Context-guided Self-Supervised Relation Embeddings

Huda Hakami[1,2] and Danushka Bollegala[1]
[1]*Department of Computer Science,The University of Liverpool,Liverpool, UK*
{h.a.hakami, danushka.bollegala}@liverpool.ac.uk
[2]*Department of Computer Science, Taif University, Saudi Arabia, Taif*
hoda.h@tu.edu.sa

***Abstract*—A semantic relation between two given words a and b can be represented using two complementary sources of information: (a) the semantic representations of a and b (expressed as word *embeddings*) and, (b) the contextual information obtained from the co-occurrence contexts of the two words (expressed in the form of lexico-syntactic patterns). Pattern based approaches suffer from sparsity while methods rely only on word embeddings for the related pairs lack of relational information. Prior work on relation embeddings have pre-dominantly focused on either one type of those two resources exclusively, except for a notable few exceptions. In this paper, we proposed a self-supervised context-guided Relation Embedding method (CGRE) using the two sources of information. We evaluate the learnt method to create relation representations for word-pairs that do not co-occur. Experimental results on SemEval-2012 task2 dataset show that the proposed operator outperforms other methods in representing relations for unobserved word-pairs.**

***Keywords*-relation embeddings; relational patterns, compositional relation representations.**

I. Introduction

Representing relations between words benefits various Natural Language Processing (NLP) tasks such as relational information retrieval [1], [2], statistical machine translation [3], question answering [4] and textual entailment [5]. For example, given a premise sentence P, *a man ate an apple*, and a hypothesis H, *a man ate a fruit*, a model that can infer the existence of *is-a* relation between *fruit* and *apple* would correctly predict that H entails P.

We consider the problem of creating a semantic representation $\boldsymbol{r}$ for the relation r that holds between two given words a and b. This problem has been approached from a *compositional* direction, where given the pre-trained word embeddings for a and b, respectively denoted by d-dimensional real vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$, the goal is to learn a relation embedding function, $f(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\theta})$, parametrised by $\boldsymbol{\theta}$ [6], [7]. Unsupervised solutions to this problem have been proposed such as ones that use a fixed operator such as the vector offset [8], or supervised approaches that implement f as a multi-layer feed forward neural network [5], [9], [7].

On the other hand, the contexts in which a and b co-occur provide useful clues regarding the relations that exist between the two related words. We call this approach *holistic relation representation* because it directly models the interaction between words as a single monolithic unit without individually considering the words that appear in the context [10], [11]. The main drawback of the holistic approaches is data sparseness. Because not every related word-pair co-occurs within a co-occurrence window even in a larger corpus, holistic relation representations fails to represent relations between words that never co-occur.

Despite the above-mentioned limitations, we argue that word embeddings and co-occurrence contexts collectively provide complementary information for the purpose of learning relational embeddings. For example, Bollegala et al. [1] observed a *duality* between word-pair and pattern-based approaches for representing relations where they refer to the former as an *intentional* definition of relation representation and the latter an *extensional* definition of relation representation. Bollegala et al. [1] used this duality to propose a sequential co-clustering algorithm for discovering relations from a corpus. Riedel et al. [12] further developed this line of research and proposed *universal schema* for representing relations, which was then used to produce relation embeddings via matrix decomposition. However, despite these prior work on relation embeddings, the two types of information sources are often used independently.

Our focus in this paper is to propose a relation representation method that uses the contextual information from a text corpus to generalise the learnt operator for unobserved word-pairs. For this purpose, we propose Context-Guided Relation Embeddings (CGREs) to represent relation between words. Specifically, CGREs are learnt using the word embeddings of related word-pairs along with the co-occurrence contexts of the word-pairs extracted from a corpus. Our experimental results on the SemEval-2012 task 2 benchmark dataset show the ability of the learnt operator to generalise to unobserved word-pairs outperforming previously proposed relational operators.

II. Related Work

As already described in the previous section, two main approaches can be identified in the literature for representing a semantic relation between two words: *holistic* and *compositional*. The holistic approach uses lexical patterns

in which the two words of interest co-occur, while the compositional approaches on the other hand attempts to represent the relation between two words from their word embeddings.

