Cross-Lingual Embeddings

Winston Wu
10/29/2019
Outline

• Motivation
• Evaluation: Bilingual Lexicon Induction
• Methods
  • Projection/Mapping
  • Shared Space
  • Pseudo-mixed
• Resources and Tools
Monolingual Embeddings

Figure from Sogaard+ 2019
Cross-Lingual Embeddings

Figure from Sogaard+ 2019
Motivation

• Words in different languages represent the same concept
  • “dog” in English, “狗” in Chinese
  • The structure of embedding spaces is similar across languages

• Applications
  • Multilingual models
  • Translation without parallel data

• Note: NMT systems implicitly learn a shared embedding space, here we explicitly learn this
Evaluation: Bilingual Lexicon Induction

- Lexical (word) translations
  - Find the best English translation of the Spanish word “perro”
  - “Best” = nearest neighbor in embedding space

- Similarity measures
  - Cosine similarity
  - Cross-domain similarity locality scaling (CSLS)
    - Hubness problem: some words are similar to many other words

\[
\text{CSLS}(Wx^s, x^t) = 2\cos(Wx^s, x^t) - r^t(Wx^s) - r^s(x^t),
\]
mean sim to target words

- Extrinsic evaluation: evaluate on a downstream task
Mapping/Projection Methods

• Run embedding method on each language separately
• Align the source space and target space
  • Possibly with a seed lexicon

Figure from Conneau+ 2018
Mapping onto Target

• Mikolov+ 2013, Dinu+ 2015, Artetxe+ 2016, ...
• Learn a matrix $W^{s\rightarrow t}$ to transform $W^{s\rightarrow t}x^s \rightarrow x^t$
  • Possibly with some constraints (Xing+ 2015, Artetxe+ 2016)
Mapping onto Shared Space

• Learn multiple matrices $W^s\rightarrow$ and $W^t\rightarrow$
• Canonical Correlation Analysis (Haghighi+ 2008, Faruqui and Dyer 2014)
• Procrustes analysis (Kementchedjiev+ 2018)
Multilingual Embeddings: Mapping

• Smith+ 2017: 90 languages
• Alaux+ 2019: 11 languages
  • No seed lexicon

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

• CCA: Ammar+ 2016: 59 languages

• Open Question: Can we do this for 1000 languages?
Pseudo-mixing

  • Translations from dictionary or word alignments

• Run your favorite monolingual embedding method

build the house  ➔  construire the house
build la maison
build the maison
Seed Lexicon

- Mikolov+ 2013, Smith+ 2017: 5000 words
- Artetxe 2017: 25 words
- Grave+ 2018, Alaux+ 2019: 0 words

- Open Questions
  - How many words do we need?
  - Is there a good set of words to use?
Resources

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

• Software to do alignment
  • https://github.com/facebookresearch/fastText/tree/master/alignment
  • https://github.com/artetxem/vecmap
  • https://github.com/facebookresearch/MUSE
What did I not talk about today?

• Sentence alignment
• Document alignment