Improving Zero-Shot Cross-Lingual Transfer Learning via Robust Training

Kuan-Hao Huang, Wasi Uddin Ahmad, Nanyun (Violet) Peng, and Kai-Wei Chang
University of California, Los Angeles

EMNLP 2021

Zero-Shot Cross-Lingual Transfer

- Learn a model f from training examples in source languages
- Apply the model f to testing examples in target languages
- Challenge: how to transfer knowledge across different languages
- Reduce the requirement of labeled data for low-resource languages

Training

Source language: English

This is a good restaurant.	Positive
The food is delicious.	Positive
The worst fried chicken in the world.	Negative
I would never come here again.	Negative

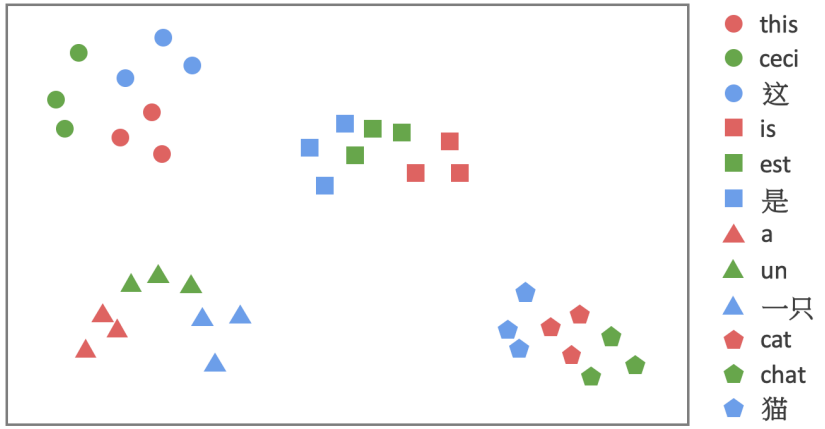
Testing

Target language: Chinese, French

这家餐厅很评价很好。	Positive
食物超级难吃。	Negative
J'aime ce resto.	Positive
Je ne viendrais plus jamais ici.	Negative

Why Zero-Shot Cross-Lingual Transfer is Possible

- Pre-trained multilingual language models learn aligned representations
  - Multilingual BERT [Devlin+ 2019], XLM-R [Conneau+ 2020]
- Words with similar meanings in different languages have similar representations [Cao+ 2020]
- This alignment makes zero-shot cross-lingual transfer become possible



The multilingual alignment is not perfect!

