BERTScore [71] and human evaluation is also carried out identifying the shortcomings of automatic evaluation methods in the field of summarization.

1.5.1 Research Contributions

Our main contributions are enlisted:-

- A *methodological framework* for utilizing deep learning based pre-trained language models trained for NLU and NLG on multiple languages for summarization of a single low resource language by reducing the size of model to fit in low resource settings.
- Creation of a *summarization dataset* in a low resource language from publicly available source in news domain chosen mainly because of its availability in multiple languages as well as ability to replicate for other low resource languages. This created dataset is the first and largest Summarization Dataset in selected low resource language i.e. Urdu.
- Experimental results demonstrating *competitive evaluation score* (ROUGE, BERT Score) of summarization model with reduced size (only monolingual vocabulary of a low resource language) finetuned on newly created dataset. Human evaluation also identified the *shortcomings of evaluation methods* in the field of summarization.

In rest of this Thesis, Chapter 2 provides an overview on evolution of summarization from early statistical, graph based approaches to machine learning and latest deep learning based approaches. Urdu Summarization framework comprising of self-attentive transformer based pre-trained language models and dataset has been described in Chapter 3

followed by Experiments and Results in Chapter 4 and Conclusion & Visualized Future Work in Chapter 5.

Chapter 2

Evolution of Automatic Summarization

Text Summarization was first focused in late 1950s where basic statistical or rule based approaches were used for summarization. These basic statistical approaches were refined with additional features and more statistical models. With the decline in between it became an area of focus again in 2000s where Document Understanding Conference (DUC)[1] and later became Summarization track in Text Analysis Conference (TAC)[2]. Major approaches of Summarization through its evolution are Statistical Approach, Graph Based Approach, Machine Learning Approach.

2.1 Statistical Approach

In early 1950, the use of term frequency (TF) was introduced for summarization for the first time [1] (generating abstract) by scoring the sentences (of technical papers and magzine articles) based on significant words derived from its frequency. Stopwords which are meaningless (a, the, is, are, there etc) was not considered. Relevant to TF, key phrases, headlines & titles, cue words and structural positions were explored later in 1969 by *Edmundson* [2]. The idea of TF was enhanced through Inverse Document Frequency (IDF) [3] to avoid biasness (repition of words vs distribution over documents) for multi-documents. IDF is a measure of how much information the word provides; either it is a common one or rare accross multiple documents. SUMMARIST [6] was proposed for

[1]DUC - https://duc.nist.gov/

[2]TAC - https://tac.nist.gov/tracks/index.html

identification of topics, fusion of concepts / topics and then either selection (extractive) or generation (abstractive) through fusion of phrases. Centroid based summarization [7] involving clustering of documents also became popular. Other techniques includes use of Latent Semantic Analysis (LSA) [9], Bayesian Model [22], Rhetorical Structure Theory (RST) [13], Maximal Marginal Relevance (MMR) [5], Conditional Random Fields (CRF) [20] etc. These statistical methods of summarization are still common with variations and additional considerations like postition, NounPhrase (NP), Named Entity Recognition (NER), cue words for events etc. These approaches are normaly favourable for extractive summarization in which usually the sentences are weighted so that they can be selected for summary. In addition, Sentence fusion and compression techniques are also in use. A simple statistical approach is illustrated in Fig. 2.1.

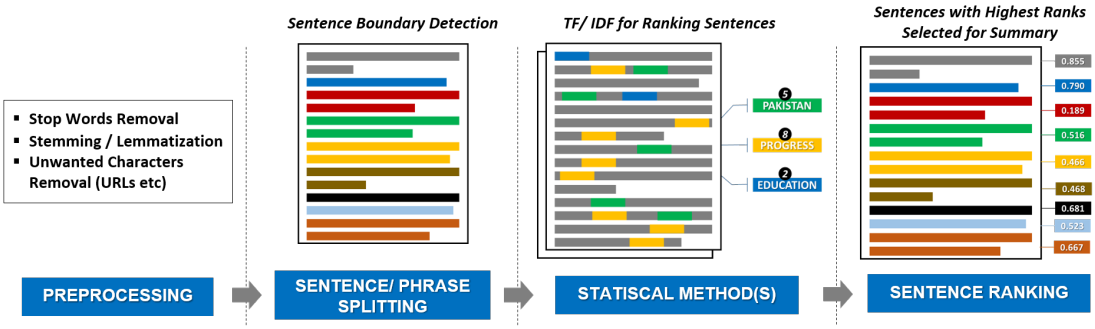


Figure 2.1: Illustration of statical approach using TF-IDF & sentence scoring

2.2 Graph Based Approach

TextRank [16] and LexRank [14] are popular graph based method derived from Google's PageRank algorithm. In both the algorithms graphs were created for sentences in documents. LexRank uses TF-IDF vectors and their cosine similarity while TextRank uses similar measure of cooccurence of words in sentences divided by sentence lengths. LexRank was only focused on summarization (can also be used for phrase extraction) basing on centrality of sentences while TextRank was demonstrated for phrase and sentence extraction with continous similarity scores as weights. LexRank is used mostly for multi-document summarization however TextRank is used for single-document summarization. Opinosis [25] was proposed for using graphs for abstractive summarization utilizing the concept of paraphrasing. *Liao et al.* [48] used the idea of Abstract Meaning Representation (AMR) graphs for summarization. Source documents were condensed

using AMR graphs and then summary was generated. Approximate Discourse Graphs (ADG) were used by G-Flow [28] for multi-document summarization in which sentences share discourse relation as graph edge. Graph based approach using TextRank and word2vec is depicted in Fig. 2.2

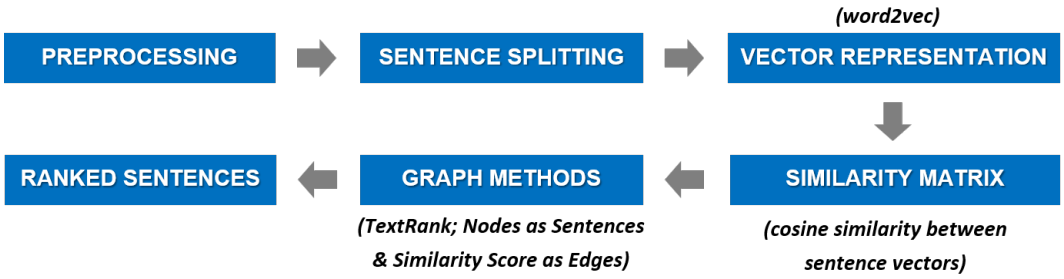


Figure 2.2: Graph based approach using TextRank & word2vec

2.3 Machine Learning

Machine Learning aproaches involved training of an algorithm which learns to perform some task without being explicity programmed. Machine Learning algorithm used in Automatic Text Summarization include simple classification, clustering, dimentionality reduction to deep learning based algorithms basing on Aritificial Neaural Networks (ANN).

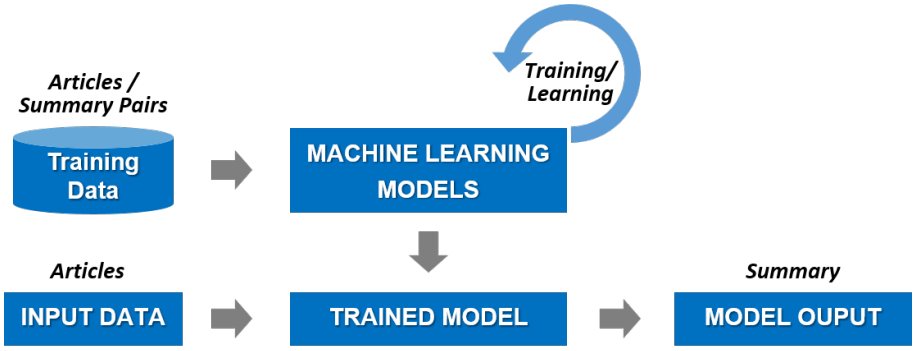


Figure 2.3: Machine Learning based approach

2.3.1 Supervised Learning

Supervised learning algorithms requires a training data with labels considered as true output which can be used to learn the loss / error while depicting an output using a

particular machine learning model; aim of which is to reduce error / loss while predicting an output from future unknown inputs. A simple machine learning approach (supervised) is shown in Fig. 2.4. Early machine learning technique for text summarization involved a binary classifier [4] which classifies a sentence to be included in summary or not. In 2001, Hidden Markov Model (HMM) [8] was used for getting the likelihood of sentence inclusion in summary. Use of maximum entropy [12] was used for sentence selection to be included in summary. *Neto* [11] used Naive Bayes to classify sentences based on various features like position, cohesion, length, title, keywords. Contrary to normal practice of sentence extractions or selection in the past, Sentence compression was explored using Decision Trees [10]. Latent Semantic Analysis (LSA) and Singular Value Decomposition (SVD) [18] was used to explore the sematic representation and generate summary respectively. Support Vector Machine (SVM) [19] was used for a query focused summarization. These algorithms were trained to predict, classify or rank the sentences as a summary sentence from all the candidate sentences in an input document. This prediction is based upon various features like centrality, position, similarity with topic words, cue words etc, entropy etc. To improve the efficiency of these algorithms various preprocessing steps also evolved with time; Part of Speech (POS) tagging (Nouns, Verbs etc), Stemming (reducing words to their base/ root word), Lemmatization (improves stemming by considering morphological analysis of the words), filtering unwanted tokens (removing stop words, URLs, punctuations etc), case sensitivity etc. Various supervised algorithm have been trained for domain specific summarization with the inclusion of ontology as well (legal documents, biomedical etc). A simple machine learning approach is shown in Fig. 2.3.

2.3.2 Unsupervised Learning

Unsupervised learning doesn't require labelled or tagged data which is considered as a true result or output for any input of a computer program to train itself or learn to produce the desired result. Without supervised training or labelled data unsupervised learning achieve subject learning through pre-designed algorithms like clustering. Early use of clustering algorithms for summarization was carried out by *Radev et al.* in 2000s [7]. Various clustering methods from centroid, density, distribution, hierarchical, fuzzy clustering are being used. Clustering is normally used with some additional method (for instance clustering is also used in LexRank [14] as mentioned in 2.2). In summarization

using clustering popular idea is to cluster the sentences basing on measure (like similarity), find centroid or centrality of the complete document (may be basing on the size of cluster - having most number of members for popular/ important content), compare all the sentence with the centroid and select summary sentences based on the similarity with the centroid because normally centroids represents the overall idea of the input document. Same idea can be employed for words and other features [24]. Scientific article were summarized using its citation network by applying clustering approach [21]. Clustering also have a relation with dimensionality reduction and topic modelling in terms of grouping a document into dimensions (if cluster is considered as a dimension) [27]. Different clustering techniques like hierarchical clustering was also used for summarization [26]. Unsupervised machine learning based approach using LDA, TF-IDF and clustering is shown in Fig. 2.4.

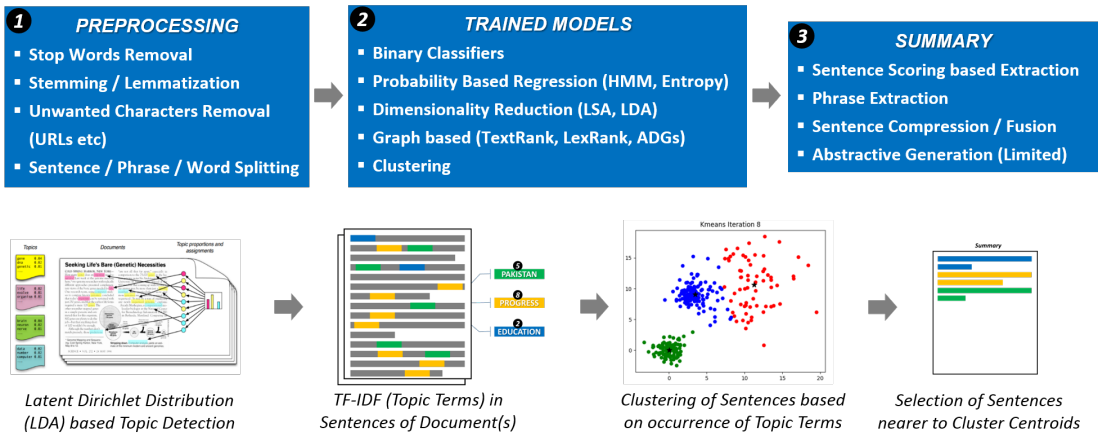


Figure 2.4: Machine Learning based approach using LDA, TF-IDF and clustering

2.4 Deep Learning

Early deep learning models like RankNet [17] algorithm in which ranking of sentences as simple probablistic cost function using artificial neural networks was proposed however these models were mostly word-based models considering them as bag of words or scaling such models to sequences but without contextual information of words in a sequence i.e. sentence(s). Phrase clustering [23] instead of words clustering was used for NER to introduce context of words. Word "Bank" in "Bank of River" and "Bank of Punjab" have different contexts which cannot be differentiated once using word clustering however when phrase clustering is being used than both will be clustered separately; however

considered as indirect method of capturing contextual information.

