# MNLP Homework-2
# Word Sense Disambiguation

**Francesco Palandra**
Department of Computer Science
Sapienza University of Rome
palandra.1849712@studenti.uniroma1.it

## Abstract

Word Sense Disambiguation is a pivotal task in Natural Language Processing aimed at determining the correct sense of a word within a specific context. Traditional approaches often overlook the potential of integrating lexical resources directly into neural networks. This report explores one approach that leverages glosses—definitions of words—as integral components of the model by creating context-gloss pairs fed into a BERT-based architecture. By fine-tuning a pre-trained BERT model with these pairs, high performance are achieved.

## 1 Introduction

Word Sense Disambiguation (WSD) represents a critical and challenging task in the field of Natural Language Processing. The objective of WSD is to accurately predict the most appropriate sense (gloss) of an ambiguous word given its context (Navigli, 2009). Initial approaches focused on leveraging contextual clues and lexical databases such as WordNet (Miller, 1994). The advent of neural networks introduced a paradigm shift, with glossBERT (Huang et al., 2019) pioneering the integration of word definitions directly into the model's learning process. Building upon the innovative strides made by glossBERT, Extractive Sense Comprehension (ESC) (Barba et al., 2021) introduces a novel approach by reframing WSD as an extractive text task, inspired by methodologies from the Extractive Reading Comprehension framework in Question Answering (QA) (Rajpurkar et al., 2016), aiming to identify the text span that most accurately reflects the target word's intended meaning. This work builds upon these methodologies, focusing on the application of gloss knowledge through a BERT-based model to refine and enhance the disambiguation process.

## 2 Methodology

In this section, the method will be presented in detail.

### 2.1 Problem Formulation

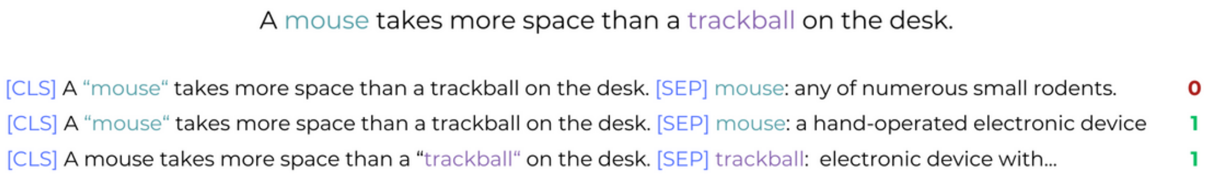
Given a sentence $S = \{w_1, w_2, ..., w_n\}$ and a subset of target words $T = \{t_1, t_2, ..., t_m\}$ where each $t_i$ is a sequence of one or more words in $S$ $t_i = \{w_j, ..w_k\}$, the challenge is to accurately associate each target word $t_i$ with the correct gloss from a set of candidate glosses $G = \{g_{t_1^1}, g_{t_1^2}, ..., g_{t_m^q}\}$.

### 2.2 GlossBERT

The core of this approach is GlossBERT, a model that utilizes the BERT architecture - a transformer-based model known for its effectiveness in capturing deep contextual relationships within text. GlossBERT enhances the traditional BERT model by incorporating gloss information directly into the training process. This is achieved by fine-tuning a pre-trained BERT model on data using sentence-gloss pairs, thus enabling the model to learn the distinctions of word senses in varied contexts.

### 2.3 Gloss Pairs Creation

A crucial part of this approach is the creation of gloss pairs, where each sentence in the dataset is paired with each possible gloss of the target words it contains. This approach effectively transforms the WSD task into a binary classification problem, where the model predicts the suitability of a gloss given the context of the sentence (see figure 1). The linearization of sentence-gloss pairs significantly increases the dataset size and so the training time, however, it simplifies the learning task for the model. Based on part-of-speech (Nathani and Joshi, 2021) compatibility, we streamline the input data by filtering the potential glosses. Additionally, weak supervision techniques, such as the insertion of special tokens around target words, have been



A mouse takes more space than a trackball on the desk.

[CLS] A “mouse“ takes more space than a trackball on the desk. [SEP] mouse: any of numerous small rodents. 0

[CLS] A “mouse“ takes more space than a trackball on the desk. [SEP] mouse: a hand-operated electronic device 1

[CLS] A mouse takes more space than a “trackball“ on the desk. [SEP] trackball: electronic device with... 1

Figure 1: Example of gloss pair, for each word to disambiguate a pair is generated with each possible candidate sense. A separator is positioned between the sentence and the gloss and the target word is wrapped with quotes to add a weak supervision. The number of entries with this approach rises in the order of 4x, resulting in more time for each epoch.

adopted as glossBERT paper suggested to further enhance model performance.

### 2.4 Gloss Extraction

The model itself does not assign the right gloss to a target word, rather it returns the degree of similarity between the sentence context and each gloss. Therefore, after processing by GlossBERT, each sentence-gloss pair is assigned a score reflecting the model's confidence in the pair's correctness. These scores are then used to rank the glosses for each target word, with the top-ranked gloss being selected as the most suitable sense. This process leverages the model's ability to evaluate the compatibility of each gloss with the given sentence context, thus achieving accurate disambiguation.

## 3 Evaluation

The evaluation of the model will be done through the F1-macro score, which, in this context, aligns with accuracy. Three distinct model configurations have been evaluated and compared to study their impact on disambiguation accuracy:

**Target-Token** A masking strategy over the target words is employed, maintaining zeros across all positions except for those of the target words. This allows the model to focus exclusively on the embeddings of the target words, averaging the embeddings in scenarios where the target comprises multiple words.

