Cross-Lingual NLU: Mitigating Language-Specific Impact in Embeddings Leveraging Adversarial Learning

Saedeh Tahery*, Sahar Kianian, Saeed Farzi
K. N. Toosi University of Technology, Shahid Rajaee Teacher Training University, Fondazione Bruno Kessler
Seyed Khandan, Tehran, Iran, Lavizan, Tehran, Iran, Trento, Italy
saedeh.tahery@email.kntu.ac.ir, sahar.kianian@sru.ac.ir, sfarzi@fbk.eu

Abstract

Low-resource languages and computational expenses pose significant challenges in the domain of large language models (LLMs). Currently, researchers are actively involved in various efforts to tackle these challenges. Cross-lingual natural language processing (NLP) remains one of the most promising strategies to address these issues. In this paper, we introduce a novel approach that utilizes adversarial techniques to mitigate the impact of language-specific information in contextual embeddings generated by large multilingual language models, with potential applications in cross-lingual tasks. The study encompasses five different languages, including both Latin and non-Latin ones, in the context of two fundamental tasks in natural language understanding: intent detection and slot filling. The results primarily show that our current approach excels in zero-shot scenarios for Latin languages like Spanish. However, it encounters limitations when applied to languages distant from English, such as Thai and Persian. This highlights that while our approach effectively reduces the effect of language-specific information on the core meaning, it performs better for Latin languages that share language-specific nuances with English, as certain characteristics persist in the overall meaning within embeddings.

Keywords: natural language understanding, adversarial learning, cross-lingual transfer learning

1. Introduction

Task completion systems, also known as Task-oriented dialogue systems (Louvan & Magnini, 2020; Zheng Zhang, Takanobu, Zhu, Huang, & Zhu, 2020), are thoroughly designed to assist users in achieving specific tasks or objectives through natural language interactions. These systems differ from typical chatbots, which engage in wide-ranging conversations (Huang, Zhu, & Gao, 2020) as they are tailored to various domains, addressing tasks such as making restaurant reservations, checking weather information, and providing customer support. Key components of these systems include recognizing user intentions and filling in relevant information slots, which are crucial for the system's understanding of human language. Natural language understanding (NLU) serves as the foundation for these systems, enabling machines to not only comprehend and respond to human language but also to enhance interactions between humans and AI. This includes chatbots, virtual assistants, as well as extending its impact to various areas such as information retrieval, sentiment analysis, and facilitating cross-cultural communication.

Despite important advancements in NLU models and techniques, such as the improved performance of large language models in intricate tasks like slot filling (Castellucci, Bellomaria, Favalli, & Romagnoli, 2019; Chen, Zhuo, & Wang, 2019; Firdaus, Ekbal, & Cambria, 2023; Ma, Ye, Yang, & Liu, 2022; Zhichang Zhang, Zhang, Chen, & Zhang, 2019), one of the major challenges in NLU is dealing with underrepresented languages, often referred to as 'rare' or 'low-resource' (Liu, Winata, Lin, Xu, & Fung, 2020; Razumovskaia, Glavaš, Majewska, Korhonen, & Vulić, 2021). These languages lack sufficient data, making it difficult to construct accurate language models.

We hold the belief that word and sentence embeddings generated by language models, such as BERT (Devlin, Chang, Lee, & Toutanova, 2019), BART (Lewis et al., 2019), and others (Sung et al., 2023), contain information that conveys meaning, alongside a complementary component that imparts language-specific features. With this premise in mind, our objective is to employ an adversarial method to extract contextual representations, hereby referred to as language-independent representations, while mitigating language-specific information without compromising the intended meaning. To achieve this goal, we explore the effect of cross-lingual transfer in NLU by introducing a model that comprises generative and discriminative sub-models in the form of a Generative Adversarial Network (GAN) (Dong et al., 2020; Goodfellow et al., 2014). The generative sub-model is responsible for providing the contextual representations, while the discriminative sub-model's role is to strip away the language-specific features from the given embeddings. The discriminative sub-model's primary function is to identify the language identity of the input utterance, while the generative sub-model endeavors to generate contextual language-independent representations from input utterances. These two sub-models engage in an adversarial interplay, each striving to outperform the other. As this adversarial process unfolds, language-independent representations are acquired that not only possess fewer language-specific features but also excel in NLU tasks.

