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

Haoyue Shi[1,2], **Xihao Wang**[1,3], **Yuqi Sun**[1,3], **Junfeng Hu**[1,4]
[1]School of Electronics Engineering and Computer Science
[2] Department of Machine Intelligence
[3] Department of Computer Science and Technology
[4] Key Laboratory of Computational Linguistics, Ministry of Education
No 5. Yiheyuan Road, Beijing China, 100871
{hyshi, victorwonder, sun_yq, hujf}@pku.edu.cn

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 multi-senses. 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. (Shi et al., 2016) used WordNet to evaluate the neighborhood similarity of each sense pair to detect such pseudo multi-senses. In this paper, a novel framework for unsupervised corpus sense tagging is presented, which mainly contains four steps: (a) train multi-sense word embeddings on the given corpus, using existing multi-sense word embedding frameworks; (b) detect pseudo multi-senses in the obtained embeddings, without requirement to any extra language resources; (c) label each word in the corpus with a specific sense tag, with respect to the result of pseudo multi-sense detection; (d) re-train multi-sense word embeddings with the pre-selected sense tags. We 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 strategy which is applied in step (a).

Keywords: multi-sense word embedding, word sense discovery, pseudo multi-sense

1. Introduction

Distributional word representations (Bengio et al., 2003; Collobert and Weston, 2008; Mnih and Hinton, 2009; Mikolov et al., 2013b) embed words into a high dimensional space, where the cosine value (or Euclidean distance) between two vectors can somehow represent the similarity of two words. Multi-sense word embedding is an intuitive extension of distributional representations of words. Most of existing works distinguish different senses of words by their contexts (Schütze, 1998; Huang et al., 2012; Pina and Johansson, 2015; Neelakantan et al., 2014; Li and Jurafsky, 2015; Cheng and Kartsaklis, 2015). Besides proposing a Chinese Restaurant Process (CRP) model for multi-sense word embeddings, Li and Jurafsky (2015) also discussed how helpful the multi-sense word embedding methods are to improve natural language understanding, and proved that they do help on some tasks. However, these methods have a common defect. Existing multi-sense word embedding models generate large amount of pseudo multi-senses (Shi et al., 2016). This phenomenon leads to the vagueness of such distributional word representations. It also has an intuitive explanation. Here shows three appearances of the word "cat(s)" from Wikipedia:

- In many countries, **cats** are believed to have nine lives, but in Italy, Germany, Greece, Brazil and some Spanish-speaking regions, they are said to have seven lives, while in Turkish and Arabic traditions, the number of lives is six.
- Female **cats** are seasonally polyestrous, which means they may have many periods of heat over the course of a year, the season beginning in spring and ending in late autumn.
- **Cats** hunt small prey, primarily birds and rodents, and are often used as a form of pest control.

The sentences present large topic shift, but we human beings can still easily determine the three "cat"s have exactly the same meaning, which is the kind of animal. In contrast, it is too easy for a context-based method to view them as different senses and learn a separate vector for each, if there is not an explicit constraint to penalize such behavior. Shi et al. (2016) detected pseudo multi-sense with the help of WordNet (Miller, 1995), and then trained a linear transformation which aims to maximize the similarity of detected pseudo multi-senses. On the transformed word-embedding space, they successfully observed improved performance on both word similarity and analogy tasks.

Considering the probable large cost obtaining external knowledge like WordNet for rare languages, we design a novel unsupervised pseudo multi-sense detection method, which evaluates neighborhood similarity of two sense using global (one-word-one-sense) word vectors. We then present a framework for sense tagging, which combines multi-sense word embedding and pseudo multi-sense detection. For tokenized corpus, our framework contains the following steps:

1. Train multi-sense word embeddings on corpus, using a specific basic (multi-sense word embedding) model.
2. Detect pseudo multi-sense in the embeddings.
3. Select sense for each instance in the corpus. Pseudo multi-sense would be tagged as one sense.

4. Re-train multi-sense word embeddings on corpus, with pre-selected sense tags in step 3.

5. Repeat step 2-4 until no pseudo multi-sense is detected.

To the best of our knowledge, we are the first to create a framework for unsupervised sense-tagging regarding elimination of pseudo multi-sense. Our framework is able to accept any context-based multi-sense word embedding method as a basic model. Moreover, for those word-based languages which are lack of studies, our framework would also shed light on automatic sense discovery, which has less redundancy than the basic multi-sense word embedding model.

