Real Multi-Sense or Pseudo Multi-Sense: An Approach to Improve Word Representation

Haoyue Shi, Caihua Li and Junfeng Hu {hyshi, peterli, hujf}@pku.edu.cn
School of Electronics Engineering and Computer Science, Peking University

Abstract

Previous researches have shown that learning multiple representations for polysemous words can improve the performance of word embeddings on many tasks. However, this leads to another problem. Several vectors of a word may actually point to the same meaning, namely pseudo multi-sense.

We introduce the concept of pseudo multi-sense, and then propose an algorithm to detect such cases. With the consideration of the detected pseudo multi-sense cases, we try to refine the existing word embeddings to eliminate the influence of pseudo multi-sense. Moreover, we apply our algorithm on previous released multi-sense word embeddings and tested it on artificial word similarity tasks and the analogy task. The result of the experiments shows that diminishing pseudo multi-sense can improve the quality of word representations. Thus, our method is actually an efficient way to reduce linguistic complexity for computer's natural language understanding.

Pseudo Multi-Sense

Polysemous words are ubiquitous in natural languages. For example, in the following two sentences, the word *star* has different meanings, leaving all metaphorical considerations aside.

- She is a famous movie **star**.
- That small **star** is brightest.

Some previous approaches have learned multiple embeddings to represent different senses of a word, discriminating the senses by their context, related syntax and topics. However, this leads to another problem. The methods may embed one sense to multiple vectors by mistake. We call such cases **pseudo multi-sense**.

Consider three different representations of word *bear* released by [2]. We show 9 nearest neighbors for each representation. Words clearly related to the domain *animals* are bolded.

- emerald, **bears**, **three-toed**, **snake**, **periwinkle**, **ruffed**, **hoopoe**, distinctive, unmistakable
- **bird**, **wolf**, arrow, **pelican**, emerald, canyon, diamond, **buck**, **deer**
- pride, lady, hide, king, gift, crane, afflict, promise, reap

We could infer that the first two representations are likely to have the same meaning, namely they two are likely to form a pseudo multi-sense case. But the last representation may have different meaning with them (real multi-sense).

Pseudo Multi-Sense Detection

Domain Similarity

To determine the domain of a sense given the multi-sense word embeddings, we can intuitively define the probability that the k^{th} sense of word w belongs to domain d as

$$P_D(w,k,d) \propto \sum_{w' \in NN(w,k)} D(p(w'),d)$$

where $NN(w,k)$ is the nearest neighbors of the k^{th} sense of word w in the given word embeddings, $p(w')$ is the protocol representation of word w', $D(p(w'),d)$ is the sum probability that domain d appears in all synsets of $p(w')$ in WordNet([3]) provided by Extended WordNet Domain. Then we can compute the domain similarity between the k^{th} and the l^{th} sense of word w by

$$Sim_D(w,k,l) = \frac{1}{n}|TopN(P_D,w,k,n) \cap TopN(P_D,w,l,n)|$$

Semantic Similarity

Analogously, we can define semantic similarity.

$$Sim_H(w,k,l) = \frac{1}{n}|TopN(P_H,w,k,n) \cap TopN(P_H,w,l,n)|$$

where

$$P_H(w,k,h) \propto \frac{1}{d(w,h)} \sum_{w' \in NN(w,k)} H(p(w'),h) \cdot \frac{1}{d(p(w'),h)}$$

Overall Similarity

We sum the domain similarity and the semantic similarity to get the overall similarity of two sense.

$$Sim(w,k,l) = Sim_D(w,k,l) + Sim_H(w,k,l)$$

With a development dataset, we set a threshold θ for regarding two senses of a word as pseudo multi-sense.

Detected Samples

Here shows nearest neighbors three "senses" of word *rock* from the released vectors in [2].

- **rock**$_1$: metal, rippling, psychedelia, bands, pop
- **rock**$_2$: sand, rocks, butte, ash, sandy, cedar
- **rock**$_3$: hip, indie, hop, reggae, roll, rock/metal

Our algorithm shows that rock$_1$ and rock$_3$ are likely to be pseudo multi-sense, while rock$_2$ has a different meaning with them.

Pseudo Multi-Sense Elimination

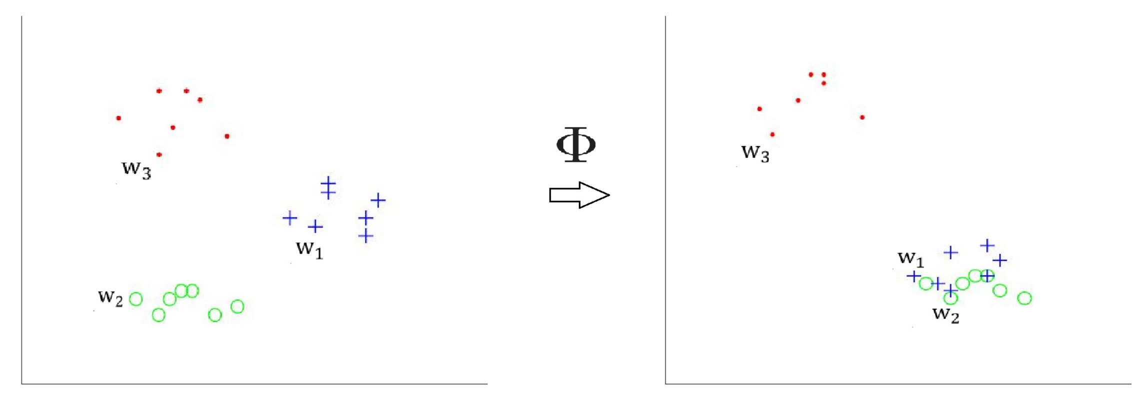

Figure 1: How the global transform matrix works. In the case, w$_1$, w$_2$ and w_3 are senses produced by previous word embedding algorithm automatically. We detect w$_1$ and w$_2$ have the same meaning in fact while w$_3$ has a different meaning.

