Multilingual Word Embeddings

Winston Wu
11/12/2020
Outline

• Motivation
• Methods
  • Projection/Mapping
  • Shared Space
  • Pseudo-mixed
• Evaluation: Bilingual Lexicon Induction
• Resources and Tools
• An Interesting Application
• Brainstorming
Motivations

- Embeddings are a semantic representation of words
- Words in different languages represent the same concept
  - “dog” in English, “狗” in Chinese, “chien” in French
  - So these words should have the same vector
- The structure of embedding spaces should be similar across languages
- We have much more monolingual text than parallel text

- Challenge: can we find a common semantic space for words in all languages?
Mapping/Projection

1. Starting with monolingual corpora in English and Spanish, separately train embeddings

Slide derived from http://mt-class.org/jhu/slides/lecture-words.pdf
2. Learn a mapping matrix $W_{S \rightarrow T}$ to minimize the Euclidean distance between each word and its translation

$$\text{cost} = \sum_i \| W_{S \rightarrow T} \cdot v_i^S - v_i^T \|$$

- Using a seed lexicon (a dictionary of known translations)
- Optimization problem, can solve with SGD
Mapping/Projection

\[ \text{cost} = \sum_i \| W_{S \rightarrow T} v_i^S - v_i^T \| \]

- Intuition: \( W_{S \rightarrow T} \) rotates and stretches the English space to match the Spanish space

- Extensions
  - Xing+ (2015): this matrix must be orthogonal (only rotation)
  - Artetxe+ (2016): mean centering

Figure from Conneau+ 2018
Mapping/Projection

• Procrustes analysis (Conneau+ 2017, Sogaard+ 2018, Grave+ 2018)
  • Similar objective

\[ W^* = \underset{W}{\text{argmin}} \|WX - Y\|_F = UV^T \]
\[ \text{s.t. } U\Sigma V^T = \text{SVD}(YX^T) \]

• Can be unsupervised (no seed lexicon)
  • Conneau: adversarial training
  • Grave: use Wasserstein distance
  • Sogaard: identical words
Seed Lexicon

• How many words do we need?
  • Mikolov+ (2013), Smith+ (2017): 5000 words
  • Artetxe+ (2017): 25 words
  • Grave+ (2018), Alaux+ (2019), Heyman+ (2019): 0 words (unsupervised)

• Which words should we use?
  • Faruqui and Dyer (2014): alignments from parallel corpus
  • Vulic and Moens (2015): translations from Wikipedia alignments
  • Artetxe+ (2017): numbers
  • Sogaard+ (2018), Zhou+ (2019): identically spelled words in both language (homographs)
  • Shi+ (2019): off the shelf bilingual dictionaries
Mapping onto a Shared Space

• **Canonical Correlation Analysis** (Haghighi+ 2008, Faruqui and Dyer 2014)
  - CCA finds projection matrices such that the projected vectors are maximally correlated
  - Example:
    - $x$ is English "dog"
    - $y$ is Spanish "perro"
    - $v$ and $w$ are projection matrices
    - let $x' = xv$ and $y' = yw$
    - $v, w = \text{CCA}(x, y) = \arg \max_{v, w} \rho(xv, yw)$
  - Seed lexicon: alignments from parallel corpora
  - Only works for two vector spaces
Mapping onto a Shared Space

• Generalized Procrustes analysis (Kementchedjhiieva+ 2018)
  • Unsupervised: does not require a seed lexicon
  • Motivation: Søgaard et al. (2018) showed that vector spaces are often far from being isomorphic
    • If they were, then an intermediate space would not be necessary
  • Objective:

\[
\arg \min_{\{T_1, \ldots, T_k\}} \sum_{i<j}^k \|T_i E_i - T_j E_j\|^2
\]

• Aligns multiple vector spaces at the same time

Generalizing Procrustes Analysis for Better Bilingual Dictionary Induction (Kementchedjhiieva+ 2018)
Multilingual Embeddings: Mapping

- Smith+ (2017): 90 languages
- Joulin+ (2018): 44 languages
- Unsupervised:
  - Chen and Cardie (2018): 6 languages
  - Alaux+ (2019): 11 languages

- Open Question: Can we do this for 1000 languages?
Multilingual Embeddings: Shared

- MultiCCA (Ammar+ 2016): 59 languages
- Unsupervised:
  - Multi-Pairwise Procrustes (Taitelbaum+ 2019): 6 languages
  - Wasserstein barycenter (Lian+ 2020): 8 languages

- Open Question: Can we do this for 1000 languages?
Translation Mixing

  - Translations from dictionary or word alignments

  build the house  →  construire the house
  build la maison  build the maison

- Run your favorite monolingual embedding method
Evaluation

• How good is our embedding space?
• Task: Bilingual Lexicon Induction (BLI)
  • Find the best English translation of the Spanish word “perro”, where “best” = nearest neighbor

• Hubness Problem: some words are the nearest neighbor of many words
  • Cross-domain similarity locality scaling (CSLS) (Conneau+ 2017)

$$CSLS(Wx^s, x^t) = 2\cos(Wx^s, x^t) - r^t(Wx^s) - r^s(x^t),$$
Monolingual Embeddings

Figure from Sogaard+ 2019
Cross-Lingual Embeddings

Figure from Sogaard+ 2019
Resources

• Good survey: Cross-Lingual Word Embeddings

• Aligned Vectors
  • https://fasttext.cc/docs/en/aligned-vectors.html

• Software to do alignment
  • https://github.com/facebookresearch/fastText
  • https://github.com/artetxem/vecmap
  • https://github.com/facebookresearch/MUSE
  • https://github.com/ccsasuke/umwe
An Interesting Application

Learning to Pronounce Chinese Without a Pronunciation Dictionary

• Data
  • Chinese Wikipedia -> remove non-Chinese text -> characters
  • Baidu Baike encyclopedia -> convert to pinyin using a dictionary

• Method
  • Train fastText embeddings on each
  • Use MUSE to map character space onto pinyin space
  • Find nearest pinyin for each character (0.5% accuracy)

• Improvements
  • Tokenize
  • When doing nearest neighbor search, prefer pinyin with same # of syllables
  • If no nearest neighbor with same # of syllables, use “de” (most common)
  • 81% accuracy!
Brainstorming

• This is related to **multi-view learning**, where the goal is to learn a representation for different “views” of the same thing
  • “dog” in English and “perro” in Spanish
  • Image and a sentence describing the image
  • A photo of a face and a sketch of the face
  • A video of someone speaking and the audio of their speech

• What are some other examples of multiple views?
• Also think about the task