ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

PARAPHRASING OF URDU TEXT USING NATURAL LANGUAGE PROCESSING

Abdul Rafay[1], Ammar Ahmad Khan[1*], Muhammad Arslan[2], Aqsa Ijaz[3]
[1]Department of Computer Science, NAMAL University, Mianwali,42250, Punjab, Pakistan
[2]Department of Information Technology, Faculty of Computer Science, Lahore Garrison University, Lahore 5400, Punjab, Pakistan
[3]Department of Computer Science and Information Technology, Superior University Lahore, Sargodha 40100, Punjab, Pakistan
[*]Corresponding Author: Ammar Ahmad Khan. Email: ammar.ahmad@namal.edu.pk
Received: *28/08/2025****Accepted:*** *20/09/2025****Published:*** *16/10/2025*

Abstract
Natural language processing is the walk through gate to interact with the computer through the natural languages which are spoken commonly. Paraphrasing of a text is basically the conversion of certain text in such a way that it's semantic or meaning doesn't change. We proposed an approach for paraphrasing of Urdu text which is a low constraint language with less data set and libraries. To deal with this process we divided our task into two sub-tasks which are i) re-ordering of the words in the sentence and ii) Changing the words with their appropriate synonym. Re-ordering of the words is done using the BART model which is a denoising sequence to sequence pre-trained model. We collected our own data set which contains the original and paraphrased Urdu sentences manually typed by the human. The BART model was trained on this data set. Bart is the bidirectional auto encoder which deals with the task of changing the order of the words along with the fill in novel spaces according to the grammar. The output is then passed to synonym replacement model which also have a separate data set which was collected by us. It contains the words with their synonyms and these words are replaced by their particular synonyms. So, we integrated both the models to get the desired result which are the paraphrasing of an Urdu text. The experiment shown that our model performed quite well, and the results were as desirable. We evaluated our model based on BLEU score. The BLEU score for the predicted text was "0.54" when compared to the human paraphrased text.
Keywords: *Natural Language Processing; eParaphrasing; Urdu text parapharasing; Text paraphrasing.*

1. Introduction

In this era of technology where everyone seems to be in a hurry and everyone wants to make their work much easier. Different people around the world are speaking different types of languages. Natural language processing (NLP) is the walk through gate to interact with the computer through the natural languages which are spoken commonly. NLP is basically the field of computer science which can be used for the interaction of machines and human languages. The main challenge for the human while interacting to the computer is the understanding of language as computer works on the basis of binary digits which is such a complex language that it is far away from the understanding of human being. So, we require some platform or framework which is capable enough to keep the interaction between the computer and human language. NLP is one of the technique which can be used for human-computer interaction. The work on NLP was started in the early 1950's when "Alan Turing" proposed an idea called "Turing test" which is now used as a rule of knowledge. He proposed this idea in an article "Processing machinery and intelligence" [1] [2] [3].

1.1.Natural language processing

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

NLP is used for the various tasks related to the human-machine interaction. ”Automatic summarizing” is one of the NLP application in which we can produce the summary of the given text document. Summary of the text is produced keeping in mind the whole concept of the text document and most often it gives the synopses of the text given to it. ”Machine translation” is also another application of NLP. This is used to automatically convert the text into one natural language to another language. This task is quite difficult as while converting the text we have to keep in mind the strong knowledge of that language with all the grammatical and semantics rules and pass that knowledge to machine. Another application is the “word segmentation”. In this technique words are segmented into different units and then categorized these units into different classes. ”Speech recognition” is used to convert the audio clips to the text. When an audio clip is given to the particular model, it represents that clip into the textual form. This machine task is done using the different NLP techniques. ”Question answering” is an NLP application in which the machine is given the question and it replies with the answer to that question. As it is a difficult task for machine to understand the question so NLP converts that question in such a way that it can be understandable by the machine. Moreover NLP is used for understanding of natural language by the machine. It is also a difficult task for the machine to find the exact meaning of the text and then apply the different grammatical rules according to the situation. ”Name Entity Recognition (NER)” uses the NLP techniques to identify the different entities in the text such as place, company or individuals. NER helps the machine to process the sentences like human beings [4] [5].

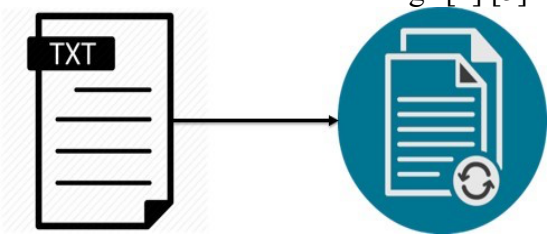

Figure 1. Paraphrasing of text

1.2.Motivation

The main motivation behind doing this work is that today where everything is being automatized and all the work is being done by machines. Using this model we can easily paraphrased any Urdu text without getting plagiarized and it will be helpful in completing the task of paraphrasing with less human efforts and time saving as world is moving to the tasks which are less time consuming. More- over, it is a sub-task for major NLP applications such as question answers, plagiarism detection, text summarizing, sentence generation, story generation, information retrieval etc. Paraphrasing is basically the conversion of certain text in such a way that it’s semantic or meaning doesn’t change. A lot of researchers have been involved in the paraphrasing of text for different languages but un- fortunately Urdu language has not been given much attention. Paraphrasing an Urdu sentence is the major problem that need to be solved. There is a lot of work done to automatize the text in the English language. The main focus of our study is that this problem should be solved using some techniques for Urdu language, as Urdu is much complicated language with respect to the other languages spoken throughout the world. Paraphrasing of any text is required in almost every field of work. There has been a lot of work done on the paraphrasing of different languages around the world e.g. English, Hindi etc. There are different types of software and websites which are helpful to do this task but this work is not applied on Urdu language till now. Another motivation behind this work is that it will be helpful for the researchers in every field working on the Urdu language throughout the world. Moreover, this can be further improved and used for designing such robots

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

which understand the Urdu sentences. This can also be used as a plagiarism detection tool to detect the plagiarized text to some extent. As this is a new technology to work on and it will help researchers to find more solutions to different problems related to the Urdu language. Moreover, as Urdu is a mother tongue of Pakistan, it is our responsibility to work on different problems related to it so that we can convince the world to adopt it.