Through experimental studies, we aim to investigate the effectiveness of mitigating language-specific features, offering a promising solution for addressing the intricate challenges within the realm of cross-lingual NLU. The results mainly show that while

* Work done while Saedeh Tahery was affiliated with University of Amsterdam (UvA).

LREC-COLING 2024, pages 4158–4163

our current approach performs well in zero-shot scenarios for Latin languages, such as Spanish, it falls short for distant languages from English, such as Thai. This opens a door for future work towards crosslinguality in NLU.

2. Related Work

Cross-lingual NLU has witnessed significant advancements, including the use of shared character embeddings (Lin, Yang, Stoyanov, & Ji, 2018; Yang, Salakhutdinov, & Cohen, 2017), multilingual pre-training (Burnyshev, Bout, Malykh, & Piontkovskaya, 2021; Ebrahimi et al., 2021; Upadhyay, Faruqui, Tür, Dilek, & Heck, 2018), cross-lingual embeddings (Plank & Agić, 2018), and shared encoders in the context of machine translation (Eriguchi, Johnson, Firat, Kazawa, & Macherey, 2018; Singla, Can, & Narayanan, 2018; Yu, Li, & Oguz, 2018). These developments hold the promise of enhancing multilingual communication; however, they also introduce challenges related to handling language-specific nuances and addressing resource limitations for low-resource languages.

Moreover, cross-lingual transfer learning (Fuad & Al-Yahya, 2022; Mrkšić et al., 2017) has proven beneficial for syntactic tasks such as part-of-speech tagging (McDonald, Petrov, & Hall, 2011; Zeman & Resnik, 2008) and dependency parsing (de Lhoneux, Bjerva, Augenstein, & Søgaard, 2018; Smith et al., 2018; Susanto & Lu, 2017; Taghizadeh & Faili, 2022). The challenges of accommodating language-specific variations and handling rare languages continue to be critical concerns in the pursuit of truly effective cross-lingual NLU.

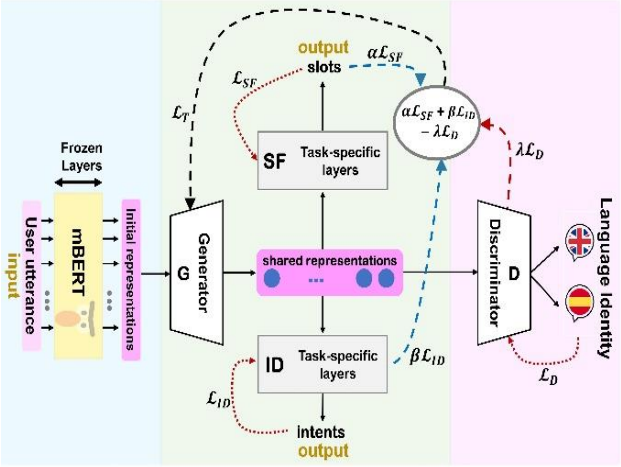
3. Methodology

We explore the effect of cross-lingual transfer in NLU by introducing a model rooted in adversarial learning through Generative Adversarial Networks (GANs). The core idea is that word and sentence embeddings from transformer-based language models like BERT and BART comprise both general meaning, semantic information and language-specific information. The goal of our model is to develop an adversarial method that retains the semantic while mitigating the language-specific details from these embeddings.

Figure 1 illustrates the core components of our model: Slot Filling (SF), Intent Detection (ID), Generator (G), and Discriminator (D), showcasing their interconnectedness. Initially, user utterance are projected using multilingual-BERT (mBERT), with its layers remaining frozen during training to maintain a lightweight model profile. Next, the initial representations go into generator G, where it works to create language-independent representations that are shared across different task-specific layers and in discriminator D.

In our proposed model, G is in charge of generating embedding vectors, while the component D's job helps remove language-specific features from these representations. Discriminator D's primary function is to determine the language identity of the input utterance, whereas G aims to create embedding vectors from input. These two components interact adversarially, each trying to outdo the other. As this competitive process unfolds, language-specific information in the embedding vectors gradually gets mitigated.

In more detail, during the adversarial learning process, the task-specific layers (SF and ID) actively participate in decoding shared representations, which, in turn, align with the generator's objective of creating contextual representations challenging the discriminator's language identification capabilities.



