NEURAL UNSUPERVISED PARAPHRASING

Team Linguisto - 47

Nevil Sakhreliya 2023201005

Darshak Devani 2023201007

Viraj Shah 2023201011

INTRODUCTION TO NLP

1

2

3

4

5

6

7

8

9

10

Introduction

In today's information-driven world, the ability to efficiently paraphrase text while retaining its original meaning is crucial for a variety of natural language processing (NLP) tasks such as machine translation, text summarization, and question-answering systems. Traditional methods for paraphrasing often rely on supervised learning, which requires large amounts of annotated data. However, the process of collecting such data can be time-consuming and expensive.

1
2
3
4
5
6
7
8
9
10

Objective

The objective of our project is to explore and develop techniques for neural unsupervised paraphrasing. Unlike supervised approaches, unsupervised methods do not require parallel corpora or labeled datasets. Instead, they leverage the inherent structure and semantics of the text to generate accurate and meaningful paraphrases.

Approaches and Implementations

1. Statistical Approach

- Implemented a statistical paraphrasing method incorporating word substitution, POS tagging, and random sampling of synonyms from WordNet.
- Utilized Flair for POS tagging to understand the grammatical structure.
- Identified specific POS tags for synonym replacement: nouns, verbs, adjectives, and interjections.

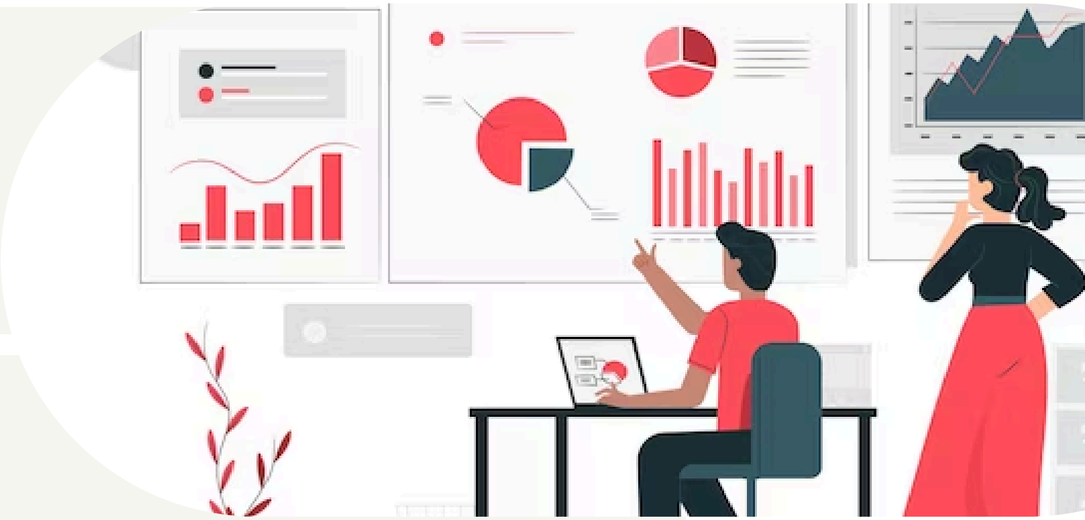
Paraphrase Generation Process

1
2
3
4
5
6
7
8
9
10

- **Synonym Retrieval**: Leveraged WordNet to obtain synonyms for words based on POS tags.
- **Combination of Sentences**: Systematically replaced words with synonymous counterparts, generating multiple paraphrased versions.
- **Random Sampling**: Employed to ensure diversity in paraphrases and reduce computational complexity.
- **Paraphrasing**: Replaced words in original sentences with selected synonyms.
- **Similarity Evaluation**: Calculated similarity scores (BERT score, BLEU score) between original and paraphrased sentences for quantitative assessment.