2. Related Work

Multi-Sense Word Embedding Neural multi-sense word embedding is a well-studied problem after the emergence of neural word embeddings (Huang et al., 2012; Pina and Johansson, 2015; Neelakantan et al., 2014; Li and Jurafsky, 2015; Cheng and Kartsaklis, 2015; Liu et al., 2015; Qiu et al., 2016). Most of them select sense for word instances with respect to their context. The producing of multi-sense word embeddings is also the procedure of word-sense discovery. In our framework, all of the models mentioned above could be applied as the basic model, which would be improved after re-training with pseudo multi-sense detection and tagging.

Different from those global word embedding models (Pennington et al., 2014; Mikolov et al., 2013b), most of the existing multi-sense word embeddings contain three types of vectors:

- *global vector.* Each word in vocabulary is embedded into a high dimensional space, aka, one word one vector.
- *context vector.* For each instance of a word in the corpus, we can compute its *instance context vector* by averaging the *global vector* of its context words. By applying SoftMax to the dot production between *instance context vector* and all given *context vector* of senses, we could obtain the probability for the word instance to be classified to each sense.
- *sense vector.* Each sense of word is embedded to the sense vector space, which is the main part of multi-sense word embeddings.

Another type of multi-sense word embeddings introduce external knowledge base for accurate sense generation (Iacobacci et al., 2015; Chen et al., 2014; Pelevina et al., 2017). However, they are somehow limited as such external knowledge may be lack for languages other than English.

From the perspective of sense definition, Flekova and Gurevych (2016) proposed a model to learn representations for supersenses of words, which performs well on the evaluation tasks.

Sense Discovery Another related work is word sense discovery (Rapp, 2003), which used co-occurrence to evaluate semantic similarity. In our work, we not only indicate senses with the method which is based on similar knowledge (Levy and Goldberg, 2014), but also improve the associate distributional word representations.

Self-Paced Learning Self-paced learning is an adaptive weight adjust technique introduced by Kumar et al. (2010). Varieties of strategies of self-paced learning have been proved efficient on matrix factorization (Jiang et al., 2015), multimedia event detection (Jiang et al., 2014; Jiang et al., 2015), multimedia search (Jiang et al., 2014), place recognition (Shi et al., 2017) and object detection (Tang et al., 2012). As far as we know, this is the first work that adapts self-paced strategy to learn word embeddings.

3. Proposed Framework

Our proposed framework contains the following five steps:

1. Train multi-sense word embeddings on corpus, utilizing a specific basic (multi-sense word embedding) model.

2. Detect pseudo multi-sense in the embeddings. Suppose we have two sense vectors $\mathbf{v}_{w,i}$ and $\mathbf{v}_{w,j}$ of the same word w. We evaluate the neighborhood similarity by

$$P_{pse}(\mathbf{v}_{w,i}, \mathbf{v}_{w,j}) \propto \sum_{\substack{\mathbf{v}_{n_i} \in kNN(\mathbf{v}_{w,i}),\\ \mathbf{v}_{m_j} \in kNN(\mathbf{v}_{w,j})}} \cos(\mathbf{v}_g(\mathbf{v}_{n_i}), \mathbf{v}_g(\mathbf{v}_{m_j})) \qquad (1)$$

where $kNN(\mathbf{v})$ indicates the k nearest neighbors set of vector $\mathbf{v}$ in the same space, $\mathbf{v}_{w,i}, \mathbf{v}_{w,j}, \mathbf{v}_{n_i}, \mathbf{v}_{m_j}$ are all *sense vectors*, $\mathbf{v}_g(\mathbf{v}_l)$ is the corresponding *global vector* of the sense vector $\mathbf{v}_l$ (multiple *sense vectors* may have the same *global vector*). To determine whether two senses $\mathbf{v}_{w,i}$ and $\mathbf{v}_{w,j}$ of the word w are pseudo multi-sense, we choose an arbitrary threshold θ. If $P_{pse}(\mathbf{v}_{w,i}, \mathbf{v}_{w,j}) > \theta$, then $\mathbf{v}_{w,i}$ and $\mathbf{v}_{w,j}$ should be treated as one sense rather than the separated two.

