Embedding Imputation with Grounded Language Information

Ziyi Yang[1], Chenguang Zhu[2], Vin Sachidananda[3], Eric Darve[1, 4]

[1]Department of Mechanical Engineering, Stanford University, [2]Microsoft Speech and Dialogue Research Group
[3]Department of Electrical Engineering, Stanford University, [4]Institute for Computational and Mathematical Engineering, Stanford University

Abstract

Due to the ubiquitous use of embeddings as input representations for a wide range of natural language tasks, imputation of embeddings for rare and unseen words is a critical problem in language processing. Embedding imputation involves learning representations for rare or unseen words during the training of an embedding model, often in a post-hoc manner. In this paper, we propose an approach for embedding imputation which uses grounded information in the form of a knowledge graph. This contrasts with existing approaches which typically make use of vector space properties or subword information. We propose an online method to construct a graph from grounded information and design an algorithm to map from the resulting graphical structure to the space of the pre-trained embeddings. Finally, we evaluate our approach on a range of rare and unseen word tasks across various domains and show that our model can learn better representations. For example, on the Card-660 task our method improves Pearson's and Spearman's correlation coefficients upon the state-of-the-art by 11% and 17.8% respectively using GloVe embeddings.

Motivations

- **OOV problem**: Pretrained words vectors, e.g. GloVe, have fixed vocabulary. Learning representations for words which are rare or unseen during training is difficult.
- **Grounded language information**, e.g. knowledge base and online encyclopedia is useful for reasoning and inference.

British = **[0.98, 0.35, 1.78, -0.62, …]**
Europe = **[2.74, -1.48, 0.39, 1.17, …]**
politics = **[1.12, 2.87, 2.45, -0.45, …]**
Brexit = **[?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?]**

WIKIPEDIA
The Free Encyclopedia

Brexit, a portmanteau of ``British'' and ``exit'', is the impending withdrawal of the United Kingdom (UK) from the European Union (EU)…

Wiktionary
The free dictionary

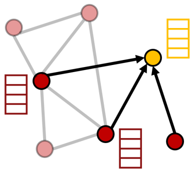
Brexit (Britain, politics): The withdrawal of the United Kingdom from the European Union.

Our Contributions

- An approach to **constructing graphical representations of entities in a knowledge base** in an unsupervised manner.
- Methods for mapping entities from a **graphical representation to the space** in which a pre-trained embedding lies.
- Experimentation on rare and unseen word datasets and a **new state-of-art performance on Card-660 dataset.**

Our Approach

- **Graph Neural Network (GNN):** GNNs can learn a representation vector for each node in the network. Node embeddings are generated by recursively **aggregating** each node's neighborhood **features**.



- At the t iteration, the information aggregation is defined as:

$$h_v^t = \mathrm{ReLU}(W^t \sum_{u \in S(v)} \frac{s_{vu} h_u^{t-1}}{C} + b^t)$$

where $S(v) = N(v) \cup \{v\}$, and the normalization constant $C = 1 + \sum_{u \in N(v)} s_{vu}$. W^t and b^t are trainable parameters.

- Grounded language information D_v for word w_v: All the words in the Wikipedia summary and the Wiktionary definition.
- Features f_v is the mean of pre-trained embeddings of words in D_v.
- Undirected Edge e_{vu} exists if the Jaccard coefficient

$$\frac{D_v \cap D_u}{D_v \cup D_u} > 0.5$$

The edge is assigned with the weight $\frac{D_v \cap D_u}{D_v \cup D_u}$

Results

- Dataset: Cambridge Rare Word Dataset (Card-660), Stanford Rare Word Similarity (RW)
- Evaluation metrics: Pearson's r and Spearman's ρ correlation
- KG2Vec achieves **SOTA** on Card-660. With ConceptNet embeddings, KG2Vec results in improvements **of 7.7%/6.7%** on r/ρ on the previous BOS. With GloVe embeddings, KG2Vec improves upon SemLand by **11%/17.8%** on on r/ρ.

Model	Missed words		Missed pairs		Pearson r		Spearman ρ	
	RW	CARD	RW	CARD	RW	CARD	RW	CARD
ConceptNet Numberbatch	5%	37%	10%	53%	53.0	36.0	53.7	24.7
+ Mimick	0%	0%	0%	0%	56.0	34.2	57.6	35.6
+ Definition centroid	0%	29%	0%	43%	59.1	42.9	60.3	33.8
+ Definition LSTM	0%	25%	0%	39%	58.6	41.8	59.4	31.7
+ SemLand	0%	29%	0%	43%	**60.5**	43.4	**61.7**	34.3
+ BoS	0%	0%	0%	0%	60.0	49.2	**61.7**	47.6
+ Node features	0.02%	7%	0.04%	12%	58.4	54.0	59.7	51.4
+ KG2Vec	0.02%	7%	0.04%	12%	58.6	**56.9**	60.1	**54.3**
GloVe Common Crawl	1%	29%	2%	44%	44.0	33.0	45.1	27.3
+ Mimick	0%	0%	0%	0%	44.7	23.9	45.6	29.5
+ Definition centroid	0%	21%	0%	35%	43.5	35.2	45.1	31.7
+ Definition LSTM	0%	20%	0%	33%	24.0	23.0	22.9	19.6
+ SemLand	0%	21%	0%	35%	44.3	39.5	45.8	33.8
+ BoS	0%	0%	0%	0%	**44.9**	31.5	**46.0**	35.3
+ Node features	0.05%	0.4%	0.01%	0.7%	43.8	36.0	45.0	37.4
+ KG2Vec	0.05%	0.4%	0.01%	0.7%	44.6	**50.5**	45.8	**51.6**

Conclusions

We introduce KG2Vec, a GNN based model for embedding imputation of OOV words which makes use of grounded language information. Using publicly available information sources like Wikipedia and Wiktionary, KG2Vec can effectively impute embeddings OOVs. KG2Vec achieves state-of-the-art results on the Card-660 dataset. Future research directions include a theoretical explanation of KG2Vec.