Understanding the Source of Semantic Regularities in Word Embeddings

Hsiao-Yu Chiang
School of Information
University of California, Berkeley, USA
hsiaoyuchiang@berkeley.edu

Jose Camacho-Collados
School of Computer Science and Informatics
Cardiff University, United Kingdom
camachocolladosj@cardiff.ac.uk

Zachary A. Pardos
Graduate School of Education
University of California, Berkeley, USA
pardos@berkeley.edu

Abstract

Semantic relations are core to how humans understand and express concepts in the real world using language. Recently, there has been a thread of research aimed at modeling these relations by learning vector representations from text corpora. Most of these approaches focus strictly on leveraging the co-occurrences of relationship word pairs within sentences. In this paper, we investigate the hypothesis that examples of a lexical relation in a corpus are fundamental to a neural word embedding's ability to complete analogies involving the relation. Our experiments, in which we remove all known examples of a relation from training corpora, show only marginal degradation in analogy completion performance involving the removed relation. This finding enhances our understanding of neural word embeddings, showing that co-occurrence information of a particular semantic relation is the not the main source of their structural regularity.

1 Introduction

The representation of words has been a long-standing task in natural language processing (NLP). The main underlying principle is known for decades, as explained by Firth (1957). This principle was based on the idea that the meaning of a word can be understood by its surrounding company (i.e., the words in its context). Most modern representation learning theory in NLP is based on this assumption, with vector representation being the most successful area to date (Turney and Pantel, 2010). More recently, low-dimensional word representations learned from text corpora using neural networks (i.e., *word embeddings*) have emerged (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017) stemming from cognitive frameworks based on distributed representation (Hinton et al., 1986; Feldman and Ballard, 1982). Neural word embeddings have been proven to contain useful information about concepts and entities, and provide a generalization boost to many NLP applications (Goldberg, 2017). Surprisingly, these representations have also been shown to exhibit linear relationships between words in the vector space, demonstrated by *analogy*. For example, Mikolov et al. (2013b) showed that a simple operation such as *king-man+woman* will result in a point near queen in the vector space[1]. These word analogies have been extensively investigated in the literature, aiming to shed light on this surprising property. However, while there has been a body of research seeking to understand how these analogies work (Arora et al., 2016; Gittens et al., 2017; Ethayarajh et al., 2019; Allen and Hospedales, 2019), and noting issues about their methodology (Linzen, 2016; Gladkova et al., 2016; Nissim et al., 2020), there has not been a specific analysis on the source of statistical cues that leads to their high performance on this task.

Concurrently, a thread of research has focused on explicitly modeling lexical relationships of word pairs in text corpora (Jameel et al., 2018; Espinosa-Anke and Schockaert, 2018; Washio and Kato, 2018; Joshi et al., 2019; Camacho-Collados et al., 2019). While these methods employ different means for learning relation vectors, they share a common initial premise: only co-occurring words in the corpus are considered[2]. While this simplified assumption works well enough in practice, providing a useful signal even in downstream NLP applications, in this paper we find that valuable information is likely lost in the process. In fact, we

[1] More information of how word analogies work can be found in Section 2.1.

[2] Some of these methods also provide tools to learn representations for out-of-vocabulary pairs (Joshi et al., 2019; Camacho-Collados et al., 2019). However, their initial vector spaces are based on co-occurring word pairs only.

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 119–131

https://doi.org/10.18653/v1/P17

find that a text corpus provides enough information to infer pairwise relations without training on any specific examples of a given relation.

In our experiments, we focus on semantic relations in particular, which are the most suited for both word and relation embedding models. Neural word embeddings are used to learn representations from text corpora, with word analogies as the evaluation mechanism to test our hypothesis. We run an extensive set of control experiments where the co-occurrence information is completely removed from the reference corpora. The results show that, with relationship instance removal, analogy performance degrades to only a limited extent, overall.[3] This finding suggests that neural embeddings do not learn lexical relation regularities from examples of the relation, but that they are still able to be inferred through the semantic featurization of individual words.

2 Related Work

2.1 Understanding analogies in word embeddings

The surprising result of Mikolov et al. (2013b), showing that word embedding can solve linear analogy problems, led to a careful investigation by researchers from different fields. A line of research proposed mathematical formalisms to try to understand the intrinsic properties of word embeddings. Arora et al. (2016) was one of the first to provide a rigorous theoretical explanation on the linear algebraic structure of word embeddings. Their formalism is based on a latent variable model that makes assumptions on the nature of the vector space. Later works rely on the notion of paraphrasing (Gittens et al., 2017; Allen and Hospedales, 2019), based on the observation that different words can be used in similar contexts interchangeably, dropping some of the previous assumptions made by Arora et al. (2016). Concurrently, other works have attempted to provide explanations of the compositional properties of distributional models through additions (Levy and Goldberg, 2014a; Paperno and Baroni, 2016; Ethayarajh et al., 2019), which lie at the core of word analogy completion.

