The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Unsupervised Interlingual Semantic Representations from Sentence Embeddings for Zero-Shot Cross-Lingual Transfer

Channy Hong,[1,2*] **Jaeyeon Lee,**[2] **Jung Kwon Lee**[2]
[1]Harvard University, [2]Superb AI Inc.
channyhong@college.harvard.edu, {jylee, jklee}@superb-ai.com

Abstract

As numerous modern NLP models demonstrate high-performance in various tasks when trained with resource-rich language data sets such as those of English, there has been a shift in attention to the idea of applying such learning to low-resource languages via zero-shot or few-shot cross-lingual transfer. While the most prominent efforts made previously on achieving this feat entails the use of parallel corpora for sentence alignment training, we seek to generalize further by assuming plausible scenarios in which such parallel data sets are unavailable. In this work, we present a novel architecture for training interlingual semantic representations on top of sentence embeddings in a completely unsupervised manner, and demonstrate its effectiveness in zero-shot cross-lingual transfer in natural language inference task. Furthermore, we showcase a method of leveraging this framework in a few-shot scenario, and finally analyze the distributional and permutational alignment across languages of these interlingual semantic representations.

Introduction

One of the greatest imbalances in the field of natural language processing (NLP) is the uneven availability of training data between high-resource and low-resource languages. For instance, while one can easily train a model for an English natural language inference (NLI) task using the widely available English NLI training data (Williams, Nangia, and Bowman 2018; Bowman et al. 2015), it is difficult to train an effective counterpart in languages such as German or Arabic, simply because there does not exist suitable labeled dataset in these languages that can be used for training. As a result, there exists substantial untapped potential for applications of NLP models in these low-resource languages, leading to increased efforts within the research community to explore cross-lingual transfer techniques (Conneau et al. 2018).

The logical first step in overcoming this imbalance is translating the training data from English to the language-of-interest, and then training a language-specific model on top of it (translate-train). On the other hand, one could also translate the test data into English at evaluation time and rely on an already trained English-based model (translate-test). However, both of these approaches would require a separate neural machine translation (NMT) system, which is not highly applicable to low-resource languages in that training these NMT systems typically require massive amounts of parallel corpora to begin with, although there has been efforts made recently to train NMT systems in an unsupervised manner (Artetxe et al. 2017; Yang et al. 2018).

Interlingual Semantic Representations. In the last approach discussed above (translate-test), English was used as the lingua franca for solving NLI tasks. Now, we take this idea of a 'medium-language' one step further. If we reduce the function of a language as merely a method of encoding semantics in the form of strings of text, we can then surmise that there exist language-agnostic semantic representations (whatever its form, or syntax, may be) that can serve as an interlingual medium. For instance, international auxiliary languages such as Esperanto, Ido, and Interlingua have been developed to serve this exact purpose.

While we are not interested in developing an actual language as such, we may be interested in training interlingual semantic representations (ISR) that can serve as a suitable medium of semantic embedding on top of which task-specific models can be trained. In fact, as the common practice within NLP is to train models on top of embeddings (Mikolov et al. 2013; Gong et al. 2018), there would not be a need for ISR to be reduced all the way back to the text level as ordinary language is usually presented to humans. The idea here is that the ISR for the English sentence "I am hungry" and French sentence "J'ai faim" would both map to proximal embedding spaces that represent the general concept of a first person singular subject expressing hunger.

With ISR as an effective embedding space for representing semantic information and assuming that an effective conversion — or translation — mechanism for encoding sentence embeddings of any language (high- and low-resource) into ISR exists, we can then easily leverage training data in whatever available language to develop models for under-resourced languages. For instance, we can train a NLI solver using the widely available English NLI training data converted to ISR. Then, this same classifier can be used

*Work done during an internship at Superb AI.

at evaluation time for any low-resource language likewise converted to ISR.

Through this architecture, we are able to stray away from relying on a certain language as being lingua franca, but rather come up with a unified interlingual framework. Furthermore, this architecture allows training to occur with training data from more than one language (since task-specific models are trained on top of semantic representations, not bound by the syntax of a specific language), when not enough training data exist within a single language.

