



## **Introduction to NLP**

### **Interim Submission Report**

**Project Title : Neural Unsupervised Paraphrasing**

**Team Name : Team Linguisto**

**Team no. : 47**

Nevil Sakhreliya - 2023201005

Darshak Devani - 2023201007

Shah Viraj Utpalbhai - 2023201011

## **Problem Statement :**

Paraphrasing is expressing a sentence using different words while maintaining the meaning. In this project teams will be implementing unsupervised approaches to generate paraphrases for Indian Languages.

## **Problem Description:**

Paraphrasing, the task of expressing the same meaning using different words or structures, is fundamental in natural language processing (NLP). Paraphrasing can improve the readability and flow of written text. It allows writers to rephrase complex or convoluted sentences into simpler, more concise language, making the text more accessible to a wider audience. Traditional methods for paraphrase generation often rely on supervised approaches, which necessitate large amounts of annotated data. However, acquiring such labeled datasets is labor-intensive and may not cover the diverse range of expressions found in natural language. In contrast, unsupervised paraphrasing aims to generate paraphrases without relying on labeled data, offering a more scalable and versatile solution.

## **Dataset:**

### **Quora Question Pairs (QQP) Dataset:**

- The QQP dataset consists of over 400,000 question pairs from the community question-answering website Quora.
- Training Set: 404,290 question pairs
- Test Set: 2,345,795 question pairs
- vocab size: 73,948 unique words.

### **Microsoft Research Paraphrase Corpus (MRPC):**

- MRPC is a corpus consisting of sentence pairs extracted from newswire articles, with each pair labeled as paraphrase or not by human annotators.
- Size: The dataset contains a total of 5,803 sentence pairs.
- Data Split: Train/test splits available.
- Vocab Size: Train: 21,608 words, Test: 12,237 words.

### **IndicCorp (Hindi)**

- The corpus is a single large text file containing one sentence per line. The publicly released version is randomly shuffled, untokenized and deduplicated.
- IndicCorp is one of the largest publicly-available corpora for Indian languages.
- The dataset contains 4.95 million Hindi news articles, comprising a total of 63.1 million sentences.



## Progress Snapshot:

- As suggested by our project guide, We implemented a statistical method for paraphrasing, leveraging techniques like word substitution, POS tagging, and random sampling of synonyms from Wordnet to generate alternate versions of sentences.
- **POS Tagging and Synonym Replacement:** POS tagging plays a crucial role in identifying the grammatical structure of sentences. We utilized Flair, a powerful NLP library, to perform POS tagging on input sentences. By tagging words with their respective POS labels, we gained insight into the syntactic structure of the text.
- We identified specific POS tags for synonym replacement, including nouns (NN, NNS, NNP, NNPS), verbs (VB, VBD, VBG, VBN, VBP), adjectives (JJ, JJR, JJS), and interjections (UH). These tags cover a wide range of lexical categories, allowing for comprehensive paraphrasing.
- **Generating Paraphrases:**
  1. **Synonym Retrieval:** Leveraging WordNet, a lexical database of English, we obtained synonyms for words corresponding to the chosen POS tags. WordNet offers an extensive repository of synonyms, enabling us to explore diverse alternatives for each word.
  2. **Combination of Sentences:** After obtaining synonyms, we systematically replaced words in the original sentence with their synonymous counterparts, generating multiple paraphrased versions. The process involved computing all possible combinations of sentences resulting from these replacements.
  3. **Random Sampling:** To ensure diversity in the paraphrases and reduce compute complexity, we employed random sampling. By selecting 20% of the total possible sentences randomly, we aimed to generate a varied set of paraphrases in each iteration.
  4. **Paraphrasing:** With the selected samples, we proceeded to paraphrase the sentences. Using the synonyms obtained from WordNet, we replaced words in the original sentences, creating rephrased versions.
  5. **Similarity Evaluation:** To evaluate the quality of paraphrases, we calculated similarity scores between the original sentence and its paraphrased versions. Metrics such as BERT score and BLEU score were employed to measure the resemblance between the sentences, providing quantitative insights into the paraphrasing process.
  6. **Conclusion:** our approach leverages Flair for POS tagging and WordNet for synonym retrieval to generate paraphrases of input sentences. By systematically replacing words and employing random sampling, we ensure diversity in the paraphrased outputs. Additionally, similarity scores offer a quantitative assessment of the resemblance between the original and paraphrased sentences, aiding in the evaluation of the paraphrasing process.

