A Survey of Cross-lingual Word Embedding Models

Sebastian Ruder, Ivan Vulić, and Anders Søgaard

Journal of Artificial Intelligence Research
Overview

• Data Usage:
  + Parallel.
  + Comparable.

• Cross-lingual Embedding Models:
  + Word-level alignment methods.
  + Sentence-level alignment methods.
  + Document-level alignment methods.
Data Usage

(a) Word, par.  (b) Word, comp.  (c) Sentence, par.  (d) Sentence, comp.  (e) Doc., comp.

The dog chases the cat.
Le chien poursuit le chat.

The dog chases the cat in the grass.

There are a lot of dogs in the park. They like to chase cats.
Les chats se relaxent. Ils fuient les chiens dès qu’ils les voient.
Word-level Alignment Methods

• Using parallel data:
  + Mapping-based approaches
  + Pseudo-multi-lingual corpora-based approaches
  + Joint methods: Bilingual language model
• Using comparable data:
  + Language grounding models
  + Comparable feature models
Mapping-based approaches

• Regression methods: \[ \Omega_{MSE} = \|WX^s - X^t\|_F^2 \]

• Orthogonal methods: constrain the transformation \( W \) to be orthogonal
  \[ W^T W = I \quad X^{t\top} X^s = U\Sigma V^{\top} \quad W = VU^T \]

• Canonical methods:
  \[ \rho(W^{s\to}x^s_i, W^{t\to}x^t_i) = \frac{\text{cov}(W^{s\to}x^s_i, W^{t\to}x^t_i)}{\sqrt{\text{var}(W^{s\to}x^s_i)\text{var}(W^{t\to}x^t_i)}} \quad \Omega_{CCA} = -\sum_{i=1}^{n} \rho(W^{s\to}x^s_i, W^{t\to}x^t_i) \]

• Margin methods:
  \[ \Omega_{MML} = \sum_{i=1}^{n} \sum_{j \neq i}^{k} \max\{0, \gamma - \cos(Wx^s_i, x^t_i) + \cos(Wx^s_i, x^t_j)\} \]
Pseudo-multi-lingual corpora-based approaches

• Using the word-level alignment of a seed bilingual dictionary to construct a pseudo-bilingual corpus by randomly replacing words in a source language corpus with their translations.
• Concatenating the source and target language corpus and replace each word that is part of a translation pair with its translation equivalent.

=> Using monolingual embedding learning methods as usual.
Bilingual language model

• Use a shared embedding matrix between two languages:
  \[ c = (c_1^T; \ldots; c_{|V_{in}|}^T) \in \mathbb{R}^{|V_{in}|d} \]

• Alignment matrix \( A \), each row corresponds to probabilities that translations are aligned to the current word.

• Cast crosslingual distributed representation induction as a multitask learning problem:
  + Treating each word \( w \) in our languages’ vocabularies as a separate task.
  + Relatedness between "tasks" are encoded in alignment matrix \( A \)

• Representations of source and target language words that are often aligned are encouraged to be similar:
  \[ L(\theta) = \sum_{l=1}^{2} \sum_{t=1}^{T^{(l)}} \log \hat{P}_{\theta^{(l)}}(w_{t}^{(l)}|w_{t-n+1:t-1}^{(l)}) + \frac{1}{2} c^T (A \otimes I_d) c \]
Grounding language in images

• Each word is associated with a set of images which are typically retrieved using Google Image Search.
• Calculating a similarity score for a pair of words based on the visual similarity of their associated image sets.
Sentence-level Alignment Methods

- Using parallel data:
  - Word-alignment based matrix factorization approaches
  - Compositional sentence models
  - Others.
- Using comparable data: similar to word-level alignment methods.
Word-alignment based matrix factorization approaches

• Intuition: If the target word is aligned with more than one source word, then its representation should be a combination of the representations of its aligned words.

• Representing the embeddings $X^s$ in the target language as the product of the source embeddings $X^s$ with the corresponding alignment matrix.

$$A^{s\rightarrow t}X^s$$

• Only learning source language embeddings:

$$\Omega_{s\rightarrow t} = \|X^t - A^{s\rightarrow t}X^s\|^2$$

$$J = \mathcal{L}_{MML}(X^t) + \Omega_{s\rightarrow t}(X^t, A^{s\rightarrow t}, X^s)$$
Compositional sentence models

• Bringing the sentence representations of aligned sentences in source and target language close to each other.

\[ y^s = \sum_{i=1}^{|sent^s|} x^s_i \]

\[ E_{dist}(sent^s, sent^t) = \| y^s - y^t \|^2 \]

• Encouraging aligned sentence representations to be more similar than each of them to each word in the other sentence.

\[ \mathcal{L} = \sum_{(sent^s, sent^t) \in \mathcal{C}} \sum_{i=1}^{k} \max(0, 1 + E_{dist}(sent^s, sent^t) - E_{dist}(sent^s, s^t_i)) \]
Other sentence-level approaches

• Bilingual autoencoder: trains an auto-encoder with language-specific encoder and decoder layers and hierarchical softmax to reconstruct from each sentence the sentence itself and its translation.

\[ J = \mathcal{L}^{s \rightarrow s}_{\text{AUTO}} + \mathcal{L}^{t \rightarrow t}_{\text{AUTO}} + \mathcal{L}^{s \rightarrow t}_{\text{AUTO}} + \mathcal{L}^{t \rightarrow s}_{\text{AUTO}} \]

• Bilingual skip-gram: adds a cross-lingual regularization term to skip-gram monolingual losses:

\[ y^s = \frac{1}{|\text{sent}^s|} \sum_{i=1}^{|\text{sent}^s|} x^s_i \]

\[ \Omega_{\text{BILBOWA}} = \sum_{(\text{sent}^s, \text{sent}^t) \in \mathcal{C}} \| y^s - y^t \|^2 \]

\[ J = \mathcal{L}^s_{\text{SGNS}} + \mathcal{L}^t_{\text{SGNS}} + \Omega \]
Document-level Alignment Methods

- Extending the sentence-level alignment methods by adding regularization terms on paragraph representations.

\[ \Omega = \sum_{(\text{sent}^s, \text{sent}^t) \in C} \alpha \| \mathbf{p}^s - \mathbf{p}^t \|^2 + (1 - \alpha) \| \mathbf{y}^s - \mathbf{y}^t \|^2 \]

\[ J = \mathcal{L}^{s}_{\text{SGNS-P}}(\mathbf{P}^s, \mathbf{X}^s) + \mathcal{L}^{t}_{\text{SGNS-P}}(\mathbf{P}^t, \mathbf{X}^t) + \Omega(\mathbf{P}^s, \mathbf{P}^t, \mathbf{X}^s, \mathbf{X}^t) \]