Unsupervised Multilingual Word Embeddings

Cornell University

Xilun Chen and Claire Cardie
{xlchen, cardie}@cs.cornell.edu

Website: http://www.cs.cornell.edu/~xlchen/
Code: https://github.com/ccsasuke/umwe

Highlights

	Supervised	Unsupervised
Bilingual	Mikolov et al. (2013) Zou et al. (2013) & many more…	Zhang et al. (2017) Conneau et al. (2017) Artetxe et al. (2017; 2018)
Multilingual	Ammar et al. (2016) Duong et al. (2017)	**This work!**

	Training Cost		
	#BWEs	time	overall
Supervised methods			
Sup-MUSE-Direct	30	$4h$	78.0
Unsupervised methods			
MUSE-Pivot	10	$8h$	77.0
MUSE-Direct	30	$23h$	77.6
Ours	**5**	**5h**	**79.3**

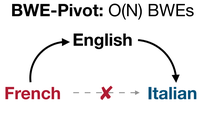
Word Translation Accuracy on 6 languages (30 pairs)

TL;DR: A multilingual generalization of the Facebook MUSE (Conneau et al., 2017) embeddings.
Try our method out if you are using MUSE to map multiple languages into a single space!

Method

Existing Methods

BWE-Pivot: O(N) BWEs

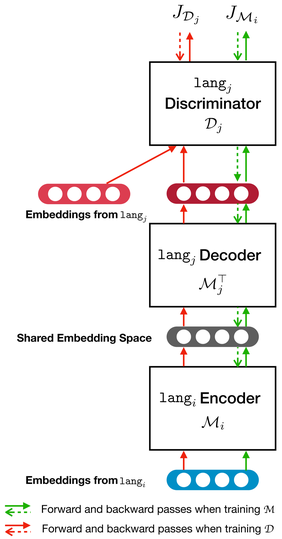


BWE-Direct: O(N²) BWEs

Learning Unsupervised Multilingual Embeddings

- N languages, each with trained monolingual embeddings as input
- Learn N-1 **orthogonal matrices** to map all languages into the same space
- Explicitly model the interaction between **all pairs of languages**
- Despite exploiting O(N²) language pairs, our method scales **linearly** with N
- Step 1: Multilingual Adversarial Training (MAT)
- Step 2: Multilingual Pseudo-Supervised Refinement (MPSR)

Multilingual Adversarial Training



Algorithm 1 Multilingual Adversarial Training

Require: Vocabulary $\mathcal{V}_i$ for each language $\texttt{lang}_i \in \mathscr{L}$. Hyperparameter $k \in \mathbb{N}$.

repeat
  ▷ $\mathcal{D}$ iterations
  for diter $= 1$ to k **do**
    $\texttt{loss}_d = 0$
    for all $\texttt{lang}_j \in \mathscr{L}$ **do**
      Select at random $\texttt{lang}_i \in \mathscr{L}$
      Sample a batch of words $x_i \sim \mathcal{V}_i$
      Sample a batch of words $x_j \sim \mathcal{V}_j$
      $\hat{x}_t = \mathcal{M}_i(x_i)$ ▷ encode to $\mathcal{T}$
      $\hat{x}_j = \mathcal{M}_j^\top(\hat{x}_t)$ ▷ decode to $\mathcal{S}_j$
      $y_j = \mathcal{D}_j(x_j)$ ▷ real vectors
      $\hat{y}_j = \mathcal{D}_j(\hat{x}_j)$ ▷ converted vectors
      $\texttt{loss}_d \mathrel{+}= L_d(1, y_j) + L_d(0, \hat{y}_j)$
    Update all $\mathcal{D}$ parameters to minimize $\texttt{loss}_d$
  ▷ $\mathcal{M}$ iteration
  $\texttt{loss} = 0$
  for all $\texttt{lang}_i \in \mathscr{L}$ **do**
    Select at random $\texttt{lang}_j \in \mathscr{L}$
    Sample a batch of words $x_i \sim \mathcal{V}_i$
    $\hat{x}_t = \mathcal{M}_i(x_i)$ ▷ encode to $\mathcal{T}$
    $\hat{x}_j = \mathcal{M}_j^\top(\hat{x}_t)$ ▷ decode to $\mathcal{S}_j$
    $\hat{y}_j = \mathcal{D}_j(\hat{x}_j)$
    $\texttt{loss} \mathrel{+}= L_d(1, \hat{y}_j)$
  Update all $\mathcal{M}$ parameters to minimize loss
  orthogonalize($\mathcal{M}$) ▷ see §3.3
until convergence