1.3.Paraphrasing for Urdu

Paraphrasing can be done in different ways like replacing the particular words with their synonyms and antonyms keeping in mind that it may not change the whole meaning of the sentence or by arranging the words of the sentence in such a way that they may change their positions as well as the meaning of the sentence remains same. For example if we take an Urdu text sentence

“ علماء اسالم نے حافظ حسین احمد کو پارٹی ترجمان کے عہدہ سے ہٹا دیا ”

Apply the paraphrasing technique on it then it can be converted into

”جمیعت علماء اسالم نے ناموار سیاستدان حافظ حسین احمد کو عہدہ ترجمانی سے ہٹا دیا

It can be seen from the example that the text is being paraphrased and rewritten keeping in mind the whole semantics and meaning of the text.

There are a different steps which needs to be followed while doing paraphrasing. The first step that needs to be done is the paragraph segmentation. Paragraph should be divided into different sentences keeping in mind the syntax of the language we are working on (in our case it’s Urdu). Then the next step is the identification of words according to the grammar rules of particular language in a sentence like noun, verb and prepositions. Then the next step is to apply the various techniques through which the paraphrasing can be done which are already discussed in the above paragraph. Then finally we combine all the works to form a sentence without changing its semantic or meanings. Then these sentence combined to become the whole paragraph. This is a hard task to achieve as Urdu is a complicated and difficult language and also this type of work has not been done before.

1.4.Research Contributions

- Our major contributions in this research work are stated below.
- Paraphrasing of low constraint language like Urdu with less data-set available which is not previously done, to the best of our knowledge.
- We make use of pre-trained language model, which were applied on English Language to shuffle the order of words in a sentence such that the semantic and the syntax of the sentence do not change. We optimized and modified the models accordingly to get the desired results for Urdu language.
- We make our own model which was responsible to change the synonyms of the words in the sentence according to the given Urdu dictionary.
- We integrated both the pre-trained model and our own model such that the shuffling and change of synonyms of the words take place and we can get the paraphrased text which is our ultimate goal.
- We collected our own word meaning data set round about two thousand and then used them to replace the words by the suitable synonym.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

- We make our own data set of original and paraphrased sentences (3500 sentences) which are to be used in pre-trained model for changing the order of words by keeping the semantic and syntax same.

2. Literature Review

A lot of work has been done in the past on the paraphrasing of text in different high and low constraint languages. In English the researchers has been moved to the tasks of text generation and story generation but some low level languages need some work to be done. In this section we will discuss some work done on English language and some co-related problems which were dealt in English language. Then we will discuss work done on the paraphrasing of text in some low constraint languages like Urdu and then in the end we will discuss some problem which were related to paraphrasing but these tasks were completed using the Urdu language.

2.1 Paraphrasing for English Language

There exists different models that are already designed for paraphrasing which works according to the different re-framing rules for the different languages. This work of paraphrasing has been done in different languages except for Urdu. Some existing work includes: Text input was given in the form of paragraph in the English language. Sentences were changed on the basis of active passive and affirmative and negative sentences keeping in mind the grammar rules of English [6].

Paraphrasing is the familiar word to every person related to any type of natural language. In [7] they conducted a comprehensive survey in which they explained the concept of automatic and sententious paraphrase generation while also explaining the importance of paraphrasing in the field of natural language processing. Moreover they discussed the recent work which has been done to automatically or manually construct the paraphrase of any text. In [8] they applied a new technique which is multi sequence alignment. They basically trained a model which learns from the certain patterns and have capacity to rewrite new sentences. Their technique was different and difficult from word or sentence level paraphrasing.

Recently different types are being trained which are performing different tasks in the field of NLP. A technique which is useful for large language models and performs well while doing the tasks of paraphrasing and text generation using the variety of texts and subjects [9]. A technique like others was proposed in which the words were converted using active-passive voice English grammar rules and positive and negative sentences in English language [10].

In this paper the authors proposed a technique for generating the sentences from disentangled semantic and syntactic spaces. They used the linearized tree sequence to train their model the syntactic information to the auto encoders. They used this technique to make the application like unsupervised paraphrase generation and syntax transfer generation. They proposed a DSS-VAE model which was extended form of the traditional VAE as they added two latent variable in it for capturing the semantic and syntactic information separately. This technique contain a lot of advantages as they can explicitly model the syntactic in the VAE which is helpful in producing more fluent sentences, more amount of encoded information and higher BLEU score for reconstruction of sentences. Moreover, they sampled and manipulate the semantic and syntactic spaces separately and then it was helpful in transferring the syntax of one sentence to the other sentence. As discussed earlier they introduced two latent variables which were responsible for capturing the semantic and syntactic information separately. Evidence lower bound is the quantity optimized in variational bayesian methods which are responsible for distribution over unobserved data and given observed

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

data. Then they build their own RNN with the gated recurrent unit. The sentence is given the specific representation and they computed the mean and variance of the sentence semantically and syntactically. Now after the information is encoded it is fed up to the decoder. The decoder make the linearized presentation of the information. As the syntax representation is more complicated and not finite categorical. So, we use the linearized tree sequence to explicitly model syntax in the latent space of VAE. Linearized tree sequence is obtained by traversing in the tree in top down order. In training the parse tree of the sentence is obtained by ZPar toolkit and its output serves as the ground truth signal. In testing as we don't need the external syntactic tree as we built our RNN to predict the linearized parse tree where each node is the different embedding. They adopted the multi task and adversarial losses to ensure that the whole decoded information is stored separately. For semantic they used the bag of words distribution using "softmax". Which basically calculates the cross entropy. Moreover, they introduced an extra model component named "adversaries" which predicts the semantic information on the basis of syntactic information and vice versa which is helpful in performing the task of syntax transfer generation. They performed different experiment to get the results of applications stated above and used different data sets like PTB data set, Quora data set and data set of 1000 non-paraphrase sentence collected manually by human. The results shown that the model DSS-VAE outperforms all the existing VAE and other models [11].

