# UniPSDA: Unsupervised Pseudo Semantic Data Augmentation for Zero-Shot Cross-Lingual Natural Language Understanding

Dongyang Li[1,2], Taolin Zhang[2], Jiali Deng[1], Longtao Huang[2], Chengyu Wang[2⋆], Xiaofeng He[1⋆], Hui Xue[2]

[1]School of Computer Science and Technology, East China Normal University [2] Alibaba Group

## Key Contributions

- In this work, we propose an **Un**supervi**s**ed **P**seudo **S**emantic **D**ata **A**ugmentation (UniPSDA) mechanism for cross-lingual natural language understanding to enrich the training data without human interventions.
- Domino Unsupervised Cluster groups languages into a hierarchical structure organized by language families to provide high-quality multilingual representations. Pseudo Semantic Data Augmentation employs the learned multilingual internal representations to address the semantic deficiencies of the training samples.
- Extensive experiments demonstrate that our model consistently improves the performance on general zero-shot cross-lingual natural language understanding tasks, including sequence classification, information extraction, and question answering.

## Introduction

**Background.** Cross-lingual representation learning transfers knowledge from resource-rich data to resource-scarce ones to improve the semantic understanding abilities of different languages. However, previous works rely on shallow unsupervised data generated by token surface matching, regardless of the global context-aware semantics of the surrounding text tokens.

**Our Work.** In this paper, we propose an **Un**supervi**s**ed **P**seudo **S**emantic **D**ata **A**ugmentation (UniPSDA) mechanism for cross-lingual natural language understanding to enrich the training data without human interventions. Specifically, to retrieve the tokens with similar meanings for the semantic data augmentation across different languages, we propose a sequential clustering process in 3 stages: within a single language, across multiple languages of a language family, and across languages from multiple language families. Meanwhile, considering the multi-lingual knowledge infusion with context-aware semantics while alleviating computation burden, we directly replace the key constituents of the sentences with the above-learned multi-lingual family knowledge, viewed as pseudo-semantic. The infusion process is further optimized via three de-biasing techniques without introducing any neural parameters.

## UniPSDA: The Proposed Model




Figure 1:Model Overview of UniPSDA.

## UniPSDA: The Proposed Model

**Domino Unsupervised Cluster:** To provide high-quality multilingual representations for performing the subsequent deep unsupervised data augmentation, we group languages into a hierarchical structure organized by language families. We perform the clustering process via the domino chain process to collect semantically similar words across different languages by comparing the embeddings themselves, a method we name Domino Unsupervised Cluster. Specifically, the domino cluster is a chain-rule process comprised of three different sequential stages: the single-language stage, the language family stage, and the multi-language stage.

**Pseudo Semantic Data Augmentation:** Considering that previous data augmentation methods focus on the surface of naive training samples, we employ the learned multilingual internal representations to address the semantic deficiencies of the training samples. Specifically, the domino clustering-enhanced ultimate multilingual representations directly replace the important positions' hidden states in training samples, as recognized by the <subject, verb, object> (SVO) structure. The potential incompatibility phenomena of inserting clustering multilingual representations may result in biased parameter learning. To further alleviate the misalignment between the replaced embeddings space and the context output space of PLMs, we introduce three de-biasing optimal transport affinity regularization techniques to make the learning process faster and more stable.

## Experiments

**Key Results.** To evaluate the effectiveness of UniPSDA, we construct extensive experiments, including sequence classification, information extraction, and question answering. Due to the space limitation, we only list the result of the text classification task here. We observe that: (1) Our approach outperforms strong baselines and nearly reaches state-of-the-art performance. (2) The performance of our method is significantly improved by leveraging the Domino Cluster to select appropriate candidates and injecting pseudo semantic knowledge into critical components of the sentences. We achieve an average accuracy of 79.3, with a particularly notable improvement for French (74.5 → 84.4) compared to the method proposed by CoSDA-ML.

| Model | en | de | zh | es | fr | it | ja | ru | Average |
|---|---|---|---|---|---|---|---|---|---|
| MLDoc | 87.2 | 71.7 | 73.5 | 65.3 | 70.2 | 65.1 | 69.8 | 56.9 | $69.9_{(\pm 0.7)}$ |
| LASER | 86.5 | 86.0 | 70.4 | 71.3 | 73.9 | 65.6 | 58.5 | 63.4 | $72.0_{(\pm 0.5)}$ |
| m-BERT | 92.1 | 74.3 | 72.5 | 67.0 | 70.5 | 61.8 | 69.7 | 61.5 | $71.2_{(\pm 0.3)}$ |
| XLM-R | 90.7 | 78.5 | 70.3 | 66.4 | 67.8 | 63.9 | 64 | 64.0 | $70.7_{(\pm 0.6)}$ |
| ZSIW | 91.3 | 82.8 | 79.6 | 71.7 | 78.1 | 67.0 | 68.5 | 64.3 | $75.4_{(\pm 0.5)}$ |
| DAP | 94.1 | 86.7 | 81.7 | 76.2 | 84.3 | 67.6 | 73.9 | 66.7 | $78.9_{(\pm 0.2)}$ |
| $SOGO_{cos}$ | 93.2 | 87.0 | 81.8 | 76.2 | 82.5 | 68.7 | 73.7 | 63.9 | $78.4_{(\pm 0.1)}$ |
| X-STA | 93.8 | 86.4 | 81.7 | 77.2 | 84.3 | 68.4 | 73.4 | 64.8 | $78.8_{(\pm 0.2)}$ |
| CoSDA-ML | 92.4 | 79.1 | 72.7 | 69.9 | 74.5 | 64.3 | 70.6 | **66.9** | $73.8_{(\pm 0.6)}$ |
| UniPSDA | **94.5** | **87.1** | **82.3** | **77.4** | **84.4** | **69.4** | **74.0** | 65.5 | $\mathbf{79.3}_{(\pm 0.2)}$ |

Table 1:General results of text classification in terms of accuracy (%) on the MLDoc dataset.

## Conclusion

In this work, we introduce UniPSDA, an unsupervised data augmentation mechanism that leverages semantic embeddings to enrich cross-lingual natural language understanding (NLU) tasks with diverse linguistic information. The Domino Unsupervised Cluster module identifies semantically similar cross-lingual content, while the Pseudo Semantic Data Augmentation module injects context-aware semantics into the training corpus. Furthermore, affinity regularization serves to minimize the representational gap between original and augmented sentences. Through extensive experimentation, our methods demonstrate superior performance relative to other strong baselines, underscoring their effectiveness in enhancing cross-lingual NLU.