That Makes Sense: Joint Sense Retrofitting from Contextual and Ontological Information

Ting-Yu Yen[1*], Yang-Yin Lee[1*], Hen-Hsen Huang[1], Hsin-Hsi Chen[1,2]
[1] Department of Computer Science and Information Engineering, National Taiwan University
Taipei, 10617 Taiwan
[2] MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
{tyyen, yylee, hhhuang}@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

ABSTRACT

While recent word embedding models demonstrate their abilities to capture syntactic and semantic information, the demand for sense level embedding is getting higher. In this study, we propose a novel joint sense embedding learning model that retrofits word representation into sense representation from contextual and ontological information. The experiment shows the effectiveness and robustness of our model that outperforms previous approaches in four public available benchmark datasets.

CCS CONCEPTS

•**Computing methodologies** → **Artificial intelligence** → **Natural language processing** → Lexical semantics

KEYWORDS

Semantic relatedness; sense embedding; joint sense retrofitting

1 INTRODUCTION

Recently, there has been an intensive research activity in studying word embedding [1]. However, most of the word embedding models use one vector to represent a word, and are problematic in some natural language processing applications that require sense level representation (e.g., word sense disambiguation). As a result, some researches try to resolve the polysemy and homonymy issue and introduce sense level embedding [2–4]. In this research, we propose a novel joint sense embedding learning algorithm that retrofits a trained word embedding using contextual and ontological information.

ACM Reference format:

T.Y. Yen, Y.Y. Lee, H.H. Huang, and H.H. Chen. 2018. That Makes Sense: Joint Sense Retrofitting from Contextual and Ontological Information. In *The 2018 Web Conference Companion (WWW 2018), April 23-27, 2018, Lyon, France*, ACM, New York, NY, 2 pages. DOI: https://doi.org/10.1145/3184558.3186906

Our proposed *joint sense retrofitting* is a post-processing method for generating low-dimensional sense embedding inspired from *sense retro* [3]. Although some studies adopt ontological information into sense embedding model [2, 3], it is the first time of employing the ontological and contextual information simultaneously. Given a trained word embedding and a lexical ontology that contains sense level relationships (e.g., synonym, hypernym, etc.), our model generates new sense vectors via constraining the distance between the sense vector and its word form vector, its sense neighbors and its contextual neighbors. In the experiment, we show that our proposed joint sense retrofitting model outperforms previous approaches in four benchmark datasets, and demonstrates robustness from two ontologies, WordNet[1] and Roget's Thesaurus[2].

2 JOINT SENSE RETROFITTING

Let $V = \{w_1, \dots, w_n\}$ be a vocabulary of a trained word embedding and $|V|$ be its size. The matrix $\hat{Q}$ will be the pre-trained collection of vector representations $\hat{q}_i \in \mathbb{R}^d$, where d is the dimensionality of a word vector. Each $w_i \in V$ is learned using a standard word embedding technique (e.g., GloVe [1]). Let $\Omega = (T, E)$ be an ontology that contains the semantic relationship, where $T = \{t_1, \dots, t_m\}$ is a set of senses and $|T|$ is total number of senses. The edge $(i, j) \in E$ indicates a semantic relationship of interest (e.g., synonym) between t_i and t_j. The edge set E can be further split into two disjoint subsets E_s and E_c. $(i, j) \in E_s$ if and only if there is more than one sense of the word form of t_j, while $(i, j) \in E_c$ if and only if t_j has only one sense. In our model, we use the word form vector to represent the neighbors of the t_is in E_c. Those neighbors are viewed as contextual neighbors as they learned from the context of a corpus. We use $\hat{q}_{t_j}$ to denote the word form vector of t_j (one should notice that $\hat{q}_{t_j}$ and $\hat{q}_{t_k}$ may map to the same vector representation even if $j \neq k$). Then the objective of our joint sense retrofitting model is to learn a new matrix $S = (s_1, \dots, s_m)$ such that each new sense vector is close to (1) its word form vertex, (2) its sense neighbors, and (3) its contextual neighbors. The objective to be minimized is:

$$\sum_{i=1}^{m}\left[\alpha_i\left\|s_i-\hat{q}_{t_i}\right\|^2+\sum_{(i,j)\in E_c}\beta_{ij}\left\|s_i-\hat{q}_{t_j}\right\|^2+\sum_{(i,k)\in E_s}\gamma_{ik}\left\|s_i-s_k\right\|^2\right] \quad (1)$$

where α, β and γ control the relative strength of the sense relations. We therefore apply an iterative updating method to the

[*] These authors contributed equally to this work.

WWW '18 Companion April 23-27, 2018, Lyon, France.

ACM ISBN 978-1-4503-5640-4/18/04.

DOI: https://doi.org/10.1145/3184558.3186906

[1] https://wordnet.princeton.edu

[2] http://www.thesaurus.com

solution of the above convex objective function [5]. Initially, the sense vectors are set to their corresponding word form vectors (i.e., $s_i \leftarrow \hat{q}_{t_i} \, \forall i$). Then in the following iterations, the updating formula for s_i would be:

$$s_i = \frac{\sum_{k:(i,k)\in E_s} \gamma_{ik} s_k + \sum_{j:(i,j)\in E_c} \beta_{ij} \hat{q}_{t_j} + \alpha_i \hat{q}_{t_i}}{\sum_{k:(i,k)\in E_s} \gamma_{ik} + \sum_{j:(i,j)\in E_c} \beta_{ij} + \alpha_i} \qquad (2)$$

Experimentally, 10 iterations are sufficient to minimize the objective function from a set of starting vectors to produce effective sense retrofitted vectors.