A. Holistic Pattern-based Approaches for Relations

An unstructured text corpus forms important resource to extract information for numerous NLP tasks such as relation extraction where the task is to identify the relation that holds between two named entities. Lexical patterns in which two related words co-occur within a corpus provide useful insights into the semantic relations that exist between those two words. Patterns for hypernym relation (i.e is-a) have been studied extensively since it plays an important role in building ontologies covering entities. For instance, Hearst's [13] patterns such as *is a*, *is a kind of* and *such as* have been used to identify the hypernym relation between words. Lexical patterns for other relations have also been studied such as Mernonymy [14] and Causial [15].

Turney [16] introduced *latent relation hypothesis* which state that word-pairs that co-occur in similar patterns tend to have similar semantic relations, and proposed Latent Relational Analysis (LRA) to represent the relation between two words using a vector. Specifically, in LRA we first create a pair-pattern matrix where the elements correspond to the number of times a pair co-occurs with a pattern [10]. Next, dimensionality reduction techniques such as Singular Value Decomposition (SVD) is applied to this pair-pattern matrix to smooth the co-occurrence data and produce low-dimensional vector representations. When evaluated on a benchmark dataset containing word analogy questions collected from scholastic aptitude tests (SATs), LRA obtains an accuracy of 56.7%, while the average high school student's accuracy on SAT word analogy questions has been 57%. Along similar lines, Jameel et al. [17] extend Global Vectors [18] for learning word embedding to learn word-pair vectors considering the 3-ways co-occurrences between the two words and the context words in which they co-occur. We collectively refer to these methods holistic representations because the pairs of words are treated as a whole rather than considered individually.

Despite the success of pattern-based approaches for representing relations, it suffers from data sparsity. To represent the relation between two words, such approaches require the two words to co-occur in a specified context in order for a pattern to be extracted. However, not every related words co-occur even in a large corpus. For example, in Figure 1 we show the co-occurrence distribution of the word-pairs in SemEval 2012 Task 2 dataset, where we count number of sentences in Wikipedia containing both words in each word-pair. In the SemEval 2012 Task 2 dataset there are $3,307$ related word-pairs, out of which 490 word-pairs never co-occur in any sentence in Wikipedia corpus, resulting in a highly sparse co-occurrence distribution as shown in Figure 1. Therefore, pattern-based holistic approaches fail to handle such unobserved but related words.

Our proposed method differs from these existing pattern-based approaches in two important ways. First, we do not require the two words to co-occur within same sentences in a corpus to be able to represent the relation between them. Second, the parametrised operator we learn generalises in the sense that it can be applied to any new word-pair or relation type, not limited to the words and relations that exist in the training data.

B. Compositional Pair-based Approaches for Relations

Prior work on word embedding learning have found that relations between words could be represented by the difference of the corresponding word embeddings (from here onwards we call it PairDiff) [19], [18]. The most popular example is: *king*$-$*man* $\approx$ *queen*$-$*woman*. We call such approaches to represent relations between words as *compositional pair-based* because the relation representation is composed using the semantic representations of the two constituent words of the related pairs. The compositional approach for relations overcomes the sparseness issues in the pattern-based methods as it relaxes the assumption that related pairs have to co-occur in the same context.

Since Mikolov et al. findings in 2013 [19], a renewed interest of exploring the relations in the semantic spaces of words has been sparked. Several recent works have targeted to evaluate different combination methods that can be applied on word embeddings to generate word-pair embeddings [20], [21], [7]. Hakami and Bollegala [20] investigated several unsupervised operators, such as vector concatenation, addition, difference and elementwise multiplication, that map the embeddings of two related words to a vector representing the relation between them.

On the other hand, recent research [22], [23], [24] have raised concerns on claims about word embedding's ability to represent relations via PairDiff. Given an analogy prediction problem in the form a is to b as c is to d, in some cases even by ignoring c it is possible to correctly predict d using the fact that d is similar to a and b individually. Roller et al. [25] showed that Hearst's patterns are more valuable for hypernym detection tasks than distributional word embeddings. Vylomova et al. [26] also showed the limitations of PaiDiff by applying it for representing semantic relations outside those in the Google dataset which were used initially for the evaluation of PairDiff. These findings suggest that in order to represent a diverse set of relation types we must combine the strengths in the holistic as well as compositional approaches, which is a motivation for our current work.

