Paraphrase Generation Model Using Transformer Based Architecture

Mosima A. Masethe, **Hlaudi D. Masethe**

Sunday O. Ojo , Pius A. Owolawi

Tshwane University of Technology
We empower people

DUT
DURBAN UNIVERSITY OF TECHNOLOGY
INYUVESI YASETHEKWINI YEZOBUCHWEPHESHE

SEFAKO MAKGATHO HEALTH SCIENCES UNIVERSITY

Abstract

In natural language processing (NLP), activities like identification and paraphrasing are crucial. In natural language processing, changing a text's style while maintaining its semantics is a computationally challenging operation that takes sentiment analysis into account. If a high degree of semantic similarity, potential grammatical faults, and sentence structure are not taken into consideration, the process of creating a paraphrase may inadvertently change the original meaning or intention of the text. Sentence diversity is increased through paraphrasing, which enhances the effectiveness of NLP tasks like question-answering and machine translation. Using particular NLP criteria, semantic similarity or relationship measures how similar words, phrases, and paragraphs are to one another conceptually and in terms of meaning. The difficulty lies in determining text similarity; whereas highly resourced languages have made significant strides, low-resource languages are lagging behind. One of the main obstacles to advancement in this field is the absence of inclusive datasets in the Sepedi language. Paraphrasing generation for Sepedi with transformer-based architecture models is investigated in this study. Using transformer architecture, the research creates a paraphrase artificial intelligence bot that modifies sentences by adding new words and sentence structures while maintaining the original semantics. Text extracts compiled from several datasets available at South African Digital Languages Resources (SADiLaR) comprise the Sepedi paraphrasing corpus. The sentences are translated both ways, from Sepedi to English. Metric assessments and human evaluations show that the transformer-based architecture experiment performs better than baseline models. BLEU and ROUGE are used to assess the model output's performance and calculate the accuracy provided.

1. Introduction

AI bot Paraphrase generation can unintentionally transform original sense or intention of the sentence, if high degree of semantic similarity and possible grammatical errors and sentence structure are not taken into considerations. The research employs transformer architectures to generate paraphrasing as a basis for development of paraphrasing AI bot. The challenge is finding a large parallel paraphrase dataset for indigenous languages, even though considerable progress has been made in resourced-rich languages, low-resourced languages are lacking behind. Lack of inclusive paraphrase datasets in the indigenous South African languages serve as a barrier to develop AI bot to generate paraphrase.

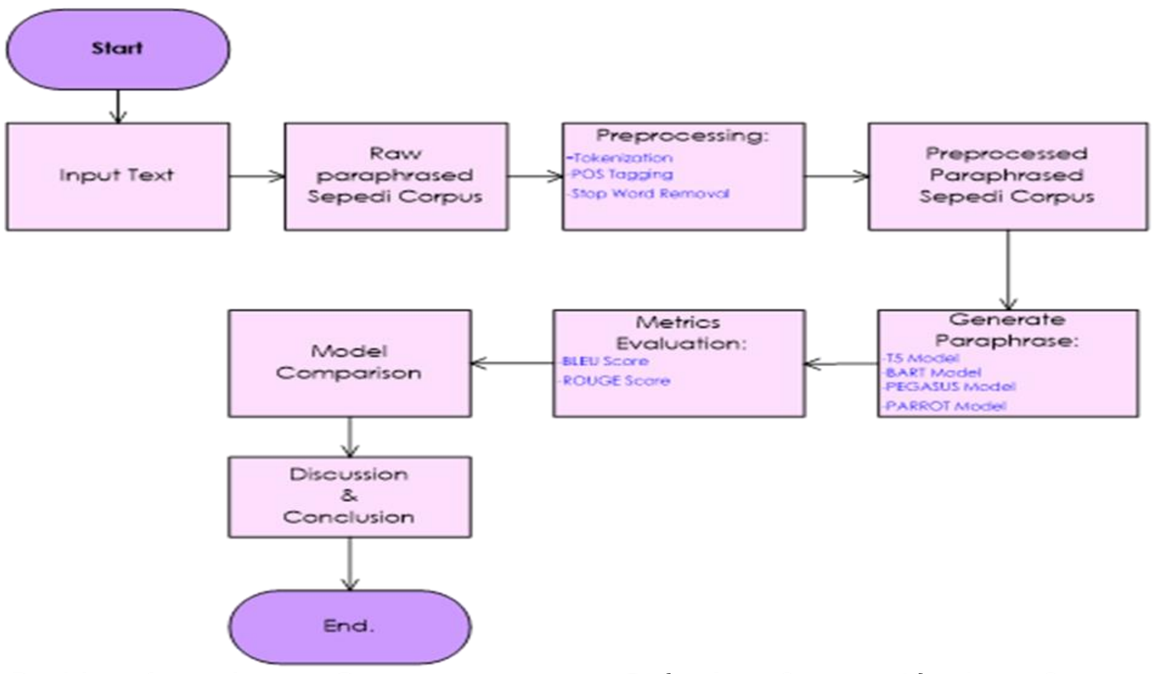
2. Related Literature

Due to lack of large parallel paraphrase datasets for Turkish language, the researchers (Bagci & Amasyali, 2021) first created several datasets and employed BERT2BERT, MT5-Base, MBart and MT5-Small transformer architectures for paraphrase generations. Furthermore, BLEU metrics score evaluation was used as well as the human evaluation factor, which found BERT2BERT outperformed other transformer methods. The researchers (Casas et al., 2021) fine-tuned a GPT-2 model to develop a system capable of transferring emotion into a paraphrase.