2.2. Paraphrasing for Low constraint Languages

There has been a lot of work done for the English language but low level languages were not given much attention. In this section we will discuss some of the work done for the low resource languages

In this Paper [12] they worked on the paraphrasing of Hindi language. They changed the words were changed on the basis of synonym and antonyms. They gave the input text as a Hindi text and then divided those paragraphs into sentences and then these sentenced were further broken in to words. Then they set some reframing rules which say that sentences were categorized in affirmative and negative category and then the words present in the sentences were replaced by the synonym or antonym words for affirmative and negative sentences respectively. The synonyms and antonyms were stored in the form of database which was used to change the words accordingly. Then the words were merged again to become the sentences and sentences when combined again make the paraphrased text. As, Hindi is low resource language so it was difficult task to make the whole data set and then apply these rules and get the appropriate results.

Experiment was performed to evaluate the syntax in the German language on the basis of the order of the words. This paper described the simple surface realization engine for the German language which is based on the SimpleNLG for English. Different types of features for the syntax and order of words in the German language are discussed. Moreover, the grammatical sense was also judged while generating the natural language sentences. In comparison to English, German language is much more complex and the order of the words are much freer than English. In this technique the sentences were created using the lexical item and the phrase specs combine in a modular way. Moreover, the canned text can be used interchangeable with non-canned text keeping in mind the realization on the basis of features while getting the final output [13].

Table 2 shows the comparative study for the different papers for different low and high resource languages and the areas in which they are focusing.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

2.3. Co-related work for Urdu language

Plagiarism detection has been a big problem and different research shows that it is very hard to detect. In [14] they constructed a paraphrase plagiarism corpus which was generated manually and some of its part was freely available for the Urdu language. This corpus was created to evaluate the Urdu plagiarism detection systems. There were volunteers which were asked to manually create a

Table 1. Paper Comparison Table

Paper Title	**Areas focused**	Language
Bollmann et al.	Syntax analysis by focusing on the order of words	German
N Sethi et al.	Paraphrasing on the basis of antonym-synonym	Hindi
Witteveen et al.	Paraphrasing and text generation	English
Yu Bao et al.	Generating Sentences from Disentangled Syntactic and Semantic Spaces	English
Muhammad Sharjeel et al.	Plagiarism detection for Urdu text	Urdu
Aqil Burney et al.	Urdu Text Summarizer using Sentence Weight Algorithm for Word Processors	Urdu

paraphrased text document on the 20 renowned personalities using their own paraphrasing skills. This corpus was realistic and natural which are used by the plagiarists as the volunteers were asked to rewrite the text by replacing the appropriate words with their antonyms, synonyms and possibly changing the structure of the sentence keeping in mind the semantics and meaning of the text. Vol- unteers were also allowed to use their own skills and can add the words, combine different sentences to make the new sentences and summarizing the text of the document. This corpus was completely nature and very close to the plagiarists do while paraphrasing the document. The size of the corpus was small but according to the authors it was the first attempt to create this type of manual corpus for the Urdu language. They also added that in the future they will be working on the increase of the corpus and then apply the state of the art plagiarism detection techniques and then report their result on their manually created corpus.

Auto summarizing of text is another tool which is quite useful and a lot of research is being done on it. In [15] they presented an auto summarizing tool in MS Word for Urdu language. The purpose of adding this tool was to summarize different articles like scientific and economical writings and sports commentaries. They proposed an algorithm in which they take the whole document as an input then label the stop words and exclude them from the content word. Content words are basically the meaning full words from the text. Then they gave the weights to the sentences on the basis of content present in them and sort them on the basis of descending order. After sorting they picked the desired number of sentences and then sort them on the basis of occurrence in the original document. Reason behind picking only the specific sentences was that as the summary of any text is

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

almost the 25 percent of the total text. This tool was successfully added to MS Word and to evaluate the accuracy of the tool twenty different types of document were given as an input to the tool. The results shows that the summary generated using this tool was easy to understand, also well-formed and most importantly it was very close to the original text documents. They also cross checked the results by human verification. The accuracy was about 80 percent after the first human checked it. As we increase the number of humans trying to verify it the accuracy decreases gradually. But according to the authors it is because every human being has his own perception related to the each document while generating the summary of any text. The results show that if the original document contained the total of 718 words the summary which was generated using this tool was almost 25 percent which is 139 words. There were 7 out of 12 lines generated using tool which were same as verified by the human which means that the similarity index was around 64 percent.