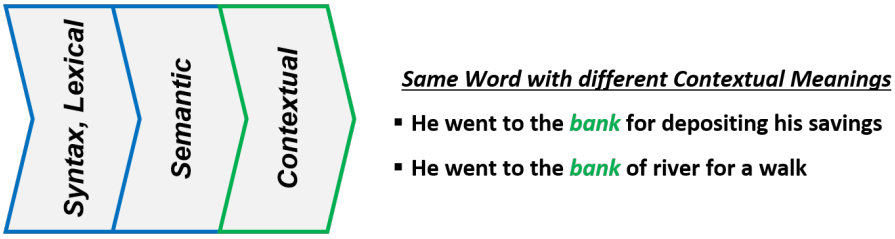


Figure 2.5: Towards Contextual Learning

2.4.1 Seq2Seq Models

Sequence to Sequence models evolved using Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) (including Long Short Term Memory (LSTM) & Gated Recurrence Unit (GRU)) to enrich syntactical and semantic analysis with contextual information.

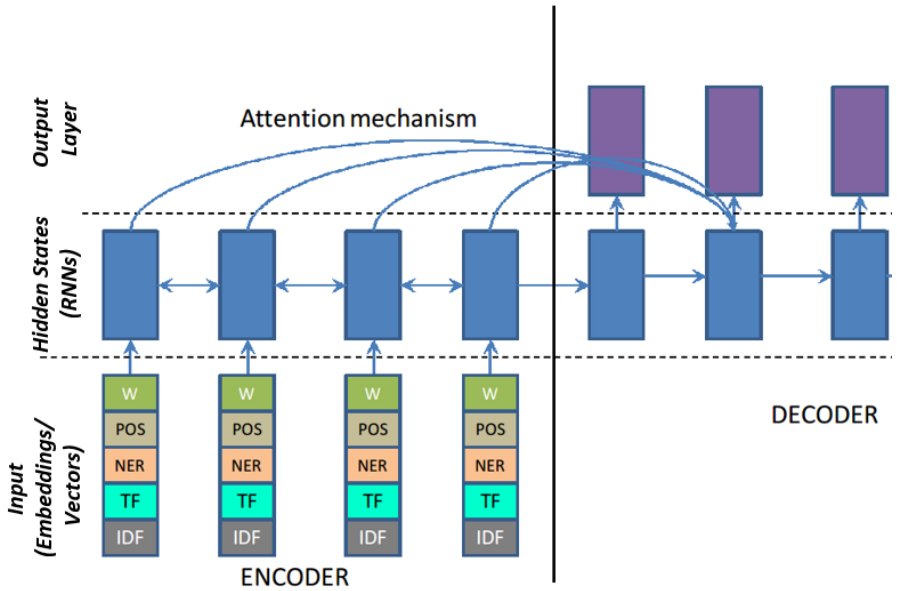


Figure 2.6: Seq2Seq based models using Recurrent Neural Networks[38].

In 2015, a local attention-based model [34] is used which generates the summary (headlines) conditioned on the input sentence by joining probabilistic model with a generation algorithm. An convolutional based model is used for encoding to learn soft alignment between input and the summary based on the context which is inspired by *Bahdanau et al.* [30] in which encoder was used to encode source sentence into a fixed length vector.

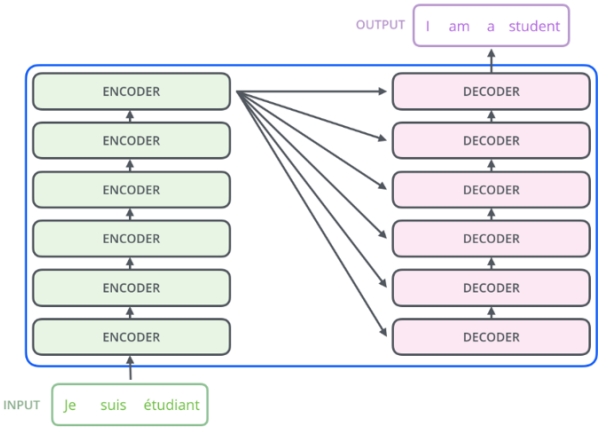
Input source in then weighted by the use of learned soft alignment. A beam-search decoder [33] is used which is a compromise between exact and greedy decoding and efficient from phrase-based machine translations in terms of computational time.

Chopra et al. [36] extended the work of *Rush et al.* [34] by replacing decoder with RNNs. In 2014, *Hu et al.* [31] and *Cheng & Lapata* [35] also made use of attentional encoder decoder RNNs inspired by *Bahdanau et al.* [30]. Encoder was a bidirectional GRU-RNN (Gated Recurrent Units) while decoder consisted of a uni-directional GRU-RNN with same hidden-state size. *Nallapati* [38] also extended the framework with proposing novel models which includes feature rich encoder (i.e. in addition to word embedding various feature embeddings; POS, NER, TF, IDF are also included in encoder input), switching generator-pointer model is used for OOV (Out of Vocabulary) words and hierarchical attention in which word level attention is further influenced by sentence level attentions with positional embedding of sentences as shown in Fig. 2.6. Selective gate network [43] was introduced to further enhance the process of distilling information for summary generation. Auto Variational Encoders [40] were proposed for latent structural modelling. Pointer-Generator [41] framework allowed use of pointers for pointing to source text to copy words which are required for summarization while also incorporating generators for retaining the ability to generate novel words not included in source text. In addition coverage was also used to keep track of what has been summarized to discourage repetitions to overcome shortcomings of previous RNN based models. Hierarchical encoder [55] based on RNN was used to capture document level dependencies / context which was used to extract sentences as well as score the remaining sentences. *Reinforcement Learning* [52] was proposed by adding saliency and entailment rewards for the output summary in training process. Bottom-up attention [46] was used to enhance content selection similar to hierarchical methods. The idea of global encoding [49] with the use of Gated Convolutional Unit was used to cater for repetitions and semantic irrelevance. A framework for retrieving candidate sentences, re-ranking on the basis of similarity and re-writing for summary generation was proposed using soft template [45].

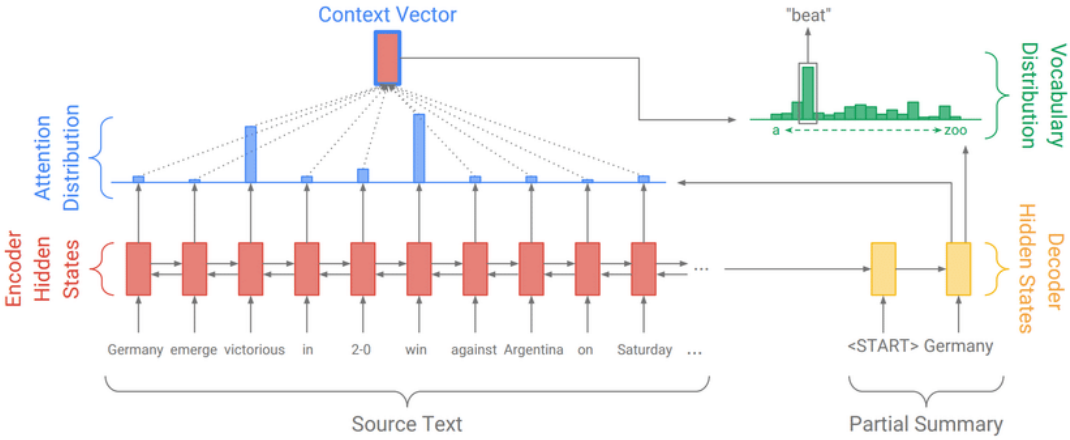
2.4.2 Transformers; Attention based Architecture

Sequence to sequence models (Seq2Seq) were based normally on LSTM, RNN, GRU or CNN with an attention mechanism for capturing dependencies between tokens (context).

RNN based models normally calculates hidden states^{ht} as a function of previous hidden state^{ht-1} for the input position^{t}. This sequential nature restricts parallelization within training examples putting memory constraints for longer sequences. Transformer model [42] was proposed entirely on drawing global dependencies between input and output without using recurrence or convolutional networks as shown in Fig. 2.7 which is called as attention mechanisms. Multi-head attention consisting of several attention layers in parallel based on scaled dot-product was used. In addition self-attention was used to reduce the computational complexity and positional encoding to keep track of order of sequence.



(a) Parallelization



(b) Attention Mechanism

Figure 2.7: Transformer based architecture. *Fig. (a) from "The Illustrated Transformer" by Jay Alammar & Fig. (b) from "Get To The Point: Summarization with Pointer-Generator Networks" [41]*

2.4.3 Transfer Learning

word2vec [29] and GloVe [32] was introduced to capture latent semantic and syntactic similarities by representing words into vector space. FastText [39] vector representation for characters n-gram derived in which words were being represented as the sum of these vectors instead of words as distinct vectors. These models are also scaled to sentences and documents however lacks contextual information over a complete sequence. ELMo (Embeddings for Language Models) [53] proposed the use of Bidirectional Language Models (BiLM) for learning contextual representation for words over an input sequence utilizing LSTM based model. *Zhang & Bowman* [71] demonstrated that LM based pre-training objective performs better than other task-specific pre-training and also for transfer learning. For a generalized LM objective ULFiT (Universal Language Model Fine-tuning) [47] was proposed which pre-trains a language model on wikipedia articles and fine-tunes it on the downstream tasks using novel techniques.

Language Models

Like humans, a computer program do have to understand and generate language for interactions based on human languages. Language models is an aproach in which some probabalistic or machine learning method is used to learn language representations. These learned representations can be used for multivarious tasks. *Transfer Learning* was enabled through learned representations of Language Models (LM) and their re-use for various downstream tasks (Summarization, Q/A, Inference etc). Transfer learning has become ubiquitous in NLP with an order of magnitude improvement in recent years. Transfer learning in NLP is not phenomenan evolved after the inception of transformers architecture based on attention mechanism. Previous learned word embeddings [32] [53] were also used as LMs however with the introduction of sequence models based on transformers, long term dependencies and contextual information is captured in a more efficient way. Generally LMs are pretrained on large unlabelled data *(self-supervised learning)* and then adapted normally referred to as fine-tuning to a target tasks using labelled data.

BERT (Bidirectional Encoder Representations for Transformers) [58] was introduced in which masked language was used for pre-training using unlabelled data (self supervised training). It introduced the concept of bidirectional pre-training unlike previous

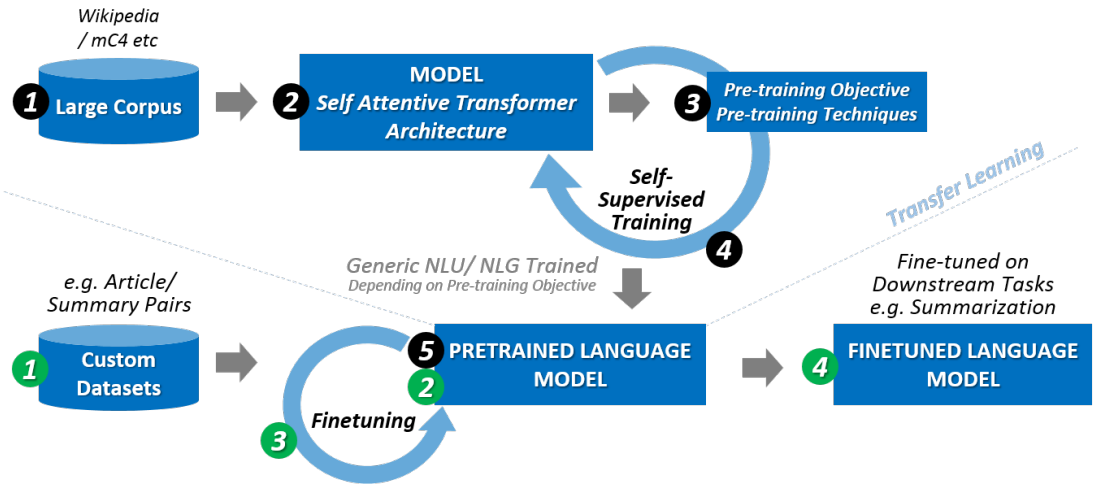


Figure 2.8: Pre-Trained Language Models *(Transfer Learning; re-useable Language Models)* & Fine-Tuning for downstream tasks

work in which either unidirectional language models pre-training was used GPT (Generative Pre-Training) [54] or concatenation of independently trained left-to-right and right-to-left LMs as shown in Fig.2.9 . BERT architecture was based on Self-Attentive Transformers [42]. BART (Bidirectional and Auto-Regressive Transformers) [62] has a similar architecture to BERT with minor changes incorporating the denoising encoder which add various arbitrary noise (masking, document rotation, token deletion, sentence permutation and text infilling) to input and unidirectional left-to-right decoder similar to GPT [54].

Use of BERT model was demonstrated for summarization [64] in which BERT encoder was used with transformer based decoder in a two stage approach; encoder is fine-tuned first with the objective of extractive summarization and secondly with the abstractive objective. BERT architecture was modified to represent sentence level embeddings to draw document level dependencies. MASS (Masked Seq to Seq pre-training) [66] was proposed inspired by BERT in which the pre-training of language model was enhanced by k masking tokens increased from k=1 in BERT. Pre-training of encoder-decoder was carried out jointly with the aim of language generation capability. Simple Encoder-Decoder is shown in Fig.2.10. BERT being a bidirectional transformer suited better for NLU tasks; to improve on language generation process (NLG), a bidirectional encoder is used for NLU however a unidirectional decoder conditioned on encoder input is used for NLG in UniLM (Unified Pre-trained LM) [59]. Various techniques are being uti-

lized using pre-trained LMs; Extraction of the candidate sentences for summary than paraphrasing it for generating summary utilizing *Reinforcement Learning* [56], Use of positional encoding to control the length of summary and improving evaluation score [67] etc.