C. Hybrid Approaches for Relations

As described earlier, holistic and compositional approaches have complementary properties when it comes to representing

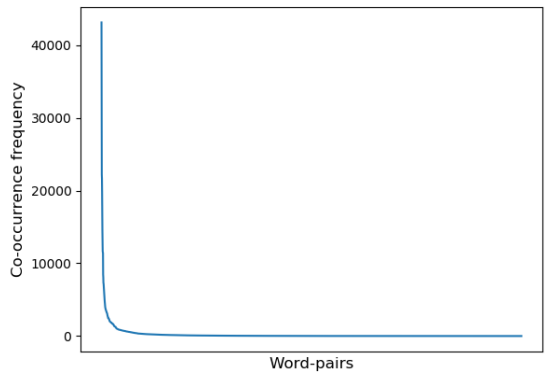

Figure 1. Co-occurrence frequency for word-pairs in SemEval-2012 task2 from Wikipedia corpus.

relations. Hybrid approaches try to balance between the data sparsity in the holistic methods and the lack of relational information in the compositional approaches. However, few recent studies have been devoted to incorporate the two types of information to improve the relation representations.

Zilah et al. [27] measure the relational similarities between word-pairs by combining heterogeneous models including distributional word embeddings and lexical patterns. In their work, the compositional method that based on PairDiff reported encouraging results for many relation types in SemEval-2012 task 2 dataset. More recently, Washio and Kato [28] proposed Neural Latent Relational Analysis (NLRA) an unsupervised relational operator that is learnt to make the compositional and holistic representations similar using a negative sampling training objective. They also found that NLRA can be used for the purpose of predicting missing dependency paths between word-pairs that don't co-occur in a corpus [9].

III. Method

Our main goal is to accurately represent relations between words. We propose to learn a parametrised operator for relations that maps a word-pair to a relation embedding considering two sources of information: (a) word embeddings of related words, and (b) the contexts in which two related words co-occur. We want the learnt operator to overcome the sparseness problem in holistic relation representations. Motivated by this, our objective is to create relation representations for word-pairs that do not co-occur or belong to unseen relations.

Given a set of related word-pairs along with their relation labels $\mathcal{D} = \{(a_i, b_i, r_i)\}_i^N$, pre-trained word embeddings that represent the semantics of words, and a text corpus, we propose a method for learning m-dimensional relation embeddings $\boldsymbol{r}_{(c,d)} \in \mathbb{R}^m$ for an unseen word-pair (c, d). Relation labels for word-pairs can be the manually annotated gold labels provided in the relational dataset such as DiffVec [26], Google [19], and BATS [29], or can be pseudo labels generated from word-pair features as described in Section III-A. Following the prior work [7], [9], [28], [5], a word-pair (a, b) is fed to a deep multilayer neural network with a nonlinearity activation for the hidden layers. The input layer of the network is the concatenation of embeddings $\boldsymbol{a}$ and $\boldsymbol{b}$ and their difference, $(\boldsymbol{a}; \boldsymbol{b}; \boldsymbol{b} - \boldsymbol{a})$. As described in Hakami and Bollegala [7], the output of the last layer of the neural network that is given by $f(a, b, \theta_f)$ is considered as a representation for a word-pair, and is passed to a fully connected softmax layer and the overall network is trained to predict the relation label for the given pair. For this purpose, we use the cross-entropy loss defined in (1) as the training objective.

$$\mathcal{J}_C = -\sum_{(a,b,r)\in\mathcal{D}} \log \quad p(r|f(a, b, \theta_f)) \tag{1}$$

Here, θ_f collectively denotes the parameters of the network.