In practice, we determine θ by the following procedure. We collect the set of pairs $S = \{(w_i, w_j)\}$, of which the neighborhood of w_i has prototype-level overlap with that of w_j. For example, if cat_0 has dog_0 in its nearest neighbors, while cat_1 has dog_1, (cat_0, cat_1) should be in S. We sort $P_{pse}(\mathbf{v}_{w,i}, \mathbf{v}_{w,j})$ for each pair of $(w_i, w_j) \in S$ in descending order, and choose the value at 90% point as θ to avoid noise. This is similar to the spy technique introduced by Liu et al. (2003). We evaluate the 20 nearest neighbors for each sense.

3. Select sense for each instance in the corpus. Pseudo multi-sense would be tagged as one sense. The existence of *context vector* gives us an easy way to compute the probability of a word instance belongs to each

sense. We compute the context vector of each instance w_i by

$$\mathbf{v}_{\mathcal{C}(w_i)} = \frac{1}{|\mathcal{C}(w_i)|} \sum_{t \in \mathcal{C}(w_i)} \mathbf{v}_g(t) \tag{2}$$

where $\mathcal{C}(w_i)$ is the context set of word instance w_i which contains prototype of words, and $\mathbf{v}_g(t)$ is the global vector of word t.

Therefore, we have

$$P(Sense(w_i) = k|\mathcal{C}(w_i)) \propto \mathbf{v}_{\mathcal{C}(w_i)} \cdot \mathbf{v}_c(w_i, k) \tag{3}$$

where $\mathbf{v}_c(w_i, k)$ is the *context vector* of the k^{th} sense of word w (prototype of w_i). This would depend on different settings of different models: in CRP model (Li and Jurafsky, 2015), the right side should be activated by sigmoid function and then multiplied by $Prob(w)$, which represents the probability of word w to appear in the corpus.

There often exists some instances we are not confident to determine which sense it should belong to, e.g. $\exists j, k, (j \neq k)$, $P(Sense(w_i) = k|\mathcal{C}(w_i))$ is similar to $P(Sense(w_i) = j|\mathcal{C}(w_i))$. To solve this problem, we apply self-paced learning strategy (Kumar et al., 2010). During each iteration, we only tag the instances with high level confidence. We reformulate the problem as the following one:

$$\min_{\mathbf{a}} \mathcal{L}_{SPL}(Emb; \lambda) = \\ \sum_{i=1}^{n} a_i (1 - \max_k P(Sense(w_i) = k|\mathcal{C}(w_i))) + f(\mathbf{a}; \lambda) \tag{4}$$

where Emb is the learned multi-sense word embedding, n is the number of word instances, $\mathbf{a} = \{0, 1\}^n$ is the weight for each instance, $f(\mathbf{a}; \lambda)$ is the self-paced learning function. Here we apply a typical binary self-paced function (Kumar et al., 2010)

$$f(\mathbf{a}; \lambda) = -\lambda ||\mathbf{a}||_1 = -\lambda \sum_{i=1}^{n} a_i \tag{5}$$

for the self-paced learning schema. At each time, we only tag those instances with $a_i = 1$. Therefore, by gradually increase λ, we can fetch more less-confident instances for our sense tagging.

4. Re-train multi-sense word embeddings on corpus, with pre-selected sense tags in step 3.

5. Repeat step 2-4 until no pseudo multi-sense is detected.

Our algorithm is summarized in Algorithm 1.

4. Evaluation

We apply Non-parametric Multi-sense Skip Gram (Neelakantan et al., 2014) as the basic model to train multi-sense word embeddings on Wikipedia Corpus (Soriano-Morales et al., 2017).

Algorithm 1 Enhanced pipeline for multi-sense word embedding

```
Require: training corpus $T$, multi-sense word embedding model $M$, self-paced function $f$ and step size $\mu$
Ensure: Multi-sense word embeddings $W$, tagged corpus $T'$
Initialize $W$ by training $M$ with $T$
Initialize $\lambda$
while not converged do
    Detect pseudo multi-sense with Eq (1), $W$
    Compute transitive closure of detected pseudo multi-sense relation
    Select confident instances in $T$ by Eq (4) and (5)
    Add sense tags for confident instances and get updated corpus $T'$, pseudo multi-senses will be given same sense tags
    Train $M$ with $T'$ to obtain updated $W$
    increase $\lambda$ by $\mu$
end while
return $W, T'$
```

Model	50d	300d
MSSG (Neelakantan et al., 2014)	49.2	57.3
NP-MSSG (Neelakantan et al., 2014)	50.9	57.5
MSSG + MT (Shi et al., 2016)	53.2	62.2
NP-MSSG + MT (Shi et al., 2016)	52.2	61.4
NP-MSSG + SPT	**58.6**	**63.7**

Table 1: Spearman rank correlation on SCWS dataset. For baselines, we evaluate the models of multi-sense skip-gram (MSSG) and non-parametric multi-sense skip-gram (NP-MSSG), as well as the combination of those with supervised pseudo multi-sense detection and matrix transformation (MT). NP-MSSG + SPT refers to our self-paced tagging model.

