Constructing High Quality Sense-specific Corpus and Word Embedding via Unsupervised Elimination of Pseudo Multi-sense

Haoyue Shi, Xihao Wang, Yuqi Sun, Junfeng Hu
{hyshi, victorwonder, sun_yq, hujf}@pku.edu.cn
School of Electronics Engineering and Computer Science, Peking University

ABSTRACT

Multi-sense word embedding is an important extension of neural word embeddings. By leveraging context of each word instance, multi-prototype version of word embeddings were accomplished to represent the multisenses. Unfortunately, this kind of context based approach inevitably produces multiple senses which should actually be a single one, suffering from the various context of a word.

In this work, we propose a novel framework for unsupervised corpus sense tagging, and evaluate our framework by training word embeddings with the obtained sense specific corpus. On the tasks of word similarity, word analogy as well as sentence understanding, the embeddings trained on sense-specific corpus obtain better results than the basic multi-sense word embeddings.

PROPOSED FRAMEWORK

Most of existing multi-sense word embeddings consist of three types of vectors:

- **global vector.** Each word in vocabulary is embedded into a high dimensional space, i.e., one word one vector.
- **sense vector.** Each sense of word is embedded to the sense vector space, which is the main part of multi-sense word embeddings.
- **context center vector.** For each instance of a word in the corpus, we can compute its context center vector by averaging the global vector of its context words. We leverage all the instance context vectors of instances with the same sense to obtain context center vectors. Each sense vector is associated with a context center vector.

Our proposed framework contains the following five steps:

1. (1) Train multi-sense word embeddings on the given corpus, using existing multi-sense word embedding frameworks.
2. (2) Detect pseudo multi-senses in the obtained embeddings without supervision.
3. (3) Select sense for each instance in the corpus and tag pseudo multi-sense as one sense.
4. (4) Retrain multi-sense word embeddings on corpus, with selected sense tags in step (3).
5. (5) Repeat step (2)-(4) until no pseudo multi-sense is detected.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (grant No.61472017, M1752015).

REFERENCES


PSEUDO MULTI-SENSE DETECTION

For a given word instance, we evaluate neighborhood similarity of its two sense vectors \( v_{w,i} \) and \( v_{w,j} \) by

\[
\text{Sim}(v_{w,i}, v_{w,j}) = \sum_{v_j \in kN(v_{w,j})} \cos(v_j(v_{w,i}), v_j(v_{w,j}))
\]

where \( kN(v) \) indicates the \( k \) nearest neighbors set of vector \( v \) in the same space, \( v_{w,i}, v_{w,j} \) are all sense vectors, \( v_j(v_k) \) is the corresponding global vector of the sense vector \( v_k \). If \( \text{Sim}(v_{w,i}, v_{w,j}) \) is larger than an arbitrarily chosen threshold \( \theta \), then \( v_{w,i} \) and \( v_{w,j} \) are considered to be one sense.

PSEUDO MULTI-SENSE ELIMINATION

For a given word instance, we find its sense according to its context instance vector and context center vectors of senses. The context vector of each instance \( w_i \) is computed by

\[
v_{C(w_i)} = \frac{1}{|C(w_i)|} \sum_{t \in C(w_i)} v_t(f)
\]

where \( C(w_i) \) is the context set of word instance \( w_i \), and \( v_t(f) \) is the global vector of word \( t \).

Therefore, we have

\[
P(\text{Sense}(w_i) = k|C(w_i)) \propto v_{C(w_i)} \cdot v_{s(w_i,k)}
\]

where \( v_{s(w_i,k)} \) is the context center vector of the \( k^\text{th} \) sense of word \( w \) (prototype of \( w_i \)).

To make tagging procedure more accurate, we apply self-paced learning strategy [2]. During each iteration, we only tag the instances with high level confidence. We reformulate the problem as the following self-paced learning function:

\[
\min_{\mathcal{L}_{\text{SPL}}} = \frac{1}{n} \sum_{i=1}^{n} a_i (1 - \max \cos(t_i, c_{w_i,j})) + \lambda(f(\alpha); \lambda)
\]

where \( n \) is the number of word instances, \( a = (0,1)^n \) is the weight for each instance. Here we apply a typical binary self-paced function [2]

\[
f(\alpha; \lambda) = -\lambda|a||1 - \lambda \sum_{i=1}^{n} a_i
\]

for the self-paced learning schema. At each time, we only tag those instances with \( a_i = 1 \).

EVALUATION & DISCUSSIONS

Word Similarity

We follow define the similarity of two instances \( w_1 \) and \( w_2 \) by

\[
\text{localSim}(w_1, w_2) = \cos(v_{s(w_1,k_1)}, v_{s(w_2,k_2)})
\]

where

\[
k_i = \text{argmax} P(\text{Sense}(w_i) = k|C(w_i))
\]

is the selected sense w.r.t. the context center vectors and the context instance vector. Different from weighted sum (i.e., soft similarity), the localSim metric helps evaluate how well the model captures the meaning for each learned sense in a "hard" way.

Sentence Understanding

We provide three baselines to further show that eliminating pseudo multi-sense is helpful to improve the performance downstream sentence understanding tasks. We feed bag-of-words features to SentEval system [3] to evaluate each type of word embeddings. Similarly to the word similarity task, we select the most probable sense for each word in the sentences.

<table>
<thead>
<tr>
<th>Model</th>
<th>50D</th>
<th>300D</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSSG</td>
<td>49.2</td>
<td>57.3</td>
</tr>
<tr>
<td>NP-MSSG</td>
<td>50.9</td>
<td>57.5</td>
</tr>
<tr>
<td>MSSG + Matrix</td>
<td>53.2</td>
<td>62.2</td>
</tr>
<tr>
<td>NP-MSSG + Matrix</td>
<td>52.2</td>
<td>61.4</td>
</tr>
<tr>
<td>NP-MSSG + SPT (Ours)</td>
<td>58.6</td>
<td>63.7</td>
</tr>
</tbody>
</table>

Table 1: The performance of vectors on SCWS dataset, with the localSim metric. Our method outperforms the basic model, as well as eliminating pseudo multi-sense by directly training a transformation matrix.

Further Discussion

In addition to presenting a framework to construct sense-specific corpus, we also show that eliminating pseudo multi-sense in multi-sense word embeddings is also helpful to natural language understanding. We also suggest that the level of pseudo multi-sense generation, which can be calculated without any external resource, can be treated as a metric to evaluate the quality of word embeddings.

Table 2: The performance of vectors on sentence understanding tasks.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSRP</th>
<th>SUBJ</th>
<th>TREC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP-MSSG</td>
<td>70.03</td>
<td>90.96</td>
<td>78.4</td>
</tr>
<tr>
<td>NP-MSSG + Matrix</td>
<td>70.55</td>
<td>91.20</td>
<td>83.2</td>
</tr>
<tr>
<td>NP-MSSG + SPT (Ours)</td>
<td>71.01</td>
<td>90.97</td>
<td>84.2</td>
</tr>
</tbody>
</table>