University of Stuttgart
Germany

# Insights from Transfer Learning Experiments with Word-in-Context and Word Sense Disambiguation Models

**Alp Mujko, Dominik Schlechtweg**

Institute for Natural Language Processing, University of Stuttgart

## Introduction & The Gap

- **Core Challenge**: Understanding contextual word meaning. We compare two related, yet typically separately studied tasks:
  1. **WiC (Word-in-Context)**: Binary classification. Does a target word have the same meaning across two contexts?
     *"She **books** a flight" vs. "I read a **book**." → 0 (Different)*
  2. **WSD (Word Sense Disambiguation)**: Multi-class. Assigning the correct sense label from a predifined inventory.
     *"She **books** a flight" → Sense 2 (**To arrange in advance**)*
- **The Research Gap**: Despite conceptual overlap, the potential benefits of **jointly training** models on both tasks, or leveraging shared representations, remain largely unexplored.

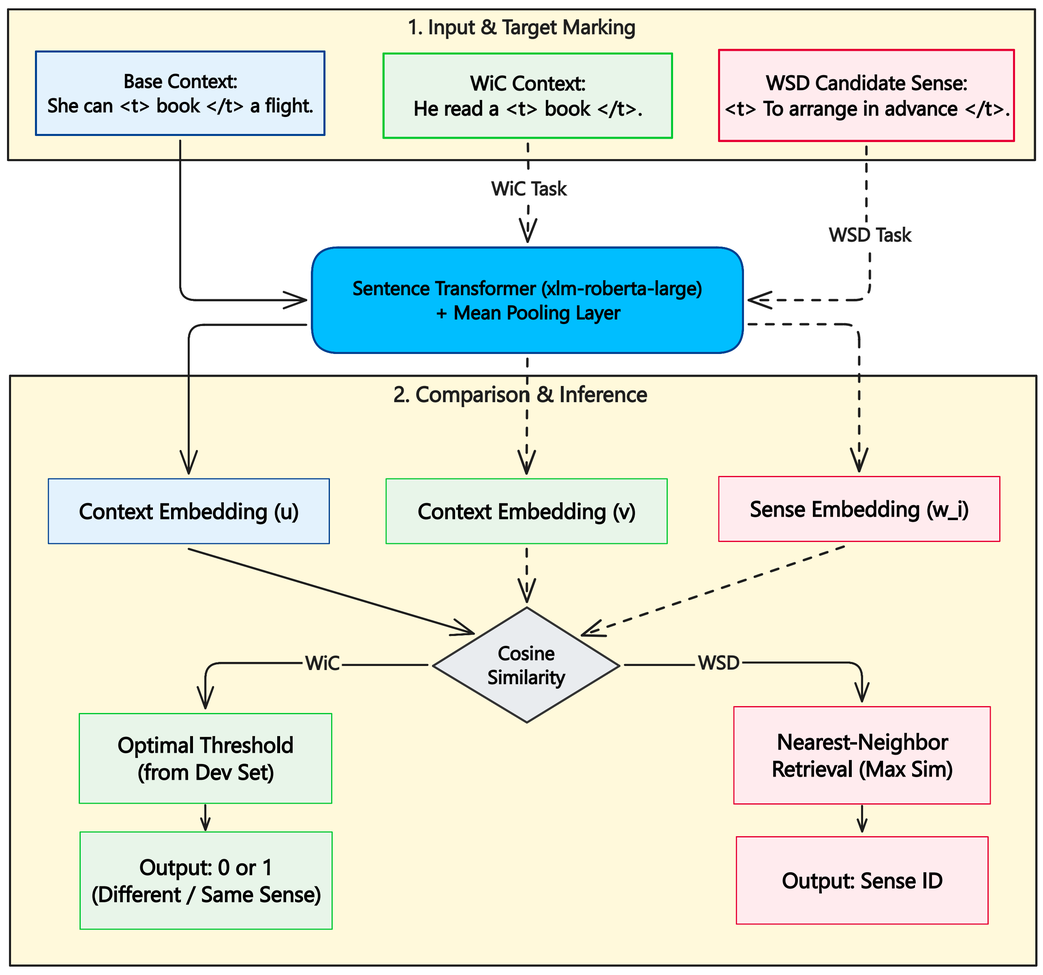
## Data & Methodology

- **WiC Datasets**: XL-Lexeme (13.4K), MCL-WiC (8K), Pil-WiC (5.4K)
  [Cassotti et al. 2023, Martelli et al. 2021, Pilehvar & Camacho-Collados 2019]
- **WSD Dataset (FEWS)**: 200K train instances; 5K zero-shot test
  [Blevins et al. 2021]
- **Architecture**: Sentence Transformer (`xlm-roberta-large`), mean pooling, Contrastive Loss.
- **Target Marking**: `<t>` and `</t>` around target words & WSD definitions to guide attention.
- **Inference**:
  - **WiC:** Cosine similarity between context embeddings, converted to a binary label using an **optimal dev-set threshold**.
  - **WSD:** Nearest-neighbor retrieval based on the highest cosine similarity between the context and candidate sense embeddings.

## Core Hypotheses

- **H1 (Joint Training):** Jointly training models on both WiC and WSD data will lead to improved, or at least non-detrimental, performance on each individual task compared to single-task training.
- **H2 (Cross-Task Transfer):** Models trained exclusively on one task can successfully generalize to the other, indicating a shared underlying semantic representation.

## Model Architecture




## Main Results: Accuracy on WiC and WSD Evaluation Sets

**Pil-WiC**

| Config | WiC | WSD |
|---|---|---|
| Pure WiC (P) | .681 | .475 |
| P + $F_{5K}$ (Low-Res) | .695 | .552 |
| P + $F_{200K}$ (High-Res) | **.714** | .665 |
| Pure WSD ($F_{200K}$) | .693 | **.670** |
| Untrained Base | .529 | .309 |
| Random Chance | .492 | .262 |

