Extending Multi-sense Word Embedding to Phrases and Sentences for Unsupervised Semantic Applications

Haw-Shiuan Chang, Amol Agrawal, Andrew McCallum

Introduction

- Previous Work:
  - Word embedding represents the input word by a set co-occurring words
  - Co-occurring word distribution might have multiple modes
- Multi-sense word embedding clusters the co-occurring words into centers
- Our goal:
  - Extending the methods to phrases and sentences
  - Do the similar thing but replacing the input word as a word sequence.
  - Senses of the input words -> Facets of the input sentence/phrases

Multi-sense Embedding [1]

- Challenges
  - Storage
    - Too many unique sentences
  - Sparse signal
    - Too few co-occurring words
  - "Out-of-vocabulary"
    - Similar sentences during testing

Main Idea

- Instead of clustering, we directly predict the cluster centers using a neural model
- Storage issue
  - Clusters are compressed in the parameters of the neural model
- Sparse signal issue
  - Clustering the co-occurring words of similar sentences
- "Out-of-vocabulary" issue
  - Can take any input sentence
- Model Design
  - What neural network architecture to use?
  - How to train end-to-end?
  - What clustering loss to use?
- Other Applications
  - Interactive Language Generation [2]
  - Distantly Supervised Relation Extraction [3]

Our Method

- Multi-facet Embedding
  - Step 1: Generate $F(I_t)$
  - Step 2: Estimate $M_{Ot}$ and $M_{Rt}$
  - Step 3: Compute Loss $L_t(F)$
  - Step 4: Fix $M_{Ot}$ and $M_{Rt}$ to do backprop

Experiments

- Multi embeddings for sentence representation is much better than single embedding
- Similar for phrase representation
- Word importance estimation using the co-occurring distribution improves various scoring functions
- More facets are better in summarization

Challenges

- Storage
  - Too many unique sentences
- Sparse signal
  - Too few co-occurring words
- "Out-of-vocabulary"
- Similar sentences during testing

Our framework

- Multi-facet Embedding
  - $F(I_t)$: Pre-trained word embedding space
  - $M_{Ot}$, $M_{Rt}$: Co-occurring word embeddings
  - $L_t(F)$: Non-negative sparse coding loss

Conclusion

- We propose a framework for learning the cooccurring distribution of the words beside a sentence or a phrase.
- Even though there are usually only a few words that co-occur with each sentence, we demonstrate that the proposed models can learn to predict interpretable cluster centers conditioned on an (unseen) sentence.

References