While these works formalize word analogies and attempt to explain how they work mathematically, our empirical analysis is focused on understanding the source of signal in corpora that affect the performance of word analogy completion, without asserting any predefined assumption. In particular, we are mostly interested in determining whether relationship pair co-occurrence in sentences is necessary in order for a word embedding to succeed at analogy completion.

2.2 Issues in word analogies

A number of publications have encountered methodological issues in the word analogy task through word embeddings. Levy and Goldberg (2014a) found that the addition operations may not be optimal, as they are reduced to three separate similarity problems that can be solved through more appropriate operations. Linzen (2016) showed that simple baselines based on nearest neighbour searches are competitive in the analogy categories proposed by Mikolov et al. (2013b). Because of this, Gladkova et al. (2016) proposed a new dataset, partially addressing some of the previous shortcomings. Other works have shown that linear relationships, while being implicit, are not directly apparent in the word embedding space, and therefore word analogies may not be the best method to retrieve this information (Drozd et al., 2016; Schluter, 2018; Bouraoui et al., 2018). Finally, Gonen and Goldberg (2019) and Nissim et al. (2020) cautioned against over-reliance on analogies as a means to uncover and correct for biases in word embeddings.

These methodological observations challenge the supremacy of analogy evaluations as the optimal proxy for downstream task performance of a word embedding. Nevertheless, they represent a valuable mechanism with which to compare the semantic regularities of two different neural embeddings. In particular, word analogies represent an ideal benchmark for our research questions, as the impact of co-occurrence statistics within word relations can be evaluated directly through analogy validation. This would not be the case for other more complicated tasks such as relation classification or extraction, which may add additional confounds.

3 Methodology

In this section we explain the experimental methodology we follow to answer our main research question. First, we briefly describe how to solve word

[3] In a few specific relations the degradation is more marked. Nonetheless, this degradation does not surpass 50% even in the most unfavorable case for accuracy, and for other metrics this degradation does not surpass 20%.

analogies using word embeddings (Section 3.1). We then explain our methodology to compile corpora to train word embeddings (Section 3.2).

3.1 Solving word analogies with neural word embeddings

The first step to solve word analogies using neural word embeddings is to first learn word vectors from an unlabeled text corpora. To do so, standard word embedding models such as Word2Vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014) or FastText (Bojanowski et al., 2017) are often used. The output of these models is a vector space where each word is represented as a single point. With these vectors, mathematical operations can then be made to solve a given analogy.

Formally, given three words (a, b and c), the task of word analogy completion consists of predicting the most appropriate word d that satisfies a is to b as c is to d. In this case, both a-b and c-d are part of the same relationship. For instance, in the case of *Paris*, *France* and *Berlin*, the word to retrieve would be *Germany*. In this case both *Paris-France* and *Berlin-Germany* belong to the *capital-of* relation. With word embeddings this can be solved with the simple vector operation[4] $\vec{b} - \vec{a} + \vec{c}$, retrieving the word whose vector is closest to that point in the space.

Word analogy completion is used as the main evaluation for our experiments.

3.2 Corpus preparation

Our main research question is whether an explicit observation of a relationship in a corpus is necessary to complete an example of that relationship via analogy. To this end, given a reference unlabeled corpus, we devise the following methodology per lexical relation type.

3.2.1 Sentence removal

First, for each relation type (e.g. *capital-of*) in a dataset, we remove all sentences from the corpus that contain word pairs belonging to the relation. This results in a modified corpus for each relation type and a respective word embedding for each relation type trained on the modified corpus. For example, for the pair *Lisbon-Portugal*, we would remove all sentences from our reference corpus where *Lisbon* and *Portugal* co-occur, and this process would be repeated for all pairs of the *capital-of* relation.

3.2.2 Sentence replacement

This setting is similar to the sentence removal strategies with the added inclusion of new sentences to replace the removed ones. In particular, for each removed sentence containing a word pair of a relation type, two sentences from a similar corpus replace it, where each sentence contains one of the words in the pair. This is arguably the most fairly comparable setting to no removal, as the number of occurrences for each word in the relation would be approximately the same with respect to the default setting. The overall number of sentences would be slightly higher, though this is negligible in comparison to the full corpus (see Section 4.1 for the specific details on the corpora used for the evaluation).