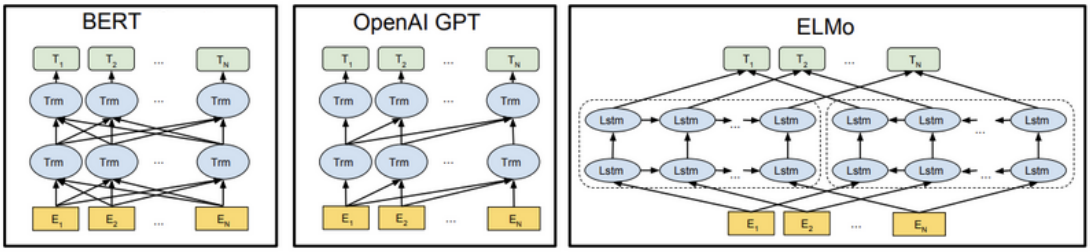


Figure 2.9: Difference between ELMO (Embeddings from Language Model)[54] using concatenation of unidirectional LSTMs, GPT (Generative Pretraining)[54] using unidirectional transformers & BERT (Bidirectional Encoder Representations for Transformers)[58] using bidirectional transformers; *Fig. from BERT*

PEGASUS (Pre-training with Extracted Gap Sentences) [70] proposed a pre-training objective for abstractive summarization with masked sentences and tokens together. XLNet [69] proposed autoregressive encoder instead of denoising autoencoder to overcome the dependency issue of masked positions integrating its ideas from Transformer-XL [57]. GSUM (Guided Summarization) [73] presented an idea of guided summarization network with both human and automatically extracted guidance signal (keywords, highlights, subject object relations) using BERT and BART. Various pre-training and fine-tuning methods have been proposed to improve results on specialized tasks like in SpanBERT [60] spans were masked instead of tokens to improve on tasks like QA. Internal working during capture of global dependencies between input and output & contextual information is normally difficult to capture however, its equally important to understand how BERT captures the linguistic information and which region of model is responsible for various NLP pipeline [68] (POS, NER, Parsing, Semantic Roles, coreference). Various fine-tuning methods have been studied as well, Rectified Adam [63] was proposed to improve variance in adaptive learning.

T5 (Text-to-Text-Transfer-Transformer)[75] goal was not to propose a new model of innovative architecture instead a deep study of all available research including pre-training, architectures, transfer approaches, datasets and other miscellaneous aspects which effects NLP tasks was carried out and efficient approaches were selected for creation of a

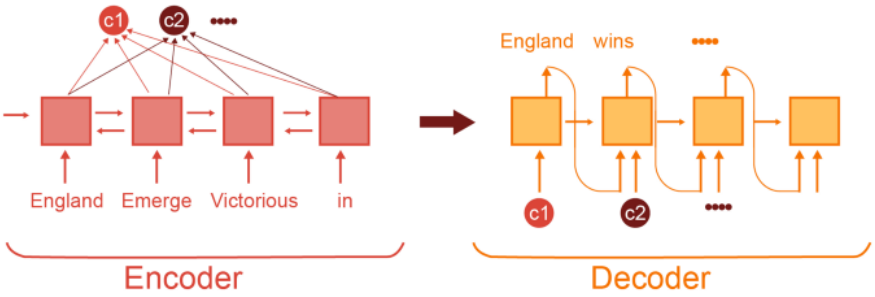


Figure 2.10: Transformer based encoder decoder with contextual information

text-to-text model. It comprised of survey of existing research in the field, their comparison, limitations and in the end utilizing the takeaways of the study and training a model which achieves state-of-the-art (SOTA) in language understanding and various downstream tasks like summarization. Baseline model was designed with an encoder and decoder. The encoder / decoder architecture was similar to BERT [58] (except that it is an encoder only model). Model was trained using "masked language modelling" and denoising objective inspired by BERT.

Multilingual Transfer Learning

Transfer learning doesn't only provides with the benefit of learning a language model to perform multiple downstream tasks however it also provides a major benefit of cross-lingual learning. Multilingual models tends to train on many languages at once by sharing subword vocabulary. Pre-training objective was extended from monolingual to crosslingual training [58], [44] & [61] with shared BPE vocabulary, evaluation of these models on various tasks showed improved results. These models based on concept of BERT presents a strong baseline however suffers with the disadvantage of under-representation of low resource languages [65]. Various models have also explored cross-lingual training with training of parallel corpus with high resource languages through automatic techniques of creating synthetic datasets to cater for problem of low resource languages.

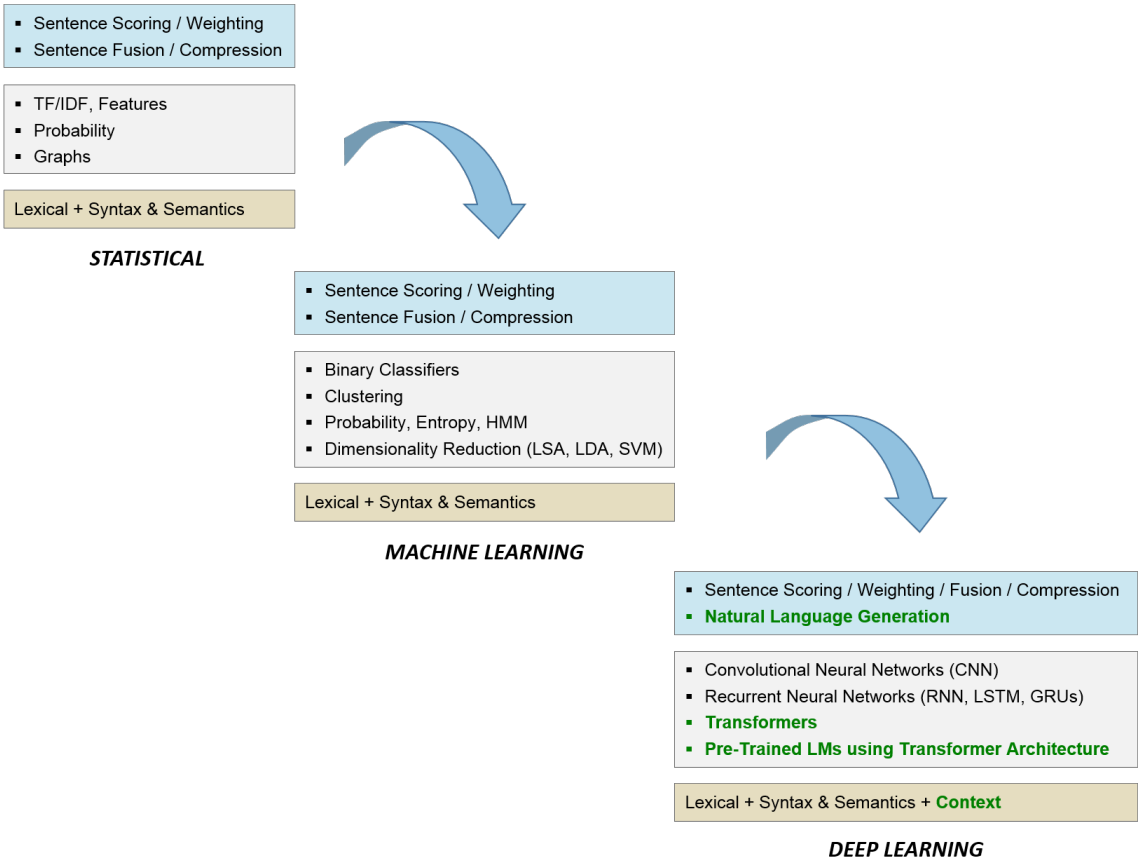


Figure 2.11: Evolution of Summarization

Chapter 3

Urdu Summarization: Models and Datasets

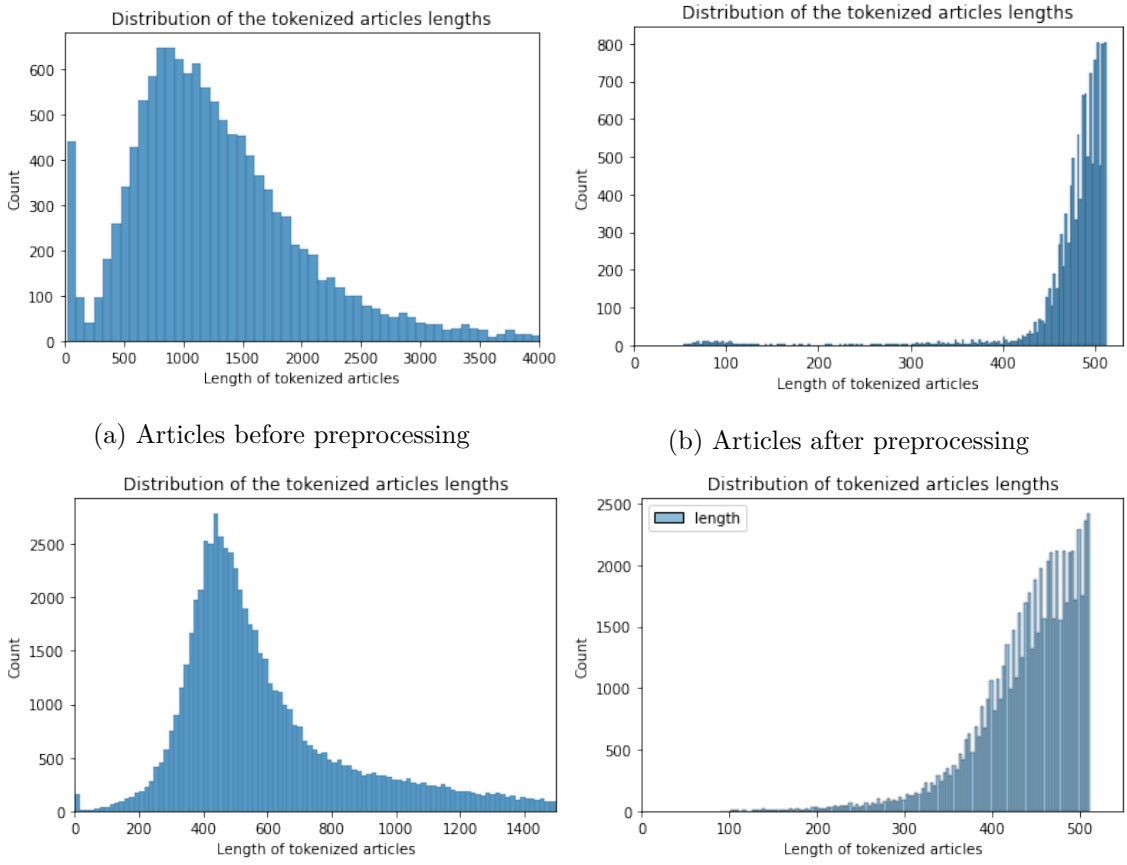
Urdu despite being a popular language doesn't have requisite resources in Automatic Summarization. There is only one dataset of 50 records available for which is not suitable for training of machine learning based algorithms. *Humayoun et al.*[37] constructed "Urdu Summary Corpus" consisted of 50 human written Urdu articles along with their summaries. These summarization were abstractive in nature and belonging to eight different news categories. Preprocessing was also carried out in the research including POS tagging, stemming & lemmatization, space segmentation etc; code and datasets made publicly available.

Noman et al. [50] utilized *Urdu Summary Corpus* for Summarization. Summarization was carried out using statistical method of sentence weight algorithm using words probability with an addition of position weights. *Ali et al.* [74] also carried out Urdu Summarization and comparison between various techniques using statistical methods (sentence weight, weighted term frequency, TextRank, distributional semantic model etc). Machine learning based embedding model for learning vocabulary on 600 articles were also utilized in sentence weight algorithm. Similar dataset of Urdu Summary Corpus was used however with an additional extractive summaries added to previous dataset of 50 records.

As per best of my knowledge no requisite dataset is available for Urdu Summarization which can be utilized for tuning models based on transformers architecture nor any research has been carried out on Urdu Summarization using machine learning methods.

3.1 Summarization Dataset

Creation of a requisite large human written document / summary dataset is an expensive and time taking task. Recently datasets have been created by utilizing the existing resources from publicly available data *(i.e. reviews and their summaries, news websites containing article / summary pairs etc)*. News domain presents a suitable choice for creation of summarization dataset being 1. publicly available and 2. easily collectable, 3. in multiple / local languages, 4. with no synthetic data as most of the news websites contains article / summary pairs written from multiple human authors. Availability of popular news platforms (in addition to local news resources) in multiple language also makes it the best possible choice for acquisition of summarization dataset.



(a) Articles before preprocessing

(b) Articles after preprocessing

(c) Articles before preprocessing

(d) Articles after preprocessing

Figure 3.1: Preprocessing: Distribution of tokenized article lengths. BBC Urdu dataset in (a) & (b), DW Urdu dataset in (c) & (d).

3.1.1 Overview & Preprocessing

Two news websites (BBC Urdu[1] and DW Urdu[2]) were selected which have 2-3 lines of short summary written by multiple writers in additional to the news articles. These two websites were scrapped for article / summary pairs. A dataset of 76.5k records having Article / Summary pairs were scrapped *(12k records from BBC Urdu and 64.5k records from DW Urdu)*. Dataset was tokenized using spaCy tokenizer[3] (word-based tokenizer), mBERT and mT5 (upto sub-word tokenizer; i.e. WordPiece) for exploration of dataset and length analysis. Statistics regarding tokenized lengths in shown in Fig.3.1. Detailed statistics of tokenized lengths and compression ratio (before / after preprocessing) are in Table 3.1 and Table 3.2 for BBC Urdu and DW Urdu dataset respectively.