$\mathcal{J}_C$ given in (1) does not consider the co-occurrence contexts. Therefore, we consider a relation representation, $g(\mathcal{P}(a, b), \theta_g)$, that encodes a set of contextual co-occurrences between a and b according to (2).

$$g(\mathcal{P}(a, b), \theta_g) = \sum_{p\in\mathcal{P}(a,b)} w(a, p, b)\boldsymbol{h}(a, p, b, \theta_h) \tag{2}$$

Here, $p \in \mathcal{P}(a, b)$ is a lexical pattern that co-occurs with a and b. We use LSTMs [30] to map a sequences of words to a fixed-length vector $\boldsymbol{h}(a, p, b, \theta_h)$. To incorporate the representativeness of a pattern of a relation, we assign a weight $w(a, b, p)$ given by (3).

$$w(a, p, b) = \frac{c(a, p, b)}{\sum_{t\in\mathcal{P}(a,b)} c(a, t, b)} \tag{3}$$

Here, c denotes the number of co-occurrences between p and (a, b).

Because the holistic and compositional methods represent the same semantic relation we require them to be close in the ℓ_2 space, captured by the constraint given by (4).

$$\mathcal{J}_{Patt} = \frac{1}{2}\,||f(a, b, \theta_f) - g(\mathcal{P}(a, b), \theta_g)||_2^2 \tag{4}$$

We would like to learn word pair embeddings that simultaneously minimise both (1) and (4). Therefore, we formulate the objective function of the proposed Context-Guided Relation Embeddings (i.e. CGRE) as a linear combination of (1) and (4) as follows:

$$\mathcal{J} = \mathcal{J}_C + \lambda\mathcal{J}_{Patt} \tag{5}$$

Here, $\lambda \in \mathbb{R}$ is a regularisation coefficient that determines the influence of the contextual patterns of the word-pairs for the learnt relational operator. After learning CGRE, we generate representations for a given word-pair (a, b) by concatenating $f_{\theta_f}(\boldsymbol{a}, \boldsymbol{b})$ and $f_{\theta_f}(\boldsymbol{b}, \boldsymbol{a})$.

A. Pseudo Relation Labels

To train CGRE, we require a dataset containing word-pairs annotated with relation labels. However, the cost of annotating word-pairs with relation labels can be high for specialised domains such as biomedical [31]. To make our proposed method self-supervised, we induce pseudo labels for word-pairs via clustering. Specifically, we cluster the PairDiff vectors of the training word pairs using k-means clustering algorithm with different k number of clusters. Because the ground truth class labels are given in DiffVec training data, we evaluate the quality of the generated clusters using the V-measure [32], which is the a harmonic mean between homogeneity and completeness of the clusters. Consistent with Vylomova et al. findings [26], we find that $k = 50$ clusters to perform well with a V-measure of 0.416.

IV. Experiments

A. Datasets

Measuring the Degrees of Prototypicality: We evaluate the relation embeddings on measuring degrees of relational similarity task using SemEval-2012 Task 2 dataset. The task is to rank word-pairs in a relation according to their degrees of prototypicality (i.e. the extent to which they exhibit the relation). The dataset has 79 relation types in total and it is split to two sets of 69 test relations and 10 train relations. Following the standard practice, we report performance on the test set and use train set for setting hyperparameters.

Training Data: We use the DiffVec dataset [26] that contains $12,458$ triples (a, b, r), where words a and b are connected by an asymmetric relation r out of 36 fine-grained relation types. We use the word-pairs set $\mathcal{D}$ of the training relations and their reverse pairs to obtain relational patterns. Word-pairs in DiffVec that also appear in SemEval test data are excluded from the training set. Following Turney [33], we extract the context of one to five words in between the two related words considering the order in which they appear in the specified context ($\mathcal{P}(a, b)$ consists of all patterns where in a occurs before b). To reduce noise, we filter out the patterns that occur between less than ten distinct word-pairs in the corpus. As a result, we obtain $5,017$ contextual patterns and the number of training triples (a, b, p) after removing out-of-vocabulary words is $158,920$[1]. We use pre-trained 300 dimensional GloVe embedding for representing words[2]. To extract co-occurrence contexts, we use the English Wikipedia corpus, which consists of ca. 337M sentences.