For both the sentence replacement and sentence removal strategies, we also experiment with a more aggressive setting. In this setting (referred to as *removal+* or *replacement+* in our experiments), all sentences containing any two words from the vocabulary of the relation[5] are removed, in the case of removal+, or replaced, in the case of replacement+. This is the most aggressive setting where no co-occurrence information of a given relation is preserved.

4 Evaluation

In this section we provide the details of our experimental setup (Section 4.1) and then, the main results of our evaluation (Section 4.2).

4.1 Experimental Setting

In the following we describe the experimental setting for all our experiments. More details and code to reproduce our experiments can be found online[6].

Text corpora. As our reference corpus we selected UMBC (Han et al., 2013), which is a diverse 3-billion-token corpus of paragraphs extracted from the web, amounting to a total of 132M sentences. In particular, we randomly select 80% of all sentences which would be the base corpus for the experiments. Then, we used the remaining

[4] Levy and Goldberg (2014a) showed that a multiplication operation can also be used to solve analogies. However, for this analysis we focused on the traditional solution of the problem proposed by Mikolov et al. (2013b).

[5] Recall from the previous subsection (Section 3.1) that each analogy instance is formed by four words.

[6] https://github.com/h-yuc/Lexical-Relations-Analogies

20% to add replacement sentences when necessary (see Section 3.2.2 for more details on the replacement setting). To complement the main results and test the generalization of our findings, we also use Wikipedia[7] (2 billion tokens and 104M sentences) for the base removal experiments.

Word embedding models. For word embedding models, we use both the CBOW and Skip-Gram variants of Word2Vec (Mikolov et al., 2013a). These neural representation learning approaches have shown to be amenable to analogy completion since their introduction (Mikolov et al., 2013b). Unlike FastText (Bojanowski et al., 2017), they do not include character information and are therefore more suitable for our experiments as pure word-based models. We use standard hyperparameters for both CBOW and Skip-Gram, with 300-dimensions and a window size of 10 in both cases. Given the difference in speed (CBOW being around five times faster to train) and the small performance difference, we considered CBOW for all our main experiments, but included Skip-Gram results in the appendix.

Validation datasets. Google (Mikolov et al., 2013b) and BATS (Gladkova et al., 2016) relation example datasets are used for analogy completion. BATS was introduced after the Google dataset to address some of its shortcomings in the number and type of relations, so the inclusion of both datasets in our experiments help give a more general overview. In particular, we focus on the semantic relations of each dataset. Table 1 shows the statistics of the relations considered in each of the datasets.

Evaluation protocol. For solving word analogies, we follow the original methodology of Mikolov et al. (2013b)[8], as explained in Section 3.1. For simplicity, we unify the evaluation setting for both datasets where we only consider a single solution. In the case of BATS, we consider the first answer as provided in the dataset, which is generally the most specific. With the expectation that performance after co-occurrence removal may degrade substantially, we report analogy completion results using recall at 1 (accuracy), 10, and 50. Recall@50, for example, reports the percentage of analogy completions in which the expected word was among the 50 nearest neighbors to the three

	Relation type	#inst	#word	Example
Google	**cap-comm-country**	506	46	Cairo, Egypt, Paris, France
	capital-world	4,524	232	Muscat, Oman, Tokyo, Japan
	currency	866	60	Europe, euro, Korea, won
	city-in-state	2,467	94	Toledo, Ohio, Dallas, Texas
	family	506	46	king, queen, man, woman
	nationality-adj.	1,599	82	Greece, Greek, Spain, Spanish
BATS	**country-capital**	2450	100	Hanoi, Vietnam, Rome, Italy
	country-language	2,450	100	Jordan, Arabic, USA, English
	UK_city-county	2,450	100	Exeter, Devon, Wells, Somerset
	name-nationality	2,450	100	Caesar, Roman, Plato, Greek
	name-occupation	2,450	100	Plato, philosopher, Dante, poet
	animal-young	2,450	100	bee, larva, ox, calf
	animal-sound	2,450	100	bee buzz frog ribbit
	animal-shelter	2,450	100	horse, stable, ant, anthill
	things-color	2,450	100	coffee, black, cream, white
	male-female	2,450	100	son, daughter, father, mother
	hypernym-animal	2,450	100	human, primate, cat, feline
	hypernym-misc	2,450	100	pastry, food, plum, fruit
	hyponym-misc	2,450	100	bag, pouch, dessert, cake
	meronym-subst.	2,450	100	bag, leather, penny, metal
	meronym-member	2,450	100	bird, flock, page, book
	meronym-part	2,450	100	day, hour, dollar, cent
	synonym-intensity	2,450	100	angry, furious, ask, beg
	synonym-exact	2,450	100	help, aid, child, kid
	antonym-gradable	2,450	100	aware, unaware, slow, fast
	antonym-binary	2,450	100	after, before, below, above

