Inducing Embeddings for Rare and Unseen Words by Leveraging Lexical Resources

Mohammad Taher Pilehvar and **Nigel Collier**
Language Technology Lab
Department of Theoretical and Applied Linguistics
University of Cambridge
Cambridge, UK
{mp792,nhc30}@cam.ac.uk

Abstract

We put forward an approach that exploits the knowledge encoded in lexical resources in order to induce representations for words that were not encountered frequently during training. Our approach provides an advantage over the past work in that it enables vocabulary expansion not only for morphological variations, but also for infrequent domain specific terms. We performed evaluations in different settings, showing that the technique can provide consistent improvements on multiple benchmarks across domains.

1 Introduction

Word representations are a core component in many natural language processing systems owing to their generalisation power, i.e., they can empower a system to share its knowledge across similar words. The prominent distributional approach to word representation (Turney and Pantel, 2010) is highly reliant on the availability of large amounts of training data and falls short of effectively modeling rare words that appear only a handful of times in the training corpus. Several efforts have been made to address this deficiency by expanding the coverage through inducing representations for rare words. Recent work has mainly focused on morphologically complex rare words has often tried to alleviate the problem by spreading the available knowledge across words that share the same morpheme (Luong et al., 2013; Botha and Blunsom, 2014; Soricut and Och, 2015). However, these techniques are unable to induce representations for words whose morphemes are not seen during training, such as infrequent domain specific terms. Importantly, the coverage issue is more evident when representations trained on abundant generic texts are applied to tasks in specific domains. As a matter of fact, the target domain can have dedicated lexical resources, such as ontologies, which are generally ignored by the distributional representation approach.

We propose a technique that exploits the knowledge encoded in lexical resources in order to expand the vocabulary of pre-trained word representations. Our approach can be applied for inducing representation not only for morphological variations but also for words whose morphemes are not seen during training, such as infrequent domain specific terms, hence giving it domain specialisation advantage. We show using different experiments that the proposed approach can provide significant improvements on multiple general and specific domain word similarity datasets.

2 Embeddings for Rare Words

The objective is to expand the vocabulary of a given set of pre-trained word embeddings $\mathcal{W}$ by adding rare words.[1] To achieve this goal, we leverage a lexical resource $\mathcal{S}$ that provides a better coverage of rare words or belongs to a specific domain and hence can be used to specialise $\mathcal{W}$ to that target domain. Our approach has two phases for inducing an embedding for a word w_r which has not been seen frequently during the training of $\mathcal{W}$ but is covered by $\mathcal{S}$. Firstly, it analyzes the lexical resource in order to extract the set of *semantic landmarks* of w_r (Section 2.1). Secondly, it induces an embedding for w_r which places the rare word in the region of the semantic space in the proximity of its semantic landmarks (Section 2.2).

Prerequisites. Our approach receives as its inputs the pre-trained word embeddings $\mathcal{W}$ and the lexical resource $\mathcal{S}$. Specifically, the resource

[1] Given their prominence, we use *embeddings* to refer to word representations in general.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 388–393,
Valencia, Spain, April 3-7, 2017.

should be viewable as a graph $\mathcal{S} = (V, E)$, where V is the set of vertices that correspond to words or concepts and E is the set of edges that denote semantic relationships between entities in V.

2.1 Extraction of semantic landmarks

The aim of this phase is to find the set of landmarks for w_r which can best indicate the proximity of semantic space in which we can position w_r. As landmarks for w_r, we take its most semantically similar words which we extract from $\mathcal{S}$ by viewing the resource as a semantic network and analyzing its structure. To this end, we use the Personalized PageRank (Haveliwala, 2002, PPR) algorithm which has been proven to be a reliable graph analysis technique in various NLP tasks, including Word Sense Disambiguation (Agirre et al., 2014) and word similarity (Ramage et al., 2009; Pilehvar and Navigli, 2015).

