Semantic Role Labeling Using Conditional Random Fields (CRF)

1. Introduction

Semantic Role Labeling (SRL) is a key task in Natural Language Processing (NLP) that assigns roles to words in a sentence

to determine their relationships within a predicate-argument structure. This report provides an in-depth analysis of a

CRF-based SRL model implemented in Python. The study explores the preprocessing, feature extraction, model training, and

evaluation techniques used in the implementation.

2. Code Overview

The code follows a structured approach for implementing SRL, consisting of the following key components:

2.1 Data Preprocessing

The dataset used is in CoNLL-U format, a tab-separated format widely used for dependency parsing and semantic annotation.

The function parse_conllu() reads the dataset and extracts key linguistic features, including:

- Lemma: The base form of a word.

- POS (Part of Speech): Grammatical category of a word.

- Head and Dependency Relation: Dependency parsing attributes.

- Predicate Information: Identifies the main verb.

- SRL Label: The semantic role assigned to each word.

2.2 Feature Engineering

Feature extraction is critical for SRL model performance. The function word2features() constructs a feature vector for

each token, capturing:

- POS Tags: Useful for syntactic information.

- Word Prefixes and Suffixes: Helps in morphological analysis.

- Sentence Position Features: Beginning-of-Sentence (BOS) and End-of-Sentence (EOS) indicators.
- Previous and Next Word Information: Contextual features for dependency modeling.

2.3 Model Development

A Conditional Random Field (CRF) model is trained using the sklearn-crfsuite package. The CRF model optimizes label

sequence predictions using:

- L1 & L2 Regularization: To prevent overfitting.
- LBFGS Optimization Algorithm: Efficient for large-scale CRF models.
- Maximum Iterations (100): Controls model convergence.

The model is trained using the parsed training dataset and evaluated using precision, recall, and F1-score metrics.

2.4 Hyperparameter Tuning & Cross-Validation

- Randomized Search CV: Finds the best values for c1 (L1 regularization) and c2 (L2 regularization).
- 5-Fold Cross-Validation: Ensures robustness and prevents overfitting.

2.5 Model Evaluation

The CRF model is evaluated using:

- Classification Report: Provides precision, recall, and F1-score per label.
- Confusion Matrix: Visualizes model performance across different role labels.
- Performance on Short vs. Long Sentences: Identifies challenges with varying sentence lengths.

3. Real-World Applications of Semantic Role Labeling

Semantic Role Labeling plays a crucial role in various real-world applications, including:

- Machine Translation: Enhancing meaning preservation across languages.
- Question Answering Systems: Improving understanding of context in user queries.
- Chatbots and Virtual Assistants: Providing more natural and context-aware responses.
- Text Summarization: Extracting key elements from long documents.

- Legal and Medical NLP: Assisting in automated document processing and summarization.

4. Conclusion

This study highlights the effectiveness of CRF for SRL tasks. The model demonstrates high accuracy in predicting common

roles while struggling with rarer semantic roles, suggesting the need for further refinement in feature engineering and

dataset balancing. Future work could explore deep learning approaches such as BiLSTM-CRF to enhance performance.

References

- Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and

Labeling Sequence Data.

- Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing. Pearson.