3 DATASETS AND EXPERIMENTAL SETUP

We downloaded four benchmark datasets from the web: MEN [6], MTurk [7], Rare Words (RW) [8] and WordSim353 (WS353) [9]. For measuring the semantic similarity between a word pair in the datasets, we adopt the sense evaluation metrics AvgSim and MaxSim [4]. We select GloVe as our pre-trained word embedding model, which is trained on Wikipedia and Gigaword-5 (6B tokens, 400k vocab, uncased, 50d vectors). In testing phase, if a test dataset has missing words, we use the average of all sense vectors to represent the missing word. Note that our reported results of vanilla sense embedding may be slightly different from other researches due to the treatment of missing words. However, within this research the reported performance can be compared due to the same missing word processing method. We adopt two ontologies in our experiment: WordNet (WN) and Roget's 21st Century Thesaurus (Roget). In WN, the relation specific weights βs and γs are set to 1.0 for synonyms and 0.5 for hypernyms or hyponyms. Unlike WN, Roget does not have the synset type. As a result, we manually built a synonym ontology from the resource. In Roget, there are three levels of synonym relationship, and we set βs and γs to 1.0, 0.6 and 0.3 for the nearest to the farthest synonyms, respectively. For each sense, α is set to the sum of all the relation specific weights βs and γs. Table 1 shows a summary of the benchmark datasets and their relationship with the ontologies. In Table 1, row 3 and row 4 are the number of words that are both listed in the datasets and the ontologies. The word counts in WN and Roget are 83,118 and 47,229, respectively.

Table 1: **Summarization of the benchmark datasets**

	MEN	MTurk	RW	WS353
Pair count	3,000	287	2,034	353
Word count	751	499	2,951	437
WN	751	444	2,502	415
Roget	705	382	2,152	411

4 RESULTS AND DISCUSSION

Table 2 shows the spearman correlation (ρ × 100) of AvgSim and MaxSim between human scores and sense embedding's scores on the benchmark datasets. We compare our proposed model (*joint*) with vanilla GloVe embedding and the sense retro model (*retro*) [3]. For vanilla GloVe, we directly compute the cosine similarity of a word pair's vectors, which can be seen as a special case of AvgSim/MaxSim. From Table 2, we find that our proposed *joint* model outperforms *retro* and GloVe in all the datasets. Interestingly, although WN is bigger and covers more words than Roget, in our model the average performance with Roget is better than WN. Surprisingly, the RW's performance declined with the WN ontology. The possible reason might be WN pays more attention to common sense words than rarely occurred words. From the viewpoint of ontology, the *retro* model's performance declines in all the datasets with the smaller Roget, showing the dependency on the ontology size. In contrast, the *joint* model performs well in both the smaller Roget and larger WN ontologies, showing the robustness of our proposed model.

Table 2: **ρ × 100 of (MaxSim / AvgSim) on test datasets**

	MEN	MTurk	RW	WS353
GloVe	65.7	61.9	30.3	50.3
retro-WN [3]	62.4 / 67.7	57.4 / 60.1	15.1 / 26.9	43.9 / 51.1
retro-Roget [3]	48.7 / 52.1	47.3 / 49.3	24.4 / 26.1	27.8 / 29.4
joint-WN	64.0 / **68.9**	57.3 / 62.1	20.1 / 28.5	47.2 / 49.6
joint-Roget	**66.5** / 67.5	**62.0 / 62.6**	**32.3 / 32.5**	**50.9 / 52.8**

5 CONCLUSIONS

In summary, we propose a novel joint sense retrofitting model that utilizes the contextual and ontological information. The sense embedding is learned iteratively via constraining the distance between the sense vector and its word form vector, its sense neighbors and its contextual neighbors. Experimentally, our proposed model outperforms previous models in four benchmark datasets. We provide the source code for the model at https://github.com/y95847frank/Joint-Retrofitting.

ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-105-2221-E-002-154-MY3, MOST-106-2923-E-002-012-MY3 and MOST-107-2634-F-002-011.

REFERENCES

[1] J. Pennington, R. Socher and C.D. Manning 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014).* 12, (2014), 1532–1543.

[2] I. Iacobacci, M.T. Pilehvar and R. Navigli 2015. SensEmbed: learning sense embeddings for word and relational similarity. *Proceedings of ACL* (2015), 95–105.

[3] S.K. Jauhar, C. Dyer and E. Hovy 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2015), 683–693.

[4] J. Reisinger and R.J. Mooney 2010. Multi-prototype vector-space models of word meaning. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (2010), 109–117.

[5] Y. Bengio, O. Delalleau and N. Le Roux 2006. Label Propagation and Quadratic Criterion. *Semi-Supervised Learning.* O. Chapelle, B. Schölkopf, and A. Zien, eds. MIT Press. 193–216.

[6] E. Bruni, N.-K. Tran and M. Baroni 2014. Multimodal Distributional Semantics. *J. Artif. Intell. Res.(JAIR).* 49, (2014), 1–47.

[7] K. Radinsky, E. Agichtein, E. Gabrilovich and S. Markovitch 2011. A word at a time: computing word relatedness using temporal semantic analysis. *Proceedings of the 20th international conference on World wide web* (2011), 337–346.

[8] T. Luong, R. Socher and C.D. Manning 2013. Better Word Representations with Recursive Neural Networks for Morphology. *CoNLL* (2013), 104–113.

[9] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman and E. Ruppin 2001. Placing search in context: The concept revisited. *Proceedings of the 10th international conference on World Wide Web* (2001), 406–414.