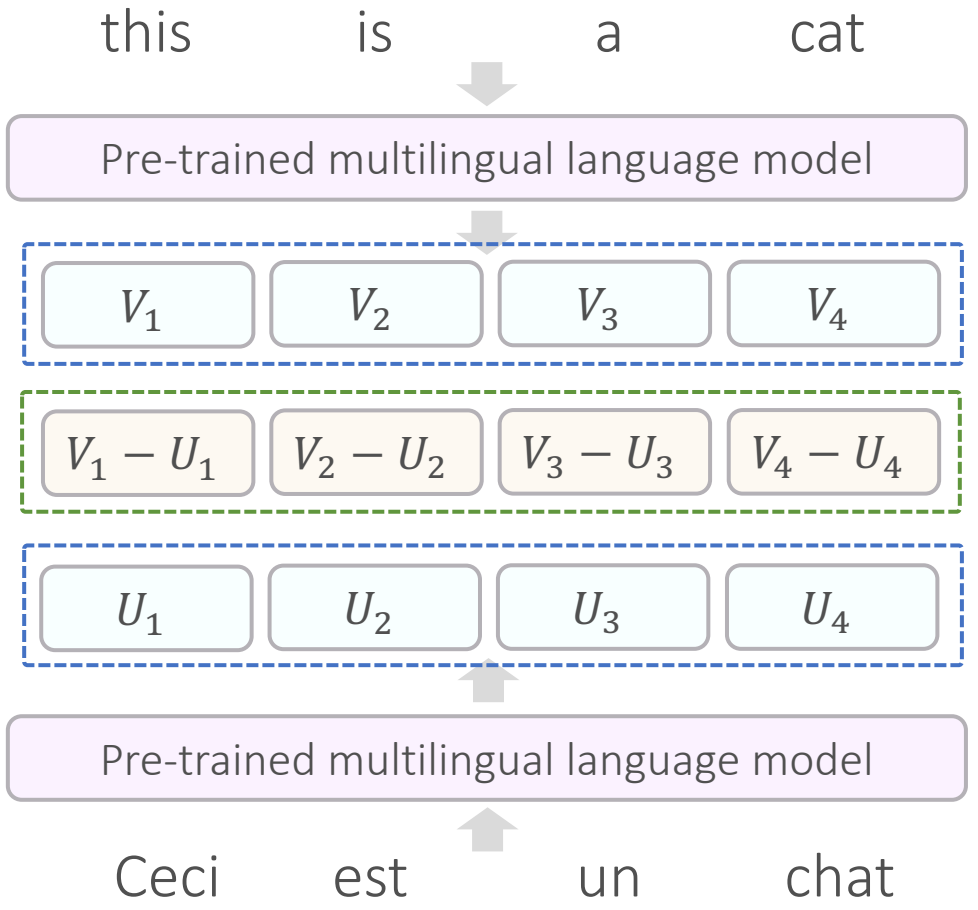
Learning Better Multilingual Alignment

- Prior studies learn a better multilingual alignment with additional resources
  - Bilingual dictionary [Cao+ 2020, Qin+ 2020, Liu+ 2020]
  - Parallel sentence pairs [Chi+ 2020, Feng+ 2020, Wei+ 2021]
- Better multilingual alignment leads to better transfer performance

Can we learn a better model without using additional resources?

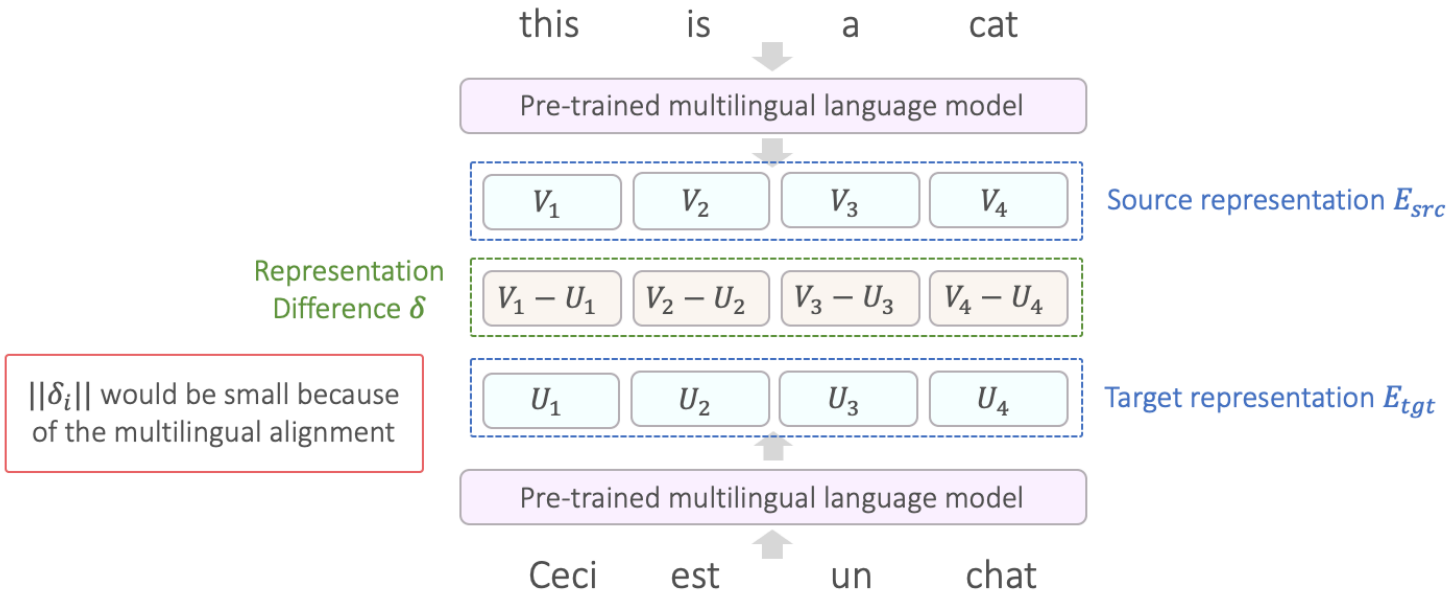
Robustness View of Zero-Shot Cross-Lingual Transfer