Preprocessing steps involved are:-

- *Multimedia* - Only text-based articles were selected to be included in the dataset excluding Multimedia Based Articles which has comparatively lesser text in articles distorting compression ratio and training of models. Search Query: **contentType=ARTICLE** in URL similar to `https://www.dw.com/search/?languageCode=ur&contentType=ARTICLE`

- *Links / URLs* - All types of links, URLs were removed. For instance, links of associated articles were removed from articles as depicted in Fig.3.2.

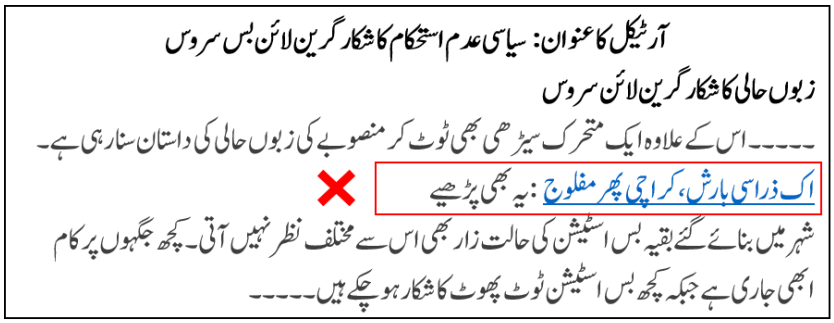


Figure 3.2: Removing Links of Associated Articles; Link Depicted in Blue Colour

- *Picture Captions* - These news website also had pictures, screenshots of tweets etc. inside articles which also had captions, picture captions as shown in Fig.3.3

[1]BBC Urdu - `https://www.bbc.com/urdu`
[2]DW Urdu - `https://www.dw.com/ur`
[3]spaCy - `https://spacy.io/`

were included in text while scrapping which breaks the coherence of text hance removed.

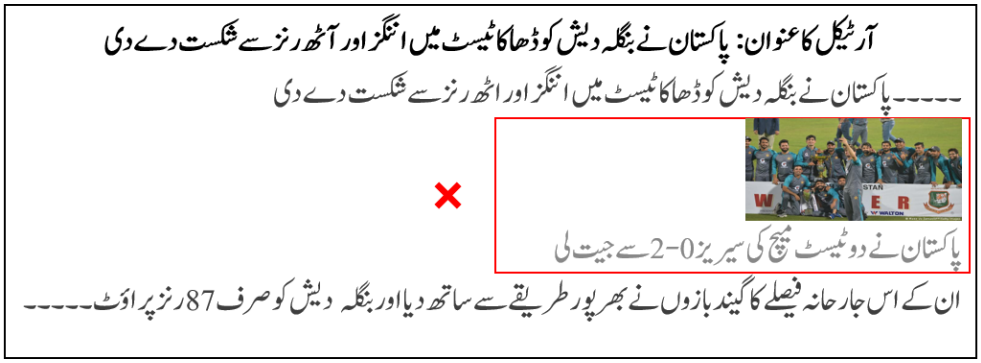


Figure 3.3: Removing Caption of Picture; Caption of Cricket Match Depicted in Grey Colour

- *Compression Ratio* - Compression Ratio was calculated for each record using tokenized length. Records having compression ratio more than 50% were removed *(i.e. 830 records)*.

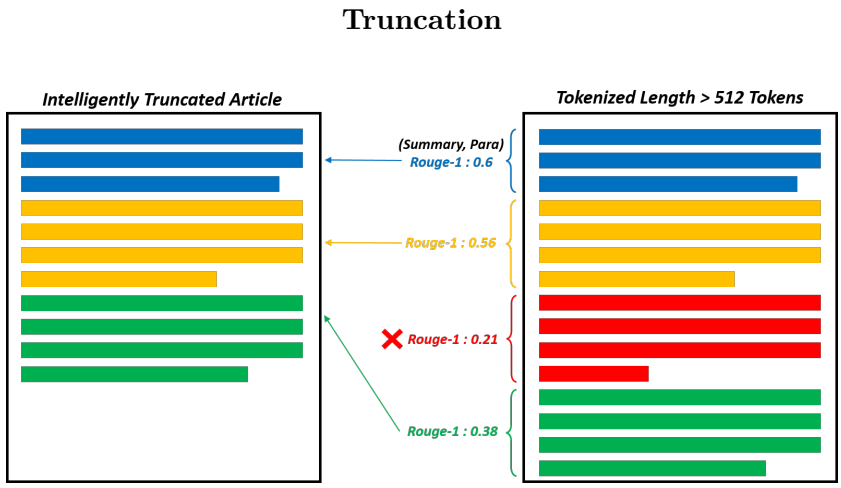


Figure 3.4: Truncation of long articles by removing paragraphs with low Rouge (Recall) scores

Lengths of articles in both datasets are larger than the input processing capability of models being utilized as shown in Fig.3.1. Tokenizers of pre-trained multilingual models based on transformers (e.g. BERT and mT5) differs from word based tokenizers (e.g. spaCy). These tokenizers implements upto sub-words tokenization therefore increasing

length of article even more then word based tokenizers. BERT based models have an input processing limit of 512 tokens whereas practically there is no limit of processing input text for mT5 however, memory consumption exponentially increases with higher length input. Articles longer than 512 token will automatically be truncated to 512 tokens resulting into removal of important information required for summarization. To cater for this limitation of models and to restrict memory consumption of mT5 based models truncation has been carried out using Rouge-1 Recall between article paragraphs and summaries as shown in Fig.3.4 and Procedure.1. It includes:-

- Initially the dataset has been scrapped not as a complete text but comprising of paragraphs as originally written by the author of news article.
- ROUGE-1 Recall score has been calculated for each paragraph in a news article as compared to the original summary.
- Only paragraphs with higher Recall score have been included in the articles. Paragraphs with lowest Recall were excluded till the time its length comes within 512 tokens.

BBC Urdu Dataset

	Before Preprocessing			After Preprocessing		
Attribute	Article	Summary	Compression %	Article	Summary	Compression %
Count	12415			12089		
Mean	1331.09	43.86	6.49	474.04	43.95	9.72
Std	908.34	12.15	12.15	61.74	12.21	4.67
Min	19.0	0.0	0.0	0.0	12.0	2.82
25%	788.0	35.0	2.54	471.0	35.0	7.42
50%	1163.0	42.0	3.73	488.0	42.0	8.88
75%	1165.0	51.0	5.51	501.0	51.0	10.84
Max	21544.0	130.0	141.67	512.0	130.0	50.0

Table 3.1: Token lengths before/after preprocessing alongwith compression ratio

Procedure 1 Truncation of Dataset using Recall Measure

Input: $article, summary$

procedure Tokenized_Length($article$)

$Encode\_Articles$ ▷ e.g.using BERT Tokenizer

for each $para \in article$ **do**

$length_{para} = len(para \in tokenized\_article)$

end for

return $length$

end procedure

if $length_{article} > 512$ **then**

procedure Score_Paragraphs($article, summary$)

for each $para \in article$ **do**

$para_{index} = i + 1$

$para_{text} = para$

$Rouge\_Score(article, summary)$

$para_{score} = Rouge - 1_{Recall}$

end for

end procedure

end if

while $length_{article} > 512$ **do**

$sorted\_paras = SortAsc(para_{score})$

$i = 0$

del $sorted\_paras_i$

$length_{article} - = length_{para_i}$

$i + = 1$

end while

$trunc\_article = SortAsc(para_{index})$

Output: $trunc\_article$

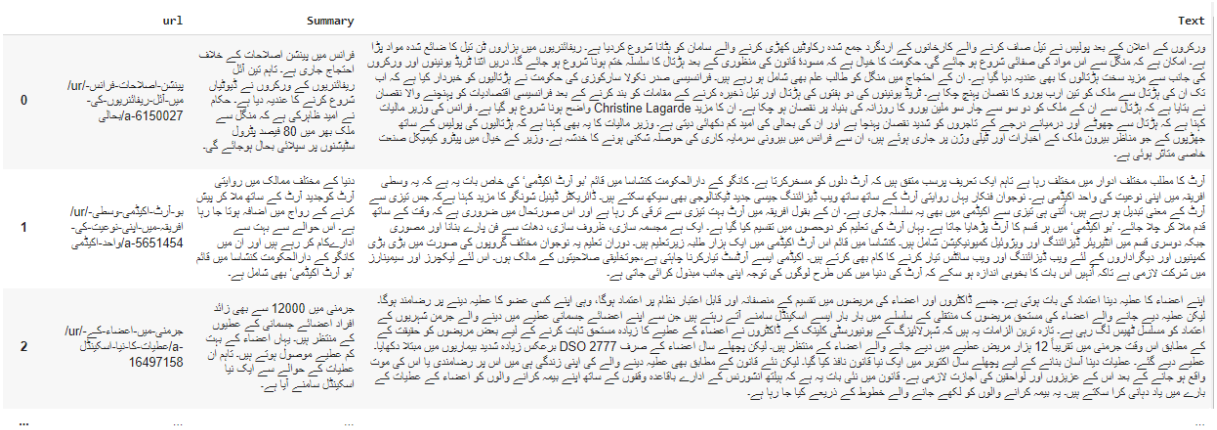
DW Urdu Dataset

Attribute	Before Preprocessing			After Preprocessing		
	Article	Summary	Compression %	Article	Summary	Compression %
Count	65044			64540		
Mean	652.33	41.18	8.52	440.54	71.84	16.73
Std	407.96	7.57	41.65	58.18	12.96	4.78
Min	0.0	1.00	0.00	0.0	17.0	4.04
25%	421.0	37.00	5.37	412.0	64.0	14.18
50%	524.0	42.00	7.64	453.0	74.0	16.40
75%	731.0	46.00	9.82	484.0	81.0	18.51
Max	6104.0	707.00	8200.0	512.0	256.0	50.0

Table 3.2: Token lengths before/after preprocessing alongwith compression ratio

3.2 Pre-trained Language Models

Pre-trained Language Models *(as depicted in Fig.2.8)* involves training of a model over a large corpus to learn the ability to understand and generate language representations. Low resource languages suffers from lack of available pre-trained language models however various popular models like BERT [58], T5 [75] [76] have been released with a scaled objective of multilingual pre-training. Multilingual models are trained over large corpus of multiple languages at a same period of time (shared vocabulary). This combined training though suffers from under-representations of language having comparatively less training data[65], still provides workable language model which can be used for various downstream tasks efficiently. Zeroshot settings of these models may not provide acceptable results on a specific task as mostly models are trained for understanding or generating language representations instead of specific tasks like summarization. However with little training (i.e. finetuning) these models can outperform various old methods of summarization (e.g. statistical). Moreover language generation capability can also be exploited for abstractive summarization which was not possible with earlier methods dependent on selecting or extracting part of content from the original input