Table 1: Statistics of the word analogy datasets used in our experiments: Number of instances (#inst) and unique words (#word).

word subtraction and addition. The inclusion of recall at different thresholds allows for a more complete overview of the performance, as the standard accuracy measure alone may not reflect the full picture (Gladkova et al., 2016; Schluter, 2018).

Training. For each relation type, we utilized three different variants of each corpora that are used to train word embeddings (i.e., the original corpus and two resulting from our removal and replacement strategies, as explained in Section 3.2). To reduce the amount of training, we only considered the default and removal strategy for Wikipedia[9], while all experiments are performed on the main UMBC reference corpus. In total, we compiled 156 different corpora, occupying 2.2TB of disk space, and learned 184 different word embedding models, totalling around 1,980 hours (around 83 full days) of model training on a high performance single node (48-core) system.

4.2 Results

As explained in the previous section, our experiments are aimed at understanding the role of co-occurring words in a given relation type (e.g.

[7] We use the Wikipedia dump of November 2016.

[8] As in the original protocol, words that are included in the instance are excluded from the nearest neighbours search.

[9] Likewise, for Skip-Gram we only considered the Google analogy dataset and a single strategy per corpus (results in the appendix).

		Sentences removed		Recall@1 (accuracy)					Recall@10					Recall@50				
		Rm/Rp	Rm+/Rp+	Def	Rm	Rp	Rm+	Rp+	Def	Rm	Rp	Rm+	Rp+	Def	Rm	Rp	Rm+	Rp+
Google	**cap.-country**	49,248	245,677	61.7	53.0	52.4	51.4	51.6	89.7	81.6	82.6	83.4	83.4	98.2	90.3	90.5	91.3	92.3
	cap.-world	79,060	452,915	49.4	31.3	32.7	30.1	33.6	82.0	65.8	67.9	65.0	67.2	91.2	79.2	80.5	78.6	80.8
	currency	4,260	145,179	8.7	5.1	6.9	5.8	6.2	37.3	32.6	34.2	33.1	34.5	63.5	58.7	59.6	58.4	60.4
	city-state	96,666	247,238	14.1	7.7	7.7	8.6	9.1	40.4	28.0	28.0	28.9	31.5	62.1	50.1	50.0	49.9	53.8
	family	450,875	2,830,852	91.1	85.0	82.0	69.8	72.9	100.0	95.3	94.7	88.1	89.7	100.0	98.4	96.8	90.5	91.1
	nation-adj	145,064	492,671	86.4	80.2	81.7	81.3	83.1	96.0	96.1	95.6	96.4	95.9	97.9	97.8	97.9	97.7	97.4
	AVERAGE	**137,529**	**735,755**	**51.9**	**43.7**	**43.9**	**41.2**	**42.7**	**74.2**	**66.5**	**67.1**	**65.8**	**67.0**	**85.5**	**79.1**	**79.2**	**77.7**	**79.3**
BATS	**country-cap**	66,988	369,463	71.6	54.7	55.8	56.7	59.6	90.9	86.9	87.3	86.1	86.5	95.1	93.0	93.1	91.6	91.2
	country-lang	21,066	649,071	26.7	24.0	24.9	16.9	19.6	60.8	56.9	57.4	47.5	52.5	74.8	73.3	73.1	62.6	69.1
	city-county	2,599	56,719	1.8	1.2	1.2	0.2	0.7	10.4	7.8	8.3	2.5	5.7	27.7	25.3	25.0	6.3	17.8
	name-nation.	12,069	1,808,784	20.6	17.5	18.1	18.9	16.1	54.7	50.7	51.6	50.6	46.9	72.5	68.9	69.6	65.1	65.9
	name-occup.	10,934	1,030,106	45.1	41.8	42.1	12.0	34.9	72.2	70.6	69.8	30.4	65.1	80.6	79.4	79.0	44.9	76.0
	animal-young	3,238	221,647	3.6	2.7	3.0	1.2	3.4	15.5	13.6	12.4	5.8	10.2	28.4	26.4	25.6	10.7	21.4
	animal-sound	1,307	75,529	3.8	2.7	3.5	0.8	2.0	12.2	9.0	11.0	2.8	6.4	20.0	15.9	17.2	4.5	11.1
	animal-shelter	11,892	569,567	3.3	1.7	1.7	0.6	1.0	16.6	12.4	12.5	3.4	9.3	32.0	26.0	25.7	8.0	18.8
	things-color	52,219	1,179,341	11.2	11.5	11.8	12.5	10.9	41.6	39.1	41.0	41.1	33.7	59.4	55.2	56.7	54.7	49.5
	male-female	108,898	445,429	48.0	42.8	43.5	38.6	40.0	70.4	69.3	68.3	63.6	64.7	79.0	77.4	77.0	73.6	72.7
	hyper-animal	3,410	93,258	9.6	8.4	7.8	1.6	4.4	41.1	37.5	38.4	9.8	29.3	67.1	63.8	64.8	24.1	52.5
	hyper-misc	15,157	456,525	4.7	4.4	4.2	1.6	3.1	23.8	22.9	22.3	14.8	20.9	43.0	41.5	40.8	28.2	37.6
	hypo-misc	32,118	274,377	9.8	10.3	9.3	8.8	10.9	55.9	54.2	55.4	49.1	55.1	81.3	79.3	79.1	73.4	79.0
	mero-subst.	68,128	1,548,273	6.5	5.4	5.3	2.9	3.7	30.7	26.0	25.8	15.6	20.8	52.5	45.2	45.4	30.5	40.7
	mero-member	257,447	4,897,322	3.1	2.9	2.8	2.3	3.0	26.9	24.7	25.3	17.9	25.9	44.7	41.8	42.1	30.4	43.4
	mero-part	30,016	364,053	2.9	2.7	2.4	2.7	2.7	29.0	27.4	27.2	23.9	26.9	49.8	48.1	48.1	42.0	46.7
	syn-intens.	40,241	1,329,093	17.9	17.7	17.4	14.8	19.0	49.6	47.1	49.0	42.3	47.2	70.2	68.4	68.9	59.8	68.3
	syn-exact	49,292	882,075	27.8	26.0	25.8	26.6	27.5	69.0	66.2	65.4	66.9	68.6	85.8	84.2	83.8	84.4	84.9
	auto-grad.	237,221	2,397,131	21.3	20.0	19.8	17.1	22.1	46.2	43.3	43.6	40.4	46.9	65.1	61.5	61.6	59.1	65.4
	auto-binary	1,648,965	42,023,189	27.6	26.2	27.6	28.6	35.1	59.2	56.1	56.7	57.4	67.2	71.1	69.9	70.0	71.8	80.6
	AVERAGE	**133,660**	**3,033,547**	**18.3**	**16.2**	**16.4**	**13.3**	**16.0**	**43.8**	**41.1**	**41.4**	**33.6**	**39.5**	**60.0**	**57.2**	**57.3**	**46.3**	**54.6**