2.4. Research Gap

As, discussed earlier that this problem has not been handled in the Urdu language as this needs to be done as a lot of research is being done around the world using the low level languages and even in Pakistan Urdu language is now being introduced as a new trend for research. The data sets for low constraint languages like Urdu is non-existent. To best of our knowledge there is no published research on paraphrasing of Urdu text till now. That is why it will be a challenging task at the present that which technique will best suit to the Urdu data set so that the best possible results of text paraphrasing can be generated.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

2.5. Problem Statement

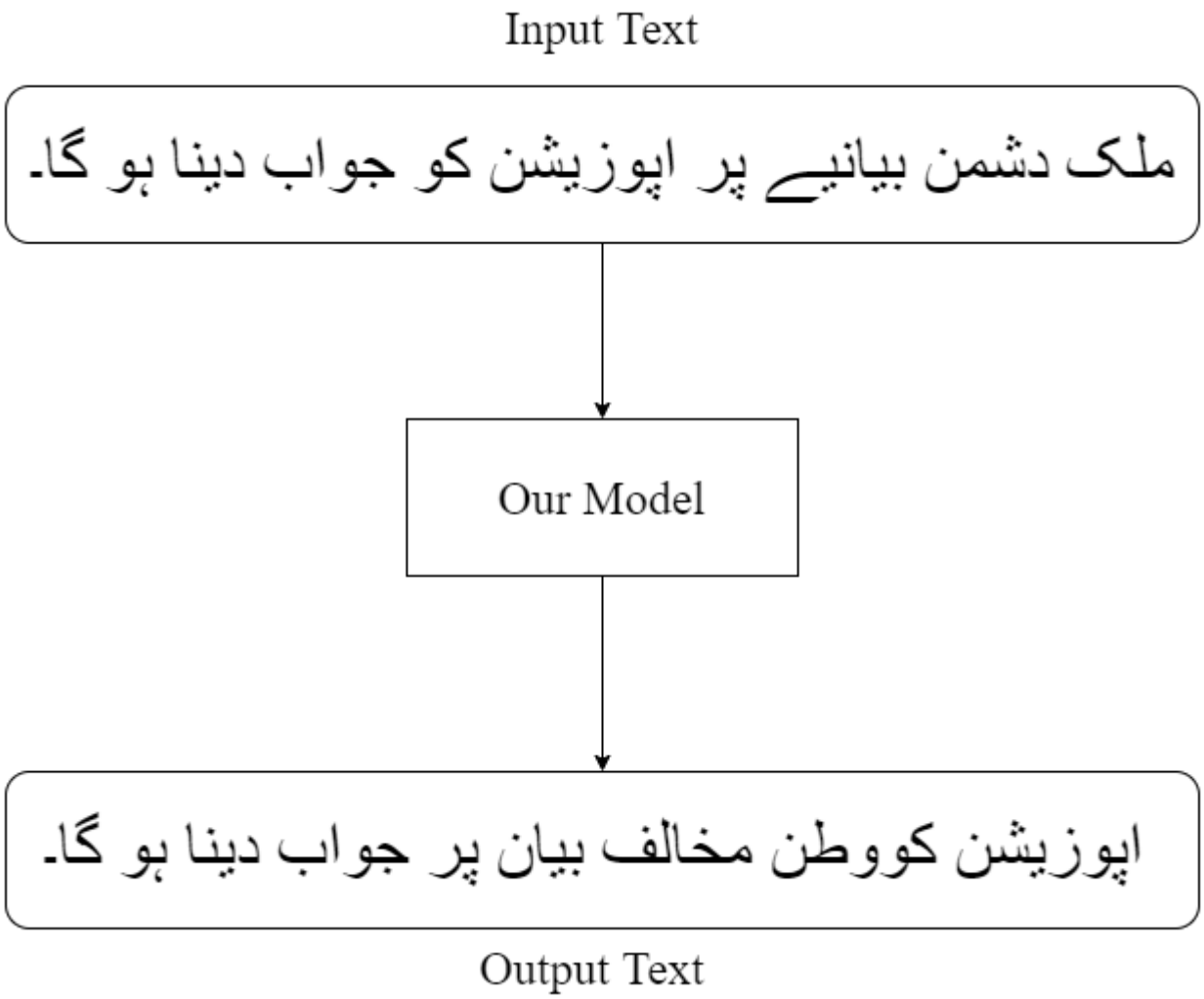

Figure 2. Problem Statement Diagram

Different types of languages are being spoken and paraphrased throughout the world, but not much work has been done in Urdu language. The concept behind this problem is to convert any type of Urdu text in such a way that it may not affect the meaning or semantic of the text. This may be helpful to convert the complex sentences to the similar ones or the either way. As shown in the example below we can see that the text input will be given to the our model and then our model will give the output as a paraphrased text which contains all the information from the original text keeping in mind the semantics and meaning of the text.

2.6. Research Objectives

In the context of this problem, we attempt to answer following research questions;

1. How will data-set be build and managed for Urdu language.
2. How synonym replacement will be done according to the context of Urdu text.
3. How will the change of order of the words be done keeping in mind that the information is not lost and the correct grammar is maintained from the original text.
4. How transfer learning can be used for training model for Urdu language?

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

5. How will we integrate the shuffling of words (semantically and syntactically) with the syn- onym replacement and get the paraphrased output of an original sentence.

3. Methodology

We proposed a technique which takes a text as an input and gives out the paraphrased text without changing the meaning of the text and contains all the information present in the original paragraph. We divided the task into two sub tasks: In the first part the order of words in original sentences are changed using the pre-trained ”BART model” keeping in mind that the semantic and syntax do not changed and in the second part the words are replaced by the synonym according to the dictionary. These both sub tasks are then integrated to form a new paraphrased sentence which contain all the information from the original sentence. The flow chart diagram is shown below in Figure 3.