	Text	Summary	url
0	ورکروں کے اعلان کے بعد پولیس نے ٹیل صاف کرنے والے کارخانوں کے اردگرد جمع شدہ رکاوٹیں کھڑی کرنے والے سامان کو ہٹانا شروع کردیا ہے۔ ریفائنریوں میں ہزاروں ٹن ٹیل کا ضائع شدہ مواد پڑا ہے۔ امکان ہے کہ منگل سے اس مواد کی صفائی شروع ہو جائے گی۔ حکومت کا خیال ہے کہ مسودۂ قانون کی منظوری کے بعد ہڑتال کا سلسلہ ختم ہونا شروع ہو جائے گا۔ دریں اثنا ٹریڈ یونینوں اور ورکروں کی جانب سے مزید سخت ہڑتالوں کا بھی عندیہ دیا گیا ہے۔ ان کے احتجاج میں منگل کو طالب علم بھی شامل ہو رہے ہیں۔ فرانسیسی صدر نکولا سارکوزی کی حکومت نے ہڑتالیوں کو خبردار کیا ہے کہ اب تک ان کی ہڑتال سے ملک کو تین ارب یورو کا نقصان پہنچ چکا ہے۔ ٹریڈ یونینوں کی دو ہفتوں کی ہڑتال اور تیل ذخیرہ کرنے کے مقامات کو بند کرنے کے بعد فرانسیسی اقتصادیات کو پہنچنے والا نقصان نے بتایا ہے کہ ہڑتال سے ان کے ملک کو دو سو سے چار سو ملین یورو کا روزانہ کی بنیاد پر نقصان ہو چکا ہے۔ ان کا مزید Christine Lagarde واضح ہونا شروع ہو گیا ہے۔ فرانس کی وزیر مالیات کہنا ہے کہ ہڑتال سے چھوٹے اور درمیانے درجے کے تاجروں کو شدید نقصان پہنچا ہے اور ان کی بحالی کی امید کم دکھائی دیتی ہے۔ وزیر مالیات کا یہ بھی کہنا ہے کہ ہڑتالیوں کی پولیس کے ساتھ جھڑپوں کے جو مناظر بیرون ملک کے اخبارات اور ٹیلی وژن پر جاری ہوئے ہیں، ان سے فرانس میں بیرونی سرمایہ کاری کی حوصلہ شکنی ہونے کا خدشہ ہے۔ وزیر کے خیال میں پیٹرول کیمیکل صنعت خاصی متاثر ہوئی ہے۔	فرانس میں پینشن اصلاحات کے خلاف احتجاج جاری ہے۔ تاہم تین آئل ریفائنریوں کے ورکروں نے ٹیوٹیاں شروع کرنے کا عندیہ دیا ہے۔ حکام نے امید ظاہر کی ہے کہ منگل سے ملک بھر میں 80 فیصد پٹرول سٹیشنوں پر سپلائی بحال ہوجائے گی۔	/ur/پینشن-اصلاحات-خرانس-میں-آئل-ریفائنریوں-کی-ہڑتال-a-6150027/بحالی
1	آرٹ کا مطلب مختلف ادوار میں مختلف رہا ہے تاہم ایک تعریف پرسب متفق ہیں کہ آرٹ دلوں کو مسخرکرنا ہے۔ کانگو کے دارالحکومت کنشاسا میں قائم 'بو آرٹ اکیڈمی' کی خاص بات یہ ہے کہ یہ وسطی افریقہ میں اپنی نوعیت کی واحد اکیڈمی ہے۔ نوجوان فنکار پہلی روایتی آرٹ کے ساتھ ساتھ ویب ڈیزائننگ جیسی جدید ٹیکنالوجی بھی سیکھ سکتے ہیں۔ ڈائریکٹر ڈینیل شونگو کا مزید کہنا ہےکہ جس تیزی سے آرٹ کے معنی تبدیل ہو رہے ہیں، اُتنی ہی تیزی سے اکیڈمی میں بھی یہ سلسلہ جاری ہے۔ ان کے بقول افریقہ میں آرٹ بہت تیزی سے ترقی کر رہا ہے اور اس صورتحال میں ضروری ہے کہ وقت کے ساتھ قدم ملا کر چلا جائے۔ 'بو اکیڈمی' میں ہر قسم کا آرٹ پڑھایا جاتا ہے۔ یہاں آرٹ کی تعلیم کو دوحصوں میں تقسیم کیا گیا ہے۔ ایک ہے مجسمہ سازی، ظروف سازی، دھات سے فن پارے بنانا اور مصوری جبکہ دوسری قسم میں انٹیریئر ڈیزائننگ اور ویژوئیل کمیونیکیشن شامل ہیں۔ کنشاسا میں قائم اس آرٹ اکیڈمی میں ایک ہزار طلبہ زیرتعلیم ہیں۔ دوران تعلیم یہ نوجوان مختلف گروپوں کی صورت میں بڑی بڑی کمپنیوں اور دیگرادروں کے لئے ویب ڈیزائننگ اور ویب سائٹس تیار کرتے کا کام بھی کرتے ہیں۔ اکیڈمی ایسے آرٹسٹ تیارکرنا چاہتی ہے جوتخلیقی صلاحیتوں کے مالک ہوں۔ اس لئے لیکچرز اور سیمینارز میں شرکت لازمی ہے تاکہ اُنہیں اس بات کا بخوبی اندازہ ہو سکے کہ آرٹ کی دنیا میں کس طرح لوگوں کی توجہ اپنی جانب مبذول کرائی جاتی ہے۔	دنیا کے مختلف ممالک میں روایتی آرٹ کوجدید آرٹ کے ساتھ ملا کر پیش کرنے کے رواج میں اضافہ ہوتا جا رہا ہے۔ اس حوالے سے بہت سے اداروں کام کر رہے ہیں اور ان میں کانگو کے دارالحکومت کنشاسا میں قائم 'بو آرٹ اکیڈمی' بھی شامل ہے۔	/ur/بو-آرٹ-اکیڈمی-وسطی-افریقہ-میں-اپنی-نوعیت-کی-واحد-اکیڈمی-a-5651454
2	اپنے اعضاء کا عطیہ دینا اعتماد کی بات ہوتی ہے۔ جسے ڈاکٹروں اور اعضاء کی مریضوں میں تقسیم کے منصفانہ اور قابل اعتبار نظام پر اعتماد ہوگا، وہی اپنے کسی عضو کا عطیہ دینے پر رضامند ہوگا۔ لیکن عطیہ دیے جانے والے اعضاء کی مستحق مریضوں کی منتقلی کے سلسلے میں بار بار ایسے اسکینڈل سامنے آتے رہتے ہیں جن سے اپنے اعضائے جسمانی عطیے میں دینے والے جرمن شہریوں کے اعتماد کو مسلسل ٹھیس لگ رہی ہے۔ تازہ ترین الزامات یہ ہیں کہ شیرلاٹینبرگ کے یونیورسٹی کلینک کے ڈاکٹروں نے اعضاء کے عطیے کا زیادہ مستحق ثابت کرنے کے لیے بعض مریضوں کو حقیقت کے برعکس زیادہ شدید بیماریوں میں مبتلا دکھایا۔ لیکن پچھلے سال اعضاء کے صرف 2777 DSO کے مطابق اس وقت جرمنی میں تقریباً 12 ہزار مریض عطیے میں دیے جانے والے اعضاء کے منتظر ہیں۔	جرمنی میں 12000 سے بھی زائد افراد اعضائے جسمانی کے عطیوں کے منتظر ہیں۔ پہلی اعضاء کے بہت کم عطیے موصول ہوئے ہیں۔ تاہم ان عطیات کے حوالے سے ایک نیا اسکینڈل سامنے آیا ہے۔	/ur/جرمنی-میں-اعضاء-کے-عطیات-کا-نیا-اسکینڈل-a-16497158
...	...	...	...

Figure 3.5: Dataset Sample

content. Summarization has been carried out in both extractive and abstractive category by first creating own dataset, later choosing available pre-trained language models which are trained on our low resource language and reducing its size to fit in low resource settings. Framework adopted to carry out our research is illustrated in Fig.3.6.

Although various models have been trained on our low resource language i.e. Urdu (for instance Fasttext; word based embeddings), however these models are word based embedding models which are not able to capture contextual information over long sequence. Pre-trained Language models in contrast due to their superiority in efficiency and accuracy are a preferable choice. Three Language Models which are pre-trained on Urdu language were available *Multilingual BERT (mBERT), T5 (mT5) and MuRIL*. These models are being used in following portion of paper for summarization using own created dataset i.e. *mBERT, MuRIL for extractive summarization and mT5 for abstractive summarization.*

3.2.1 Extractive Summarization

Extractive Summarization has been carried out by using mBERT based models which are trained on large corpus for NLU (not requiring NLG; utilized in abstractive summarization). These models are used as a zeroshot settings without finetuning to enhance the scope of experimentation by exploring various available options. Moreover, these BERT based models were used for finding the contextual embeddings which is similar to the pre-training objective of these models i.e. NLU. Sentence embeddings encoded by these models were later clustered using k-means clustering using Cosine similarity

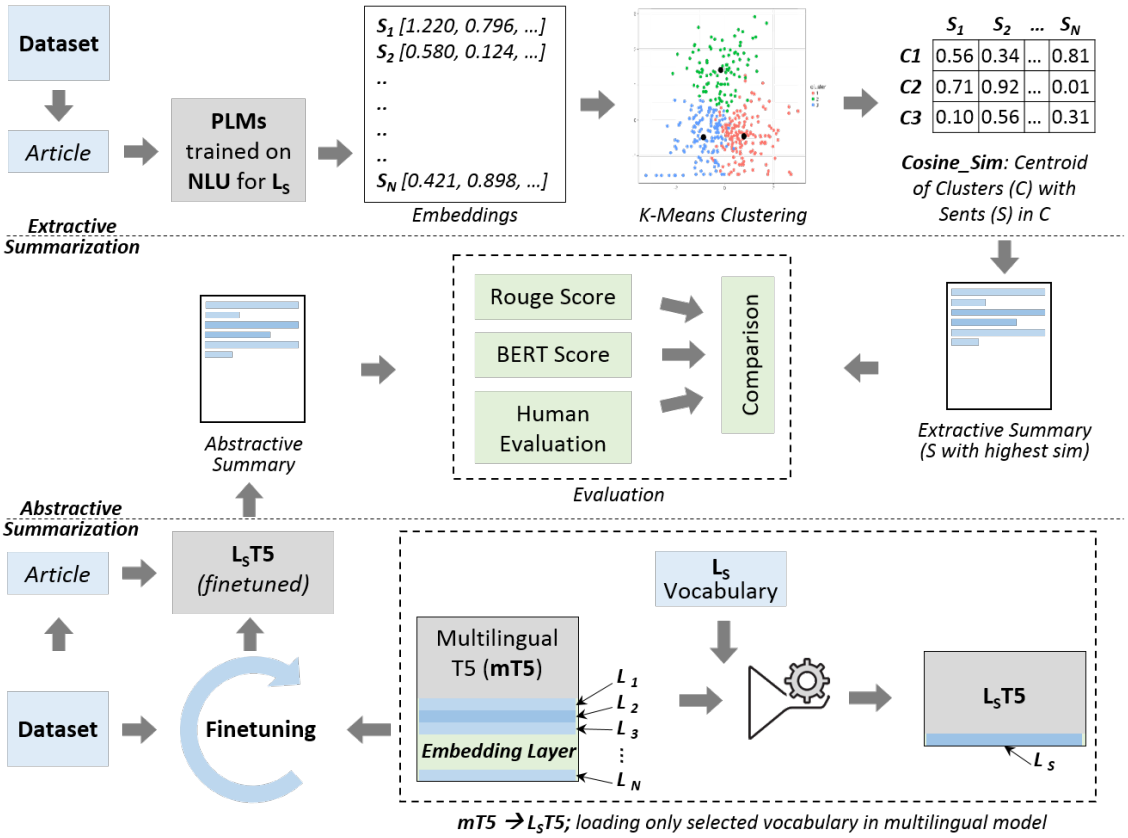


Figure 3.6: Adopted Summarization Framework. Ls = Selected Low Resource Language, PLM = Pre-trained Language Models

measure. Sentence are ranked considering the proximity to Centroid of the cluster and selected as summary sentences being top on the rank. mBERT based models included mBERT (trained over 104 languages; base - 110M parameters, $\approx 681M$ size), MuRIL [77] (trained over 17 Indian languages; base - 236M parameters, $\approx 909M$ size, large $\approx 1.89G$ size), Geotrend/BERT (monolingual Urdu version of mBERT with 48% reduced size i.e. $\approx 354M$ and reduced memory utilization considering low resource settings).

3.2.2 Abstractive Summarization

mT5 (Multilingual T5)[76] was selected for Abstractive Summarization task which is trained over mC4[4] covering 101 languages following similar recipe as T5. mT5 has a pre-training objective of NLG but this generation capability is not particularly trained for generating summaries hence needs finetuning. mT5 has 5 checkpoints (small, base, large,

[4]Multilingual C4 - https://www.tensorflow.org/datasets/catalog/c4

XL, XXL). Due to the extensive size of large checkpoints requiring requisite memory during finetuning, mT5-small & mT5-base were selected for experimentation. Memory consumption was further reduced by loading only monolingual vocabulary in a multilingual model which reduced the parameters from embedding layers of the model retaining same efficiency inspired by *Abdaoui et al.* [72]. Monolingual vocabulary comprised of 40k tokens collected from *1M Urdu News Dataset*[5] & own created dataset in comparison with 250k tokens of mT5-base. Size of model was reduce to 44.78% of its original size *(mT5-base:2.17GB → Urdu T5; urT5-base:1.04GB).*

3.2.3 Evaluation

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

Evaluation of summarization is normally carried out through Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[15]. The main idea of ROUGE is to calculate terms overlaps between the original gold summary which is normally written by human and generated or predicted summary by the model. Basic evaluation measure in ROUGE is ROUGE-N in which N is N-gram overlap statistics including Precision, Recall and F-Measure. ROUGE Evaluation has inherent issues not restricted to preprocessing steps involved before the evaluation phase but also for abstractive summarization. Abstractive summarization includes words / phrases which are not included in original gold / reference summaries but are generated innovatively to fit into the context of the sequence being generated. ROUGE evaluation depending upon N-gram co-occurrences/ overlap becomes contrary to the very concept of abstractive summarization. Moreover it considers a sequence as bag-of-words which takes out contextual information and its dependencies over the complete sequence. There may be cases where a summary is evaluated as a good quality with high evaluation score however in human evaluation it may scored as inferior and vice versa.

BERTScore

Most of the evaluation methods proposed earlier were based on exact matching of N-grams like ROUGE for summarization, METEOR (Automatic Machine Translation

[5] 1M Urdu News Classification Dataset - DOI: 10.17632/834vsxnb99.3 https://data.mendeley.com/datasets/834vsxnb99/3

Evaluation System), BLEU (BiLingual Evaluation Understudy) for machine translation etc. After the release of pre-trained language models which were successfully demonstrated to capture contextual information in sequence(s). To overcome the shortcomings of exact word matching, BERTScore[71] was introduced recently which instead of exactly matching the N-grams, calculates similarity between the contextualized token embeddings. By considering the similarity between contextual token embeddings, paraphrasing as well as dependencies between words were also catered for, which was not considered by metrics like ROUGE. This improvement in the contextual aspects of evaluation doesn't necessarily means BERTScore will correctly identify the high quality summaries due to its inherent dependency on BERT model and its learning of language representation along with its inherent shortcomings.

Human Evaluation

To overcome the impediments of automated evaluation as discussed earlier and absence of a unanimous standard. Few generated summaries were evaluated by human which had their primary language as Urdu. Human evaluation has been carried out on 20 x summaries generated by each model / dataset as depicted in Section **??**, Table 4.4 & Table 4.5. To correctly validate the evaluation results of already used metrics and to verify the quality of summaries, various summaries have been selected; ranked higher, lower and in mid-range.

