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.

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: $\Omega_{\text{MSE}} = \|\mathbf{W}\mathbf{X}^s - \mathbf{X}^t\|_F^2$

- Orthogonal methods: constrain the transformation $\mathbf{W}$ to be orthogonal

$$\mathbf{W}^\top\mathbf{W} = \mathbf{I} \qquad \mathbf{X}^{t\top}\mathbf{X}^s = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \qquad \mathbf{W} = \mathbf{V}\mathbf{U}^\top$$

- Canonical methods:

$$\rho(\mathbf{W}^{s\to}\mathbf{x}_i^s, \mathbf{W}^{t\to}\mathbf{x}_i^t) = \frac{\text{cov}(\mathbf{W}^{s\to}\mathbf{x}_i^s, \mathbf{W}^{t\to}\mathbf{x}_i^t)}{\sqrt{\text{var}(\mathbf{W}^{s\to}\mathbf{x}_i^s)\text{var}(\mathbf{W}^{t\to}\mathbf{x}_i^t)}} \qquad \Omega_{\text{CCA}} = -\sum_{i=1}^{n}\rho(\mathbf{W}^{s\to}\mathbf{x}_i^s, \mathbf{W}^{t\to}\mathbf{x}_i^t)$$

- Margin methods:

$$\Omega_{\text{MML}} = \sum_{i=1}^{n}\sum_{j\neq i}^{k}\max\{0, \gamma - \cos(\mathbf{W}\mathbf{x}_i^s, \mathbf{x}_i^t) + \cos(\mathbf{W}\mathbf{x}_i^s, \mathbf{x}_j^t)\}$$

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:

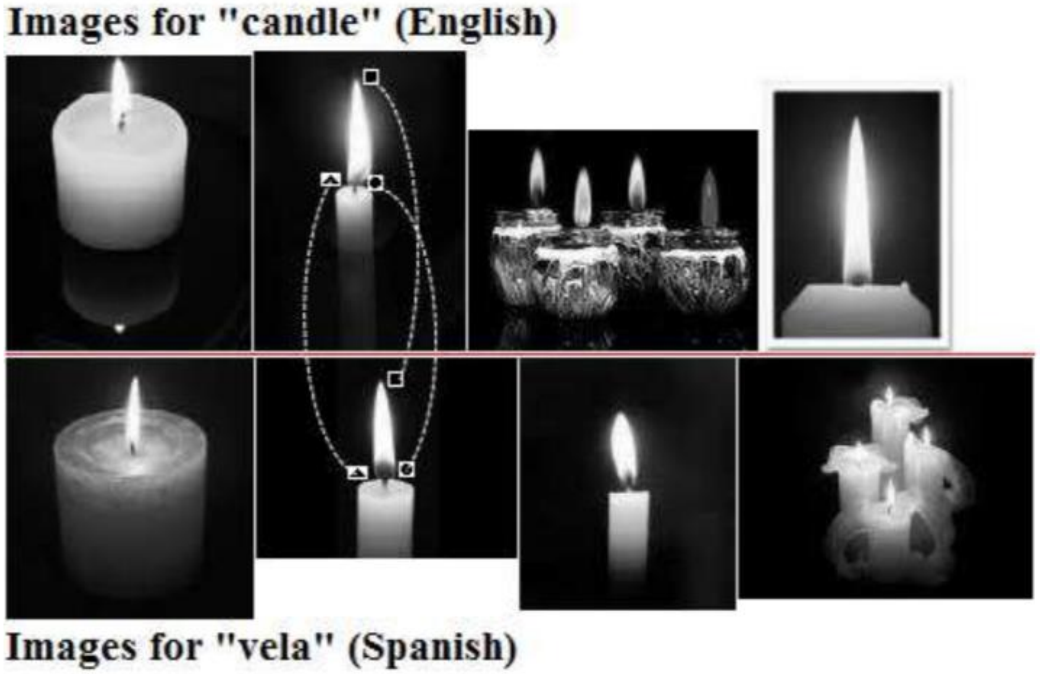
$$c = (c_1^\top; \ldots; c_{|V_{in}|}^\top)^\top \in \mathbb{R}^{|V_{in}|d}$$

- 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:

$$L(\theta) = \sum_{l=1}^{2} \sum_{t=1}^{T^{(l)}} \log \hat{P}_{\theta^{(l)}}(w_t^{(l)} | w_{t-n+1:t-1}^{(l)}) + \frac{1}{2} c^\top (A \otimes I_d) c$$

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 $\mathbf{X}^s$ in the target language as the product of the source embeddings $\mathbf{X}^s$ with the corresponding alignment matrix.

$$\mathbf{A}^{s\to t}\mathbf{X}^s$$

- Only learning source language embeddings:

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

$$J = \underbrace{\mathcal{L}_{\mathrm{MML}}(\mathbf{X}^t)}_{1} + \underbrace{\Omega_{s\to t}(\underline{\mathbf{X}^t}, \underline{\mathbf{A}^{s\to t}}, \mathbf{X}^s)}_{2}$$

Compositional sentence models

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

$$\mathbf{y}^s = \sum_{i=1}^{|sent^s|} \mathbf{x}_i^s$$

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

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

$$\mathcal{L} = \sum_{(sent^s, sent^t) \in \mathcal{C}} \sum_{i=1}^{k} \max(0, 1 + E_{dist}(sent^s, sent^t) - E_{dist}(sent^s, s_i^t))$$

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:

$$\mathbf{y}^s = \frac{1}{|sent^s|} \sum_{i=1}^{|sent^s|} \mathbf{x}_i^s$$

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

$$J = \mathcal{L}^s_{\text{SGNS}} + \mathcal{L}^t_{\text{SGNS}} + \Omega$$

Document-level Alignment Methods

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

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

$$J = \mathcal{L}^s_{\text{SGNS-P}}(\mathbf{P}^s, \mathbf{X}^s) + \mathcal{L}^t_{\text{SGNS-P}}(\mathbf{P}^t, \mathbf{X}^t) + \Omega(\mathbf{P}^s, \mathbf{P}^t, \mathbf{X}^s, \mathbf{X}^t)$$