D: binary classifier (MLP), G: language-invariant generator (Bi-LSTM),
SF: sequence-tagger (Bi-LSTM + Softmax), ID: multi-class classifier (MLP)

Figure 1: An overview of the model. The generator and discriminator work in opposition, striving to craft language-independent representations for NLU tasks.

The discriminator D is updated through $\mathcal{L}_D$,

$$\mathcal{L}_D = \mathbb{E}_{x\sim X_{src}}[\log D(G(x)] + \mathbb{E}_{x\sim X_{aux}}[\log(1 - D(G(x))]],$$

where X_{src} and X_{aux} respectively denote the source and auxiliary data. The generator is updated using a total loss ($\mathcal{L}_T$), which is a combination of the loss originating from the tasks with a positive coefficient and the loss propagated from the discriminator ($\mathcal{L}_D$) with a negative coefficient,

$$\mathcal{L}_T = \alpha * \mathcal{L}_{SF} + \beta * \mathcal{L}_{ID} - \lambda * \mathcal{L}_D.$$

This balance between task-specific objectives and the adversarial component allows the generator to iteratively refine its language-independent representations.

Our proposed model utilizes a training approach that combines annotated data from a high-resource language (e.g., English) and unannotated data from a low-resource language (e.g., Spanish), treating the latter as auxiliary. This approach employs a zero-shot setting where the auxiliary data is used solely to determine its language identity, without utilizing any labels. A warm-up phase precedes the adversarial learning phase, during which the G, SF, and ID components are jointly trained using English validation data. This initial training equips the model components with knowledge from a high-resource language, enhancing their adaptability in the subsequent main phase, where a *k*-step training GAN is employed, sampling a half-batch of data from the

source language (X_{src}) and the other half from the auxiliary data (X_{aux}).

In essence, the approach aims to train models effectively by incorporating both high-resource and low-resource language data, providing a foundation in the high-resource language that benefits their adaptability to low-resource languages.

4. Experimental study

The main research question revolves around the capabilities of our approach when confronted with various linguistic contexts. To tackle this query comprehensively, we begin by elaborating on the dataset used, the configuration and parameters employed, and subsequently proceed to assess the performance of the model.

4.1 Data

The experiments are conducted using the Facebook-multilingual dataset (Schuster, Gupta, Shah, & Lewis, 2019). Table 1 presents statistics for this publicly available dataset. It comprises three languages (English, Spanish, and Thai) covering three domains: alarm, reminder, and weather. Moreover, it contains a total of 12 intent and 11 slot types.

To enhance the diversity of our exploration, we study the model's impact on Italian as another Latin language, as well as on Persian as a non-Latin language. The Italian and Persian datasets are constructed by means of stratified random sampling from English data, matching the statistics of Spanish and Thai, respectively. Subsequently, they are translated into the target language using the Google Translate machine[1] and aligned using Simalign (Sabet, Dufter, Yvon, & Schütze, 2020).

Data	Train	Validation	Test
English (EN)	30521	4181	8621
Spanish (ES), Italian (It)	3617	1983	3043
Thai (Th), Persian (FA)	2156	1235	1692

Table 1: Summary statistics for the Facebook-Multilingual Dataset and the translated versions.

4.2 Configuration and setting parameters

All the experiments are performed in a Google Colab[2] environment, utilizing the PyTorch framework[3]. Here, we provide an overview of our experimental setup, including the specific parameters used in our models.

Initial Embeddings: To generate initial embeddings, the 'BERT multilingual base model', accessible through the Hugging Face Transformers library[4], is employed. The maximum sequence length is set to 37, and the padding strategy is applied.

GAN Configuration: The generator is constructed as a bidirectional LSTM (Bi-LSTM) comprising two layers, each with a hidden-size of 256. The discriminator is designed as a three-layer Multi-Layer Perceptron (MLP), featuring 256, 128, and 1 neurons in its respective layers. To mitigate overfitting, a dropout rate of 0.1 is incorporated into the discriminator. Binary Cross Entropy is chosen as the loss function, and Adam is selected as the optimizer for the GAN models, with a learning rate of 3e-4.

Multi-task Configuration: A BiLSTM with the same architecture as the generator is implemented for SF. The learning rate is distinct at 0.01, distinguishing its role in the SF process. The discriminator is a three-layer MLP with 64, 64, and 12 neurons in its respective layers and with dropout of 0.1. Cross Entropy loss is utilized as the loss function for SF and ID, complemented by the stochastic gradient descent (SGD) optimizer.