To create the evaluation process easy only gold reference summaries and generated summaries were presented to the evaluators to avoid reader's biasses and also prevent their disinterest in reading long articles. Ranking of the summaries have been carried out considering two factors on the scale of 0 *(considered as Lowest)* to 5 *(considered as Highest)* considering the reference summary as gold standard and true in all aspects:-

- Accuracy / Relevance - Information conveyed in summary predicted by the model is accurate, consistent and relevant as conveyed in original reference summary written by human author.
- Coherence - Ability to convey information with continuity and linked ideas and language together to form coherent, well formulated and connected sentences; as conveyed in original reference summary written by human author.

3.2.4 Experiments

Experimentation was carried out using Google's Colaboratory (Colab) platform being accessible freely without specialized environment setup (pursuing the aim of low resource summarization). Google's Colab is a free development environment based on Jupyter notebook environment that runs on a cloud supporting collaborative developments. It supports popular ML libraries and offers a limited amount of GPU (i.e. $\approx 12GB$). Free usage has other limitations including session usage time, inactivity time, background execution limitations etc.

Chapter 4

Experimental Results

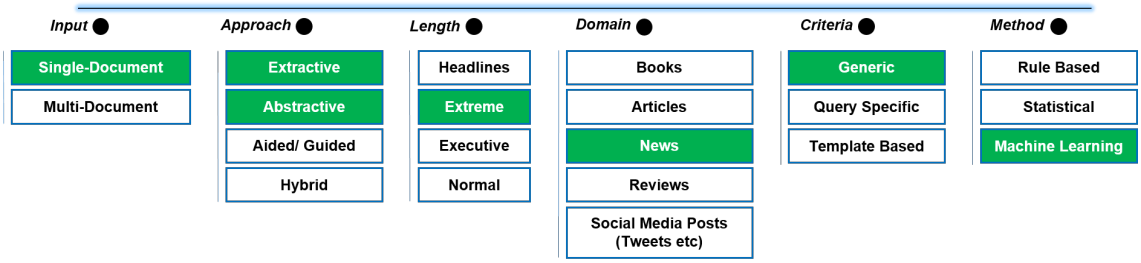


Figure 4.1: Adopted Categories of Summarization

Extractive Summarization was carried out on BBC Urdu and DW Urdu dataset separately, results of which are shown in Table 4.1 and 4.2.

4.1 MuRIL; Extractive Summarization

mBERT [58] is trained over 104 languages including language under research i.e. Urdu with the largest Wikipedia. MuRIL [77] is a BERT based model pre-trained on 17 Indian languages which contained translated and transliterated documents as well for cross lingual training from Wikipedia, Common Crawl, PMINDIA and Dakshina. MuRIL was used for *extractive summarization only* whose evaluation score shows only minor difference with mBERT despite its larger size.

Abstractive Summarization was carried out on a dataset of 72k comprising of combined BBC Urdu and DW Urdu datasets. Training was carried out up to 5 epochs with a batch size of 4 & gradient accumulation of 8. Testing was carried out on joint BBC, DW Urdu dataset as well as on separate subsets. Results are shown in Table 4.3

Extractive Summarization: BBC Urdu Dataset

Model	Rouge-1	Rouge-2	Rouge-L	BERT Score
mBERT-base	39.595	23.504	33.308	74.595
mBERT-base (Trunc)	47.983	31.598	41.905	77.605
MuRIL-base	39.303	23.417	33.209	74.33
MuRIL-base (Trunc)	47.032	30.767	40.975	77.164
MuRIL-large	40.732	24.347	34.367	74.95
MuRIL-large (Trunc)	48.745	32.354	42.627	77.826
Geotrend-BERT-base	39.576	23.455	33.279	74.571
Geotrend-BERT-base (Trunc)	47.985	31.596	41.901	77.603

Table 4.1: Rouge F Score & BERT Score

Extractive Summarization: DW Urdu Dataset

Model	Rouge-1	Rouge-2	Rouge-L	BERT SCORE
mBERT-base	30.616	9.821	21.127	71.517
mBERT-base (Trunc)	34.194	12.184	23.819	72.634
MuRIL-base	30.198	9.638	21.148	71.505
MuRIL-base (Trunc)	33.29	11.658	23.408	72.233
MuRIL-large	30.946	9.978	21.462	71.6
MuRIL-large (Trunc)	34.085	12.238	23.94	72.601
Geotrend-BERT-base	30.597	9.803	21.01	71.504
Geotrend-BERT-base (Trunc)	34.225	12.22	23.845	72.644

Table 4.2: Rouge F Score & BERT Score

Abstractive Summarization

Model	F-Score (Rouge-1)	Precision (Rouge-1)	BERT Score
urT5-base (without finetuning)	19.54	21.77	58.42
mT5-small	36.43	37.37	73.36
urT5-small	36.39	37.41	73.43
urT5-base	39.92	44.14	75.07
urT5-base (50% epochs)	40.03	44.32	75.1
urT5-base (50% dataset)	39.13	43.47	74.77
urT5-base (50% dataset + 50% epochs)	38.03	42.64	74.27
urT5-base BBC Urdu	46.35	52.12	77.0
urT5-base DW Urdu	36.91	40.4	74.17

Table 4.3: Rouge F Score, Precision & BERT Score

4.2 Training; Abstractive Summarization

- *Dataset* - Larger dataset with more training examples improves the models ability of summarization which is already well known. However fewer training examples upto a threshold should be sufficient for satisfactory results for low resource summarization as lesser difference in evaluation is observed as compared to reduction in training data (i.e. 0.89 for reducing 30k taining data from 72k).

- *Training Epochs* - Finetuning has been carried out for 5 epochs however evaluation has also been carried out for 2.5 epochs (50%). Though there is a minor difference in evaluation score but it was observed that more training doesn't necessarily means high evaluation or efficiency of trained model.

- *Abstractive vs Extractive* - Automated evaluations are usually comprising of term overlap or contextual similarity of terms overlap. In relatively simple summaries, extractive nature is favourable for high evaluation scores currently adopted. However in reality datasets are more complex in nature which favours for abstractive summarization to convey information presented in a complex input text. Abstrac-

tive summarization also have comparative results however more complex models with more quality datasets are a prerequisite for being generic & understanding complex text e.g. sarcasm, idioms etc).

- *Zeroshot Evaluation* - mT5 based models are trained for NLG however not particularly for summary generation tasks therefore as expected results of usage of these models without finetuning are quite low.

4.3 Truncation

BERT based models used in extractive summarization limits processing of input documents up to 512 tokens. These model automatically discards remaining text of input document resulting into loss of important information. Truncation has been carried out using Recall between article text paragraphs and summary text as explained in section 3.1.1 to cater for the limitations of BERT based models. Evidently 512 truncated versions of both datasets have high evaluation score in *extractive summarization*. In *abstractive summarization* 512 truncated version was only used both for training and testing to cater for the memory utilization of mT5 based models. Theoratically mT5 doesn't have input processing limit however memory utilization exponentially increase with large input text.

4.4 Geotrend / mT5 - urT5

Due to absence of models for low resource languages, multilingual models are a suitable alternative to be used for monolingual purpose. These multilingual models can be efficiently utilized for monolingual task by reducing the shared vocabulary of model to monolingual vocabulary as proposed by *Andaoui et al.*[72] in Geotrend/BERT models. *Andaoui et al.* proposed loading only monolingual vocabulary in a multilingual model as most of the parameters of these multilingual models are in embedding layers. By reducing the vocabulary; input / output embeddings of the model are reduced. As a result size of the model and it memory utilization is reduced retaining almost the same efficiency as of original multilingual model. In both *extractive and asbtractive summarization* no distinct difference in evaluation is observed as compared to actual multilingual model.

4.5 Extractive vs Abstractive Summaries

In *extractive summarization* evaluation score of BBC Urdu dataset is comparatively high as compared to that of DW Urdu dataset. Summaries with top evaluation score were analyzed and found that BBC Urdu has number of summaries having maximum terms extracted from articles hence increasing term overlap resulting into high automated evaluation scores. In *abstractive summarization* training was carried out by joint dataset however same effect was observed while evaluating BBC Urdu and DW Urdu separately. BBC Urdu evaluation score was comparatively high however effect was reduced as compared to extractive summarization results because of capability of mT5 based models to generate abstract summaries instead of selecting sentences from input text.

4.6 Human Evaluation

Human Evaluation has been carried out to cater for the deficiencies of automated evaluations and to verify the results and findings of the experimentations. Evaluation has been performed by 10 Human evaluators whose primary language is urdu and are educated enough to understand the criteria set-up for evaluation of summaries as described in section 3.2.3. Results of evaluation score of 20 summaries from each dataset are shown in Table 4.4 and Table 4.5 respectively. Evaluation has been carried out for same summaries for extractive as well as abstractive summarization to draw a fair comparison. Major findings after carrying out human evaluation on few summaries are *(reference to selected summaries in Fig. 4.2)*:-

- *M*ost of the simple summaries were evaluated with similar high score with non distinguishable differences.
- Complex summaries have variations in different evaluations methods in both extractive and abstractive summarization.
  - Few summaries were scored 100% as per human evaluation however due to total word count used in Rouge score ranked a little less e.g. summary ser 1.
  - Summaries sometimes convey the same meaning / facts however conveyed using different words (similar to abstractive summarization) are ranked lower

than mean Rouge score but relatively quite higher as per human evaluation.

  - Few summaries were ranked very low by Rouge score however ranked higher than average score by both BERT Score & human evaluation e.g. summary ser 5; Extractive Summary's Rouge score is 0.10 as compared to 0.67 BERT Score and 0.74 rank given by human evaluator. These summaries mostly used concepts related to same idea.
  - Few summaries were ranked lower by Rouge score due to lesser word overlap, ranked higher by BERT Score due to semantic similarity of words revolving around same idea however ranked comparatively lower by human evaluation because summary as whole was enable to present the same information as conveyed by actual summary.

- Human Evaluation also have biasses / difference of opinion for the standard of summary. However neglecting the issue of personal preferences, considering the current models and evaluation methods one can have a fair idea that certain type of complex text may not have accurate automatic evaluation scores.

BBC Urdu: Comparison between Evaluation Results

	Abstractive			Extractive		
Ser	Rouge F1	BERT Score	Human	Rouge F1	BERT Score	Human
1	0.71	0.86	1.00	0.85	0.94	1.00
2	0.79	0.92	0.93	0.64	0.82	0.64
3	0.31	0.73	0.82	0.50	0.75	0.70
4	0.31	0.69	0.83	1.00	1.00	1.00
5	0.78	0.92	0.97	1.00	1.00	1.00
6	0.75	0.75	0.86	0.34	0.73	0.82
7	0.38	0.75	0.76	0.35	0.75	0.72
8	0.30	0.73	0.84	0.32	0.75	0.65
9	0.66	0.86	0.71	0.33	0.69	0.67
10	0.16	0.66	0.29	0.42	0.74	0.88
11	0.28	0.69	0.72	0.32	0.70	0.75
12	0.37	0.71	0.75	0.34	0.68	0.59
13	0.80	0.92	0.94	0.26	0.69	0.79
14	0.29	0.76	0.78	0.39	0.77	0.95
15	0.26	0.71	0.68	0.18	0.66	0.46
16	0.16	0.64	0.58	0.10	0.62	0.48
17	0.16	0.65	0.52	0.11	0.66	0.54
18	0.28	0.71	0.82	0.11	0.69	0.64
19	0.17	0.68	0.50	0.12	0.67	0.64
20	0.15	0.67	0.49	0.12	0.67	0.51

Table 4.4: BBC Urdu: Rouge F Score, BERT Score & Human Evaluation for Extractive & Abstractive Summarization

DW Urdu: Comparison between Evaluation Results

	Abstractive			Extractive		
Ser	Rouge F1	BERT Score	Human	Rouge F1	BERT Score	Human
1	0.63	0.85	0.72	0.81	0.93	0.83
2	0.37	0.73	0.76	0.85	0.89	1.00
3	0.61	0.82	0.96	0.87	0.92	1.00
4	0.65	0.83	0.84	0.89	0.93	1.00
5	0.57	0.79	0.88	0.98	0.98	1.00
6	0.43	0.75	0.72	0.38	0.74	0.71
7	0.47	0.74	0.86	0.46	0.75	0.88
8	0.90	0.96	1.00	0.67	0.88	0.92
9	0.47	0.77	0.69	0.35	0.73	0.66
10	0.53	0.83	0.69	0.56	0.80	0.66
11	0.28	0.72	0.58	0.20	0.71	0.62
12	0.52	0.79	0.65	0.25	0.70	0.61
13	0.59	0.81	0.88	0.25	0.69	0.72
14	0.29	0.72	0.79	0.26	0.71	0.88
15	0.52	0.81	0.75	0.24	0.72	0.7
16	0.13	0.68	0.67	0.06	0.64	0.6
17	0.26	0.70	0.63	0.08	0.65	0.41
18	0.27	0.73	0.59	0.09	0.68	0.36
19	0.41	0.69	0.74	0.09	0.59	0.4
20	0.27	0.73	0.89	0.10	0.67	0.74

Table 4.5: DW Urdu: Rouge F Score, BERT Score & Human Evaluation for Extractive & Abstractive Summarization