Thus far, the most prominent effort made on tackling this problem involves producing shared embeddings via sentence alignment training between languages, which requires significant parallel corpora. However, given that low-resource languages are lacking in parallel corpora by definition, we are naturally interested in training an ISR-producing encoder in a fully unsupervised manner, for maximum scalability. In this paper, we propose a novel method of implementing unsupervised training of ISR on top of sentence embeddings, for pan-lingual transfer of learning done via high-resource language(s) to applications in low-resource language(s).

Main Contribution. We summarize our main contribution as three-fold:

- We describe our unsupervised method of leveraging easily-attainable monolingual corpora for training an ISR encoder that can convert fixed sentence embeddings from any language into ISR.
- We demonstrate the effectiveness of this framework in zero-shot cross-lingual transfer, showcase a few-shot scenario in which training data from a high-resource language can be used to augment training done on a low-resource language, and finally analyze distributional and permutational alignment of ISR converted from multiple languages.
- We provide the code implementations for training the ISR encoder as well as the task-specific classifier on top of ISR (github.com/ChannyHong/ISREncoder).

Related Works

Benchmark for Zero-Shot Transfer. Reflecting the recent trend of increased interest in cross-lingual studies, the multilingual XNLI benchmark (Conneau et al. 2018) has been proposed as a comprehensive suite of NLI evaluations for non-English languages, by providing high-quality translation (by certified translators) of the MNLI development (2500 examples) and test datasets (5000 examples) into 14 different languages. The MNLI dataset itself is shown to be a comprehensive benchmark for the full complexity of the English language (Williams, Nangia, and Bowman 2018), offering examples from multiple genre domains which makes it a substantially more difficult task than the previous Stanford NLI benchmark (Bowman et al. 2015). Thus, the XNLI benchmark in turn represents a challenging body of development and test datasets in 15 different languages.

Cross-Lingual Sentence Alignment. One idea for tackling this problem is by training an encoder that embeds sentences in any language to ISR in such a way that semantically equivalent sentences across languages would map to proximal embedding spaces via sentence alignment training.

For example, a multilingual sentence encoder is trained using large amounts of parallel corpora by aligning parallel sentences via minimization of distance between their languages (Artetxe and Schwenk 2018). Similarly, a multi-task dual-encoder model (Chidambaram et al. 2018) applied similar sentence alignment as the bridging translation task of their multi-step encoder training procedure.

While these efforts have shown substantial results, sentence alignment techniques are ultimately limited by the extent to which parallel corpora is available.

Distribution Alignment. On the other hand, UG-WGAN (Aghajanyan, Song, and Tiwary 2019) approached this problem by aligning the overall distribution shape of embedding spaces of each language by minimizing the Wasserstein distance amongst them. While this approach does not theoretically guarantee permutational alignment of sentences within distribution as the previous methods that leverage parallel corpora do, it nonetheless provides an interesting direction in which this problem can be addressed in an unsupervised manner.

In general, other ideas and methods of distribution alignment can be derived from generative adversarial networks (Goodfellow et al. 2014) such as WGAN (Arjovsky, Chintala, and Bottou 2017) and domain adaptation methods such as gradient reversal layer (Ganin and Lempitsky 2015).

Generative Adversarial Networks. In fact, the concept of distribution alignment is analogously used within the realm of computer vision. Notably, the idea of generative adversarial networks (GAN) have been successfully applied in image generation, wherein the generator is induced by adversarial loss minimization to generate outputs within a distribution that matches the target distribution, thus fooling the discriminator (Goodfellow et al. 2014).

Furthermore, CycleGAN (Zhu et al. 2017) demonstrates significant results in image-to-image translation only using unpaired sets of images as training data, by introducing reconstruction — or cycle consistency — loss to further minimize the possibility of permutational mismatch within matching distribution. Likewise, the applicability of adversarial and reconstruction loss within NLP is shown by He et al. (2016), whereby a NMT system is trained only using an unpaired set of monolingual corpora by including adversarial game and back-translation within training.

Finally, Choi et al. (2018) extends the dual approach taken by Zhu et al. (2017) to n-way, by further introducing mask vectors and classification loss, thus inducing the generator to generate images of correct target domain out of n (not binary) domains.

