Unsupervised Multilingual Word Embeddings (Chen and Cardie, 2018)

Overview

- Proposes a framework to learn Multilingual Word Embeddings (MWEs)
- Exploits relations between all language pairs
- Performs in $O(N)$ time, where N is the number of languages

Related Work

- Most use cross-lingual supervision, some sort of parallel corpora
- pivot-BWEs: mapping all languages individually into a target language space (training Bilingual Word Embeddings, N times)
  - Does not capture relations between all language pairs
- BWE-Direct: training embeddings for all language pairs.
  - Computational Complexity: $O(N^2)$

Solution

- Maps all monolingual embeddings into a shared space via a two-step algorithm:
  - Multilingual Adversary Training (MAT)
  - Multilingual Pseudo-Supervised Refinement (MPSR)
- Outperforms both pivot-BWE and BWE-Direct
- $O(N)$ complexity

Solution (cont.)

Overview of the Algorithm

- MAT takes monolingual word embeddings and aligns them on the target embedding space
- MPSR takes the solution provided by MAT and improves it using dictionaries of highly confident word pairs for every language pair

Definitions for the Architecture

For each language $l \in L$ (where L is the set of languages considered), we take the embedding E_l that is in the embedding space S_l

- The models learn:
  - Encoder M_l into target space T s.t. $M_l : S_l \rightarrow T$
  - Decoder M_l^{-1}, so $M_l^{-1} : T \rightarrow S_l$

Encoders M_l are all orthogonal linear matrices

Definitions for the Architecture (cont.)

Language classifiers D_l: a binary classifier with a sigmoid layer on top, trained to identify how likely it is a vector is from space S_l

Multilingual Adversary Training

Overview

- Setup an adversarial relation between D_l and M_l
- Stimulates M_l to learn a shared multilingual embedding space
  - So that D_l cannot predict if the vector is genuine or converted from another language

Multilingual Adversary Training (cont.)

Language Discriminators

- For random pair (l_i, l_j) convert vector from S_i to S_j (using M_{l_i}, $M_{l_j}^{-1}$ and via T)
- Objective: confuse D_j, update it
- Formally, objective function:

  $$J_d = E_{i-L} E_{x_i - S_i, x_j - S_j} (L_d(1, D_j(x_j)) + L_d(0, D_j(M_j^T M_i x_i)))$$

Multilingual Adversary Training (cont.)

Training M

- Pick words and embed into target space
- Based on loss, update parameters of M
- Formal objective function of M:

$$J_{M_i} = E_{j-L} E_{x_i - S_i, x_j - S_j} (L_d(1, D_j(M_j^T M_i x_i)))$$

For both iterations, the Loss function L_d is cross entropy loss.

Multilingual Adversary Training (cont.)

Other improvements and optimizations

- l_i and l_j can be the same language (adversarial autoencoder is formed, shown to be beneficial)
- Instead of random sampling throughout, the external iteration loops through all languages to ensure updation of all language discriminators at least once

Multilingual Pseudo-Supervised Refinement

Overview

- MAT gives reasonable quality embeddings, but not SOTA
- May be due to noisy training signals from D
- Improvement: Induce a dictionary of *highly confident word pairs* for each language pair, and use this

Multilingual Pseudo-Supervised Refinement (cont.)

Building dictionary
For a language pair (l_i, l_j), $Lex(l_i, l_j)$ is constructed from mutual nearest neighbours between $M_i E_i$ and $M_j E_j$, among most frequent 15K words of both languages.

Multilingual Pseudo-Supervised Refinement (cont.)

Algorithm

- Sample x_i, x_j from languages l_i, l_j
- Embed into t_i, t_j
- Update M given the loss
- Formal objective:

  $$J_r = E_{(i,j)-L^2} E_{(x_i,x_j)-Lex(i,j)}(L_r(M_i x_i, M_j x_j))$$