Table 2: UMBC corpus word analogy results using CBOW with five different configurations: Default (Def), Remove (Rm), Replace (Rp) and their more aggressive counterparts removing all pairwise co-occurrences (Rm+ and Rp+).

capital-of) as they pertain to analogy completion. Table 2 shows the main set of results of the CBOW model on the UMBC corpus. As can be observed, and as expected, the default model (i.e., no co-occurrence removal) trained on the original corpus provides the highest analogy completion results, overall. However, less expected is the low magnitude decrease in performance of the experiments involving co-occurrence removal. The default experiments performed analogies with 51.9% accuracy (R@1), on average, compared to 42.7% with the most aggressive replace plus (Rp+) strategy on the Google dataset, and 18.1% vs. 16.0%, respectively, on the BATS dataset. Figure 1 shows the average decrease in performance of the replace strategy (Rp) per relation type as a percentage of the default performance. For R@1, the decrease in performance is lower than 10% for the majority of relations. When considering R@10 and R@50, this decrease in performance is even less pronounced, which suggests that the main geometrical features of the space were largely preserved (a more visualization of the space is presented in Section 5.2). The *animal-shelter* significant decrease in performance is a special case as the performance of the default model was very low to start with (3.3%), which highlights the difficulty to model that particular relationship via word analogies. The same could be attributed to the *city-state* (Default accuracy of 14.1%), which is the relation with the second highest decrease in performance for R@1.