Proposed Method

We describe our proposed method of training an encoder enc that produces interlingual semantic representations (ISR)

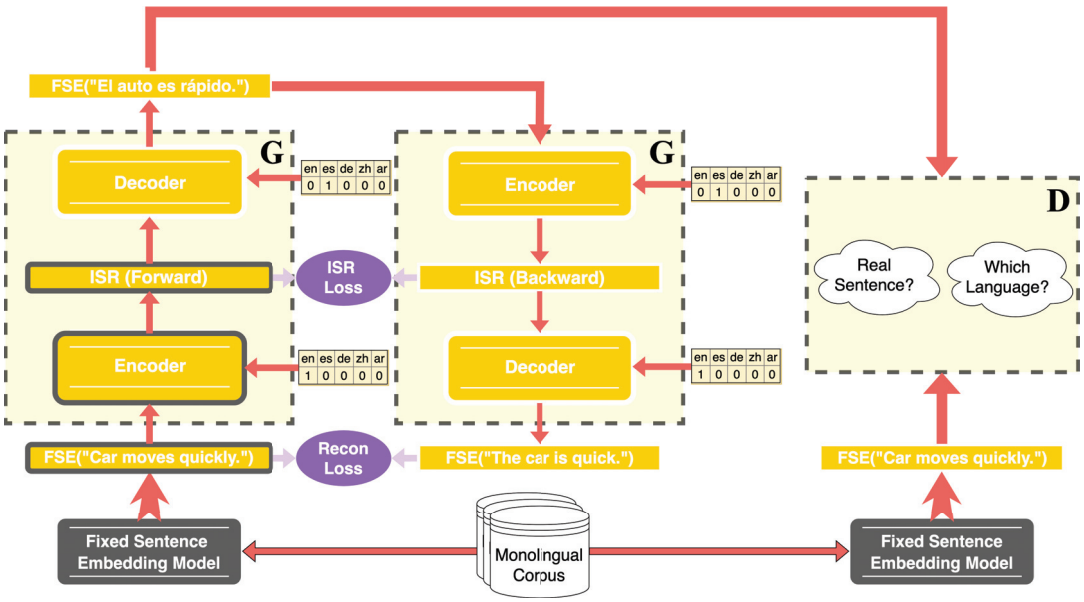


Figure 1: Overview of our proposed method of training an encoder that produces ISR, which map semantically equivalent sentences to proximal embeddings. Note that once the encoder is sufficiently trained for zero-shot transfer, task-specific models would be trained on top of the fixed forward-translation ISR.

that can be used for cross-lingual zero-shot transfer tasks. We outline the entire training procedure in Figure 1.

To achieve this, we train a single Generator G — consisting of aforementioned enc plus a decoder dec — that generates sentences in target domains given real sentences in their original domains. Within the same model, we jointly train a Discriminator D that performs the following two tasks: distinguishing whether a given sentence is real or generated, plus classifying the domain of a given sentence (both real and generated).

Initially, enc is fed a sentence embedding with its original domain label, producing the ISR of the sentence. Then, dec is fed the ISR along with a randomly selected target domain label (which includes the original domain), producing the generated ('translated') sentence embedding. Note that a single encoder and a single decoder are used for every language used in training. The D is fed both the original (real) sentence and the generated sentence embeddings, distinguishes whether each sentence embedding is real or generated, and classifies its domain (language).

Our framework extends the multi-domain translation architecture (Choi et al. 2018) within NLP by taking advantage of the fact that the nature of cycle consistency serendipitously stages a set of parallel sentence embeddings at each training step. Thus, we introduce our ISR consistency loss, which minimizes the distance between ISR of both forward-translation and backward-translation directions, thereby inducing permutational alignment by performing an ISR-variant of sentence alignment training.