Let k be the corresponding vertex of w_r in $\mathcal{S}$. We estimate the PPR distribution $\mathbf{x}^{\mathrm{T}}$ for this vertex. This distribution can be seen as a column vector $(n \times 1)$ whose cells denote the semantic association of their corresponding vertices to k. To compute $\mathbf{x}^{\mathrm{T}}$, we first construct a row-stochastic transition matrix $\mathbf{P}_{n \times n}$ where $n = |V|$ and cell $\mathbf{P}_{ij}$ denotes the probability of shifting from vertex i to vertex j within a single step of random walk. This probability is equal to 0 if there is no semantic relation between these two vertices and, otherwise, equal to the inverse of the total number of edges that connect vertex i to other vertices in the network (under the assumption that all edges are equally likely to be taken in a random walk). We can then obtain the PPR distribution $\mathbf{x}^{\mathrm{T}}$ by solving the eigenvector problem $\mathbf{x}^{\mathrm{T}}\mathbf{P} = \mathbf{x}^{\mathrm{T}}$ (Langville and Meyer, 2004). This computation has traditionally been performed using the power method: $\mathbf{x}^{(t)\mathrm{T}} = \alpha\, \mathbf{x}^{(t-1)\mathrm{T}}\mathbf{P} + (1-\alpha)\mathbf{v}_k^{\mathrm{T}}$, where $\mathbf{v}_k^{\mathrm{T}}$ is a column vector in which all the probability mass is assigned to the cell corresponding to vertex k and α is the scaling factor which is usually set to 0.85 (Langville and Meyer, 2004). Once $\mathbf{x}^{\mathrm{T}}$ was computed we can sort its elements according to their probabilities and obtain the list of most semantically similar words to vertex k, i.e., semantic landmarks for word w_r.

2.2 Embedding induction

Let $\mathcal{L}_r$ be the sorted list of semantic landmarks for w_r and $\mathbf{d}(x)$ be an embedding for word x in the space of $\mathcal{W}$. We adopt the approach of Pilehvar and Collier (2016a) and induce an embedding for w_r in the same semantic space using the following formula:

$$\hat{\mathbf{d}}(w_r) = \theta\, \mathbf{d}(w_r^0) + \frac{1}{\mathcal{L}_r} \sum_{i=1}^{|\mathcal{L}_r|} \mathrm{e}^{-i} \mathbf{d}(l_{i,r}). \quad (1)$$

where $l_{i,r}$ is the i^{th} word in $\mathcal{L}_r$. The formula computes an embedding for w_r which maps the word to the weighted centroid of its semantic landmarks. The exponential weighting assigns more importance to the top words in the list which are semantically more representative of w_r. Note that $\mathbf{d}(w_r^0)$ is the initial embedding for w_r. We include this in our formulation in order to extend the application of our approach from induction only to *embedding enrichment*, where we tend to improve an unreliable embedding $\mathbf{d}(w_r^0)$ obtained for a rare word by leveraging knowledge encoded in the lexical resource, and to *domain adaptation*, where the semantics of $\mathbf{d}(w_r^0)$ are adapted to a target domain by using domain specific landmarks that are extracted from a lexical resource in that domain. Parameter θ adjusts the contribution of initial embedding. Setting the parameter to zero reduces the formulation to that of inducing an embedding for an unseen word. In the next section, we discuss how the parameters were set in our experiments.

3 Experiments

As evaluation framework, we used word similarity. To verify the ability of the approach in inducing embeddings in both general and specific domains, we carried out two different experiments.

Embeddings. We used three different pretrained word embeddings: (1) GLOVE embeddings trained by Pennington et al. (2014) on Wikipedia and Gigaword 5 (vocab: 400K, dim: 300), (2) W2V-GN, Word2vec (Mikolov et al., 2013) trained on the Google News dataset (vocab: 3M, dim: 300), and (3) W2V-250K, the same Word2vec embeddings with a vocabulary of 250K most frequent words. We opted for these embeddings mainly for their popularity but we note that the proposed approach is equally applicable to any other vector representation.

Parameters. In experiments, whenever we had access to frequency statistics in the training data, we considered words with frequency $< 10K$ as rare and induced their representations along with

	Vanilla			+Induction		
	OOV	r	ρ	OOV	r	ρ
GLOVE	11%	34.9	34.4	0%	38.6	39.7
W2V-250K	34%	31.0	25.9	0%	44.2	47.5
W2V-GN	9%	43.8	45.3	0%	48.3	50.5

Table 1: Spearman ($\rho \times 100$) and Pearson ($r \times 100$) correlation performance of our approach when using three different embeddings on the RW dataset.

unseen words. We also limit the size of $\mathcal{L}_u$ to the top 25 words for faster computation. Also, we set θ in formula 1 to one in order to assign equal weights to the initial embedding $\mathbf{d}(w_r^0)$, whenever available, and to the one induced based on the knowledge extracted from the lexical resource. We did not perform any tuning on these parameters. Notably, θ can be set based on the reliability of $\mathbf{d}(w_r^0)$, for instance according to the frequency of w_r^0 in the training corpus. We leave the tuning of these and the evaluation of other word vectors to future work.

3.1 General domain setting

As our general domain evaluation benchmark we used the Stanford Rare Word (RW) similarity dataset (Luong et al., 2013) which is a suitable framework for evaluating the performance of representation induction techniques. The dataset comprises 2034 word pairs, 173 of which have at least one of their words not covered in our highest coverage embeddings, i.e., W2V-GN with a vocabulary size of 3 million words. As our general domain lexical resource, we opted for WordNet (Fellbaum, 1998), the community's *de-facto* standard English lexical resource.