- Consider a pair of parallel sentences
  - "this is a cat" in English and "Ceci est un chat" in French



$||\delta_i||$ would be small because of the multilingual alignment

Robustness View of Zero-Shot Cross-Lingual Transfer



- If $f(E_{tgt}) = f(E_{src})$, transfer is successful
- Otherwise, we have

$$f(E_{tgt}) = f(E_{src} + \delta) \neq f(E_{src})$$

where $||\delta_i||$ is small

Compared to Definition of Adversarial Perturbations

- Definition of adversarial perturbations [Goodfellow+ 2014, Alzantot+ 2018]
  - Given an instance x and a model h, the adversarial perturbation is Δ such that

$$h(\tilde{x}) = h(x + \Delta) \neq h(x)$$

where $||\Delta||$ is small

- Failure case of zero-shot cross-lingual transfer
  - Given a source representation E_{src} and a target representations E_{tgt}

$$f(E_{tgt}) = f(E_{src} + \delta) \neq f(E_{src})$$

where $||\delta_i||$ is small

Robust training against adversarial perturbations
can help zero-shot cross-lingual transfer!

Robust Training — Adversarial Training

- Normal training

$$\min_{f} \sum_{(x,y)\in X_{src}} \mathcal{L}(f(x), y)$$

- Adversarial training [Ebrahimi+ 2018, Dong+ 2021, Zhou+ 2021]
  - Find the most effective perturbation

$$\min_{f} \sum_{(x,y)\in X_{src}} \max_{||\delta_i||\leq\epsilon} \mathcal{L}(f(x+\delta), y)$$

Robust Training — Randomized Smoothing (Random Perturbation)