3.1 Datasets

3.1.1 Words/Synonyms Data Set

As, there is no Urdu dictionary available on the internet and very few amount of information is present on the internet for the Urdu to Urdu meanings. So, it was a difficult tasks to find the meanings of words which were identified and then stored in the data-set so that they can used to perform the task of paraphrasing. The words and the meanings are typed manually with the help of human. The words and their meaning can be used two-way like both the words can be used corresponding to each other to achieve the task of changing the text either to the simple one or to the complicated one. The data set contains around two thousand word/synonyms .A small glimpse of words-synonyms data-set is shown in table 2:

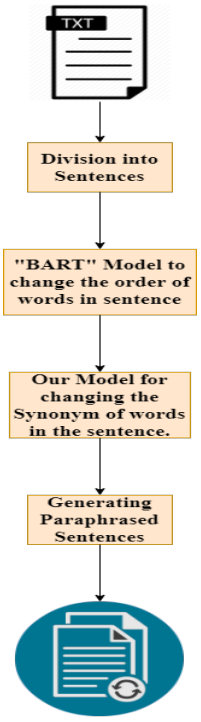

Figure 3. Flow Chart for Proposed Technique

3.1.2. Original and Paraphrased Sentences Data Set

This data set include some sentences which were manually paraphrased by human. This data set is basically the data which is to be use by the BART Model to change the order of the sentences. The model is trained on this data set which includes around thirty five hundred original and their

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

paraphrased sentences. It was hard task to do as each sentence was manually paraphrased which took a lot of time. A small glimpse of original and paraphrased sentences data-set is shown in table 3:

Table 2. Words-Synonyms Data-Set

Word	Synonym1	Synonym2	Synonym3
عظمت	عزت		
عظیم	اعلی	کبیر	
سلطنت	ریاست	مملکت	
مجبور	بے کس	بے بس	لاچار
خطاب	لقب	تسمیہ	عرفیت
تاریخ	روداد	ماضی کا مطالعہ	

Table 3. Original and Paraphrased Sentences Data-Set

Original	Paraphrased
نریندر مودی کو کیسے مل سکتے ہیں؟	نریندر مودی جی سے کیسے مل سکتا ہوں؟
سیاست میں کریئر کیسے شروع کریں ؟	سیاست میں کریئر کیسے شروع کرتے ہیں؟
کیا بھارت اور پاکستان کے درمیان جنگ ہو سکتی ہے؟	کیا پاکستان اور بھارت کے درمیان جنگ ہوگی؟
اگر ڈولن ٹرمپ سردر بن جاتا ہے تو کیا ہوگا؟	کیا ہوگا اگر صدر کے لیے ڈولن ٹرمپ جیتیں گے؟

3.2. Algorithm for the Methodology

1. The words (W) present in the original text are searched one by one so that they can be find in word-meaning corpus.
2. $\boldsymbol{W} \in \{w_1, w_2, w_3, \ldots, w_n\}$
3. Then the words are replaced by the meanings (WM) present in the corpus.
4. $\boldsymbol{WM} \in \{wm_1, wm_2, wm_3, \ldots, wm_n\}$
5. Then entities (E) are identified from the original text using the data set.
6. $\boldsymbol{E} \in \{e_1, e_2, e_3, \ldots, e_n\}$
7. Now the sentences are passed to get the semantics (S) of each sentence.
8. $\boldsymbol{S} \in \{s_1, s_2, s_3, \ldots, s_n\}$
9. The semantics of the sentence along with entities information and word meaning replacement is collected to generate paraphrased sentences (PS).
10. $\boldsymbol{PS} = \{(wm_1, e_1, s_1), (wm_2, e_2, s_1), (wm_3, e_3, s_3), \ldots, (wm_n, e_n, s_n)\}$

3.3. Implementation Detail

3.3.1. Synonym Replacement

The sentence are segmented into words. The next step after the segmentation of words is to identify the words whose meanings are present in the data-set. As, in the figure 4 we can see that there are words and corresponding to them there exist different meanings of the same words which can be used according to the suitable situation. The synonym of the word is chosen randomly for now with an option for the end user to change it with the other synonym if he want to. Then after the synonyms of the words are changed we rearrange the words keeping in mind that the meaning or semantic of the sentence doesn't changes.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