For our enc to produce the ISR within this framework, we assign the following losses to our optimizer:

Adversarial Loss. For G to generate (translate) sentence embeddings that are indistinguishable from real sentence embeddings, we adopt the following adversarial losses:

$$\begin{aligned} L_{adv_D} = & -E_{x_{l_s}}[\log D_{src}(x_{l_s})] \\ & + E_{x_{l_s},l_s,l_t}[\log(D_{src}(G(x_{l_s},l_s,l_t)))] \\ L_{adv_G} = & -E_{x_{l_s},l_s,l_t}[\log(D_{src}(G(x_{l_s},l_s,l_t)))] \end{aligned} \quad (1)$$

where G generates sentence embeddings given a real sentence embedding (of a source language) and a target domain (target language), while D tries to distinguish between real and generated sentence embeddings (D_{src} refers to the probability distribution of D distinguishing an input sentence embedding as being real). In essence, G tries to generate sentence embeddings that D would distinguish as being real, while D tries to correctly distinguish between the two.

For our implementation, we update the above adversarial losses by adopting the widely-used Wasserstein GAN objective with gradient penalty (Gulrajani et al. 2017) to D's loss function:

$$\begin{aligned} L_{adv_D} = & -E_{x_{l_s}}[D_{src}(x_{l_s})] \\ & + E_{x_{l_s},l_s,l_t}[D_{src}(G(x_{l_s},l_s,l_t))] \\ & + \lambda_{gp}E_{\hat{x}}[(\|\nabla_{\hat{x}}D_{src}(\hat{x})\|_2 - 1)^2] \\ L_{adv_G} = & -E_{x_{l_s},l_s,l_t}[D_{src}(G(x_{l_s},l_s,l_t))] \end{aligned} \quad (2)$$

Domain Classification Loss. For G to generate sentence embeddings into the correct target domain (i.e. 'translate sentences into correct target language') we adopt the following domain classification loss:

$$L^r_{cls} = E_{x_{l_s}, l_s}[-\log D_{cls}(l_s|x_{l_s})]$$
$$L^g_{cls} = E_{x_{l_s}, l_s, l_t}[-\log D_{cls}(l_t|G(x_{l_s}, l_s, l_t))] \quad (3)$$

where D is trained to correctly classify the domain (i.e. language) of real sentence embeddings, while G is trained to generate (translate) sentence embeddings that would be correctly classified by such D.

Reconstruction Loss. While the adversarial losses motivate G to produce real-like sentence embeddings and the classification losses likewise motivate G to produce sentence embeddings in the target domain label, we are yet to ensure that the semantic information is preserved during the generation (translation) process. To address this issue, we introduce the following reconstruction (cycle consistency) loss:

$$L_{rec} = E_{x_{l_s}, l_s, l_t}[d(x_{l_s}, G(G(x_{l_s}, l_s, l_t), l_t, l_s))] \quad (4)$$

where $d(\cdot, \cdot)$ represents the distance measure between embeddings. Here, G is used twice, first to generate sentence embedding from original domain to target domain, then to reconstruct the original sentence embedding from target domain back to the original domain. G is trained to minimize the difference between the original and reconstructed sentence embeddings, thereby motivating the preservation of semantic information across the training flow.

ISR Consistency Loss. For our encoder to produce ISR that maps semantically parallel sentence embeddings from different domains into practical equivalence, we take advantage of the fact that we are essentially generating a set of parallel sentence embeddings during training, and that the semantic information is preserved throughout the training flow via the aforementioned reconstruction loss. In order for us to capture the language-agnosticity of our ISR, we introduce the following ISR consistency loss:

$$L_{isr} = E_{x_{l_s}, l_s, l_t}[d(enc(x_{l_s}, l_s), enc(G(x_{l_s}, l_s, l_t), l_t))] \quad (5)$$

where we are minimizing the distance between ISR of the forward-translation and that of the backward-translation directions.

Note that by ensuring that our enc maps semantically parallel sentence embeddings from two different domains (i.e. original and target domains) to practical equivalence, the random combinations of domain pair-wise matching (since target label is randomly selected during training) effectuates domain distribution matching of every language used in training. Furthermore, the granularity of distribution matching via difference minimization between ISR from semantically-parallel sentence pairs ensures permutational synchronization of ISR beyond the naive distribution matching, as discussed earlier.

Full Loss Functions. We combine the adversarial, domain classification, reconstruction, and ISR consistency losses for both D and G to define the full loss functions as the following:

$$L_D = L_{adv_D} + \lambda_{D_{cls}} L^r_{cls}$$
$$L_G = L_{adv_G} + \lambda_{G_{cls}} L^g_{cls} + \lambda_{rec} L_{rec} + \lambda_{isr} L_{isr} \quad (6)$$

where each loss is prefixed with its corresponding λ indicating their relative importance during training.

