Evaluating Cross-lingual Sub-word Embeddings

Ali Hakimi Parizi, Paul Cook
Department of Computer Science, University of New Brunswick
ahakimi@unb.ca, paul.cook@unb.ca

Cross-lingual Word Representations

Word embeddings model the distribution of words based on their surrounding words

Cross-lingual word embeddings create a shared space for embeddings in two languages

Enable knowledge to be transferred between languages

For tasks such as:

- POS Tagging
- Language Modeling
- Dependency Parsing

In the case of out-of-vocabulary (OOV) words, however, no information is available. This could be particularly problematic for low-resource languages

Goal ⟶ Can we overcome this probelm using sub-word embeddings?

Methodology

A mapping based approach;

- Find a mapping between source and target vector space

$$min_W \|AW - B\|_F$$

A is the embedding matrix of the first language, B is the embedding matrix of the second language and W is the transformation matrix

- Requirements:
  - Two monolingual corpora, one for each language
  - Bilingual dictionary
- To solve the OOV problem, fasttext is employed to form word representations based on their sub-words.

- Evaluation

  Bilingual lexicon induction For OOV words in the source language and in-vocabulary in the target language

  Accuracy @k is selected as the evaluation metric.

Training data and Dataset

- Six languages are considered for the experiments. English, Spanish, German, Finnish, Russian and Japanese.
- Embeddings are trained on the raw data from Wikipedia
- The source language is considered, a low-resource language. To simulate the situation for each language Embeddings are formed over 100M tokens
- Test set is extracted from Panlex.

Results

Language	Method	% Accuracy English source @1	@5	@10	English target @1	@5	@10
Finnish	OOV	1.49	3.55	4.97	2.43	5.67	7.74
	BL	0.46	-	-	0.27	-	-
	IV	20.29	30.25	48.16	47.11	64.77	71.01
German	OOV	2.35	5.60	7.35	3.16	8.07	10.77
	BL	2.06	-	-	0.81	-	-
	IV	44.79	66.51	73.13	51.62	69.54	73.58
Japanese	OOV	0.45	1.61	2.17	0.67	1.73	2.33
	BL	0.13	-	-	1.19	-	-
	IV	25.30	40.25	44.79	27.60	44.36	49.93
Russian	OOV	2.11	5.14	6.85	3.86	9.19	12.07
	BL	0.09	-	-	0	-	-
	IV	33.91	53.51	59.67	46.58	66.04	70.54
Spanish	OOV	6.09	10.99	13.43	3.69	8.20	10.68
	BL	3.56	-	-	2.34	-	-
	IV	62.88	79.31	83.58	61.53	77.61	82.31

Low-resource Language Experiments

- One truly low-resource language is also considered, **Cherokee**
- Pre-trained word embeddings are used.
  - Size of embeddding matrix : **7034**
  - Number of training instances: **1309**
  - Number of test instances: **1472**

- **Accuracy@1 : 1.11%**
- **Accuracy@5 : 2.38%**
- **Accuracy@10 : 3.66%**

- The accuracy@1 for the copy baseline is **0.08%**

Conclusions and Future Work

- A novel bilingual lexicon induction task in which we identify translations for OOV words
- Sub-word embeddings provide information for identifying translations of OOV words
- This is the case for Cherokee, a morphologically-rich low-resource language

- Future work
  - Expand the evaluation to include other strategies for forming cross-lingual embeddings
  - Learn crosslingual embeddings that incorporate knowledge of sub-words during training