Impact of CLIP Multimodal Embeddings on Zero-Shot Cross-Lingual SLU Robustness in CoNLL-2002 Evaluation

Assignee Research

June 19, 2026

Abstract

Spoken language understanding (SLU) typically includes two subtasks: intent detection and slot filling. Currently, it has achieved great success in high-resource languages, but it still remains challenging in low-resource languages due to the scarcity of labeled training data. Hence, there is a growing interest in zero-shot cross-lingual SLU. Despite of the success of existing zero-shot cross-lingual SLU models, most of them neglect to achieve the mutual guidance between intent and slots. To address this issue, we propose an Intra-Inter Knowledge Distillation framework for zero-shot cross-ling

1 Introduction

This paper examines: I^{2}KD-SLU: An Intra-Inter Knowledge Distillation Framework for Zero-Shot Cross-Lingual Spoken Language Understanding. Research question: What is the impact of incorporating multimodal embeddings from CLIP on the robustness of zero-shot cross-lingual SLU models like I^{2}KD-SLU when evaluated on the CoNLL-2002 dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

10 papers retrieved. 15 claims extracted; 11 independently verified. Quality review score: 7.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The MultiATIS++ dataset consists of 18 intents and 84 slots for each language.	✓	0.18
The MultiATIS++ dataset includes human-translated data for six languages: Spanish, German, Chinese, Japanese, Portuguese	✓	0.23
The mBERT model used in the experiment has N = 12 attention heads and M = 12 transformer blocks.	✓	0.19
Hyperparameters are selected by searching a combination of batch size and learning rate within the candidate sets {4, 8,	✓	0.29
The parameters α, β, λ, γ are set to 0.9, 0.1, 0.7, and 0.3 in Eq.10, respectively.	✓	0.16
The Adam optimizer is used with β1 = 0.9 and β2 = 0.98.	×	0.14
The learning rate decreases proportionally to the inverse square root of the step number after the warm-up phase.	✓	0.25
The model that achieves the highest overall accuracy on the dev set is selected and evaluated on the test set.	✓	0.18
All experiments are conducted on an Nvidia Tesla-V100 GPU.	✓	0.19
I2KD-SLU achieves the new state-of-the-art performance on the MultiATIS++ dataset, obtaining a significant improvement o	✓	0.31
The I2KD-SLU model achieves an average intent accuracy of 92.91% on the MultiATIS++ dataset.	×	0.14
The I2KD-SLU model achieves an average slot F1 score of 87.32% on the MultiATIS++ dataset.	×	0.09
The I2KD-SLU model achieves an overall accuracy of 87.32% on the MultiATIS++ dataset.	×	0.10
The I2KD-SLU model correctly identifies the intent 'atisgroundservice' and the slot 'B-cityname I-cityname' for the Engl	✓	0.23
The I2KD-SLU model correctly identifies the intent 'atisgroundservice' and the slot 'B-cityname I-cityname' for the Germ	✓	0.29

References

- http://arxiv.org/abs/2201.05729v3
- http://arxiv.org/abs/2310.02594v1
- http://arxiv.org/abs/2507.07104v2