Results


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



```
Top Peraphrases :
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
```

2. Generation Using Pre-trained Model

- Utilize unsupervised paraphrasing with pre-trained GPT-2 models.
- Employ specific pre-trained models for English and Hindi texts.
- Approach is robust to domain shifts as it doesn't rely on labeled data.
- Treat paraphrasing as sentence reconstruction from corrupted input.
- Form Source (S) by removing stop words from original sentence.
- Original sentence serves as Target (T).
- Use GPT-2 to generate Target given Source, estimating P(T | S).

Paraphrase Generation Process

1
2
3
4
5
6
7
8
9
10

Data Preprocessing:

1. Stop Words Removal :

- **English:** Utilize NLTK Stop Words
- **Hindi:** Implement Custom Stop Words List for Effective Filtering

2. Special Character Removal :

- Eliminate Symbols such as ",", """, etc. to Ensure Clean Data for Further Processin

Paraphrase Generation Process

Sentence Corruption Techniques:

For English Text:

1. **Random Shuffle:**
   - Reorder the given input sentence randomly for variation.

2. **Synonym Replacement:**
   - Randomly select 40% of the words from the sentence.
   - Utilize WordNet library to find synonyms for selected words.
   - Replace each selected word with a randomly chosen synonym to introduce diversity.

Paraphrase Generation Process

Sentence Corruption Techniques:

For Hindi Text:

1. **Random Shuffle:**
   - Shuffle the given input sentence multiple times for enhanced variation.

2. **Random Word Deletion:**
   - Randomly remove certain words from the original sentence to induce alterations and generate diverse outputs.

Paraphrase Generation Process

1 2 3 4 5 6 7 8 9 10

Sentence Corruption Techniques:

- **Input Transformation:**
  - Convert sentences into structured format:
    - **<BeginingOfSentence>** Corrupted Sentence **<SeperaterToken>** Original Sentence **<EndOfSentence>**.

- **Inference Process:**
  - Remove stop words.
  - Provide modified input to model for paraphrase generation.

- **Output Generation:**
  - Generate multiple paraphrases (e.g., 20) for test sentence.
  - Each paraphrase captures different interpretation or wording.

Results

1 2 3 4 5 6 7 8 9 10 11

English Corrupted Sentence

```
------------------------------------------------------------
input:  if we see something in our dreams and it happens to come out true after few days, what does that mean?

corrupted:  dreams happens come out true after days, mean?

input:  how do you send a private message to someone you’re following on quora?

corrupted:  send private message following quora ?
```

Hindi Corrupted Sentence

```
Original Sentence :  अपने लक्ष्य को हासिल करने के लिए हमें कभी भी हार नहीं माननी चाहिए।
corrupted :  हमें कभी हासिल नहीं हार । लक्ष्य चाहिए के लिए को अपने माननी करने भी

Original Sentence :  मैं अपनी मां के साथ बाजार के लिए निकला
corrupted :  के निकला मैं लिए साथ मां के बाजार अपनी
```

Results

English Paraphrasing BERT Score

```
input:  how do you send a private message to someone you’re following on quora?

corrupted:  send private message following quora ?

Op 1 :  How do I post a private message that I have not following in quora?  Score: 0.8226505517959595

Op 2 :  When can I post a private message by following on not quora?  Score: 0.8057623505592346

Op 3 :  How can I message a private message following on quora?  Score: 0.8053717613220215

Op 4 :  Can you time or send a private message to someone following that person on Quora?  Score: 0.80044

Op 5 :  Why are you private in a message following quora?  Score: 0.7868369817733765

Op 6 :  How to send private message when following on quora?  Score: 0.7769818902015686

Op 7 :  What should I put in my private message following quora?  Score: 0.7722793817520142

Op 8 :  How can i notify my friends regarding my private messages on quora?  Score: 0.7692236304283142

Op 9 :  Can I follow my private message on Quora?  Score: 0.7407433390617371

Op 10 :  Can I messages be after I have left private message?  Score: 0.7263333201408386
```

