UNIVERSITY OF PIRAEUS

NATURAL LANGUAGE PROCESSING

21/06/2025

Exams June Project

6TH Semester

Full Name and AM:

Marina Ioannou π22211

Ilia Vatti π22208

Table of Contents

NLP Project

– Introduction

Semantic reconstruction plays a vital role in the field of Natural Language Processing (NLP), especially when transforming ambiguous, grammatically incorrect, or poorly structured texts into semantically accurate, fluent, and well-structured language. Such tasks are critical in machine translation, automated editing, intelligent tutoring systems, and human-computer communication.

In this study, we explore multiple approaches to semantic reconstruction using both rule-based methods and modern transformer-based models. We further apply embedding-based similarity analysis to evaluate the preservation of meaning between the original and rewritten texts.

– Methodology

We divide the methodology into three parts (A, B, C), each addressing a different level of reconstruction and evaluation.

- **Part A: Sentence-Level Manual Reconstruction (Custom Model)**

Two original, grammatically incorrect sentences were selected from the original texts. A rule-based reconstruction was manually designed, applying syntactic correction and fluency enhancement using:

- Subject–Verb agreement
- Lexical substitutions
- Verb tenses and active voice
- Removal of redundancy or ambiguity

Example:

- Original: `I am very appreciated the full support of the professor.`
- Rewritten: `I really appreciate the full support of the professor.`

- **Part B: Full Text Reconstruction via 3 Pipelines**

We applied three different automatic reconstruction methods on the full original texts.

1. Custom Rule-Based Rewriter: Implements hard-coded grammar correction and phrasing logic using `spaCy` and simple string rewriting.
2. HuggingFace T5 Pipeline: Uses the `t5-base` model for paraphrasing via HuggingFace Transformers.
3. TextAttack Paraphraser: Applies transformer-based paraphrasers (e.g., BART, T5) using the TextAttack API.

Each version of the text was evaluated for semantic accuracy and fluency.

- **Part C: Evaluation Using Similarity Metrics**

For evaluation, we used:

- Cosine Similarity: Calculates semantic closeness between vector representations.
- Word Embeddings:
    - Word2Vec
    - GloVe
    - FastText
    - BERT (SentenceTransformer)
- Dimensionality Reduction: PCA and t-SNE for visual analysis.

All techniques were implemented in **Python** with **`scikit-learn`**, **`gensim`**, **`sentence-transformers`** and **`matplotlib`**.

– Experiments & Results

Examples (Before/After)

Sentence A:

- Before: `I got this message to see the approved message.`

- After (T5): `I received the approved message.`

Sentence B:

- Before: `Hope you too, to enjoy it as my deepest wishes.`

- After (Rule-based): `I hope you also enjoy the festival – my best wishes.`

Cosine Similarity Results

Method	Text 1 Similarity	Text 2 Similarity
Custom Rule-Based	0.9977	1.0
HuggingFace T5	0.9912	0.9944
TextAttack Paraphraser	0.9074	0.948

Text1:

Cosine Similarity (token-level avg.): 0.9986

Text2:

Cosine Similarity (token-level avg.): 0.9992

Visualizations

- PCA and t-SNE showed that embeddings of the rewritten texts clustered closely with the original texts, especially for T5-based paraphrasing.
- BERT-based embeddings maintained semantic alignment more accurately than GloVe or Word2Vec.

– Discussion

Embedding Performance

- BERT embeddings captured contextual semantics better than static embeddings like GloVe or Word2Vec.
- Cosine similarity aligned well with perceived semantic quality.

Challenges

- Handling vague or grammatically broken inputs was especially difficult for rule-based approaches.
- Maintaining tone and nuance in paraphrasing without distortion is non-trivial.
- Some transformer outputs introduced overly simplified or altered meanings.

Automation with NLP Models

- Transformers like T5 can effectively automate semantic reconstruction.
- Rule-based methods are useful for predictable structures but do not generalize.

Comparative Analysis

- HuggingFace's T5 yielded the best fluency and coherence.
- TextAttack offered similar performance but slightly less control.
- Rule-based methods were interpretable but limited in flexibility.

– Conclusion

This study demonstrates that semantic reconstruction of unstructured texts is achievable using modern NLP models. While rule-based methods offer simplicity and control, transformer models provide high-quality paraphrasing at scale. Evaluation using embeddings and cosine similarity supports that meaning can be preserved effectively in many automated approaches.

– References

- TextAttack – https://github.com/QData/TextAttack
- Sentence-Transformers – https://www.sbert.net/
- Preprocess – https://github.com/dimitris1pana/nlp_lab_unipi/blob/6eee9d138a952fd06d7a4029b22611de62e1d651/lab2/textPreprocessing.ipynb
- Embeddings visualization – https://github.com/dimitris1pana/nlp_lab_unipi/blob/6eee9d138a952fd06d7a4029b22611de62e1d651/Neuron2transformer/1.linearSimple.ipynb

Link for GitHub:

https://github.com/Mar1na04/fysiki_glossa