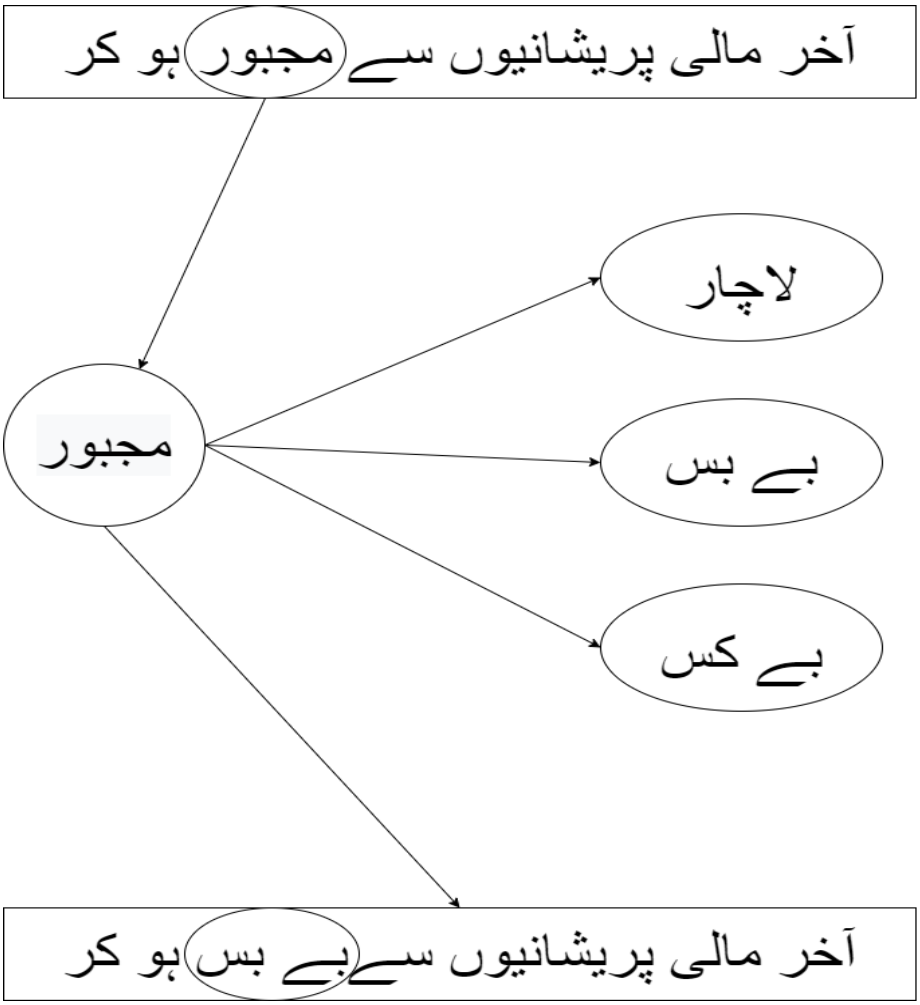

Figure 4. Synonym Replacement Example

3.3.2. Changing the Order of Words

Moving to next task which is the changing of the words present in the sentence. Before moving to this implementation we firstly explain that if we change the order of words it may change the meaning or disturb the grammar of the sentence. For this reason we used a"Denoising Sequence-to-Sequence Pre-training model known as BART".

3.4. BART Model

BART is a denoising auto-encoder which is used for pre training of sequence to sequence model. It is a multi-language model which basically works as bidirectional encoder. It is used to reconstruct the original text. This model performed well for randomly shuffling the order of words in the sentences and also filling the novel spaces according to the grammar rules. This model was basically designed for text generation but it can be used for other tasks like comprehension tasks, question answering and summarizing tasks of the text [16].

We trained the BART model on our data set of Urdu text which contains the original and para-phrased sentences. We also did some small changes in the model to attain the results for Urdu. Like for Urdu we used right to left encoding and decoding to get the desired results. This model performed well and we got the results which were desired. A small example of how the re-ordering and the fill in spaces according to the grammar rules is as follow in figure 5:

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

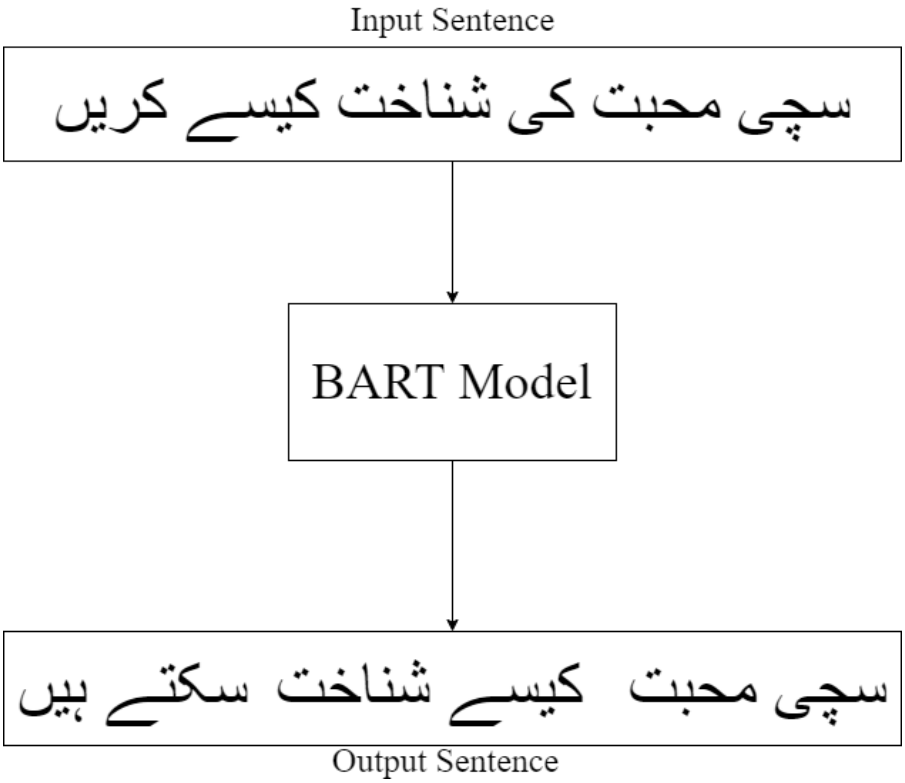

Figure 5. Shuffle of words and fill in according to grammar using BART

Integrating Synonym Replacement and Changing of Order of Words

In this step we used the output of the BART model and then give that output to our synonym replacement model. As, earlier we told that paraphrasing of the text is dependent on the shuffling of text along with the synonyms replacements keeping in mind that the semantic and syntax of the sentence do not changes. We just integrated both the models to get the desired results for Urdu paraphrasing. Figure 4.4 shows an example of working of the integration model which is as follows:

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

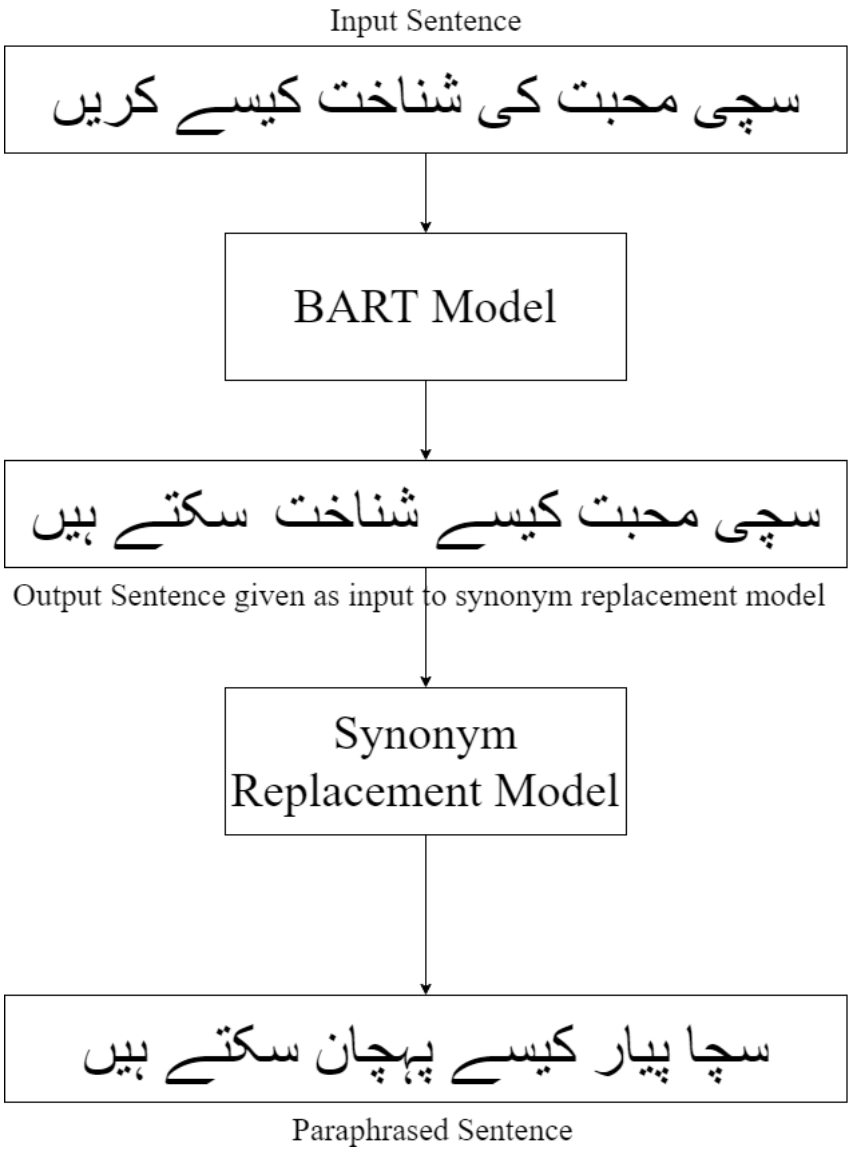

Figure 5. Paraphrased Sentence after Integration

4. Results and Evaluation

4.1.Results

We have discussed about how the problem of paraphrasing can be handled. We have discussed about how the models are being implemented using the collected data set and what are their outputs and how these outputs are being used by the other model and then integrate to get the desired results. Table 4 shows the inputs and outputs of our model. It shows that how our model performed well while dealing the Urdu text using the pre-trained model and then getting the desired output by integrating it with synonym replacement model. There were total of thirty five hundred human paraphrased sentences out of which thirty three hundred were used as training data and the other two hundred were used as a testing data. The results are shown below:

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Table 4. Input and the paraphrased output predicted by our model

Input	انگریزی کو جلد ہی بہتر بنانے کا طریقہ کیا ہے
Output	انگریزی کو جلد ہی بہتر کیسے بنانےسکتے ہیں
Input	وائٹ جرم" کا تصور کیا ہے
Output	وائٹ جرم" کیا ہے
Input	ہے" اور "ہیں" کے درمیان کیا فرق ہے
Output	کیا فرق ہے، ہے اور" ہیں" کے مابین
Input	کمپیوٹر زبانوں کو جاننے کا بہترین طریقہ کیا ہے
Output	کسی بھی کمپیوٹر کی زبان سیکھنے کا بہترین طریقہ کیا ہے
Input	کیا بھارتی میڈیا با صالحیت ہے
Output	کیا اپ کو لگتا ہے کہ بھارتی میڈیا باصالحیت ہے

4.2.Evaluation

For the evaluation purposes we used the “Bilingual Evaluation Understudy Score (BLEU)” metric. It is a metric which is used to evaluate the output text or sentence with respect to the original text. The values of BLEU score varies from 0.0 to 1.0 where the perfect mismatch gives the value 0.0 and a perfect match gives 1.0. We can use the NLTK library for python to calculate the BLEU score for the sentences or documents [17]. For our experimentation we compare the predicted sentence with the paraphrased sentence which was already present in our data set. So, basically we compared the human generated paraphrased sentences with the model predicted paraphrased sentence. The BLEU score for our model was **”0.54”** which according to the scale used for evaluating an output with respect to BLEU score is **”Very high quality, adequate, and fluent translations**. In addition we cross checked the paraphrased text i.e. output of the model manually, taking the human help.

We trained different models on our data set. We trained DyNet, T5 and BART model. After calculating the BLEU scores for all the models we compared them and we can see from the figure 6 that BART model gave the best BLEU score.

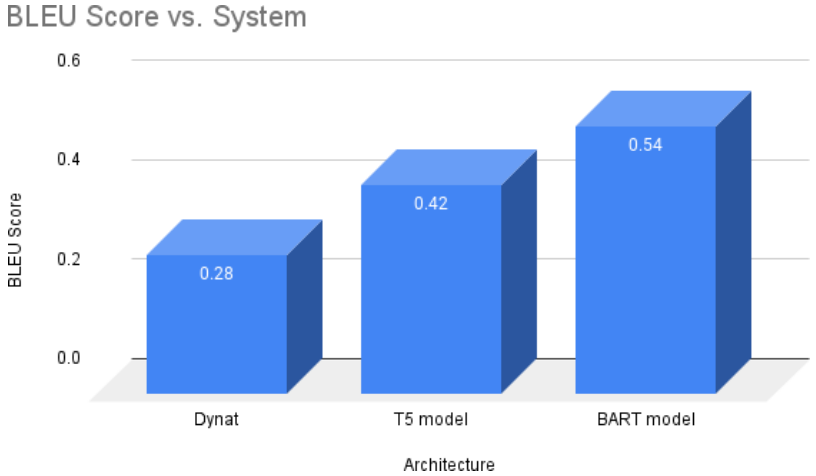

Figure 6. BLEU score comparison for different models

According to the BLEU score ranking the score 0.8 and above is very close to the human accuracy so we compared our models BLEU score with human accuracy considering it 0.8 and we

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

can see that there is the difference of 0.26 which is not a big score. So, that we can say that our model gave good results.