**MCL-WiC**

| Config | WiC | WSD |
|---|---|---|
| Pure WiC (M) | .889 | .477 |
| M + $F_{8K}$ (Low-Res) | .890 | .575 |
| M + $F_{200K}$ (High-Res) | **.891** | .663 |
| Pure WSD ($F_{200K}$) | .860 | **.670** |
| Untrained Base | .653 | .309 |
| Random Chance | .479 | .262 |

**XL-Lexeme**

| Config | WiC | WSD |
|---|---|---|
| Pure WiC (X) | .790 | .472 |
| X + $F_{13K}$ (Low-Res) | **.792** | .609 |
| X + $F_{200K}$ (High-Res) | .782 | .669 |
| Pure WSD ($F_{200K}$) | .754 | **.670** |
| Untrained Base | .585 | .309 |
| Random Chance | .508 | .262 |

Table 1: *WiC and WSD test accuracies. Abbreviations: P = Pil-WiC, M = MCL-WiC, X = XL-Lexeme, F = FEWS. Subscripts indicate WSD training size. Joint training is shown with limited (Low-Res) vs. full (High-Res) WSD data. Combined (C) models omitted for brevity.*

## Key Takeaways & Implications

- **Low-Resource Synergy (H1)**
  1. **Boost for small datasets**: Joint training with downsampled WSD subsets reliably improves both WiC and WSD accuracy.
  2. **Practical Implication**: Merging cross-task data is highly effective when annotated resources are scarce.
- **The Limit of Multi-Tasking (H1)**
  1. **Task Interference**: When large-scale WSD data (200K) is available, adding the WiC objective slightly degrades WSD performance.
  2. **Implication**: Once a primary task reaches data saturation, auxiliary tasks can act as a distraction.
- **Strong Cross-Task Generalization (H2)**
  1. **Shared Foundations**: Models trained exclusively on a single task significantly outperform random and untrained baselines on the alternate task.
  2. **Asymmetric Transfer (WSD → WiC)**: A model trained *only* on 200K WSD instances generalized so well it outperformed an in-domain WiC model (.693 vs .681 on Pil-WiC).
  3. **The Semantic Mechanism**: Explicit sense-anchoring (WSD) provides richer, more robust semantic supervision than relative context-matching (WiC).

References

Blevins, T., Joshi, M., & Zettlemoyer, L. (2021). FEWS: Large-scale, low-shot word sense disambiguation with the dictionary. In *Proceedings of EACL*.
Cassotti, P., Siciliani, L., de Gemmis, M., Semeraro, G., & Basile, P. (2023). Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of ACL*.
Martelli, F., Kalach, N., Tola, G., & Navigli, R. (2021). SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of SemEval*.
Pilehvar, M. T., & Camacho-Collados, J. (2019). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL*.