Experimental Setting

For our experiments, we evaluate the effectiveness of our framework in zero-shot transfer performances on non-English NLI tasks, with English as the designated high-resource language. For zero-shot evaluations, we intentionally chose two languages that are linguistically near (Spanish, German), and two that are linguistically distant (Chinese, Arabic). For our few-shot scenario, we chose Chinese based on availability of NLI training examples.

Datasets

XNLI. The primary metric we used for zero-shot transfer evaluations was the XNLI benchmark. The XNLI benchmark provides 2500 development examples and 5000 test examples that are translated by humans via the One Hour Translation service and further curated by Conneau et al. (2018). Furthermore, XNLI provides the machine-translated versions of 392,702 MNLI training examples into each 14 languages, translated by NMT systems that saw parallel corpora during their own trainings. We report our accuracies measured from performance against the XNLI test set.

CNLI. For our few-shot scenario, we use the Chinese NLI (CNLI) dataset[1] which includes 90,000 training examples, 10,000 development examples, and 10,000 test examples, all in Chinese. We again report our accuracies measured against the test set.

Monolingual Corpora. For our monolingual corpora, we used *WikiExtractor*[2] to extract the publicly available Wikipedia dumps for English, Spanish, German, Chinese, and Arabic. We then performed clean-up on the sentences as necessary. Subsequently, we randomly sampled 400,000 sentences for our use.

Zero-Shot Evaluation

We used the publicly available BERT-Base multilingual cased model (Devlin et al. 2019) – pre-trained unsupervisedly on masked language modeling and next sentence prediction tasks – as our meaningful representations of text. Note that BERT only uses monolingual corpora during its own training, which fits the unsupervised training scenario of our main model. An attribute not explicitly intended by its pre-training objectives, BERT has been shown to be surprisingly effective in zero-shot cross-lingual transfer by itself (Pires, Schlinger, and Garrette 2019), and we seek to build our framework on top of it and demonstrate stronger performance than this already solid baseline.

[1] http://www.cips-cl.org/static/CCL2018/call-evaluation.html
[2] https://github.com/attardi/wikiextractor

Model Type	en	es	de	zh	ar
Supervised (parallel corpora, separate model for each language)					
NMT + BERT (Devlin et al. 2019) (Translate-Train)	63.8	61.4	60.7	60	56.9
X-CBOW (Conneau et al. 2018)	64.5	60.7	61.0	58.8	57.5
Multi-Task en-es (Chidambaram et al. 2018)	70.2	65.2	-	-	-
Multi-Task en-de (Chidambaram et al. 2018)	71.5	-	65.0	-	-
Multi-Task en-zh (Chidambaram et al. 2018)	69.2	-	-	62.8	-
Unsupervised (monolingual corpora, single model)					
BERT (Devlin et al. 2019)	63.8	57.1	51.9	53.4	50.2
BERT + ISR (Ours)	**65.4**	**60.4**	**58.8**	**58.4**	**55.4**

Table 1: XNLI evaluations for two near (Spanish, German) and two distant (Chinese, Arabic) languages with English as the base language. Translate-Train entails a supervised NMT system that translates training examples into respective languages, while models by Conneau et al. (2018) and Chidambaram et al. (2018) are supervisedly pre-trained multilingual encoders with classifier on top, trained with English training examples only. Finally, the baseline model (BERT) and our main model are unsupervisedly pre-trained encoders frozen with task-specific classifier on top, again trained with English training examples only.

For our fixed sentence embeddings, we used the bert-as-service library (Xiao 2018) and specifically its default settings, which produces a 768-dimensional BERT sentence embedding (BSE) for each sentence string.

Translate-Train. We provide evaluation on the translate-train method as previously discussed. We train separate NLI solvers directly on top of BSE for each language, using machine-translated Spanish, German, Chinese, and Arabic training examples provided by Conneau et al. (2018). We then run evaluation of the models on XNLI test examples of each language. Note that this method utilizes parallel corpora during the training of the NMT systems and therefore cannot be directly compared against our unsupervised model.