B. Baselines

We compare the proposed method with unsupervised compositional operators PairDiff and Concatenation (Concat) for the given pre-trained word embeddings. We also compare against the supervised Multi-class Neural Network Penultimate Layer (MnnPL) method proposed by Hakami and Bollegala [7]. Specifically, MnnPL learns a relation classifier using a relation labelled word-pairs and does not use contextual patterns (corresponds to $\lambda = 0$).

We compare the proposed CGRE with NLRA using the contextual patterns provided by the original authors [28]. Because we are interested in relation representation methods that can generalise to word-pairs that *do not* co-occur in the corpus, we re-train NLRA using the same training data that we used for our proposed method such that NLRA doe not observe the word-pairs in SemEval dataset. LRA requires all word-pairs to be represented using lexical patterns extracted from the co-occurrence contexts. Because we strictly focus on evaluating relation representations for word-pairs without using their contextual patterns, LRA is excluded from the evaluations. Following Washio and Kato [28], we also evaluate the performance of each learnt relation representation method when it is combined with PairDiff. Simply, we average the scores of a learnt method and the PairDiff score for each target word-pair.

Table I
Average MaxDiff accuracy and Spearman correlation for the 69 test relations in SemEval 2012 Task 2.

Method	**MaxDiff**	**Correlation**
PairDiff	43.48	0.31
Concat	41.67	0.29
NLRA	42.32	0.29
NLRA+PairDiff	44.35	0.33
MnnPL($\lambda = 0$)	43.75	0.31
MnnPL+PairDiff	45.42	0.35
CGRE-Gold	44.87	0.34
CGRE-Gold+PairDiff	**45.92**	**0.37**
CGRE-Proxy	44.34	0.34
CGRE-Proxy+PairDiff	45.49	0.36

C. Implementation Details

For a given word-pair (a, b), we compose their embeddings $\boldsymbol{a}$ and $\boldsymbol{b}$ using a multi-layer feedforward neural networks with 3 hidden layers followed by the batch normalization and a $tanh$ nonlinearity function. All the word vectors were first normalised to unit ℓ_2 length before feeding them to the neural net. The size of the hidden layers are set to 300. We did not update the input word embeddings during training to preserve their distributional regularity. A unidirectional LSTM with a 300 dimensional hidden state is used to encode the contextual patterns. AdaGrad [34] with mini-batch size 100 is used to learn the parameters of the proposed operator. All parameters are initialised by uniformly sampling from $[-1, +1]$ and the initial learning rate is set to 0.1. The best model was selected by early stopping using the MaxDiff accuracy on the SemEval train set.

[1]Code, training data and pre-trained models will be publicly available upon paper acceptance.

[2]http://nlp.stanford.edu/data/glove.6B.zip

Table II
AVERAGE MAXDIFF AND SPEARMAN CORRELATION FOR EACH MAJOR RELATION IN THE TEST SET OF SEMEVAL 2012-TASK2. THE VALUES BETWEEN PARENTHESES INDICATE THE PERFORMANCE OF A METHOD COMBINED WITH PAIRDIFF.

	MaxDiff				**Correlation**			
Relation	PairDiff	MnnPL	CGRE-Gold	CGRE-Proxy	PairDiff	MnnPL	CGRE-Gold	CGRE-Proxy
CLASS-INCLUSION	48.50	**52.00** (51.60)	51.40 (51.67)	50.45 (49.35)	0.375	0.519 (**0.537**)	0.533 (0.516)	0.515 (0.462)
PART-WHOLE	43.50	41.33 (43.36)	39.61 (42.80)	43.35 (**44.38**)	0.287	0.245 (0.288)	0.228 (0.292)	0.314 (**0.321**)
SIMILAR	41.26	36.20 (41.15)	40.02 (40.82)	**41.68** (41.10)	0.252	0.186 (0.260)	0.245 (**0.286**)	0.280 (0.282)
CONTRAST	33.72	38.57 (38.73)	**40.21** (38.44)	36.39 (36.67)	0.113	0.160 (0.202)	0.209 (**0.226**)	0.157 (0.171)
ATTRIBUTE	46.32	44.84 (47.23)	46.19 (**47.97**)	45.44 (47.83)	0.410	0.351 (0.409)	0.396 (**0.444**)	0.387 (0.437)
NON-ATTRIBUTE	39.11	42.45 (41.82)	42.41 (42.79)	**43.00** (41.85)	0.209	0.264 (0.265)	0.287 (0.279)	**0.313** (0.274)
CASE RELATIONS	46.49	49.53 (49.57)	**52.04** (51.67)	49.46 (50.21)	0.383	0.425 (0.467)	**0.475** (0.466)	0.419 (0.445)
CAUSE-PURPOSE	44.43	44.17 (46.89)	47.57 (**48.59**)	47.74 (48.17)	0.343	0.332 (0.384)	0.422 (**0.436**)	0.400 (0.404)
SPACE-TIME	49.48	45.53 (48.50)	48.62 (**50.21**)	45.36 (49.79)	0.422	0.373 (0.433)	0.432 (**0.455**)	0.385 (0.437)
REFERENCE	41.92	45.94 (**47.84**)	41.32 (44.74)	41.52 (45.74)	0.303	0.323 (**0.377**)	0.212 (0.323)	0.295 (0.375)