The BLEU, METEOR and TER evaluation metrics were used to evaluate the paraphrasing features of the developed system. The developed system was able to generate emotion paraphrasing with some grade of success, however failed to perform state-of-the art models in paraphrasing-related metrics. Computational inference of BERT and XLNet was compared using Microsoft Research Paraphrase Corpus (MPRC)by the researchers (Li et al., 2020), both models exhibit similar computational characteristics, although XLNet preceded with a better benchmark score.

3. Research Methodology

Paraphrase Generation Methodology Flowchart



Problem formulation: Given an input text S={s_1, s_2, …,s_j } where S denotes a sentence and D denotes the description for paraphrase generation, where

$$D = \{d_1, d_2, \dots, d_j\} \qquad P_\theta(D|S) = \prod_{i=1}^{D} P_\theta(S_i|S_{<i}, D) \qquad (1)$$

The research employs a language model which is a probabilistic model that forecasts the next token in a sequence given preceding tokens, which learns probability of the occurrence of an input sentence based on the text seen during training, represented by the equation (Singh, 2023):

$P(w\_1\hat{}T)=P(w\_t\ |w\_1\hat{}(t-1)$

Where w_t is the $t\hat{}th$ token with sub-sequences $w\_i\hat{}j=(w\_i,\dots, w\_j)$

4. Experiment -Dataset

The corpus is assembled from South African Digital Languages Resources (SADiLaR) Sepedi datasets with 3 000 sentences. The length of the sentence ranged between 4 and 20 words with an average sentences of 8 words. The dataset was preprocessed removing stop words, characters such as question mark, exclamation mark, commas, etc. from the sentences. The researchers created a CSV file with two columns after preprocessing with source and paraphrased sentence

Source Sentence in Sepedi	Paraphrase Sentence
Kgopelo ya gago e tla ditelega ge o ka se romele ditokomane ka moka tšeo di hlokegago	Ge eba ao romele ditokomane ka moka ka nako, kgopelo ya gago etla emišiwa

5. Evaluation Metrics

The research evaluates performance of the NLP transformer models using equation (2)-(4)

$$ROUGE-N = \frac{\sum_{S\in\{CandidateSummary\}} \sum_{n-gram\in S} Count_{match}(n-gram)}{\sum_{S\in\{ReferenceSummaries\}} \sum_{n-gram} Count(n-gram)} \qquad (2)$$

$$Geometric\ Average\ Precision\ (N) = exp(\sum_{i=1}^{N} w_i \log p_i) = \prod_i^N p_i^{w_i} \qquad (3)$$

And compute Brevity Penalty(BP), as an offset to penalize short sentences, using equation: $BLEU\ (N) = \prod_i^N p_i^{w_i} * BP$ (4)

6. Research Results

Model		Input Text	BLUE Score	ROUGE	Precision	Recall	F-Measure
BART	SADiLaR assembled Sepedi Paraphrase Corpus	Text 1	0.62	Rouge 1	0.74	1.0	0.85
				Rouge 2	0.72	1.0	0.84
				Rouge L	0.74	1.0	0.85

Model		Input Text	BLUE Score	ROUGE	Precision	Recall	F-Measure
PEGASUS	SADiLaR assembled Sepedi Paraphrase Corpus	Text 1	0.43	Rouge 1	0.53	0.67	0.59
				Rouge 2	0.39	0.5	0.44
				Rouge L	0.42	0.53	0.47
		Text 2	0.58	Rouge 1	0.63	0.8	0.71
				Rouge 2	0.5	0.64	0.56
				Rouge L	0.58	0.73	0.64
		Text 3	0.41	Rouge 1	0.52	0.67	0.59
				Rouge 2	0.39	0.5	0.44
				Rouge L	0.42	0.53	0.47

Model		Input Text	BLUE Score	ROUGE	Precision	Recall	F-Measure
T5	SADiLaR assembled Sepedi Paraphrase Corpus	Text 1	0.68	Rouge 1	0.95	1.0	0.97
				Rouge 2	0.89	0.94	0.95
				Rouge L	0.95	1.0	0.97

Model	Input Text	Paraphrased Sentence
PARROT	It is a task that has been attempted for many decades using statistical and rules-based approaches	('this task has been attempted for decades using statistical and rules-based approaches', 26)

7. Discussions & Conclusions

The research presented machine-generated paraphrased paragraphs using **BART, PEGASUS, T5** and **PARROT** transformer models for generating paraphrases. The input text originated from **SADiLaR** assembled **Sepedi Paraphrase Corpus** database. The research serve as a foundation to propose a paraphrasing technique for morphologically low resourced languages. The research evaluated four different transformer-based language models trained on Sepedi Paraphrase data for paraphrase generation NLP task. BLEU seems to have produced good score to badly generated paraphrase, which is flawed. Synonyms are not taken into considerations for the n-grams unless included in the references. **ROUGE-1, ROUGE-2, ROUGE-L** performed well in evaluating single short text documents generated paraphrasing.

References