**CLS-Token** An alternative approach utilizes the BERT [CLS] token, which is designed to aggregate the overall sentence context for classification tasks. This method examines the efficacy of leveraging this aggregated context representation for WSD.

**CLS-Token-WS** This variant explores the influence of weak supervision by inserting special tokens around the target word. The hypothesis made in glossBERT is that such contextual markers can enhance the model's ability to select the correct gloss based on the sentence's context.

### 3.1 Coarse Grained vs Fine Grained

The evaluation extends beyond model configurations to incorporate the granularity of gloss disambiguation, distinguishing between coarse-grained and fine-grained approaches. Coarse-grained disambiguation groups related senses into clusters, simplifying the task by reducing the number of distinct classes. In contrast, fine-grained disambiguation treats each sense as a separate class, demanding a more refined understanding from the model. While primary evaluation focuses on coarse-grained disambiguation, the model's training leveraged fine-grained gloss distinctions to enhance its understanding.

## 4 Results

The comparative analysis 1 of the three model configurations—Target-Token, CLS-Token, and CLS-Token-WS—reveals insightful trends in their performance. The Target-Token model demonstrates a high level of accuracy, achieving a remarkable score of 0.93 on the test set, underscoring the efficacy of focusing on the target embeddings. Conversely, the CLS-Token model, while still performing great, attains a slightly lower accuracy of 0.90 in the test phase. Lastly, the CLS-Token-WS model surpasses the others in the test set, with an accuracy of **0.95**, attributed to the enhanced model comprehension by indicating the target word's position

| **Model** | Train | Val | Test |
|---|---|---|---|
| Target | 0.981 | 0.924 | 0.931 |
| CLS-Token | 0.983 | 0.926 | 0.904 |
| CLS-Token-WS | **0.985** | **0.927** | **0.946** |

Table 1: Accuracy of the models in the train, validation and test set

and embedding it within quotes. This improvement suggests the significant impact of positional awareness on model performance. Furthermore, the rapid achievement of high accuracy from the first epoch across all models indicates swift learning and stability, emphasizing the models' efficiency in grasping the task's complexities with minimal training.

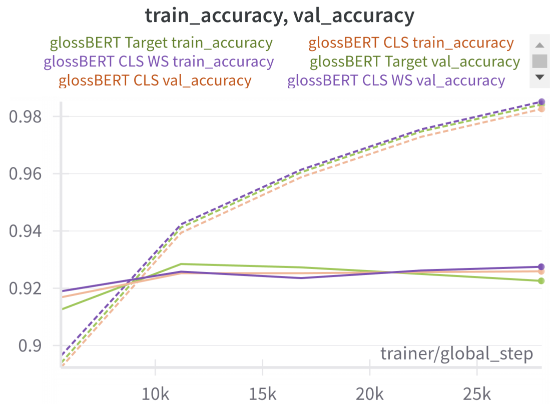


Figure 2: Validation and Train accuracy of the three models: Target, CLS-Token, CLS-Token-WS

## 5 Conclusion

This report explores Word Sense Disambiguation (WSD) through a BERT-based approach, emphasizing the integration of glosses for enhanced understanding. The study investigates various model configurations, with the CLS-Token-WS model showing the most promise by effectively utilizing positional cues and quotations for improved disambiguation accuracy. The findings suggest that precise embeddings and contextual framing are crucial for WSD performance. Future directions may include exploring more complex contextual integrations and a deeper study of the definitions of the glosses.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| embeddings | BERT | optimizer | Adam |
| mode | CLS | lr | 1e-5 |
| WS | True | weight decay | 0.01 |
| epochs | 5 | dropout | 0.1 |
| batch size | 16 | glosses | fine |

Table 2: List with the best hyperparameters found

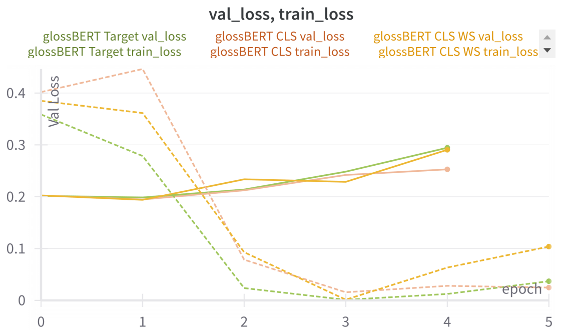


Figure 3: Validation and Train loss of Target, CLS-Token, CLS-Token-WS model

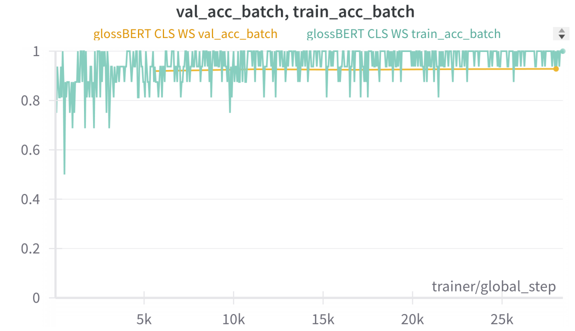


Figure 4: Accuracy in each batch of train and validation for the CLS-Token-WS model

## References

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.*

Bharti Nathani and Nisheeth Joshi. 2021. Part of speech tagging for a resource poor language : Sindhi in Devanagari script using HMM and CRF. In *Proceedings* *of the 18th International Conference on Natural Language Processing (ICON)*, pages 611–618, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.