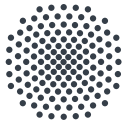


University of Stuttgart
Germany

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

April 3, 2026

Introduction

- **Core Challenge**: Understanding word meaning in context is essential for NLP (MT, IR, text processing).
- **Two Primary Tasks**:
  - **Word-in-Context (WiC)**: Binary classification: Does a word have the same meaning in two different sentences?
  - **Word Sense Disambiguation (WSD)**: Multi-class classification: Assigning a specific sense label from a predefined inventory.
- **The Gap**: Despite conceptual overlap, the benefits of **joint training** and **shared representations** between these tasks remain largely unexplored.
- **Aim**: Systematically investigate the relationship between WiC and WSD using a unified sentence transformer architecture.

Task Definitions

Word-in-Context (WiC)

"She can **book** *a flight"* vs *"He read a* **book***"* $\rightarrow$ **Output: 0** (Different)

Word Sense Disambiguation (WSD)

"She can **book** *a flight"* + Sense Inventory: (1) A written work; (2) To arrange in advance $\rightarrow$ **Output: 2**

- Both tasks are treated as a **sequence-comparison problem**:
- WiC compares context vs. context.
- WSD compares context vs. gloss definition.

Hypotheses

- **Hypothesis 1 (Joint Training)**:
  - Training on both WiC and WSD data simultaneously will lead to improved or non-detrimental performance compared to single-task training.
- **Hypothesis 2 (Cross-Task Generalization)**:
  - Models trained exclusively on one task can generalize effectively to the other, indicating shared semantic representations.

Datasets

Note: All experiments strictly operate on the* **English subsets *of these datasets.*

Word-in-Context (WiC)

Low-Resource

- **Pil-WiC**: ≈ 5.4k Pairs (Pilehvar & Camacho-Collados, 2019)
- **MCL-WiC**: ≈ 8k Pairs (Martelli et al., 2021)
- **XL-Lexeme**:≈ 13.4k Pairs (Cassotti et al., 2023)

Word Sense Disamb. (WSD)

High-Resource

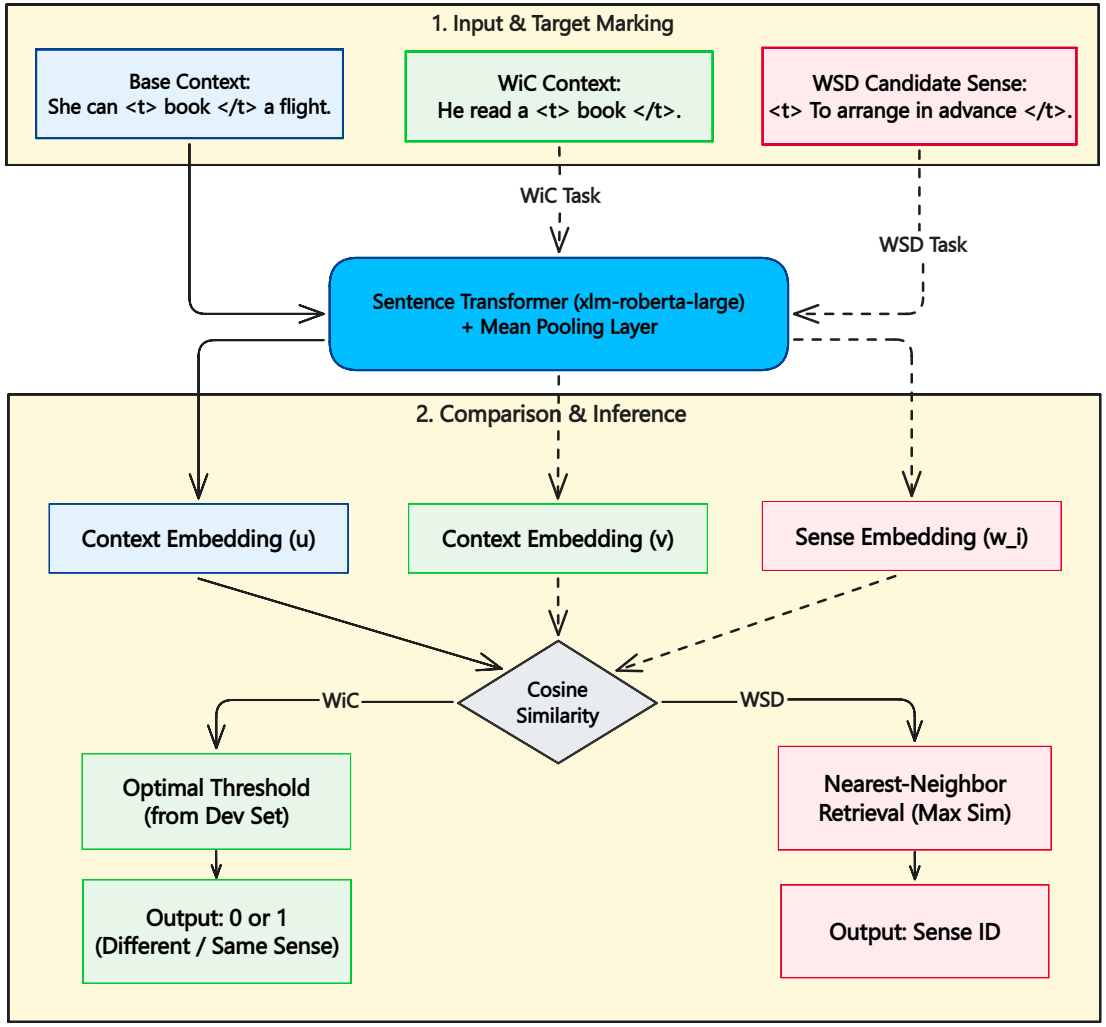
- **FEWS (Train)**: ≈ 200k Pairs (Blevins et al., 2021)
- **FEWS (Test)**: Zero-Shot 5K instances
- **Downsampled Subsets**: Reduced to WiC sizes (5k, 8k, 13k) for joint training.