D. Experimental Results

Table I shows the macro-averaged MaxDiff accuracy and Spearman correlations for the 69 test relations in the SemEval2012 Task 2 dataset. Our proposed method (GCRE) achieved the best results on both evaluation metrics when combined with PairDiff. CGRE trained using pseudo labels (CGRE-Proxy) can successfully reach the performance of CGRE trained using the gold labels in the DiffVec dataset (CGRE-Gold). This is encouraging because it shows that GCRE can be trained in a self-supervised manner, without requiring manually labelled data. Overall, for all the methods, adding the relational similarity scores from PairDiff improves the performance of ranking the word-pairs, which confirm the complementary properties between the two approaches when it comes to representing relations. As seen in Table I, NLRA performs poorly when it is trained on DiffVec using patterns extracted for the word-pairs in DiffVec and tested on SemEval[3]. This shows that NLRA is unable to generalise well to the relations in the SemEval dataset, not present in the DiffVec dataset.

To evaluate the performance for different relation types, we breakdown the results for the 10 major relations in the 69 SemEval test set as presented in Table II. By incorporating contextual patterns when training CGRE, we obtain better performance in 8 out of the 10 test relations in terms of MaxDiff and Spearman correlation. These improvements are statistically significant according to a paired t-test ($p < 0.01$). MnnPL reports the best accuracy and correlation for CLASS-INCLUSION and REFERENCE relations (either without or with the addition of PairDiff).

V. CONCLUSION

We consider the problem of representing relations between words. Specifically, we proposed a method that uses the contextual patterns in a corpus to improve the compositional relation representation using word embeddings of the related word-pairs. For this purpose, we proposed a parametrised relational operator using the contexts where two words co-occur in a corpus and require that holistic representation to be similar to a compositional representation computed using the corresponding word embeddings. Experiments on measuring degrees of relational similarity between word pairs show that we can overcome the sparsity problem of the holistic pattern-based approaches for relations.

[3]The accuracy of NLRA when its trained on pattern extracted using word pairs in the entire SemEval dataset is 45.28%, which is similar to the result reported in the original paper.

REFERENCES

[1] D. T. Bollegala, Y. Matsuo, and M. Ishizuka, "Relational duality: Unsupervised extraction of semantic relations between entities on the web," in *Proceedings of the 19th international conference on World Wide Web*. ACM, 2010, pp. 151–160.

[2] M. J. Cafarella, M. Banko, and O. Etzioni, "Relational web search," in *WWW Conference*, 2006.

[3] P. Nakov123, "Improved statistical machine translation using monolingual paraphrases," in *ECAI 2008: 18th European Conference on Artificial Intelligence, July 21-25, 2008, Patras, Greece: Including Prestigious Applications of Intelligent Systems (PAIS 2008): Proceedings*, vol. 178. IOS Press, 2008, p. 338.

[4] S. Yang, L. Zou, Z. Wang, J. Yan, and J.-R. Wen, "Efficiently answering technical questions-a knowledge graph approach." in *AAAI*, 2017, pp. 3111–3118.

