Word Sense Disambiguation (WSD)

1. Neural WSD using Pre-trained Language Models

Word Sense Disambiguation (WSD), the task of identifying the correct meaning of a word in context, is a longstanding challenge in Natural Language Processing (NLP). With the rise of transformer-based models such as BERT, RoBERTa, and DeBERTa, WSD systems have seen significant performance improvements due to their strong contextual understanding.

However, while these models perform well on general NLP tasks, their effectiveness on fine-grained sense-level distinctions remains underexplored. There is also a lack of clarity on how different transformer architectures compare in the context of WSD.

This project aims to evaluate and compare the performance of selected pre-trained transformer models on WSD tasks in English. The study will involve fine-tuning these models using standard datasets such as SemCor and ALL-Words corpora, and measuring their accuracy in distinguishing word senses in various contexts.

The goal is to identify the most suitable models and fine-tuning strategies for real-world WSD applications, and to better understand the strengths and limitations of neural language models in handling semantic ambiguity.

2. Multilingual or Cross-lingual WSD for Low-resource or Code-mixed Text

Code-mixed languages such as Singlish (Sinhala-English) are widely used in digital communication in Sri Lanka. These languages present unique challenges due to frequent code-switching, non-standard grammar, and limited annotated data. Traditional monolingual NLP models often fail to generalize in such settings.

This project focuses on Cross-lingual Word Sense Disambiguation using multilingual transformer models such as XLM-RoBERTa. The goal is to leverage cross-lingual knowledge to improve WSD performance in low-resource languages like Sinhala and code-mixed text like Singlish.

The approach involves fine-tuning multilingual models on English WSD datasets and evaluating their ability to generalize in zero-shot or few-shot scenarios. Synthetic or semi-supervised methods may also be used to generate data for evaluation.

The objective is to explore the potential of multilingual models in understanding semantic ambiguity across languages and contribute to better support for underrepresented linguistic communities in NLP research.

3. Knowledge-augmented WSD

Transformer-based models like BERT have achieved impressive results in contextual understanding but lack access to structured external knowledge. Lexical resources such as WordNet, BabelNet, and ConceptNet offer valuable semantic relationships that can improve WSD accuracy and interpretability.

This project proposes a hybrid approach to Word Sense Disambiguation that combines transformer embeddings with external knowledge bases. The idea is to enhance the model's predictions by incorporating graph-based information such as word relationships, definitions, and semantic hierarchies.

The research will involve representing context using pre-trained language models, integrating knowledge features from resources like WordNet, and experimenting with graph-based techniques or feature fusion strategies.

The goal is to improve disambiguation accuracy and provide more interpretable results, especially in cases where context alone is insufficient to resolve meaning. This project aims to bridge the gap between symbolic and neural approaches in modern NLP systems.