4.1. Word Similarity

Stanford Contextual Word Similarity (SCWS) dataset (Huang et al., 2012) is a reliable and professional dataset to estimate the performance of word embeddings, especially multi-sense word embeddings. It contains 2,003 pairs of words together with the context.

In our experiments, we follow Neelakantan et al. (2014) to define the similarity of two instances w_1 and w_2 by

$$localSim(w_1, w_2) = \\ \cos(\mathbf{v}_{w_1, Sense(\mathcal{C}(w_1))}, \mathbf{v}_{w_2, Sense(\mathcal{C}(w_2))}) \tag{6}$$

where

$$Sense(\mathcal{C}(w_i)) = \operatorname*{argmax}_k P(Sense(w_i) = k|\mathcal{C}(w_i)) \tag{7}$$

is the sense index chosen by $\mathcal{C}(w_i)$ and Eq (3). Table 1 shows that our framework outperforms not only the basic model, but also the linear (matrix) transformation one proposed by Shi et al. (2016).

4.2. Word Analogy

To evaluate how embeddings capture intuitive pairwise word relations, Mikolov et al. (2013a) released analogy task, which contains 19,544 quadruples in total. Each quadruple contains four words A, B, C, D, where word A is similar to word B in the same way as word C is similar to D, for instance, (Berlin, Germany, Paris, France). Among all, there are 5 classes of semantic quadruples and 9 classes of syntactic ones.

For the evaluation of multi-sense word embeddings, we follow the method proposed by Shi et al. (2016): for given multi-sense word embeddings, if there is a index quadruple (i, j, k, l) for word quadruple (A, B, C, D) s.t. $v_{A,i} - v_{B,j} + v_{D,l}$ is most similar to $v_{C,k}$, then we treat (A, B, C, D) is a correct case for the model. In addition, considering the symmetric property of the quadruples, there are totally four linear combinations for each quadruple to be evaluated, and we treat the quadruple as a positive case if one of them is satisfied.

Model	Semantic	Syntactic
MSSG (50d)	75.8	85.2
NP-MSSG (50d)	74.6	80.7
MSSG + MT (50d)	**77.5**	**88.0**
NP-MSSG + MT (50d)	75.6	82.3
NP-MSSG + SPT (50d)	76.2	86.1
NP-MSSG (300d)	83.9	89.0
NP-MSSG + MT (300d)	**85.9**	**90.2**
NP-MSSG + SPT (300d)	85.3	89.0

Table 2: Accuracy on word analogy task.

According to Table 2, we see our self-paced tagging framework ensures that the learned word embeddings keeps the semantic and syntactic relations well, although it performs not as well as the one with matrix transformation (Shi et al., 2016).

4.3. Sentence Understanding

We evaluate the quality of word embeddings with the SentEval system (Conneau et al., 2017; Kiela et al., 2017). In the evaluation, bag of words (BoW) is fed to the system as sentence features. We report the accuracies of two models on three different tasks in Table 3. The tasks are paraphrase detection (MSRP; Dolan et al. (2004)), subjectivity detection (SUBJ; Pang and Lee (2004)), and question classification (TREC; Voorhees and Buckland (2003)). On the tasks of MSRP and TREC, self-paced tagging with respect to elimination of pseudo multi-sense improves word embeddings at a non-trivial level.

4.4. Case Study

We show the k-nearest neighbors of some words in Table 4. We clearly see that the self-paced tagging model can not only eliminate the pseudo multi-senses (Norway, star, algorithm), but also keep the real multi-senses (star).

Model	MSRP	SUBJ	TREC
NP-MSSG	70.03	90.96	78.4
NP-MSSG + MT	70.55	**91.20**	83.2
NP-MSSG + SPT	**71.01**	90.97	**84.2**

Table 3: Evaluation result on transfer learning tasks.