Finally, Table 3 shows experimental results for the default, remove (Rm) and remove plus (Rm+) strategies for models trained on the Wikipedia corpus. The results are slightly higher overall, given the clean, consistent, and topically comprehensive nature of the corpus. The overall difference between default and removal strategies was similar to that of UMBC (reminder that the remove plus strategy consisted of removing all sentences where any two words from a given relation co-occur). No co-occurrence removal (Def) vs. removal plus (Rm+) had analogy completion accuracies of 59% vs. 52.5%, respectively, on the Google dataset, and

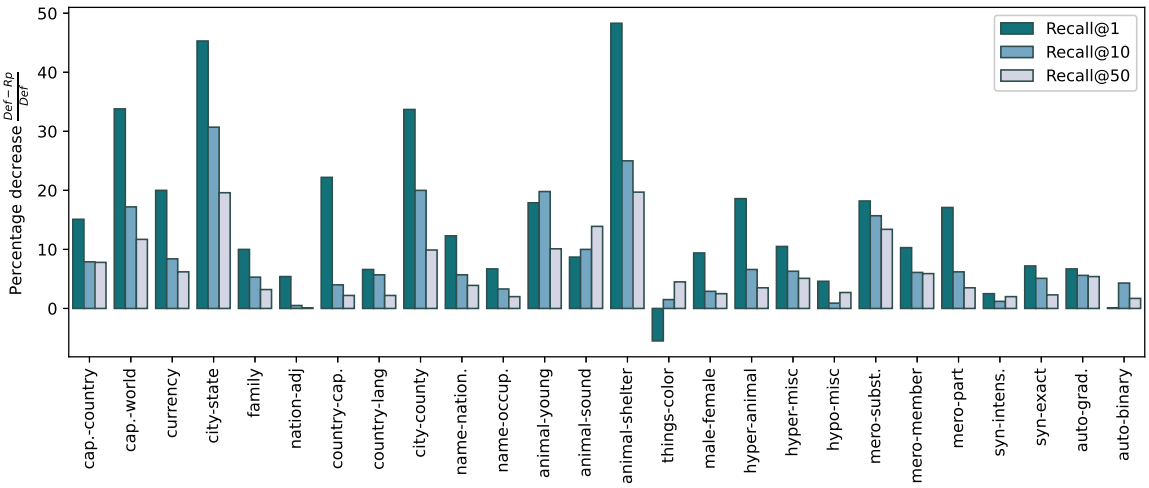

Figure 1: Average decrease in performance (%) of the CBOW model trained on the original UMBC corpus with respect to the replace strategy (Rp) per relation type (Google and BATS datasets).

21.7% vs. 20.3% on the BATS dataset.

5 Analysis

In this section we aim to better understand the results presented in the previous section. In particular, we analyze to what extent the performance drops that exist are correlated with frequencies of words in the relationship (Section 5.1). We also compare, through principal component analysis (PCA) visualization, the structure of the relation with the highest point decrease, before and after co-occurrence removal (Section 5.2).

5.1 Correlation with word frequency

A natural question that may arise when looking at the results is whether word frequency has any influence on the performance drop of co-occurring word pairs. For instance, one may wonder if getting a high-quality word embedding, which is generally achieved when word frequency is sufficiently high in the corpus, is enough to compensate for the lack of sentences with words forming a certain relation. Or, alternatively, if the relative frequency of co-occurring words in a relation has an effect on the final embedding, as this would mean that frequency is a necessary condition for learning robust semantic regularities. To answer these questions, we computed the correlation between word and pair frequencies in a word analogy instance and the performance drop. For the frequency indicator, we computed two numbers, H_{ind} and H_{pair}:

1. Harmonic mean[10] between all individual words in the analogy instance (H_{ind}) was computed as follows:

$$H_{ind} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \quad (1)$$

where x_i is the frequency of a word in the given word analogy instance and n is the number of words, i.e., four in the case of individual words in word analogies. For example, the harmonic mean of the four words, *king* (220,958 occurrences in UMBC), *queen* (52,262), *man* (751,262), and *woman* (296,915) would be 141,048.

2. The relative pairwise frequency (H_{pair}) was computed as the previous number divided by the harmonic mean of the number of sentences where two words of a pair co-occur (i.e., H_{co}):

$$H_{pair} = \frac{H_{co}}{H_{ind}} = \frac{2 \cdot p_1 \cdot p_2}{(p_1 + p_2)} \cdot \frac{1}{H_{ind}} \quad (2)$$

where p_i corresponds to the frequency of a relation pair in a word analogy instance. This number can give a better indication of how relevant the co-occurrence information of a given word pair is. Following the previous example, the relative pairwise frequency H_{pair} of the instance composed of *king-queen* (5,498 joint co-occurrences in UMBC) and *man-woman* (36,189) is 0.068.