- Normal training

$$\min_f \sum_{(x,y)\in X_{src}} \mathcal{L}(f(x), y)$$

- Randomized smoothing [Cohen+ 2019]
  - Consider the expectation case

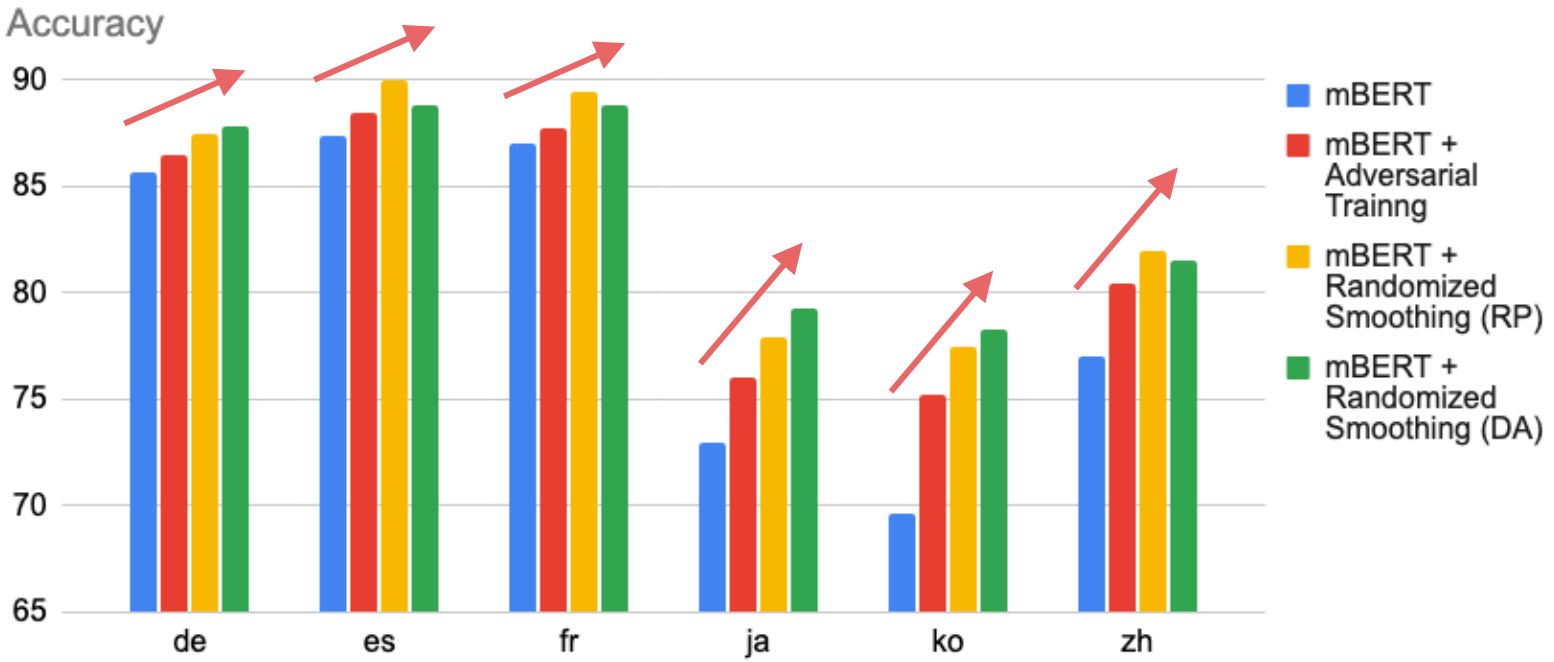
$$\min_f \sum_{(x,y)\in X_{src}} \mathbb{P}_\delta(\mathcal{L}(f(x+\delta), y))$$

Robust Training — Randomized Smoothing (Data Augmentation)

- Randomized smoothing by data augmentation with synonyms [Ye+ 2020]
- Every word is replaced by one of its synonym (including itself)
  - Original example
    - This restaurant looks beautiful and its food is great.
  - Augmented examples
    - The restaurant looks pretty and its food is great.
    - This restaurant looks beautiful and its food is good.
    - This restaurant looks pretty and its food is nice.
    - …
- Train a smooth model with augmented data

Experimental Results on PAWS-X

- (sentence1, sentence2) → are paraphrase or not
- Transfer from English to other languages

