NEURAL UNSUPERVISED PARAPHRASING

Team Linguisto

Nevil Sakhreliya 2023201005

Darshak Devani 2023201007

Viraj Shah 2023201011

INTRODUCTION TO NLP

1

2

3

4

5

6

7

8

9

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

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.

Progress Snapshot

1
2
3
4
5
6
7
8
9

- 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

- **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
```

Future Scope

1
2
3
4
5
6
7
8
9

Experiment with the Encoder-Decoder Model for Paraphrasing

- Utilization of encoder-decoder architecture
- Encoding captures semantic essence
- Decoder generates paraphrased versions while preserving the original meaning
- Strategic loss propagation to maintain creativity and diversity

Leveraging Pre-trained Transformer Models

- Harnessing capabilities of pre-trained transformer models like BERT or GPT-2
- Selection of sentences and identification of POS tags
- Masking out 15-20% of words for diversity
- Inputting masked sentences into transformer models
- Generating diverse paraphrases leveraging contextual understanding
- Alternative approach: creating labeled dataset for seq2seq training
- Use of trained seq2seq model for paraphrase generation

Thank you 🙂