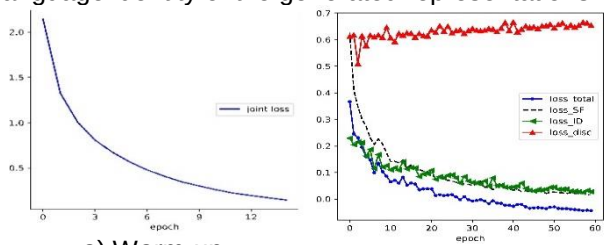
Additional Parameters: The batch size is set to 64, k to 3, α and β to 0.5, and λ to 0.1.

Furthermore, our approach boasts a significant feature in its lightweight design, with approximately 7 million trainable parameters in total, representing a substantial reduction compared to the 110 million parameters found in standard BERT models.

4.3 Results

Convergence: First of all, let's discuss the behavior of the model and the convergence of the GAN. Figure 2(a) shows a diminishing trend in the joint loss during the warm-up phase as the number of epochs increases. Figure 2(b) also depicts the loss values for all components during adversarial learning with respect to the number of epochs.

Since the generator is updated using the total loss, it naturally decreases with the passage of epochs. This reduction in loss is also mirrored in the loss values for SF and ID. The primary aim of GAN training in this context is to achieve an equilibrium where the generator can generate language-independent representations. However, it is important to note that a reduction in the discriminator's loss is not the main objective, as some fluctuations in the discriminator loss may be observed during training. Generally, the discriminator performs well in the initial stages but gradually encounters challenges in distinguishing the language identity of the generated representations.



a) Warm-up b) Adversarial learning

Figure 2: The training loss trends over epochs.

Zero-shot adaptation: Table 2 presents the results for the Spanish language in a zero-shot scenario in terms of accuracy and F1 score, highlighting the significant superiority of the proposed method over the baseline approaches. It reports the mean values of the micro-average across five runs and shows that the variance values approach zero, indicating the model's high stability.

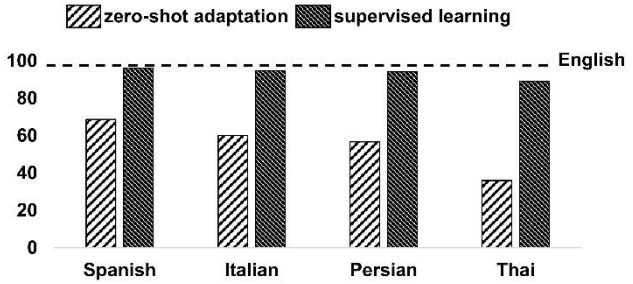
While the model performs well for Spanish, our investigation on Thai reveals less promising results than the Spanish experiments, with 40% accuracy for

[1] https://translate.google.com/
[2] https://colab.research.google.com/
[3] https://pytorch.org/
[4] https://huggingface.co/

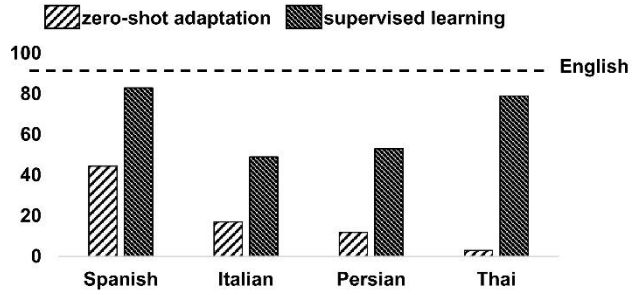
ID and less than 5% F1 for SF. This disparity can be attributed to the significant linguistic differences between Thai and English, as Thai is a non-Latin language morphologically distant from English. These findings underscore the potential challenges our model faces when dealing with non-Latin languages. Thus, there is room for improvement in handling non-Latin scripts in zero-shot scenarios.

Model	ID (Acc.)	SF (F1)
CL. XLU embd.◆	36.94	17.50
CL.CoVe◆	37.13	5.35
CL. muti CoVe◆	53.34	22.5
CL. multi CoVe w/ auto◆	53.89	19.25
Zero-shot SLU●	46.64	15.41
Ours	**68.74†**	**44.45†**
(*variance*)	(*7e-5*)	(*2e-3*)

Table 2: Results for Spanish on the Facebook multilingual dataset utilizing Spanish auxiliary data. †: significant results with p-value < 1-e5. ◆: (Schuster et al., 2019), ●: (Upadhyay et al., 2018).



a) Intent Detection



b) Slot Filling

Figure 3. Evaluation results on multiple languages. The dotted line represents the performance of supervised learning over English.