Baseline (BERT). For our baseline, we train an English-based NLI solver directly on top of BSE, using the MNLI English training examples. We then run evaluation of the model on the XNLI English test examples (equivalent to the MNLI test examples) and on zero-shot transfers against XNLI Spanish, German, Chinese, and Arabic test examples.

Main Model (ISR). For our main model, we trained our encoder with easily attainable monolingual corpora in English, Spanish, German, Chinese, and Arabic, per training procedures described in the previous section. After sufficient training, we froze our encoder. Then, we trained a NLI solver on top of ISR, converted from MNLI English training examples by our trained encoder. We perform evaluations on zero-shot transfer against XNLI Spanish, German, Chinese, and Arabic test examples, each converted to ISR.

Few-Shot Evaluation

Mixed Language NLI Classifier. For our few-shot scenario, we devise a situation in which the language-of-interest lacks sufficient training examples on its own for training a NLI model, and thus requires augmentation from a high-resource language. We leverage the unique characteristic of our framework which trains task-specific models on top of semantic representations (unbounded by language-specific syntax), which allow training examples from mixture of languages to be utilized. We run evaluations on the CNLI test set using varying number of CNLI training examples for training, specifically 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, and 10000 examples.

We first train models directly on top of BSE of the Chinese CNLI training examples (similar to our baseline model from the zero-shot evaluations) then evaluate their performances against the CNLI test set. Then, we also train models on top of ISR converted by our frozen ISR encoder (same main model from the zero-shot evaluations), but this time augment each variation in the number of CNLI training example used with the full MNLI English training examples. By comparing the two methods, we are able to analyze how much high-resource language data augmentation improves performance by and at how many language-of-interest training examples do improvements from such data augmentation become trivial.

Implementation Details

For every method described above, the task-specific NLI solver consist of a single feed-forward layer followed by a three-way classification layer feeding as input the concatenation of the premise and the hypotheses embeddings. We now enumerate the specific implementation details of our main model.

Assuming the language-like nature of our ISR, we set the dimension of our ISR to be equal to that of BSE. When feeding an input sentence and its domain label to the encoder, we simply concatenated the domain label one-hot vector (i.e. [1,0,0,0,0] represents 'English' whilst [0,0,0,1,0] represents 'Chinese') to the 768-dimensional BSE. Likewise, when feeding ISR and destination domain label to the decoder, we again concatenated the domain label one-hot vector to the 768-dimensional ISR.

For our Generator, we used 2 upsampling layers, 4 feed-forward layers, and 2 downsampling layers for each encoder and decoder (16 layers total). The Discriminator of our model used 2 shared upsampling layers, 1 shared feed-forward layer, 2 shared downsampling layers, and 1 classi-

fication layer for each adversarial loss and domain classification loss (6 layers total). Our model is implemented in Tensorflow.

In order to help us decide when to halt training of our encoder, we performed mid-training evaluations on Discriminator's sentence classification task and also attached an English-based baseline NLI solver atop sentence embeddings generated into English (as their target domain). We stopped training of our encoder when the generator seemed to be reasonably functional in generating sentences of correct target domain (classification task accuracy) without losing the semantics of the original sentence (English NLI solving accuracy).

We found that $\lambda_{G_{cls}}$ of 10 worked best for training ISR (while holding all other λ to 1). We adopt the L2 distance as our distance measure between embeddings. Training took about 60 hours on a single Tesla T4 GPU.

Results and Discussion

ISR for Zero-Shot Evaluation

We report the results of the zero-shot evaluations of our main model in Table 1, alongside those of the baseline model (BERT), and evaluations on models utilizing parallel corpora including translate-train and pre-trained encoders from related works (Chidambaram et al. 2018; Conneau et al. 2018).

Our main model demonstrates significant improvements in zero-shot transfer in comparison to the baseline model, showing an average of slightly over 4% increase in accuracy from the baseline. Furthermore, our model shows comparable results to the translate-train method which utilizes parallel corpora, trailing by an average of around 1.5%.