Multilingual Pseudo-Supervised Refinement

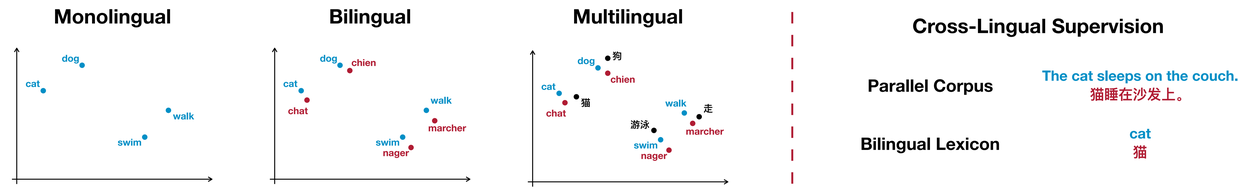
Algorithm 2 Multilingual Pseudo-Supervised Refinement

Require: A set of (pseudo-)supervised lexica of word pairs between each pair of languages $\texttt{Lex}(\texttt{lang}_i, \texttt{lang}_j)$.

repeat
  loss $= 0$
  for all $\texttt{lang}_i \in \mathscr{L}$ **do**
    Select at random $\texttt{lang}_j \in \mathscr{L}$
    Sample $(x_i, x_j) \sim \texttt{Lex}(\texttt{lang}_i, \texttt{lang}_j)$
    $t_i = \mathcal{M}_i(x_i)$ ▷ encode x_i
    $t_j = \mathcal{M}_j(x_j)$ ▷ encode x_j
    loss $\mathrel{+}= L_r(t_i, t_j)$ ▷ refinement loss
  Update all $\mathcal{M}$ parameters to minimize loss
  orthogonalize($\mathcal{M}$) ▷ see §3.3
until convergence

- MAT does a good job for more frequent words but may produce noisier signals for rare words.
- MPSR anchors on the more accurately predicted relations between frequent words to improve performance on full vocabulary.
- $\texttt{Lex}(\texttt{lang}_i, \texttt{lang}_j)$ is constructed using mutual nearest neighbors among 15k most frequent words

Cross-Lingual Word Embeddings



Cross-Lingual Supervision

Parallel Corpus	The cat sleeps on the couch. 猫睡在沙发上。
Bilingual Lexicon	cat 猫

Experiments

	en-de	en-fr	en-es	en-it	en-pt	de-fr	de-es	de-it	de-pt	fr-es	fr-it	fr-pt	es-it	es-pt	it-pt
Supervised methods with cross-lingual supervision															
Sup-BWE-Direct	73.5	81.1	81.4	77.3	79.9	73.3	67.7	69.5	59.1	82.6	83.2	78.1	83.5	87.3	81.0
Unsupervised methods without cross-lingual supervision															
BWE-Pivot	74.0	82.3	81.7	77.0	80.7	71.9	66.1	68.0	57.4	81.1	79.7	74.7	81.9	85.0	78.9
BWE-Direct	74.0	82.3	81.7	77.0	80.7	73.0	65.7	66.5	58.5	83.1	83.0	77.9	83.3	87.3	80.5
MAT+MPSR	**74.8**	**82.4**	**82.5**	**78.8**	**81.5**	**76.7**	**69.6**	**72.0**	**63.2**	**83.9**	**83.5**	**79.3**	**84.5**	**87.8**	**82.3**