Norway	
NP-MSSG	Denmark, Troms, Sogn, Hedmark Denmark, Sweden, Finland, Iceland Denmark, Sweden, Finland, Netherlands Denmark, Austria, Germany, Belgium
+ SPT	Denmark, Norwegian, Sweden, Trondheim

star	
NP-MSSG	stars, wars, alongside, beetlejuice stars, award, eagle, two-time supergiant, constellation, aurigae
+ SPT	stars, movie, superstar, MVP supergiant, stars, g5v, white_main

algorithm	
NP-MSSG	hash, algorithms, quick_sort, recursive algorithms, optimization, public-key
+ SPT	algorithms, computation, iteratively

Table 4: Case study of k nearest neighbors. Each row refers to a learned "sense".

5. Conclusion

In this work, we present a novel framework for corpus sense tagging with unsupervised elimination of pseudo multi-sense, utilizing any multi-sense word embedding model as the basic model. By applying the proposed self-paced tagging strategy, we could improve the quality of multi-sense word embeddings. Experiments have shown the efficiency of our framework.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (grant No.61472017, M1752015). The authors would like to thank Xian Wu for his help on experimental environment.

7. Bibliographical References

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *JMLR*, 3(Feb).

Chen, X., Liu, Z., and Sun, M. (2014). A unified model for word sense representation and disambiguation. In *EMNLP*.

Cheng, J. and Kartsaklis, D. (2015). Syntax-aware multi-sense word embeddings for deep compositional models of meaning. In *Proc. EMNLP*.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proc. ICML*.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proc. EMNLP*.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proc. COLING*.

Flekova, L. and Gurevych, I. (2016). Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proc. ACL*.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proc. ACL*.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). Sensembed: Learning sense embeddings for word and relational similarity. In *Proc. ACL*.

Jiang, L., Meng, D., Mitamura, T., and Hauptmann, A. G. (2014). Easy samples first: Self-paced reranking for zero-example multimedia search. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM.

Jiang, L., Meng, D., Zhao, Q., Shan, S., and Hauptmann, A. G. (2015). Self-paced curriculum learning. In *AAAI*, volume 2.

Kiela, D., Conneau, A., Jabri, A., and Nickel, M. (2017). Learning visually grounded sentence representations. *arXiv preprint arXiv:1707.06320*.

Kumar, M. P., Packer, B., and Koller, D. (2010). Self-paced learning for latent variable models. In *NIPS*.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *NIPS*.

Li, J. and Jurafsky, D. (2015). Do multi-sense embeddings improve natural language understanding? In *Proc. EMNLP*.

Liu, B., Dai, Y., Li, X., Lee, W. S., and Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. In *ICDM*.

Liu, Y., Liu, Z., Chua, T.-S., and Sun, M. (2015). Topical word embeddings. In *AAAI*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *NIPS*.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Mnih, A. and Hinton, G. E. (2009). A scalable hierarchical distributed language model. In *NIPS*.

Neelakantan, A., Shankar, J., Passos, A., and McCallum, A. (2014). Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proc. EMNLP*.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. ACL*.

Pelevina, M., Arefyev, N., Biemann, C., and Panchenko, A. (2017). Making sense of word embeddings. *arXiv preprint arXiv:1708.03390*.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proc. EMNLP*.

Pina, L. N. and Johansson, R. (2015). A simple and efficient method to generate word sense representations. In *Proceedings of Recent Advances in Natural Language Processing*.

Qiu, L., Tu, K., and Yu, Y. (2016). Context-dependent sense embedding. In *Proc. EMNLP*.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the ninth machine translation summit*.

Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics*, 24(1).

Shi, H., Li, C., and Hu, J. (2016). Real multi-sense or pseudo multi-sense: An approach to improve word representation. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*.

Shi, H., Chen, J., and Hauptmann, A. G. (2017). Joint saliency estimation and matching using image regions for geo-localization of online video. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*.

Tang, K., Ramanathan, V., Fei-Fei, L., and Koller, D. (2012). Shifting weights: Adapting object detectors from image to video. In *Advances in Neural Information Processing Systems*.

Voorhees, E. M. and Buckland, L. (2003). Overview of the trec 2003 question answering track. In *TREC*, volume 2003.

8. Language Resource References

Soriano-Morales, Edmundo-Pavel and Ah-Pine, Julien and Loudcher, Sabine. (2017). *Syntactically Annotated Wikipedia Dump*. ISLRN 958-233-248-056-1.