Unsupervised Multilingual Word Embeddings

Xilun Chen and Claire Cardie
(xichen, cardie)@cs.cornell.edu

Method

Learning Unsupervised Multilingual Embeddings

- \( N \) languages, each with trained monolingual embeddings as input
- Learn \( N-1 \) orthogonal matrices to map all languages into the same space
- Explicitly model the interaction between all pairs of languages
- Despite exploiting O(N) language pairs, our method scales linearly with \( N \)
- Step 1: Multilingual Adversarial Training (MAT)
- Step 2: Multilingual Pseudo-Supervised Refinement (MPSR)

Multilingual Pseudo-Supervised Refinement

- **Optimization:**
  - DIRECT: Directly optimize the direct translation score
  - PIVOT: Use a pivot language

Cross-Lingual Word Embeddings

- **Cross-Lingual Supervision**
  - Parallel Corpus
  - Bilingual Lexicon

Cross-Lingual Word Similarity

- Dataset from SemEval-2017 Shared Task
- Evaluates how well the similarity in the cross-lingual embedding space corresponds to a human-annotated semantic similarity score
- Luminoso and NASARI have access to EuroParl and OpenSubtitles2016 parallel corpora

Experiments

- **Word Translation**
  - 6 languages: English, German, French, Spanish, Italian, Portuguese
  - Word translation retrieved as nearest neighbors in embeddings space

Baselines

- **BWE-Pivot:**
  - Map each language independently from and to English
  - In total 2(N-1) MUSE (Conneau et al., 2017) BWEs
  - Use English as a pivot for word translation: e.g. de -> en -> fr

- **BWE-Direct:**
  - Learn N(N-1) MUSE BWEs for each language pair

- **Sup-BWE-Direct:**
  - Learn N(N-1) Supervised BWEs for each language pair
  - Each pair uses 5k labeled word pairs for training