[10] We decided to use the harmonic mean because it is generally more robust to outliers (e.g., a highly frequent word) than the usual arithmetic mean.

		Sentences removed		Recall@1 (accuracy)			Recall@10			Recall@50		
		Rm	Rm+	Def	Rm	Rm+	Def	Rm	Rm+	Def	Rm	Rm+
Google	**capital-country**	162,671	521,038	61.9	54.6	55.5	93.3	89.5	90.3	97.6	94.9	96.1
	capital-world	320,106	1,102,291	66.8	48.1	51.4	92.9	80.6	82.1	97.2	90.3	91.1
	currency	4,327	223,588	31.4	28.5	25.3	60.9	58.2	51.3	77.9	78.4	71.0
	city-in-state	344,196	659,539	19.7	12.6	15.2	54.0	34.8	39.4	72.3	55.7	58.5
	family	375,904	1,012,935	78.1	73.3	72.3	94.3	92.9	89.7	95.7	94.5	93.9
	nationality-adjective	437,316	1,381,200	95.9	94.9	95.5	98.5	98.6	97.6	99.8	99.7	98.9
	AVERAGE	**274,087**	**816,765**	**59.0**	**52.0**	**52.5**	**82.3**	**75.8**	**75.1**	**90.1**	**85.6**	**84.9**
BATS	**country-capital**	267,095	885,924	83.8	79.5	78.7	91.1	90.0	88.9	93.6	91.8	91.3
	country-language	82,842	527,339	28.6	25.3	23.9	62.0	58.5	58.9	72.7	70.9	71.2
	UK_city-county	39,780	198,005	17.0	9.1	9.7	43.5	34.0	34.3	59.0	52.3	51.8
	name-nationality	31,980	390,864	31.4	29.5	30.7	56.0	53.9	56.5	67.8	65.2	69.1
	name-occupation	24,798	136,447	39.9	35.9	34.7	63.0	59.5	58.3	72.1	68.8	67.8
	animal-young	7,021	135,815	5.4	2.4	2.4	20.0	12.5	10.5	33.1	24.4	22.3
	animal-sound	2,921	121,416	5.3	2.8	1.7	15.2	10.5	7.6	24.2	18.6	14.5
	animal-shelter	23,832	262,708	2.0	1.3	0.9	8.7	5.4	4.6	17.5	12.1	10.7
	things-color	47,298	445,312	14.5	15.0	15.2	39.9	38.6	42.8	55.6	52.7	56.2
	male-female	400,035	1,387,664	51.8	47.4	46.5	77.6	73.7	71.6	84.5	82.4	80.7
	hypernyms-animals	11,723	77,632	20.1	16.8	14.4	57.6	52.7	47.4	75.4	71.0	66.6
	hypernyms-misc	16,825	164,097	7.4	6.7	6.2	33.1	29.9	30.3	55.8	53.0	51.9
	hyponyms-misc	90,811	594,541	12.4	12.4	10.9	52.2	51.2	49.4	72.7	70.6	69.8
	meronyms-substance	70,896	510,291	7.8	6.7	6.9	28.9	24.2	24.9	48.7	43.4	41.4
	meronyms-member	774,473	4,259,786	10.4	9.1	12.0	33.3	28.2	35.6	49.9	44.7	53.0
	meronyms-part	91,041	539,371	6.7	6.1	5.5	34.0	31.5	31.7	56.5	52.0	51.4
	synonyms-intensity	44,149	624,662	15.6	14.9	14.4	42.9	40.4	40.7	61.3	60.7	58.7
	synonyms-exact	83,651	1,308,183	23.2	20.7	25.4	52.6	50.0	54.4	69.1	67.1	70.2
	antonyms-gradable	282,330	1,233,454	21.6	18.5	22.3	49.0	46.3	49.6	67.1	64.6	66.5
	antonyms-binary	1,726,724	21,329,952	29.4	26.1	44.6	57.7	55.4	77.2	72.2	70.7	85.6
	AVERAGE	**206,011**	**1,756,673**	**21.7**	**19.3**	**20.3**	**45.9**	**42.3**	**43.8**	**60.4**	**56.8**	**57.5**

Table 3: Wikipedia corpus word analogy results using CBOW with three different configurations: Default (Def), Remove (Rm), and its more aggressive setting removing all pairwise co-occurrences (Rm+)

With respect to performance drops, for each analogy completion instance we considered the ranks[11] of the correct completion words in both the default and replace settings and computed the difference. Table 4 shows the results of the correlation results in the Google analogy dataset. Not surprisingly, the correlation between the individual frequency of the words in an instance and the rank difference is negative in all relation types except for one (*nationality-adjective*). However, the correlation is rather weak, as the addition of new sentences compensate for the initial removal, even if the sentences are of a different kind. As for the relative frequency of pairs in the instance, the correlation is positive as expected. In this case, the signal is higher than with the individual frequency case, especially in the *family* relationship. Overall, this experiment shows a level of support for our initial premises on the effect of relative pair frequencies, but further research would be necessary to understand other reasons behind the performance drop.