Ser	Extractive Summarization	Abstractive Summarization	Abstractive			Extractive		
			Rouge F1	BERT Score	Human	Rouge F1	BERT Score	Human
1	ماہرین کے مطابق پاکستانی عوام نئی کمپنیوں کی کاروں میں دلچسپی لے رہے ہیں لیکن ان گاڑیوں پر اعتماد پیدا ہونے اور ملک بھر میں ان کے خریدار بننے میں ابھی وقت لگے گا۔							
	ماہرین کے مطابق نئی گاڑیاں بنانے والی کمپنیوں میں لوگ دلچسپی لے رہے ہیں لیکن ان پر اعتماد پیدا ہونے اور ملک بھر میں ان کے خریدار بننے میں ابھی وقت لگے گا۔	ماہرین کے مطابق نئی گاڑیاں بنانے والی کمپنیوں میں لوگ دلچسپی لے رہے ہیں لیکن ان پر اعتماد پیدا ہونے اور ملک بھر میں ان کے خریدار بننے میں ابھی وقت لگے گا۔ سرکاری خبر رساں ایجنسی اے پی پی نے لاہور میں گاڑیوں کے ایک ڈیلر احمد علی سے بات کی ہے۔	0.71	0.86	1.00	0.85	0.94	1.00
2	اس نئی پالیسی سے ٹیکنالوجی کے شعبے سے وابستہ اعلیٰ ہنر افراد، بچوں کی دیکھ بھال کرنے والے گھریلو ملازمین، اجرتی ملازمین اور اعلیٰ عہدیدار متاثر ہوں گے۔							
	امریکی صدر ڈونلڈ ٹرمپ نے کچھ گرین کارڈز جاری کرنے کے وقفے میں توسیع کر دی ہے جبکہ 2020 کے اختتام تک غیر ملکی ملازمین کے لیے ویزوں کو بھی معطل کر دیا ہے۔ وائٹ ہاؤس کا کہنا ہے کہ اس نئی پالیسی سے تقریباً 525000 افراد متاثر ہوں گے۔	امریکی صدر ڈونلڈ ٹرمپ نے کچھ گرین کارڈز جاری کرنے کے وقفے میں توسیع کر دی ہے جبکہ 2020 کے اختتام تک غیر ملکی ملازمین کے لیے ویزوں کو بھی معطل کر دیا ہے۔	0.28	0.71	0.82	0.11	0.69	0.64
3	الزائمر پر تحقیق کرنے والے سائنس دانوں کی جانب سے تجویز پیش کی گئی ہے کہ اس بیماری کو درست انداز سے سمجھنے کے لیے یادداشت کی کمزوری اور فکری گراوٹ جیسی علامات کی بجائے جسمانی تبدیلیوں کو سمجھا جائے۔							
	ماہرین کے مطابق الزائمر کی وجہ سے جسم میں زبردست اور تیز رفتار طبعی تبدیلیاں رونما ہوتی ہیں اور اس بیماری کو ان تبدیلیوں سے سمجھا جائے، تو زیادہ بہتر ہو گا۔ اس کے علاوہ اعصاب اور دماغی خلیات کی موت کے شواہد ہیں اور تیسرا عنصر دماغ کی سرگرمی کی رفتار میں کمی ہے، جو سوچنے اور سمجھنے کی صلاحیت کا باعث بنتی ہے۔	سائنس دانوں نے ایک تازہ رپورٹ میں تجویز دی ہے کہ الزائمر کو تین مختلف عناصر سے تشخیص کیا جائے، جس میں پہلا تو اعصاب اور دماغی خلیات کی موت کے شواہد ہیں اور دوسرا عنصر دماغ کی سرگرمی کی رفتار میں کمی ہے۔	0.29	0.72	0.79	0.26	0.71	0.88
4	عرب اسپرنگ کے آغاز کو چھ ماہ گزرنے کے بعد مصر میں جمہوریت کی جنگ لڑنے والوں کے اندر پائی جانے والی خوش فہمی کافی حد تک دور ہونا شروع ہو گئی ہے۔							
	وہاں عبوری ملٹری کونسل اپوزیشن کے خلاف اُسی انداز اور انہیں طریقوں سے نمٹ رہی ہے جو حسنی مبارک کی حکومت نے اختیار کیے ہوئے تھے۔ فروری کی گیارہ تاریخ کو مصری عوام قاہرہ کی سڑکوں پر نکل کر حسنی مبارک کی حکومت کے خاتمے کا جشن منا رہے تھے۔	مصر میں فوج نے اپوزیشن کے خلاف اُسی انداز اور انہیں طریقوں سے نمٹ رہی ہے جو حسنی مبارک کی حکومت نے اختیار کیے ہوئے تھے۔ عوام کے جمہوریت کے خواب چکنا چور ہو گئے ہیں۔ چھ ماہ بعد وہ تمام جوش و خروش اور خوشی کا عالم ناپید ہو چکا ہے۔	0.26	0.70	0.63	0.08	0.65	0.48
5	انسانی حقوق کی نگرانی کرنے والی بین الاقوامی تنظیم ایمنسٹی انٹرنیشنل نے مصر کی جیلوں کی ناگفتہ بہ صورتحال پر گہری تشویش کا اظہار کیا ہے۔							
	رپورٹ کے مطابق قیدیوں کو غیر صحت بخش کھانا دیا جاتا ہے اور انہیں اندھیرے اور غیر ہوادار سیلز میں رکھا جاتا ہے۔ رپورٹ میں یہ بھی کہا گیا ہے کہ مصر میں مہلک کورونا وبا کے خلاف جنگ میں یکساں حکمت عملی اختیار نہیں کی گئی ہے۔	ایمنسٹی انٹرنیشنل نے کہا ہے کہ مصر میں کورونا وبا کے خلاف جنگ میں یکساں حکمت عملی اختیار نہیں کی گئی ہے۔ رپورٹ کے مطابق یہ صورتحال قیدیوں کو غیر ضروری کرب اور تکلیف پہنچانے کا سبب بن رہی ہے۔	0.27	0.73	0.89	0.10	0.67	0.74

Figure 4.2: Selected samples from summaries undergone human evaluation; Actual Summary in top merged row, Extractive & Abstractive Summarization and Evaluation Scores in bottom row

Chapter 5

Conclusion and Future Research

NLP has evolved significantly due to recent inception of transformer-based architecture comprising of Deep Learning based Artificial Neural Networks. Automatic Summarization is comparatively complex downstream task under NLP umbrella due to various factors like difference of opinion regarding importance of information, absence of unanimous evaluation standards etc. Moreover in Abstractive Summarization, new words are utilized which are not present in the vocabulary of document to be summarized. This property of abstractive summarization presents an endless possibilities for summarization making it difficult for automatic evaluation. Despite the challenges latest transformer based models have proven their efficiency even in Automatic Summarization (both extractive and abstractive).

Considering lack of research in low resource summarization a dataset has been created from publicly available source which can be replicated for any low resource language. Utilizing the newly created dataset and available multilingual models, a framework was adopted for utilizing multilingual models for monolingual purpose efficiently with comparative evaluation results in a low resource development environment which is easily available. Evaluation results are comparative to SOTA (state-of-the-art) results on relatively similar dataset of XSUM (Extreme Summarization) [51] in high resources language English i.e. Rouge-1 score of 47.21, 45.14 and 38.81 claimed by PEGASUS [70], BART [62] and BERTSumExtAbs [64] respectively. However these results cannot be compared truly as evaluation results are dependent on datasets and no dataset or previous research is proposed in Urdu language which can be compared to this research. Research including newly created dataset has been made have also been made available

online[1] which can be utilized for future experiments. Few of the future areas of research which demands exploration are:-

- *Datasets* - Creation of quality datasets including multi-domain datasets (comprising of variety of sources, news, reviews, books, lectures etc) and cross lingual parallel datasets specifically to tackle problems of low resource languages.
- *Models* - Multilingual models (universal) which are capable of tackling the problem of under-representation of low resource languages and able to understand more complex lingua .
- *Low Resource* - Most of the available multilingual models are generally trained over multiple tasks of NLU and NLG with large number of parameters resulting into larger sizes with more memory consumption. Modular approach towards models may be explored where model while retaining the generalization of NLU and NLG tasks may be able to utilize modules / layers necessary for specific downstream task resulting into lesser resource utilization. One such technique of loading only monolingual vocabulary is used however models itself don't provide such flexibility.
- *Evaluation* - Available Evaluation methods for summarization are lacking research as compared to NLU and NLG tasks. These methods have inherent issues which are already quite frequently being discussed for high resource languages. Authenticity & verification of these evaluation methods for low resource languages and their global applicability for cross lingual purpose is altogether another avenue of research which still lacks progress.

[1] https://www.huggingface.com/mbshr

Bibliography

[1] H. P. Luhn. "The Automatic Creation of Literature Abstracts". In: *IBM Journal of Research and Development* 2.2 (1958), pp. 159–165. DOI: 10.1147/rd.22.0159.

[2] H. P. Edmundson. "New Methods in Automatic Extracting". In: *J. ACM* 16.2 (1969), pp. 264–285. ISSN: 0004-5411. DOI: 10.1145/321510.321519. URL: https://doi.org/10.1145/321510.321519.

[3] Karen Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of Documentation* 28 (1972), pp. 11–21.

[4] Julian Kupiec, Jan Pedersen, and Francine Chen. "A Trainable Document Summarizer". In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '95. Seattle, Washington, USA: Association for Computing Machinery, 1995, pp. 68–73. ISBN: 0897917146. DOI: 10.1145/215206.215333. URL: https://doi.org/10.1145/215206.215333.

[5] Jaime Carbonell and Jade Goldstein. "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries". In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '98. Melbourne, Australia: Association for Computing Machinery, 1998, pp. 335–336. ISBN: 1581130155. DOI: 10.1145/290941.291025. URL: https://doi.org/10.1145/290941.291025.

[6] Eduard Hovy and Chin-Yew Lin. "Automated Text Summarization and the SUMMARIST System". In: *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998*. TIPSTER '98. Baltimore, Maryland: Association for Computational Linguistics, 1998, pp. 197–214. DOI: 10.3115/1119089.1119121. URL: https://doi.org/10.3115/1119089.1119121.

[7] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. "Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies". In: NAACL-ANLP-AutoSum '00. Seattle, Washington: Association for Computational Linguistics, 2000, pp. 21–30. DOI: 10.3115/1117575.1117578. URL: https://doi.org/10.3115/1117575.1117578.

[8] John M. Conroy and Dianne P. Oleary. "Text Summarization via Hidden Markov Models". In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '01. New Orleans, Louisiana, USA: Association for Computing Machinery, 2001, pp. 406–407. ISBN: 1581133316. DOI: 10.1145/383952.384042. URL: https://doi.org/10.1145/383952.384042.

[9] Yihong Gong and Xin Liu. "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis". In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '01. New Orleans, Louisiana, USA: Association for Computing Machinery, 2001, pp. 19–25. ISBN: 1581133316. DOI: 10.1145/383952.383955. URL: https://doi.org/10.1145/383952.383955.

[10] Kevin Knight and Daniel Marcu. "Summarization beyond sentence extraction: A probabilistic approach to sentence compression". In: *Artificial Intelligence* 139.1 (2002), pp. 91–107. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(02)00222-9. URL: https://www.sciencedirect.com/science/article/pii/S0004370202002229.

[11] Joel Larocca Neto, Alex A. Freitas, and Celso A. A. Kaestner. "Automatic Text Summarization Using a Machine Learning Approach". In: *Advances in Artificial Intelligence*. Ed. by Guilherme Bittencourt and Geber L. Ramalho. Springer Berlin Heidelberg, 2002, pp. 205–215. ISBN: 978-3-540-36127-5.

[12] Miles Osborne. "Using maximum entropy for sentence extraction". In: *Proceedings of the ACL-02 Workshop on Automatic Summarization*. Phildadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002, pp. 1–8. DOI: 10.3115/1118162.1118163. URL: https://aclanthology.org/W02-0401.

[13] Radu Soricut and Daniel Marcu. "Sentence Level Discourse Parsing using Syntactic and Lexical Information". In: *Proceedings of the 2003 Human Language Tech-*

nology Conference of the North American Chapter of the Association for Computational Linguistics. 2003, pp. 228–235. URL: https://aclanthology.org/N03-1030.

[14] Gunes Erkan and Dragomir R. Radev. "LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization". In: *J. Artif. Int. Res.* 22.1 (2004), pp. 457–479. ISSN: 1076-9757.

[15] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[16] Rada Mihalcea and Paul Tarau. "TextRank: Bringing Order into Text". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2004, pp. 404–411. URL: https://aclanthology.org/W04-3252.

[17] Christopher Burges et al. "Learning to Rank using Gradient Descent". In: *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*. 2005, pp. 89–96. DOI: 10.1145/1102351.1102363.

[18] Jen-Yuan Yeh et al. "Text summarization using a trainable summarizer and latent semantic analysis". In: *Information Processing and Management* 41.1 (2005). An Asian Digital Libraries Perspective, pp. 75–95. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2004.04.003. URL: https://www.sciencedirect.com/science/article/pii/S0306457304000329.

[19] Maria Fuentes, Enrique Alfonseca, and Horacio Rodriguez. "Support Vector Machines for Query-focused Summarization trained and evaluated on Pyramid data". In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 57–60. URL: https://aclanthology.org/P07-2015.

[20] Dou Shen et al. "Document Summarization Using Conditional Random Fields". In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. IJCAI'07. Hyderabad, India: Morgan Kaufmann Publishers Inc., 2007, pp. 2862–2867.