## Output of statistical method for paraphrasing:

```
Enter a sentence: I want to purchase chocolate
['PRP', 'VBP', 'TO', 'VB', 'NN']
['I', 'want', 'to', 'purchase', 'chocolate']
```

Plain Text ▾

 Copy  Caption ...

```
Top Paraphrases :
I require to buy cocoa
I wishing to buy cocoa
I privation to buy cocoa
I need to buy drinking_chocolate
I need to buy coffee
I need to buy burnt_umber
I need to buy hot_chocolate
I desire to buy hot_chocolate
I desire to buy burnt_umber
I desire to buy deep_brown
```

```
Enter a sentence: I went to market with my brother to buy fruits
['PRP', 'VBD', 'IN', 'NN', 'IN', 'PRP$', 'NN', 'TO', 'VB', 'NNS']
['I', 'went', 'to', 'market', 'with', 'my', 'brother', 'to', 'buy', 'fruits']
```

```
Top Paraphrases :
I get to market_place with my Brother to purchase fruit
I conk_out to market_place with my Brother to purchase fruit
I live_on to marketplace with my Brother to purchase fruit
I go to market_place with my Brother to purchase fruit
I pass to marketplace with my Brother to purchase fruit
I decease to marketplace with my Brother to purchase fruit
I cash_in_one's_chips to market_place with my Brother to purchase fruit
I buy_the_farm to commercialize with my Brother to purchase fruit
I get to commercialise with my Brother to purchase fruit
I locomote to market_place with my Brother to grease_one's_palms fruit
```

## **Future Scope:**

- **Experimenting with an encoder-decoder model to generate paraphrasing text:**

1. we will try the utilization of an encoder-decoder architecture. Here, input sentences undergo encoding to capture their semantic essence. Subsequently, the decoder utilizes this encoded information to generate paraphrased versions while maintaining fidelity to the original sentence's meaning.
2. Central to our approach is the strategic propagation of loss from the decoder concerning the original sentence. By minimizing this loss to a predefined threshold, we aim to ensure the creativity and variability of the model's outputs. This threshold-based approach prevents the generation of overly similar paraphrases, thus fostering diversity in the paraphrased text.

- **Leveraging Pre-trained Transformer Models and Seq2Seq Training**

1. Our proposed approach involves harnessing the capabilities of pre-trained transformer-based models such as BERT or GPT-2 for the generation of unsupervised paraphrases. This strategy presents a promising avenue for expanding the diversity and quality of paraphrased text.
2. We intend to leverage the robustness of pre-trained transformer models in understanding and generating natural language. Our methodology begins by selecting sentences from the corpus and identifying their parts of speech (POS) tags. Subsequently, we target the most relevant words corresponding to these POS tags and mask them out. Approximately 15-20% of the words in each sentence will be masked to facilitate the generation of diverse paraphrases.
3. With the masked sentences prepared, we will input them into pre-trained transformer models. These models possess the ability to understand contextual nuances and generate coherent text. By feeding the masked sentences to these models, we anticipate the generation of new, paraphrased versions.
4. In an alternative approach, we can utilize the output sentences as paraphrases or create a new labeled dataset consisting of pairs of original sentences and their corresponding generated paraphrases. This labeled dataset can then be used to train a sequence-to-sequence (seq2seq) model. By training on this dataset, the seq2seq model learns to generate paraphrases effectively. Subsequently, we can leverage this trained seq2seq model as a paraphrase generator for future tasks.
5. The utilization of pre-trained transformer-based models for unsupervised paraphrase generation represents a significant advancement in natural language processing. By incorporating masked words and leveraging the contextual understanding of these models, we aim to generate diverse and contextually relevant paraphrases. Whether directly using model-generated paraphrases or training a seq2seq model on labeled datasets, our approach promises to enhance the quality and diversity of paraphrased text.