Cross-lingual exploration: Conducting supervised learning with labeled data for target languages, by adjusting the Discriminator coefficient to zero, enables us to define an upper bound for our zero-shot scenario. Figure 3 illustrates that zero-shot adaptation attains more than 70% of supervised learning in the ID task and more than 50% for the SF task on Spanish data. The results are generally better for ID than SF across all languages. The more complex the task, the harder it is to match the performance of supervised approaches. A similar trend is observed for Italian and Persian, both of which are generated through translation. However, the results decrease drastically for Thai, especially for SF, probably due to its nature, such as the absence of spaces between words, making it necessary to understand the context to determine word boundaries.

Ablation Study: We conduct experiments with and without the discriminator, training exclusively on labeled English data as shown in Table 3. This clearly provides insights into the impact of the discriminator in finding contextual language-independent representations.

		difference
ID (Acc.)	63.12	**5.62**
SF (F1)	31.00	**13.45**

Table 3: Ablation study of cross-lingual transfer by eliminating the discriminator component. (Target Language: Spanish)

Analysis of Utterance Length: To assess the impact of utterance length, we divide the test set for the target language into two subsets based on whether the utterances are below or above the average length. The findings presented in Table 4 reveal that, while the impact of length on ID is not particularly significant, the model works better for SF when handling shorter utterances. As we did not observe substantial differences in terms of utterance length for English and Thai, related results are excluded from Table 4.

Language	Task	*Below average*	*Above average*
Spanish (ES)	ID	69.71	71.02
	SF	46.51	38.14
Italian (It)	ID	51.92	56.75
	SF	19.24	17.06
Persian (FA)	ID	56.70	57.39
	SF	14.14	11.80

Table 4: Results across languages based on the average utterance length in terms of accuracy and F1 for ID and SF, respectively.

5. Conclusion

Cross-lingual NLU is challenging due to the time-intensive data collection, especially for low-resource languages. Leveraging adversarial learning, this paper presents a cross-lingual transfer approach, providing an effective means to mitigate language-specific features from contextual representations. Since contextual embeddings generated by language models encompass intertwined linguistic and semantic features, the model's performance in the discriminator role heavily relies on the quality of the initial representations they establish. This explains why the model excels for Spanish, a Latin-script language, while yielding negative results for low-quality embeddings such as Thai, a non-Latin language. It's worth noting that zero-shot results for English do not deteriorate when compared to supervised learning, demonstrating the effectiveness of language-independent representations. Considering our desire to assess the model's generalizability, we conduct an analysis on Italian and Persian, both of which are entirely generated by machines. However, we acknowledge that there are

some unavoidable errors imposed by automatic translation and alignment. The need for a model that works across a wide range of languages leads us to work on more robust discriminators in the future to better extract the language-specific features from embeddings less affected by their initial quality.

6. Bibliographical References

Burnyshev, P., Bout, A., Malykh, V., & Piontkovskaya, I. (2021). *InFoBERT: Zero-shot approach to natural language understanding using contextualized word embedding.* Paper presented at the Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021).

Castellucci, G., Bellomaria, V., Favalli, A., & Romagnoli, R. (2019). Multi-lingual intent detection and slot filling in a joint BERT-based model. *arXiv preprint arXiv:1907.02884*.

Chen, Q., Zhuo, Z., & Wang, W. (2019). Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

de Lhoneux, M., Bjerva, J., Augenstein, I., & Søgaard, A. (2018). Parameter sharing between dependency parsers for related languages. *arXiv preprint arXiv:1808.09055*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.* Paper presented at the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Dong, X., Zhu, Y., Zhang, Y., Fu, Z., Xu, D., Yang, S., & De Melo, G. (2020). *Leveraging adversarial training in self-learning for cross-lingual text classification.* Paper presented at the Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.

Ebrahimi, A., Mager, M., Oncevay, A., Chaudhary, V., Chiruzzo, L., Fan, A., . . . Meza-Ruiz, I. (2021). Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

Eriguchi, A., Johnson, M., Firat, O., Kazawa, H., & Macherey, W. (2018). Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Firdaus, M., Ekbal, A., & Cambria, E. (2023). Multitask learning for multilingual intent detection and slot filling in dialogue systems. *Information Fusion, 91*, 299-315.