Results

Hindi Paraphrasing BERT Score

Original Sentence : अगर आपका आज परिश्रम से बढ़िया है, तो आपका कल स्वयं आप बेहतर होगा।
0: अगर समय हो तो आज बेहतर तो होकर बिना समय खर्च ही आपका परिश्रम बढ़िया स्वयं करिए होगा । 0.8821431398391724
1: बेहतरीन कल, तो आज आपका स्वयं परिश्रम से आपको बढ़िया होगा आज आपका होगा, कल तो परिश्रम बढ़िया से । 0.8882477283477783
2: आपका तो कल, आपका होगा और आज ही बढ़िया स्वयं परिश्रम से होगा तो बेहतर बात । 0.9005407094955444
3: आज का समय बेहतर, आपका होगा परिश्रम आज तो तब बात बेहतर कल आपका स्वयं तो नहीं होगी । 0.8811371326446533
4: आपका होगा, कल है, तो आज आपका स्वयं बेहतर परिश्रम से आपको बढ़िया आज कल बिताया होगा । 0.9377497434616089
5: कल है, आपका बेस्ट परिश्रम और होगा स्वयं तो आज आपका ही बेहतर आज तो सुख ही आपका सूख बढ़िया होगा । 0.7610171437263489
6: आज केदारनिया, कल हो आपका परिश्रम से आपकी स्वयं बढ़िया कल सफलता केदारनिया होगा । 0.8636866807937622
7: आज होगा तो आपका बेहतर कल, आज मेरा ही स्वयं आज परिश्रम आपका होगा । 0.9103513360023499
8: कल है, स्वयं आपका बेहतर स्वजन परिश्रम होगा और आपका हजारों से ज्यादा ब्लॉगजगत बढ़िया होगा । 0.8279228210449219
9: यदि आज स्वयं तो सपनों, क्षितिज, बुन्दे परिश्रम से भी आपका बढ़िया होगा तो बढ़िया परिस्रम । 0.8395029902458191
10: आपका बेहतर आज, काम करिये तो वो होगा तो होगा आपका तो स्वयं से भी परिश्रम बढ़िया होगा, सारे तीन प्रयत्त करिये । 0.8657934665679932
11: जब आपका कल है, आपका परिश्रम होगा बेहतर तो स्वयं से ही आज बढ़िया स्वयं करियर प्राप्त होगा । 0.904374897480011
12: पहली तो आजकल तो आपका मेहनत से आपका दूसरी बेहतर स्वयं ही परिश्रम होगा । 0.8676429986953735
13: आज काम स्वयं पाठक होगा तो बेहतर हालातों परिश्रम हाथ से लगा तो आपका होगा बढ़िए, कल स्वयं ही करेंगे । 0.8347238898277283
14: कल, आपका बेहतर स्वयं कार्य, परिश्रम, आज तो से भविष्य आपका बढ़िया स्वार्थ होगा । 0.9057255983352661
15: आज आपका समय हास्य स्वयं से आपका होगा, कल कल परिश्रम आपका बढ़िया स्वयँ परिधान करेंगे । 0.8861396312713623
16: कल है, कल यानि आज आपका वेबसाइट होगा सारा बेहतर ही स्वयं से परिश्रम करेंगे आपका परिचय होगा । 0.875823974609375
17: कई बार, आज आपका भविष्य आपका होगा, तो परिश्रम से बेहतर स्वयं होगा..! 0.8603755235671997
18: तो कल है, आपका स्वयं बेहतर आज तो परिवर्तन आपका बेहतर ही से होगा बढ़िया से आपका होगा । 0.8716936111450195
19: स्वयं होगा आपका आज बेहतर ये होगा कि कल भी आपका परिश्रम से होगा जो ज्यादा । 0.9032458662986755

Analysis

- A statistical method is less efficient than fine-tuning with pre-trained models.
- Statistical method focuses on word substitution without considering sentence structure.
- Fine-tuning utilizes pre-trained models and reconstructs sentences for paraphrases.
- Limited fine-tuning of GPT2 due to resource constraints, resulting in less accurate results.
- English pre-trained models have fewer parameters and perform better in certain contexts compared to Hindi models.

Thank you 🙂