Key Dynamic

Evaluating the synergy and interference between fundamentally different resource scales.

Methodology & Architecture

- **Model**: SBERT based on `xlm-roberta-large`.
- **Target Marking**: `<t>...</t>` around target words and glosses.
- **Inference**: Cosine similarity threshold for WiC; Nearest-Neighbor retrieval for WSD.



Key Results: WiC & WSD Accuracies

	Pil-WiC		MCL-WiC		XL-Lexeme	
Configuration	**WiC**	**WSD**	**WiC**	**WSD**	**WiC**	**WSD**
Random Base	.492	.262	.479	.262	.508	.262
Untrained Base	.529	.309	.653	.309	.585	.309
Pure WiC	.681	.475	.889	.477	.790	.472
Joint (Low-Res WSD)	.695	.552	.890	.575	**.792**	.609
Joint (High-Res WSD)	**.714**	.665	**.891**	.663	.782	.669
Pure WSD (200k)	.693	**.670**	.860	**.670**	.754	**.670**

Table 1: Comparison of training configurations.

Joint Training Excels in Low-Resource Settings (H1)

- **WiC Performance**:
  - Joint training with small WSD subsets **consistently improves** WiC accuracy.
  - Pil-WiC accuracy rose from .681 to .714 (highest) when using joint training.
- **WSD Performance**:
  - Joint training helps **only in low-resource settings**.
  - In the large-data regime (200k), auxiliary WiC data becomes a **distraction**, slightly degrading WSD scores.
- **Verdict**: H1 is **partially supported**. Benefits depend on the primary task's data saturation point.

Asymmetric Transfer: WSD Generalizes Strongly (H2)

- **WSD** $\rightarrow$ **WiC**: Extremely strong transfer.
  - A model trained *only* on FEWS 200k outperforms the in-domain Pil-WiC model (.693 vs .681).
- **WiC** $\rightarrow$ **WSD**: Successful but less powerful.
  - WiC-only models score $\sim$.47-.48, significantly beating random (.26) and untrained (.31) baselines.
- **Asymmetry Mechanism**: WSD provides **richer semantic supervision** because it anchors words to explicit external definitions, whereas WiC only matches relative contexts.

Conclusion

- **Shared Representations**: WiC and WSD are deeply related; models learn context-sensitive lexical semantics applicable to both.
- **Practical Guidance**:
  - **Low-resource**: Always use joint training; cross-task data is a high-value supplement.
  - **High-resource**: Auxiliary tasks may act as noise once the primary task is saturated.
- **Key Takeaway**: The explicit "sense-anchoring" in WSD creates more robust and generalizable representations than the binary matching in WiC.

Future Work

- Test alternative architectures (e.g., cross-encoders).
- Investigate prompt optimization in Large Language Models (LLMs) using mixed WiC/WSD examples.
- Broaden dataset coverage to multilingual and diverse domain sources.

References I

Blevins, T., Joshi, M., & Zettlemoyer, L. (2021). FEWS: Large-scale, low-shot word sense disambiguation with the dictionary. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume.* Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.eacl-main.36/` doi: 10.18653/v1/2021.eacl-main.36

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 the 61th annual meeting of the association for computational linguistics.* Association for Computational Linguistics.

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 the 15th international workshop on semantic evaluation (semeval-2021).* Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.semeval-1.3` doi: 10.18653/v1/2021.semeval-1.3

Pilehvar, M. T., & Camacho-Collados, J. (2019). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers).* Association for Computational Linguistics. doi: 10.18653/v1/N19-1128