[21] Vahed Qazvinian and Dragomir R. Radev. *Scientific Paper Summarization Using Citation Summary Networks*. 2008. arXiv: 0807.1560.

[22] Aria Haghighi and Lucy Vanderwende. "Exploring Content Models for Multi-Document Summarization". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 362–370. URL: https://aclanthology.org/N09-1041.

[23] Dekang Lin and Xiaoyun Wu. "Phrase clustering for discriminative learning". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 2009, pp. 1030–1038.

[24] Pei-ying Zhang and Cun-he Li. "Automatic text summarization based on sentences clustering and extraction". In: *2009 2nd IEEE International Conference on Computer Science and Information Technology*. 2009, pp. 167–170. DOI: 10.1109/ICCSIT.2009.5234971.

[25] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. "Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions". In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, 2010, pp. 340–348. URL: https://aclanthology.org/C10-1039.

[26] Dingding Wang and Tao Li. "Document Update Summarization Using Incremental Hierarchical Clustering". In: CIKM '10. Toronto, ON, Canada: Association for Computing Machinery, 2010, pp. 279–288. ISBN: 9781450300995. DOI: 10.1145/1871437.1871476. URL: https://doi.org/10.1145/1871437.1871476.

[27] Charu C. Aggarwal and Cheng Xiang Zhai. *Mining Text Data*. Springer Publishing Company, Incorporated, 2012. ISBN: 1461432227.

[28] Janara Christensen et al. "Towards Coherent Multi-Document Summarization". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2013, pp. 1163–1173. URL: https://aclanthology.org/N13-1136.

[29] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781.

[30] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2014. arXiv: 1409.0473.

[31] Baotian Hu et al. "Convolutional Neural Network Architectures for Matching Natural Language Sentences". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 2042–2050.

[32] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

[33] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. "Sequence to Sequence Learning with Neural Networks". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 3104–3112.

[34] Alexander M. Rush, Sumit Chopra, and Jason Weston. "A Neural Attention Model for Abstractive Sentence Summarization". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 379–389. DOI: 10.18653/v1/D15-1044. URL: https://aclanthology.org/D15-1044.

[35] Jianpeng Cheng and Mirella Lapata. "Neural Summarization by Extracting Sentences and Words". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2016, pp. 484–494. DOI: 10.18653/v1/P16-1046. URL: https://aclanthology.org/P16-1046.

[36] Sumit Chopra, Michael Auli, and Alexander M. Rush. "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Compu-

tational Linguistics, 2016, pp. 93–98. DOI: 10.18653/v1/N16-1012. URL: https://aclanthology.org/N16-1012.

[37] Muhammad Humayoun et al. "Urdu Summary Corpus". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC'16*. European Language Resources Association (ELRA), 2016, pp. 796–800. URL: https://aclanthology.org/L16-1128.

[38] Ramesh Nallapati et al. "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond". In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2016, pp. 280–290. DOI: 10.18653/v1/K16-1028. URL: https://aclanthology.org/K16-1028.

[39] Piotr Bojanowski et al. *Enriching Word Vectors with Subword Information*. 2017. arXiv: 1607.04606.

[40] Piji Li et al. "Deep Recurrent Generative Decoder for Abstractive Text Summarization". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017, pp. 2091–2100. DOI: 10.18653/v1/D17-1222. URL: https://aclanthology.org/D17-1222.

[41] Abigail See, Peter J. Liu, and Christopher D. Manning. *Get To The Point: Summarization with Pointer-Generator Networks*. 2017. arXiv: 1704.04368.

[42] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[43] Qingyu Zhou et al. "Selective Encoding for Abstractive Sentence Summarization". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017, pp. 1095–1104. DOI: 10.18653/v1/P17-1101. URL: https://aclanthology.org/P17-1101.

[44] Mikel Artetxe and Holger Schwenk. "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond". In: *CoRR* abs/1812.10464 (2018). arXiv: 1812.10464. URL: http://arxiv.org/abs/1812.10464.

[45] Ziqiang Cao et al. “Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018, pp. 152–161. DOI: 10.18653/v1/P18-1015. URL: https://aclanthology.org/P18-1015.

[46] Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. “Bottom-Up Abstractive Summarization”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018, pp. 4098–4109. DOI: 10.18653/v1/D18-1443. URL: https://aclanthology.org/D18-1443.

[47] Jeremy Howard and Sebastian Ruder. “Fine-tuned Language Models for Text Classification”. In: *CoRR* abs/1801.06146 (2018). arXiv: 1801.06146. URL: http://arxiv.org/abs/1801.06146.

[48] Kexin Liao, Logan Lebanoff, and Fei Liu. *Abstract Meaning Representation for Multi-Document Summarization*. 2018. arXiv: 1806.05655.

[49] Junyang Lin et al. “Global Encoding for Abstractive Summarization”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2018, pp. 163–169. DOI: 10.18653/v1/P18-2027. URL: https://aclanthology.org/P18-2027.

[50] Aslam Muhammad et al. “EUTS: Extractive Urdu Text Summarizer”. In: *2018 Seventeenth Mexican International Conference on Artificial Intelligence (MICAI)*. 2018, pp. 39–44. DOI: 10.1109/MICAI46078.2018.00014.

[51] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. “Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization”. In: *CoRR* abs/1808.08745 (2018). arXiv: 1808.08745. URL: http://arxiv.org/abs/1808.08745.

[52] Ramakanth Pasunuru and Mohit Bansal. *Multi-Reward Reinforced Summarization with Saliency and Entailment*. 2018. arXiv: 1804.06451.

[53] Matthew E. Peters et al. “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long*

Papers). Association for Computational Linguistics, 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

[54] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[55] Qingyu Zhou et al. *Neural Document Summarization by Jointly Learning to Score and Select Sentences*. 2018. arXiv: 1807.02305.

[56] Sanghwan Bae et al. "Summary Level Training of Sentence Rewriting for Abstractive Summarization". In: *CoRR* abs/1909.08752 (2019). arXiv: 1909.08752. URL: http://arxiv.org/abs/1909.08752.

[57] Zihang Dai et al. "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context". In: *CoRR* abs/1901.02860 (2019). arXiv: 1901.02860. URL: http://arxiv.org/abs/1901.02860.

[58] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805.

[59] Li Dong et al. "Unified language model pre-training for natural language understanding and generation". In: *Advances in Neural Information Processing Systems* 32 (2019).

[60] Mandar Joshi et al. "SpanBERT: Improving Pre-training by Representing and Predicting Spans". In: *CoRR* abs/1907.10529 (2019). arXiv: 1907.10529. URL: http://arxiv.org/abs/1907.10529.

[61] Guillaume Lample and Alexis Conneau. "Cross-lingual Language Model Pretraining". In: *CoRR* abs/1901.07291 (2019). arXiv: 1901.07291. URL: http://arxiv.org/abs/1901.07291.

[62] Mike Lewis et al. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. arXiv: 1910.13461.

[63] Liyuan Liu et al. "On the Variance of the Adaptive Learning Rate and Beyond". In: *CoRR* abs/1908.03265 (2019). arXiv: 1908.03265. URL: http://arxiv.org/abs/1908.03265.

[64] Yang Liu and Mirella Lapata. *Text Summarization with Pretrained Encoders*. 2019. arXiv: 1908.08345.

[65] Telmo Pires, Eva Schlinger, and Dan Garrette. "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://aclanthology.org/P19-1493.

[66] Kaitao Song et al. *MASS: Masked Sequence to Sequence Pre-training for Language Generation*. 2019. arXiv: 1905.02450.

[67] Sho Takase and Naoaki Okazaki. "Positional Encoding to Control Output Sequence Length". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 3999–4004. DOI: 10.18653/v1/N19-1401. URL: https://aclanthology.org/N19-1401.

[68] Ian Tenney, Dipanjan Das, and Ellie Pavlick. "BERT Rediscovers the Classical NLP Pipeline". In: *CoRR* abs/1905.05950 (2019). arXiv: 1905.05950. URL: http://arxiv.org/abs/1905.05950.

[69] Zhilin Yang et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *CoRR* abs/1906.08237 (2019). arXiv: 1906.08237. URL: http://arxiv.org/abs/1906.08237.

[70] Jingqing Zhang et al. "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization". In: *CoRR* abs/1912.08777 (2019). arXiv: 1912.08777. URL: http://arxiv.org/abs/1912.08777.

[71] Kelly W. Zhang and Samuel R. Bowman. *Language Modeling Teaches You More Syntax than Translation Does: Lessons Learned Through Auxiliary Task Analysis*. 2019. arXiv: 1809.10040.

[72] Amine Abdaoui, Camille Pradel, and Gregoire Sigel. "Load What You Need: Smaller Versions of Multilingual BERT". In: *CoRR* abs/2010.05609 (2020). arXiv: 2010.05609. URL: https://arxiv.org/abs/2010.05609.

[73] Zi-Yi Dou et al. "GSum: A General Framework for Guided Neural Abstractive Summarization". In: *CoRR* abs/2010.08014 (2020). arXiv: 2010.08014. URL: https://arxiv.org/abs/2010.08014.

[74] Ali Nawaz et al. “Extractive Text Summarization Models for Urdu Language”. In: *Information Processing and Management* 57.6 (2020), p. 102383. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2020.102383. URL: https://www.sciencedirect.com/science/article/pii/S0306457320308785.

[75] Colin Raffel et al. “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *J. Mach. Learn. Res.* 21.140 (2020), pp. 1–67.

[76] Linting Xue et al. “mT5: A massively multilingual pre-trained text-to-text transformer”. In: *CoRR* abs/2010.11934 (2020). arXiv: 2010.11934. URL: https://arxiv.org/abs/2010.11934.

[77] Simran Khanuja et al. “MuRIL: Multilingual Representations for Indian Languages”. In: *CoRR* abs/2103.10730 (2021). arXiv: 2103.10730. URL: https://arxiv.org/abs/2103.10730.

[78] Eberhard et al. *Ethnologue: Languages of the World. Twenty-fifth edition.* SIL International. 2022. URL: https://www.ethnologue.com/guides/ethnologue200.

Urdu Summarization using Pre-Trained Language Models

by Raja Mubashir Munaf

Submission date: 25-Jul-2022 09:29PM (UTC+0500)
Submission ID: 1875067496
File name: 5b_-_Thesis_Report_-_For_Turnitin_Check.pdf (1.92M)
Word count: 11138
Character count: 62226

Urdu Summarization using Pre-Trained Language Models

ORIGINALITY REPORT

4%	2%	2%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	export.arxiv.org Internet Source	<1%
2	Submitted to Punjab Technical University Student Paper	<1%
3	Submitted to University of Newcastle Student Paper	<1%
4	"Experimental IR Meets Multilinguality, Multimodality, and Interaction", Springer Science and Business Media LLC, 2021 Publication	<1%
5	ayaka14732.github.io Internet Source	<1%
6	Submitted to Manchester Metropolitan University Student Paper	<1%
7	Submitted to University of Southampton Student Paper	<1%
8	res.mdpi.com Internet Source	<1%

9	ai6abuja.github.io Internet Source	<1 %
10	"Artificial Intelligence Trends in Intelligent Systems", Springer Science and Business Media LLC, 2017 Publication	<1 %
11	Nadeem Khan Jadoon, Waqas Anwar, Usama Ijaz Bajwa, Farooq Ahmad. "Statistical machine translation of Indian languages: a survey", Neural Computing and Applications, 2017 Publication	<1 %
12	HanQi Jin, Yue Cao, TianMing Wang, XinYu Xing, XiaoJun Wan. "Recent advances of neural text generation: Core tasks, datasets, models and challenges", Science China Technological Sciences, 2020 Publication	<1 %
13	logiclux.com Internet Source	<1 %
14	1library.net Internet Source	<1 %
15	Submitted to SASTRA University Student Paper	<1 %
16	Submitted to Indian Institute of Science, Bangalore Student Paper	<1 %

17	acikbilim.yok.gov.tr Internet Source	<1 %
18	vdoc.pub Internet Source	<1 %
19	Submitted to Higher Education Commission Pakistan Student Paper	<1 %
20	scholar.colorado.edu Internet Source	<1 %
21	iugspace.iugaza.edu.ps Internet Source	<1 %
22	trepo.tuni.fi Internet Source	<1 %
23	www.coursehero.com Internet Source	<1 %
24	Aldin Kovačević, Dino Kečo. "Chapter 21 Bidirectional LSTM Networks for Abstractive Text Summarization", Springer Science and Business Media LLC, 2022 Publication	<1 %
25	Ercan Canhasi, Igor Kononenko. "Multi-document summarization via Archetypal Analysis of the content-graph joint model", Knowledge and Information Systems, 2013 Publication	<1 %

26	José Ángel González Barba. "Attention-based Approaches for Text Analytics in Social Media and Automatic Summarization", Universitat Politecnica de Valencia, 2021 Publication	<1 %
27	Muhammed Abd-Elnaby, Marco Alfonse, Mohamed Roushdy. "Classification of breast cancer using microarray gene expression data: A survey", Journal of Biomedical Informatics, 2021 Publication	<1 %
28	huggingface.co Internet Source	<1 %
29	www.michael-waibel.de Internet Source	<1 %
30	"Chinese Computational Linguistics", Springer Science and Business Media LLC, 2019 Publication	<1 %
31	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2019 Publication	<1 %
32	"Advances in Information Retrieval", Springer Science and Business Media LLC, 2021 Publication	<1 %