Fuad, A., & Al-Yahya, M. (2022). Cross-Lingual Transfer Learning for Arabic Task-Oriented Dialogue Systems Using Multilingual Transformer Model mT5. *Mathematics, 10*(5), 746.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems, 27*, 2672–2680.

Huang, M., Zhu, X., & Gao, J. (2020). Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS), 38*(3), 1-32.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., . . . Zettlemoyer, L. (2019). *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.* Paper presented at the 58th Annual Meeting of the Association for Computational Linguistics.

Lin, Y., Yang, S., Stoyanov, V., & Ji, H. (2018). *A multi-lingual multi-task architecture for low-resource sequence labeling.* Paper presented at the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Liu, Z., Winata, G. I., Lin, Z., Xu, P., & Fung, P. (2020). *Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Louvan, S., & Magnini, B. (2020). *Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey.* Paper presented at the Proceedings of the 28th International Conference on Computational Linguistics.

Ma, Z., Ye, J., Yang, X., & Liu, J. (2022). *Hcld: A hierarchical framework for zero-shot cross-lingual dialogue system.* Paper presented at the Proceedings of the 29th International Conference on Computational Linguistics.

McDonald, R., Petrov, S., & Hall, K. (2011). *Multi-source transfer of delexicalized dependency parsers.* Paper presented at the Proceedings of the 2011 conference on empirical methods in natural language processing.

Mrkšić, N., Vulić, I., Séaghdha, D. Ó., Leviant, I., Reichart, R., Gašić, M., . . . Young, S. (2017). Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics, 5*, 309-324.

Plank, B., & Agić, Ž. (2018). Distant supervision from disparate sources for low-resource part-of-speech tagging. *arXiv preprint arXiv:1808.09733*.

Razumovskaia, E., Glavaš, G., Majewska, O., Korhonen, A., & Vulić, I. (2021). Crossing the Conversational Chasm: A Primer on Multilingual Task-Oriented Dialogue Systems. *arXiv preprint arXiv:2104.08570*.

Sabet, M. J., Dufter, P., Yvon, F., & Schütze, H. (2020). SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. *arXiv preprint arXiv:2004.08728*.

Schuster, S., Gupta, S., Shah, R., & Lewis, M. (2019). *Cross-lingual transfer learning for multilingual task oriented dialog.* Paper presented at the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota.

Singla, K., Can, D., & Narayanan, S. (2018). *A multi-task approach to learning multilingual representations.* Paper presented at the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).

Smith, A., Bohnet, B., de Lhoneux, M., Nivre, J., Shao, Y., & Stymne, S. (2018). 82 treebanks, 34 models: Universal dependency parsing with multi-treebank models. *arXiv preprint arXiv:1809.02237*.

Sung, M., Gung, J., Mansimov, E., Pappas, N., Shu, R., Romeo, S., . . . Castelli, V. (2023). *Pre-training intent-aware encoders for zero-and few-shot intent classification.* Paper presented at the 2023 Conference on Empirical Methods in Natural Language Processing.

Susanto, R. H., & Lu, W. (2017). *Neural architectures for multilingual semantic parsing.* Paper presented at the Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).

Taghizadeh, N., & Faili, H. (2022). Cross-lingual transfer learning for relation extraction using universal dependencies. *Computer Speech & Language, 71*, 101265.

Upadhyay, S., Faruqui, M., Tür, G., Dilek, H.-T., & Heck, L. (2018). *(Almost) zero-shot cross-lingual spoken language understanding.* Paper presented at the 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP).

Yang, Z., Salakhutdinov, R., & Cohen, W. W. (2017). Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*.

Yu, K., Li, H., & Oguz, B. (2018). *Multilingual seq2seq training with similarity loss for cross-lingual document classification.* Paper presented at the Proceedings of The Third Workshop on Representation Learning for NLP.

Zeman, D., & Resnik, P. (2008). *Cross-language parser adaptation between related languages.* Paper presented at the Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages.

Zhang, Z., Takanobu, R., Zhu, Q., Huang, M., & Zhu, X. (2020). Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences, 63*(10), 2011-2027.

Zhang, Z., Zhang, Z., Chen, H., & Zhang, Z. (2019). A joint learning framework with BERT for spoken language understanding. *IEEE Access, 7*, 168849-168858.