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

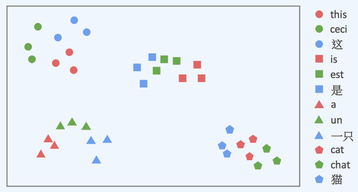
Kuan-Hao Huang, Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang
Computer Science Department, University of California, Los Angeles

Zero-Shot Cross-Lingual Transfer Learning

- 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

Why Zero-Shot Cross-Lingual Transfer is Possible

- Pre-trained multilingual language models learn aligned multilingual representations
  - E.g., multilingual BERT and XLM-R
- Those words with similar meanings in different languages have similar representations



- This multilingual alignment makes zero-shot cross-lingual transfer become possible

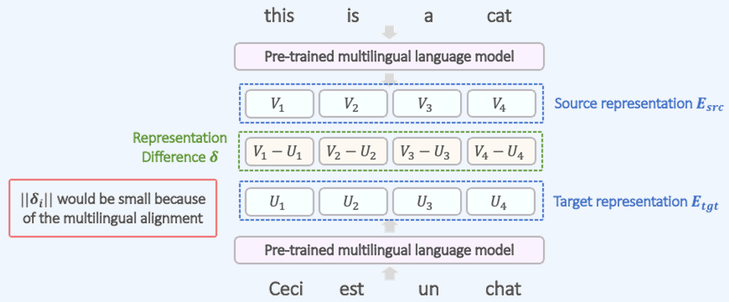
Learning Better Multilingual Alignment

- Prior studies learn a better multilingual alignment with additional resources
  - Bilingual dictionary
  - Parallel sentence pairs
- Better multilingual alignment leads to better transfer performance
- Can we learn a better model without using additional resources?

Connection to Adversarial Perturbations

Consider an English-French translation pair

- "this is a cat" and "Ceci est un chat"



When transferring from English to French

- 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

Definition of adversarial perturbations

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

where $||\Delta||$ is small

Robust Training

- Adversarial training

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

- Randomized smoothing (random perturbation)

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

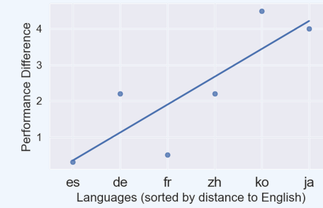
- Randomized smoothing (data augmentation)
  - The food in this restaurant is pretty good
  - The food in the restaurant is very good
  - The food in this restaurant is pretty great
  - The food in the restaurant is very nice

Zero-Shot Cross-Lingual Transfer on PAWS-X

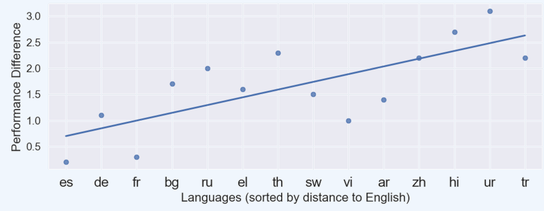
Model	en	de	es	fr	ja	ko	zh	avg.
mBERT*	94.0	85.7	87.4	87.0	73.0	69.6	77.0	82.0
mBERT (reproduce)	93.7	85.4	88.2	87.8	75.3	74.2	79.1	83.4
mBERT-ADV	93.7	86.5	88.5	87.8	76.1	75.3	80.4	84.0
mBERT-RS-RP	**94.5**	87.4	**90.0**	**89.5**	77.9	77.5	**82.0**	**85.5**
mBERT-RS-DA	93.5	**87.8**	88.8	88.8	**79.3**	**78.3**	81.5	85.4

What Languages are Improved More

- Use lang2vec to calculate the distance between languages



Languages with larger distances to source languages have larger performance improvement



Transfer for Generalized Setting

- (sent1, sent2) → are paraphrase or not
- Sent1 and sent2 are in different languages