Results. Table 1 lists the performance of our approach on the RW dataset. Results are shown for the three initial embeddings. For each of these we report the percentage of uncovered (OOV) words in the initial set ("Vanilla") as well as that after the induction of new embeddings to expand the vocabulary ("+Induction"). We observe that, irrespective of the utilized embeddings, our approach provides consistent improvements according to both evaluation measures. The improvement is highest for W2V-250K that has the smallest vocabulary size, highlighting the ability of our approach in effective vocabulary expansion.

We also benchmark our system against three other representation induction techniques (cf. Section 4) that have reported performance on the RW dataset. Results are shown in Table 2.[2] To have a fair comparison, in this setting we used a $500d$ set of embeddings trained by the Skipgram model (Mikolov et al., 2013) on the Wikipedia corpus (Shaoul and Westbury, 2010), similarly to Soricut and Och (2015). The table also shows results on RG-65 (Rubenstein and Goodenough, 1965), which is a standard dataset with relatively high frequency words, to provide a baseline for comparing the relative quality of the initial embeddings prior to any induction. We can see that our approach outperforms all the comparison work, particularly that of Soricut and Och (2015) which uses the same initial embeddings. This underlines the effectiveness of our approach in inducing embeddings for morphologically complex rare words.

3.2 Specific domain setting

As was mentioned before, our approach provides domain specialisation advantage in that it can be used to induce embeddings not only for morphologically complex forms but also for domain specific terms for which no subword information might be available in the training corpus. We evaluated the ability of our approach in specialising general domain embeddings to the medical domain which provides a challenging benchmark with its extensive terminology. We performed experiments on UMNSRS (Liu et al., 2012) and MayoSRS (Pakhomov et al., 2011) which are two standard word similarity datasets for the domain.

Lexical resource. We used Medical Subject Headings[3](MeSH) as our medical lexical resource. MeSH is a medical thesaurus that was created mainly for the purpose of indexing journal articles in the domain. As of December 2016,

[2]For this experiment, we show Spearman ρ results only as none of the comparison work reported Pearson correlation.

[3]https://www.nlm.nih.gov/mesh/

Approach	RW		RG-65	
	OOV	ρ	OOV	ρ
Botha and Blunsom (2014)	NA	30.0	NA	41.0
Luong et al. (2013)*	0%	34.4	0%	65.5
Soricut and Och (2015)*	0%	41.8	0%	75.1
Our approach*	0%	**43.3**	0%	75.1
Number of pairs	2034		65	

Table 2: Evaluation results on the RW dataset (and on RG-65 as baseline). Systems marked with * are trained on the same corpus.

		Vanilla			+Induction		
		OOV	r	ρ	OOV	r	ρ
Mayo	GLOVE	16%	11.1	11.6	11%	36.7	26.1
	W2V-250K	41%	1.2	2.9	21%	27.8	20.1
	W2V-GN	12%	15.5	14.0	10%	18.4	10.9
UMN	GLOVE	17%	31.6	24.4	6%	38.2	33.6
	W2V-250K	38%	11.8	3.2	13%	27.8	20.1
	W2V-GN	17%	25.8	21.5	7%	32.8	32.4

Table 3: Evaluation results on two biomedical word similarity datasets: MayoSRS (101 pairs) and UMNSRS (566 pairs).

the thesaurus comprises 25,186 headings that are arranged in a hierarchical structure, covering 75% and 38% of unique words in the UMNSRS and MayoSRS datasets, respectively.

Results. Table 3 shows the results on the two domain specific datasets. On both datasets and for all the three embeddings, our approach provides considerable raise in vocabulary coverage which results in significant performance improvements according to both evaluation measures. This highlights the effectiveness of our approach in inducing representations for terms such as *rhonchi*, *osteophyte*, and *cardura* for which no subword information is available in the training data. It is important to note that none of the comparison work, which generally focus on morphologically complex words, can induce representations for such terms. This advantage enables us to train embeddings in general domain, for which text are available abundantly, and specialise them to specific domains for which large amounts of training data might not be available. We also note that our system did not provide full coverage of the words in the two datasets, missing several words which were not included in MeSH, e.g., *dysguesia*, *hemetemesis* and *ceftiaxone*. This can be substantially improved by using larger medical ontologies, such as SNOMED CT[4]. We leave this to future work.