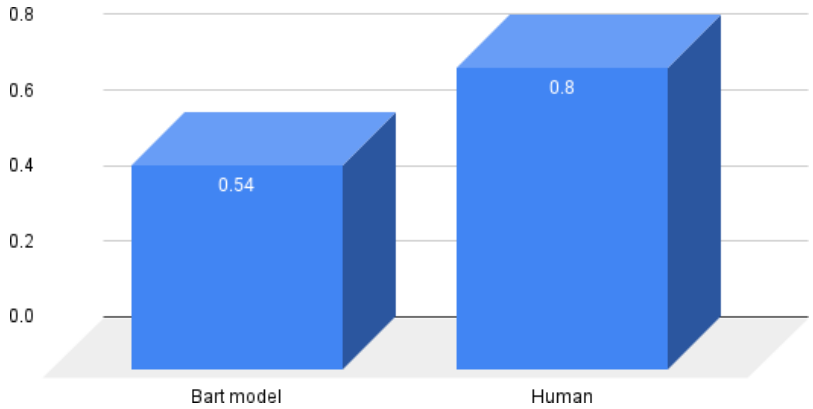

Figure 7. Comparison of Human and BART model with respect to BLEU score

5. Conclusion

Paraphrasing of a text is a sub task of major NLP applications like question answers, plagiarism detection, text summarizing, sentence generation, story generation, information retrieval etc. Para- phrasing of low constraint languages is given less attention. We proposed a technique which deal with the problem of paraphrasing of Urdu Text. As, this is the first technique to the best of our knowledge which deals with Urdu paraphrasing. We proposed a synonym replacement model and integrated it with a pre-trained model BART. The problem is divided into two sub tasks. In the first task we use BART model to change the order of words and fill the novel spaces according to the grammar rule such that the semantic and the syntax of the original sentence remains same. In the second task we passed the information of the first task to replace the words with the particular synonyms present in the sentence. We integrated both the models to get the desired output which was one of the major contribution of our research. We also collected the data set of words/synonyms by our own which contains around two thousand words. Moreover, we also collected the data set which contain the original sentences and the paraphrased sentence typed manually by human. This data set was used to train the BART model and generate the new sentences for Urdu. We evaluated our model on the basis of BLEU score and Human cross checking. The results shown that the BLEU score for our model was"**0.54"** which is very good score for a new problem. Though it doesn't reach to the human level quality fully but some of the sentences were generated very nearer to the human paraphrased sentences. This research will open new gates for solving the other NLP tasks for the Urdu language.

Funding: Please add: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

References

[1] Dan W Patterson. *Introduction to artificial intelligence and expert systems*. Prentice-hall of India, 1990.

[2] Jacob Perkins. *Python text processing with NLTK 2.0 cookbook*. Packt Publishing Ltd, 2010.

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[4] Mark Lutz. *Python Pocket Reference: Python In Your Pocket*. " O'Reilly Media, Inc.", 2014.

[5] NLPApplications, howpublished = http://www.dmi.unict.it/~faro/tesi.php?req=02, note = Accessed: 2020-11-11.

[6] Vishu Madaan, Prateek Agrawal, Nandini Sethi, Vikas Kumar, and Sanjay Kumar Singh. A novel approach to paraphrase english sentences using natural language processing. *Int. J. Control Theory Appl*, 9(11):5119–5128, 2016.

[7] Nitin Madnani and Bonnie J Dorr. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387, 2010.

[8] Regina Barzilay and Lillian Lee. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. *arXiv preprint cs/0304006*, 2003.

[9] Sam Witteveen and Martin Andrews. Paraphrasing with large language models. *arXiv preprint arXiv:1911.09661*, 2019.

[10] Vikas Kumar and Prateek Agrawal. *Reframing of English Sentences Using NLP*. PhD thesis, Lovely Professional University, 2015.

[11] Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xinyu Dai, and Jiajun Chen. Generating sentences from disentangled syntactic and semantic spaces. *arXiv preprint arXiv:1907.05789*, 2019.

[12] Nandini Sethi, Prateek Agrawal, Vishu Madaan, and Sanjay Kumar Singh. A novel approach to paraphrase hindi sentences using natural language processing. *Indian J. Sci. Technol*, 9(28):1–6, 2016.

[13] Marcel Bollmann. Adapting simplenlg to german. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 133–138, 2011.

[14] Sharjeel Muhammad, Paul Edward Rayson, and Rao Muhammad Adeel Nawab. Uppc-urdu paraphrase plagiarism corpus. 2016.

[15] Aqil Burney, Badar Sami, Nadeem Mahmood, Zain Abbas, and Kashif Rizwan. Urdu text summarizer using sentence weight algorithm for word processors. *International Journal of Computer Applications*, 46(19):38–43, 2012.

[16] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[17] A Gentle Introduction to Calculating the BLEU Score for Text in Python, howpublished=https://machinelearningmastery.com/calculate-bleu-score-for-text-python/#:~:text=the. Accessed: 2020-11-11.