	de-en	fr-en	es-en	it-en	pt-en	fr-de	es-de	it-de	pt-de	es-fr	it-fr	pt-fr	it-es	pt-es	it-pt
Supervised methods with cross-lingual supervision															
Sup-BWE-Direct	72.4	**82.4**	82.9	76.9	**80.3**	69.5	68.3	67.5	63.7	85.8	87.1	84.3	87.3	91.5	81.1
Unsupervised methods without cross-lingual supervision															
BWE-Pivot	72.2	82.1	83.3	**77.7**	80.1	68.1	67.9	66.1	63.1	84.7	86.5	82.6	85.8	91.3	79.2
BWE-Direct	72.2	82.1	83.3	**77.7**	80.1	69.7	68.8	62.5	60.5	86	87.6	83.9	87.7	92.1	80.6
MAT+MPSR	**72.9**	81.8	**83.7**	77.4	79.9	**71.2**	**69.0**	**69.5**	**65.7**	**86.9**	**88.1**	**86.3**	**88.2**	**92.7**	**82.6**

(a) Detailed Results

	Training Cost		*Single Source*						*Single Target*						
	#BWEs	time	en-xx	de-xx	fr-xx	es-xx	it-xx	pt-xx	xx-en	xx-de	xx-fr	xx-es	xx-it	xx-pt	**Overall**
Supervised methods with cross-lingual supervision															
Sup-BWE-Direct	$N(N-1)$	$4h$	78.6	68.4	79.2	81.6	80.0	80.2	79.0	68.5	82.3	82.1	78.9	77.1	78.0
Unsupervised methods without cross-lingual supervision															
BWE-Pivot	$2(N-1)$	$8h$	79.1	67.1	77.1	80.6	79.0	79.3	**79.1**	67.8	81.6	81.2	77.2	75.3	77.0
BWE-Direct	$N(N-1)$	$23h$	79.1	67.2	79.2	81.7	79.2	79.4	**79.1**	67.1	82.6	82.1	78.1	77.0	77.6
MAT+MPSR	$N-1$	$5h$	**80.0**	**70.9**	**79.9**	**82.4**	**81.1**	**81.4**	**79.1**	**70.0**	**84.1**	**83.4**	**80.3**	**78.8**	**79.3**

(b) Summarized Results

Table 1: Multilingual Word Translation Results for English, German, French, Spanish, Italian and Portuguese. The reported numbers are *precision@1* in percentage. All systems use the nearest neighbor under the CSLS distance for predicting the translation of a certain word.

Word Translation

- 6 languages: English, German, French, Spanish, Italian, Portuguese
- Word translation retrieved as nearest neighbors in embeddings space

Baselines

- **BWE-Pivot:**
  - Map each language independently from and to English
  - In total 2(N-1) MUSE (Conneau et al., 2017) BWEs
  - Use English as a pivot for word translation: e.g. de -> en -> fr
- **BWE-Direct:**
  - Learn N(N-1) MUSE BWEs for each language pair
- **Sup-BWE-Direct:**
  - Learn N(N-1) Supervised BWEs for each language pair
  - Each pair uses 5k labeled word pairs for training

Cross-Lingual Word Similarity

- Dataset from SemEval-2017 Shared Task
- Evaluates how well the similarity in the cross-lingual embedding space corresponds to a human-annotated semantic similarity score
- Luminoso and NASARI have access to EuroParl and OpenSubtitles2016 parallel corpora

	en-de	en-es	de-es	en-it	de-it	es-it	en-fa	de-fa	es-fa	it-fa	**Average**
Supervised methods with cross-lingual supervision											
Luminoso	**.769**	**.772**	**.735**	**.787**	**.747**	**.767**	.595	.587	.634	.606	.700
NASARI	.594	.630	.548	.647	.557	.592	.492	.452	.466	.475	.545
Unsupervised methods without cross-lingual supervision											
BWE-Pivot	.709	.711	.703	.709	.682	.721	.672	.655	.701	.688	.695
BWE-Direct	.709	.711	.703	.709	.675	.726	.672	.662	.714	.695	.698
MAT+MPSR	.711	.712	.708	.709	.684	.730	**.680**	**.674**	**.720**	**.709**	**.704**

Table 2: Results for the SemEval-2017 Cross-Lingual Word Similarity task. Spearman's ρ is reported. Luminoso (Speer and Lowry-Duda, 2017) and NASARI (Camacho-Collados et al., 2016) are the two top-performing systems for SemEval-2017 that reported results on all language pairs.