[5] M. Joshi, E. Choi, O. Levy, D. S. Weld, and L. Zettlemoyer, "pair2vec: Compositional word-pair embeddings for cross-sentence inference," *arXiv preprint arXiv:1810.08854*, 2018.

[6] H. Hakami, K. Hayashi, and D. Bollegala, "Why does pairdiff work? - a mathematical analysis of bilinear relational compositional operators for analogy detection," in *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, 2018.

[7] H. Hakami and D. Bollegala, "Learning relation representations from word representations," 2018.

[8] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations." in *Proceedings of HLT-NAACL*, 2013, pp. 746–751.

[9] K. Washio and T. Kato, "Filling missing paths: Modeling co-occurrences of word pairs and dependency paths for recognizing lexical semantic relations," *arXiv preprint arXiv:1809.03411*, 2018.

[10] P. D. Turney, "Measuring semantic similarity by latent relational analysis," *arXiv preprint cs/0508053*, 2005.

[11] R. Snow, D. Jurafsky, and A. Y. Ng, "Learning syntactic patterns for automatic hypernym discovery," in *Advances in neural information processing systems*, 2005, pp. 1297–1304.

[12] S. Riedel, L. Yao, A. McCallum, and B. M. Marlin, "Relation extraction with matrix factorization and universal schemas," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 74–84.

[13] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1992, pp. 539–545.

[14] R. Girju, A. Badulescu, and D. Moldovan, "Learning semantic constraints for the automatic discovery of part-whole relations," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003, pp. 1–8.

[15] E. MARSHMAN, "The cause-effect relation in a biopharmaceutical corpus: English knowledge patterns," in *Terminology and knowledge engineering*, 2002, pp. 89–94.

[16] J. Bigham, M. L. Littman, V. Shnayder, and P. D. Turney, "Combining independent modules to solve multiple-choice synonym and analogy problems," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 2003, pp. 482–489.

[17] S. Jameel, Z. Bouraoui, and S. Schockaert, "Unsupervised learning of distributional relation vectors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2018, pp. 23–33.

[18] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[20] H. Hakami and D. Bollegala, "Compositional approaches for representing relations between words: A comparative study," *Knowledge-Based Systems*, vol. 136, pp. 172–182, 2017.

[21] K. Gábor, H. Zargayouna, I. Tellier, D. Buscaldi, and T. Charnois, "Exploring vector spaces for semantic relations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1814–1823.

[22] O. Levy, S. Remus, C. Biemann, and I. Dagan, "Do supervised distributional methods really learn lexical inference relations?" in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 970–976.

[23] T. Linzen, "Issues in evaluating semantic spaces using word analogies," *arXiv preprint arXiv:1606.07736*, 2016.

[24] A. Rogers, A. Drozd, and B. Li, "The (too many) problems of analogical reasoning with word vectors," in *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, 2017, pp. 135–148.

[25] S. Roller, D. Kiela, and M. Nickel, "Hearst patterns revisited: Automatic hypernym detection from large text corpora," *arXiv preprint arXiv:1806.03191*, 2018.

[26] E. Vylomova, L. Rimell, T. Cohn, and T. Baldwin, "Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning," *arXiv preprint arXiv:1509.01692*, 2015.

[27] A. Zhila, W.-t. Yih, C. Meek, G. Zweig, and T. Mikolov, "Combining heterogeneous models for measuring relational similarity," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 1000–1009.

[28] K. Washio and T. Kato, "Neural latent relational analysis to capture lexical semantic relations in a vector space," *arXiv preprint arXiv:1809.03401*, 2018.

[29] A. Gladkova, A. Drozd, and S. Matsuoka, "Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't." in *Proceedings of the NAACL Student Research Workshop*, 2016, pp. 8–15.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] P. Patel, D. Davey, V. Panchal, and P. Pathak, "Annotation of a large clinical entity corpus," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2033–2042.

[32] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007.

[33] P. D. Turney, "The latent relation mapping engine: Algorithm and experiments," *Journal of Artificial Intelligence Research*, vol. 33, pp. 615–655, 2008.

[34] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.