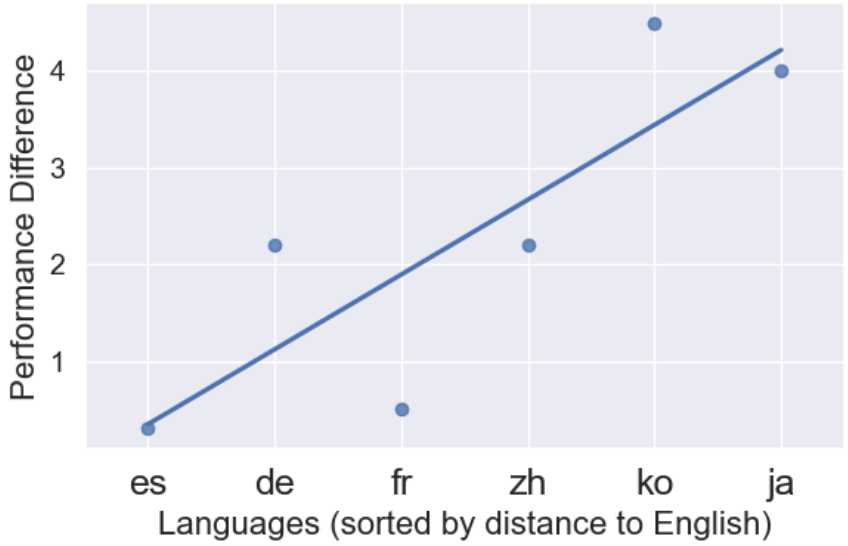


Robust training improves the performance of zero-shot cross-lingual transfer

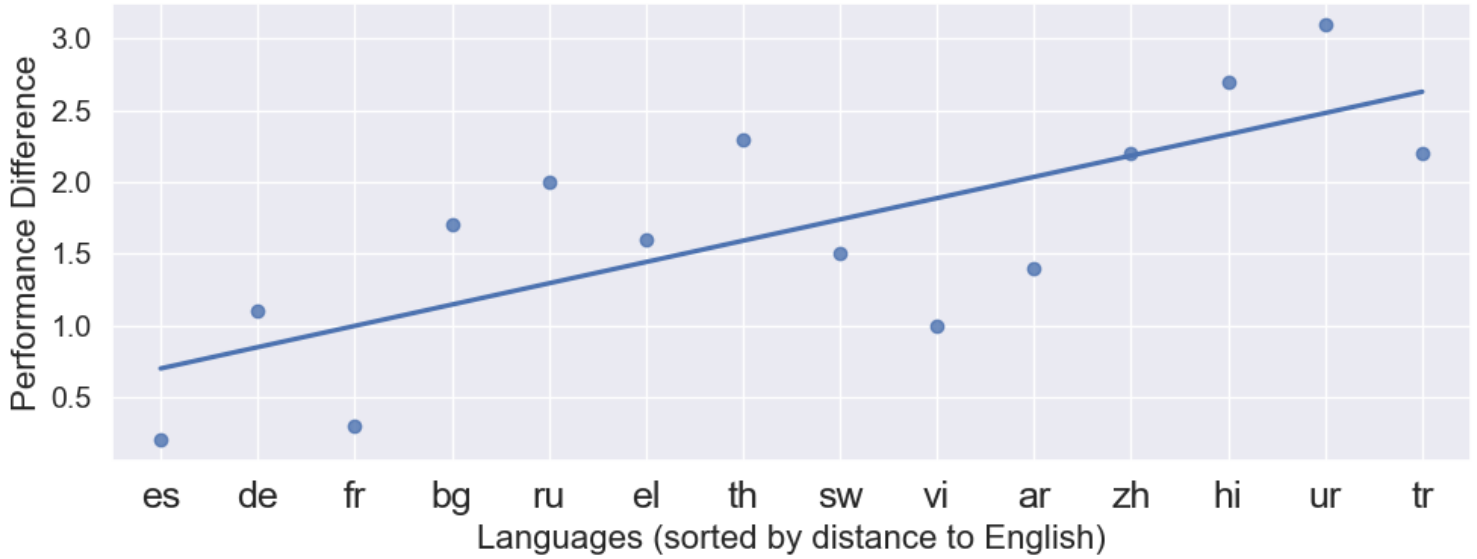
Randomized smoothing is more helpful than adversarial training

What Languages are Improved More

- Use lang2vec to calculate the distance between languages [Littell+ 2017]