It should be further noted that consistent with the results shown by Pires, Schlinger, and Garrette (2019), our baseline model (BERT) demonstrates already strong zero-shot transfer performance. Finally, it is interesting to note that the XNLI English evaluation yielded stronger results for our main model than it did for the baseline model (by 1.6%), indicating that ISR provide richer representation of semantics on top of which NLI solver can be trained, in comparison to the syntax-bound language-specific embeddings of the baseline model.

ISR for Few-Shot Evaluation

We also report the results of our few-shot scenario in Figure 2, where it is shown unsurprisingly that high-resource language data augmentation leads to significant performance boost as shown by the gap between CNLI only and CNLI + MNLI. From 14.6% difference at 10 CNLI training examples used, the performance boost via data augmentation converges to around 1% at 10,000 CNLI training examples used. While more CNLI training examples used improve model performance in general, the gap holds at around 1% until the full 90,000 CNLI training examples (not shown in Figure 2); thus, we qualify the extent of the performance boost from leveraging our framework to around 10,000 training examples available in the language-of-interest.

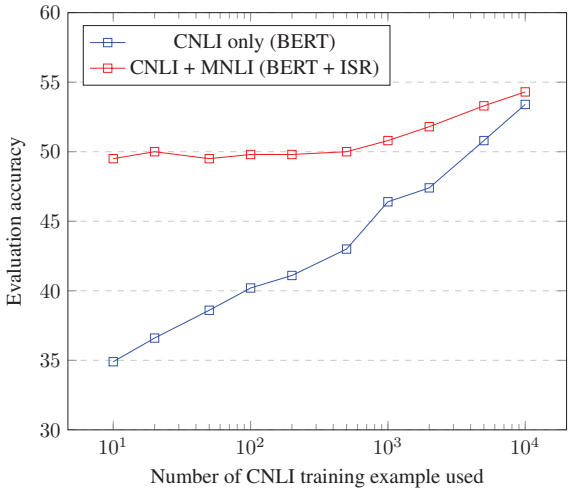


Figure 2: CNLI evaluations of our few-shot scenario, comparing the performance of a model trained on CNLI training examples only against a model augmented with high-resource language training examples via our proposed framework.

Ablation Studies

To evaluate the significance of the various losses defined by our main model, we run a series of ablation studies. We additionally trained 3 separate models ec absent in one of ISR consistency loss ($\lambda_{isr} = 0$), classification loss ($\lambda_{D,G_{cls}} = 0$), or reconstruction loss ($\lambda_{rec} = 0$); the rest of the experiment setting was held exactly the same as was for the training of our main model.

We report the results of the three additional models' zero-shot evaluations in Table 2. While $\lambda_{isr} = 0$ and $\lambda_{D,G_{cls}} = 0$ models show trailing but comparable results to the main model, $\lambda_{rec} = 0$ model fails to produce meaningful ISR for training of a NLI solver to take place.

Model Type	en	es	de	zh	ar
BSE (Baseline)	63.8	57.1	51.9	53.4	50.2
ISR ($\lambda_{isr} = 0$)	65.2	57.9	55.0	55.8	50.4
ISR ($\lambda_{D,G_{cls}} = 0$)	60.1	56.1	52.6	51.2	50.0
ISR ($\lambda_{rec} = 0$)	37.6	36.3	36.0	38.0	37.4
ISR	**65.4**	**60.4**	**58.8**	**58.4**	**55.4**

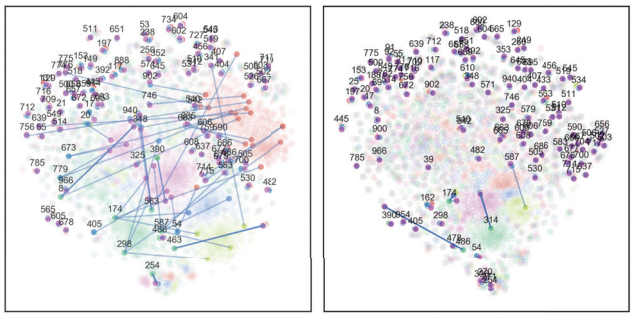
Table 2: XNLI zero-shot evaluations of models absent each in ISR consistency loss, classification loss, or reconstruction loss.