Having the existing word embeddings, assume that we have a detected pseudo multi-sense group $G = \{w_{k_1}, w_{k_2}, ..., w_{k_n}\}$, in which $w_{k_1}, w_{k_2}, ..., w_{k_n}$ are senses of word w, taking the same meaning. Thus, we can find a representative vector for the group. Let $v_s(w,k_i)$ be the corresponding vectors of w_{k_i}, and $v_r(G)$ be the representative vector for the group G. Such vector $v_r(G)$ can be randomly chosen from $\{v_s(w,k_1), v_s(w,k_2), ..., v_s(w,k_n)\}$, or simply the mean vector of them. Other methods to compute $v_r(G)$ are also worth trying if reasonable.

Assume there is a transition matrix, by which for all pseudo multi-sense group G, $\forall w_{k_i} \in G$, $v_{w_{k_i}}$ can be projected to $v_r(G)$. In other words, we suggest that there exists a global matrix Φ, for any given pseudo multi-sense group $G = \{w_{k_1}, w_{k_2}, ..., w_{k_n}\}$ and its representative vector $v_r(G)$, we have

$$v_r(G) = \Phi * v_s(w,k_i), \forall w_{k_i} \in G, \forall G$$

Thus we can eliminate pseudo multi-sense by applying Φ to the original vector space.

Experiments

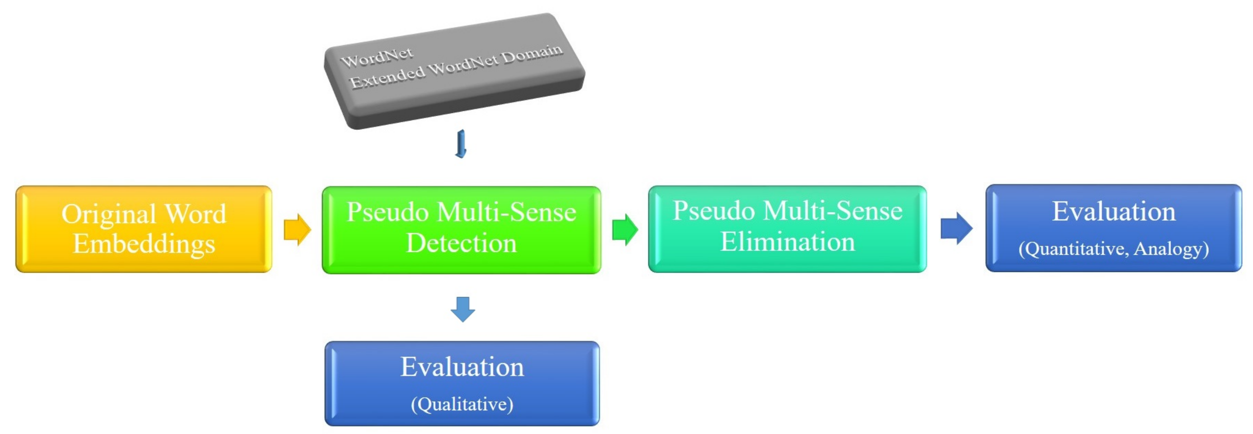

Figure 2: The pipeline of our experiment.

Local Word Similarity

We evaluate our algorithm by applying it to the word vectors released by Huang et al.([1]) and Neelakantan et al.([2]), and comparing the word similarity on SCWS dataset([1]). The experiment shows that our algorithm gains a significant improvement on localSim metric, which is defined as

$$localSim(w,w') = s(v_s(w,k), v_s(w',k'))$$

where $k = \text{argmax}_i\, P(w,c,i)$, $k' = \text{argmax}_{i'}\, P(w',c',i')$.

Original Vectors (Model)	localSim orig	trans
Huang et al.	26.1	**37.6**
MSSG 50d	49.2	**53.2**
MSSG 300d	57.3	**62.2**

Table 1: The performance of the original vectors and that of the vectors after transformation, on localSim metric.

Analogy

Analogy task evaluate the performance of word vectors on quadruples (a,b,c,d) with the equation $v(a) - v(b) = v(c) - v(d)$. The quadruples can be (Beijing, China, Paris, France), or (great,greater,tough,tougher) etc. The experiment shows that our pseudo multi-sense eliminating algorithm improves the vectors' performance on analogy task.

Model	Semantic		Syntactic	
	orig	trans	orig	trans
Huang et al.	52.8	**53.5**	53.5	**56.1**
MSSG 50d	75.8	**77.5**	85.2	**88.0**
MSSG 300d	92.0	**93.1**	93.3	**94.5**

Table 2: Evaluation result on analogy task.

References

[1] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *ACL 2012*

[2] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *ACL 2014*

[3] George A Miller Wordnet: a lexical database for English. 1995, Communications of the ACM.

[4] Aitor GonzÃąlez, German Rigau, and Mauro Castillo. A graph-based method to improve wordnet domains. In *ICITPCL 2012*

Acknowledgement

This work is supported by the National Natural Science Foundation of China (grant No.61472017, M1552004).