PAWS-X

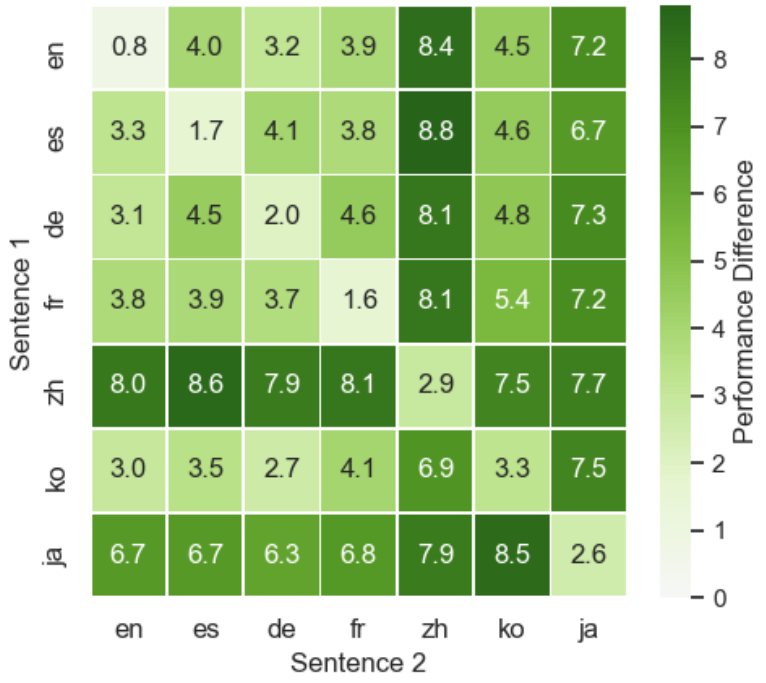


XNLI

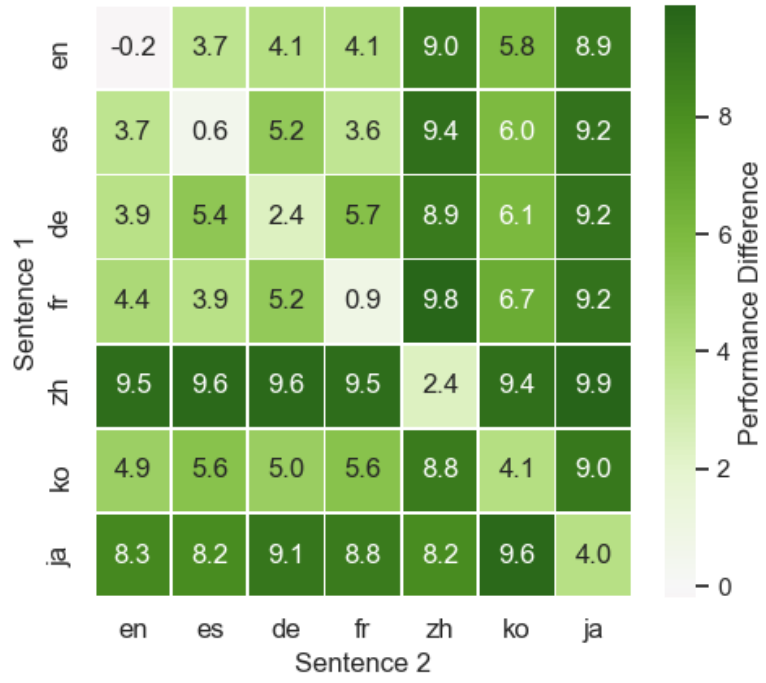
Languages with larger distances to source languages have larger performance improvement

Experimental Results on PAWS-X (Generalized Setting)

- (sentence1, sentence2) → are paraphrase or not
- Sentence1 and sentence2 are in different languages



Performance Gap
(Randomized Smoothing (RP) - Baseline)



Performance Gap
(Randomized Smoothing (DA) - Baseline)

Robust training leads to much more improvement when sentence1 and sentence2 in different languages

Interesting future work:
How to handle code-switching input sequences

Conclusion

- We draw connections between adversarial perturbations and the failure cases of zero-shot cross-lingual transfer
- We propose to use robust training methods to train models that can tolerate some noise in representations
- Experimental results demonstrate that robust training can improve the zero-shot cross-lingual transfer, especially in the generalized setting

Code is available at
https://github.com/uclanlp/Robust-XLT

Thank You!