Now, in order to visually discern the role of each loss for the ISR, we sampled 1000 parallel sentences for each English, Spanish, German, Chinese, and Arabic from the XNLI development set, and plotted each encoder's ISR embeddings using t-SNE (Figure 3).

As expected, both the baseline (BERT) and the $\lambda_{D,G_{cls}} = 0$ model mapped the sentences to the locality of each language, while the $\lambda_{isr} = 0$ model and the main model displayed extensive distribution matching amongst all lan-

(a) BERT (Baseline) (b) ISR ($\lambda_{D,Gcls}$=0) (c) ISR (λ_{isr}=0) (d) ISR

Figure 3: t-SNE visualization of 1000 parallel sentences from the English, Spanish, German, Chinese, and Arabic XNLI development sets. Colors correspond to languages: English as teal, Spanish as blue, German as yellowish green, Chinese as purple, and Arabic as red.



(a) ISR (λ_{isr}=0) (b) ISR

Figure 4: t-SNE visualization of the sentences from Figure 3, with a subset of sentences highlighted and edges drawn between semantically parallel sentences.

guages. However, the $\lambda_{isr} = 0$ model seems to display more pronounced localities within matching distribution, suggesting poor permutational alignment between semantically parallel sentences.

To further illustrate this last observation, we select a subset of sentences and draw edges to their semantically parallel counterparts in order to visualize the permutational alignment of both the $\lambda_{isr} = 0$ model and the main model (Figure 4). The subset was selected based on sentences whose other-language counterparts were closer than any other sentence of that language. By noting the marked increase in the average distances between semantically parallel sentences, we argue that the ISR consistency loss contributes significantly to the permutational alignment of ISR.

Conclusion

In this work, we propose an entirely unsupervised method of training interlingual semantic representations on top of sentence embeddings for zero-shot cross-lingual transfer. Through this architecture, we demonstrate how low-resource languages can benefit from zero-shot or few-shot transfer from learning done via training examples in high-resource language(s). Furthermore, we present a series of analyses that outline the significance of each component of our training procedure. Although just a small step, we hope that our work opens the door in a novel, scalable direction in which this problem of lack of data in low-resource language can be addressed; and in that spirit, the code implementations we have used in this experiment and the instructions for running them have been made public (github.com/ChannyHong/ISREncoder).

References

Aghajanyan, A.; Song, X.; and Tiwary, S. 2019. Towards language agnostic universal representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4033–4041. Florence, Italy: Association for Computational Linguistics.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In Precup, D., and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 214–223. International Convention Centre, Sydney, Australia: PMLR.

Artetxe, M., and Schwenk, H. 2018. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond.

Artetxe, M.; Labaka, G.; Agirre, E.; and Cho, K. 2017. Unsupervised Neural Machine Translation. 1–12.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Chidambaram, M.; Yang, Y.; Cer, D.; Yuan, S.; Sung, Y.-H.; Strope, B.; and Kurzweil, R. 2018. Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model. 1–14.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks

for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8789–8797.

Conneau, A.; Rinott, R.; Lample, G.; Williams, A.; Bowman, S.; Schwenk, H.; and Stoyanov, V. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2475–2485. Brussels, Belgium: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. volume 2, 1180–1189.

Gong, C.; He, D.; Tan, X.; Qin, T.; Wang, L.; and Liu, T.-Y. 2018. Frage: Frequency-agnostic word representation. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc. 1334–1345.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*, 2672–2680. Curran Associates, Inc.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*, 5767–5777.

He, D.; Xia, Y.; Qin, T.; Wang, L.; Yu, N.; Liu, T.-y.; and Ma, W.-y. 2016. Dual Learning for Machine Translation. Number Nips, 1–9.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc. 3111–3119.

Pires, T.; Schlinger, E.; and Garrette, D. 2019. How multilingual is multilingual bert? *CoRR* abs/1906.01502.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. Association for Computational Linguistics.

Xiao, H. 2018. bert-as-service. https://github.com/hanxiao/bert-as-service.

Yang, Z.; Chen, W.; Wang, F.; and Xu, B. 2018. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 46–55. Melbourne, Australia: Association for Computational Linguistics.

Zhu, J. Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. volume 2017-Octob, 2242–2251.