4 Related Work

Recent research on representation induction for rare words has mainly focused on the case of infrequent morphological variations (Alexandrescu and Kirchhoff, 2006) and has tried to address the problem by resorting to information available for subword units. A morphological analyzer, such as Morfessor (Creutz and Lagus, 2007), is usually used in a pre-processing step to break inflected words into their morphological structures. Representations are then induced for morphologically complex words from their morphemes either by combining recursive neural networks (Luong et al., 2013) or using log-bilinear language models (Botha and Blunsom, 2014). Lazaridou et al. (2013) induced embeddings for complex words by adapting phrase composition models, whereas Soricut and Och (2015) automatically constructed

[4] http://www.ihtsdo.org/snomed-ct

a morphological graph by exploiting regularities within a word embedding space. In the latter case, the representations were inferred by analyzing morphological transformations in the graph. Also related to our work is the retrofitting (Faruqui et al., 2015) of pre-trained embeddings by exploiting semantic lexical resources. Despite being effective in improving the representations for seen words, the retrofitting approaches are generally unable to induce new embeddings to address the unseen words problem. Cotterell et al. (2016) designed an extension of the retrofitting procedure that uses morphological resources to generate vectors for forms not observed in the training data.

A common strand in all these works is that they assume that the training corpus covers the morpheme or other morphological variations of an unseen word. As a result, they fall short of modelling words whose morphemes are not seen during training. The proposed model is different in that it can induce embeddings not only for inflected forms and derivations, but also for words whose morphemes are not seen during the training. In (Pilehvar and Collier, 2016b), we proposed a model that exploited Wikipedia articles in order to adapt a set of pre-trained embeddings to a specific domain. Here, we extend that model and apply it to the task of vocabulary expansion for rare and unseen words.

5 Conclusions and Future Work

An approach was proposed for inducing embeddings for rare words on the basis of the knowledge extracted from external lexical resources. We showed using different experiments that the approach is effective in addressing the rare word problem for morphologically complex words in the general domain as well as for specialising a pre-trained set of embeddings to the medical domain. As future work, we plan to experiment with larger lexical resources and representations, such as that of Camacho-Collados et al. (2016), and perform evaluations on other domains. We also intend to extend the model to handle less structured resources, such as the Paraphrase Database (Ganitkevitch et al., 2013).

Acknowledgments

The authors gratefully acknowledge the support of the MRC grant No. MR/M025160/1 for PheneBank.

References

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based Word Sense Disambiguation. *Computational Linguistics*, 40(1):57–84.

Andrei Alexandrescu and Katrin Kirchhoff. 2006. Factored neural language models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 1–4, New York City, USA.

Jan A. Botha and Phil Blunsom. 2014. Compositional Morphology for Word Representations and Language Modelling. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, Beijing, China.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Ryan Cotterell, Hinrich Schütze, and Jason Eisner. 2016. Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1651–1660, Berlin, Germany.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):3:1–3:34.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia.

Taher H. Haveliwala. 2002. Topic-sensitive PageRank. In *Proceedings of the 11th International Conference on World Wide Web*, pages 517–526, Honolulu, Hawaii, USA.

Amy N. Langville and Carl D. Meyer. 2004. Deeper inside PageRank. *Internet Mathematics*, 1(3):335–380.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositionally derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 1517–1526, Sofia, Bulgaria.

Ying Liu, Bridget T. McInnes, Ted Pedersen, Genevieve Melton-Meaux, and Serguei Pakhomov. 2012. Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, umls and wordnet. In *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium*, pages 363–372.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop at International Conference on Learning Representations*, Scottsdale, Arizona.

Serguei V.S. Pakhomov, Ted Pedersen, Bridget McInnes, Genevieve B. Melton, Alexander Ruggieri, and Christopher G. Chute. 2011. Towards a framework for developing semantic relatedness reference standards. *Journal of Biomedical Informatics*, 44(2):251–265, April.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543, Doha, Qatar.

Mohammad Taher Pilehvar and Nigel Collier. 2016a. De-conflated semantic representations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1680–1690, Austin, Texas, November. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Nigel Collier. 2016b. Improved semantic representation for domain-specific entities. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 12–16, Berlin, Germany, August. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Roberto Navigli. 2015. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128.

Daniel Ramage, Anna N. Rafferty, and Christopher D. Manning. 2009. Random walks for text semantic similarity. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 23–31, Suntec, Singapore.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

C. Shaoul and C. Westbury. 2010. The Westbury Lab Wikipedia Corpus. `http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html`. Accessed: 2016-11-10.

Radu Soricut and Franz Och. 2015. Unsupervised morphology induction using word embeddings. In *Proceedings of NAACL-HLT*, pages 1627–1637, Denver, Colorado.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.