[11] We only considered the position of the fifty first nearest neighbours. If the correct word was not among the fifty nearest neighbours, 51 was used as the position, which would be equivalent to a wrong answer.

	Frequency		**Drop**	**Correlation**	
	Ind	**Pair**		H_{ind}	H_{pair}
cap.-country	26,111	2,002	15.1	-0.239	-0.114
cap.-world	4,328	434	33.8	-0.128	+0.188
currency	3,815	52	20.0	-0.092	+0.041
city-state	14,226	1,121	45.3	-0.012	+0.192
family	58,197	5,617	10.0	-0.065	+0.315
nation-adj	30,076	2,238	5.4	+0.063	+0.021
AVERAGE	**22,792**	**1,911**	**22.9**	**-0.079**	**+0.107**

Table 4: Pearson correlations between frequency (average among all instances in the dataset) and performance drop between the default and replace (Rp) corpora.

5.2 Visualization

In this section, we present visualizations of the word pairs from one lexical relation, before and after co-occurrence removal, in order to gain insight into the effect of removal on the learned structure of the space. In particular, we selected the relation from the Google datasets with the largest raw

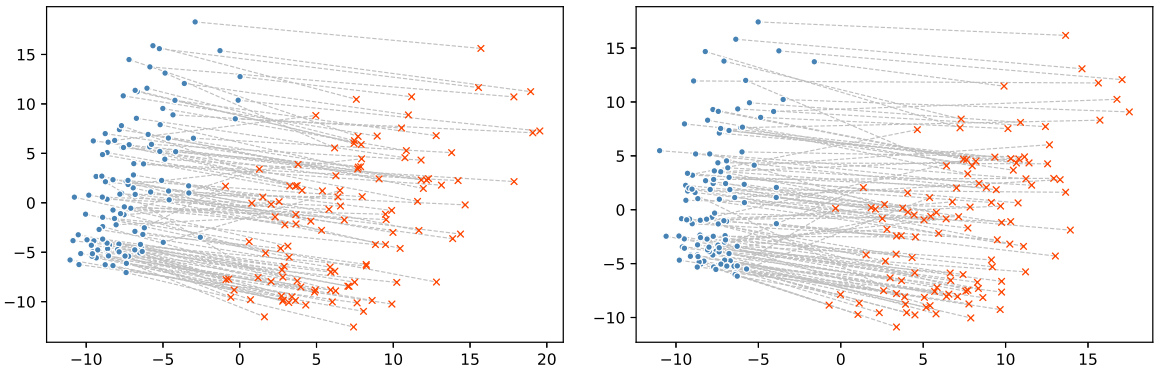

Figure 2: *Capital-world* vectors projected by PCA. All models trained on different subsets of UMBC: default (left) and remove (right). Red marks (x) correspond to *countries* and blue dots, *capitals*.

performance point drop[12]. Figure 2 shows the two principal components of word pairs in the *capital-world* relation for both the default (Def) and remove (Rm) settings. A reminder from Section 3.2.1, the remove setting involves removing all sentences where both words from a relation co-occur, without any replacement. As can be seen, even in the case of a relation with a R@1 performance gap as high as 36.6%, the linear relations are largely preserved in the word embedding space. This can also be supported by the fact that the performance drop is much smaller for R@10 and R@50 (see Figure 1), suggesting that the correct completion word is still somewhere near in the space. For example, for the instance *Australia, Canberra, Spain*, the correct word *Madrid* is found as the fifth nearest neighbours in the Rm vector space, with the top two words being *Barcelona* and *Valencia* (other large Spanish cities). Intuitively, this error does not affect the overall representation of the relation in the vector space, as those words were already in a similar linear relationship in the default model.

6 Discussion

6.1 The source of semantic regularities

In this paper, we find that neural word embeddings (i.e., Word2Vec models) do not require observation of instances of the relation (e.g., Madrid is the *capital of* Spain) in order to maintain nominal accuracy in relation completion tasks. We believe this is the first time such an observation has been made, empirically, using natural language, though it has been observed in neural embeddings trained on non-linguistic data (Pardos and Nam, 2020).

In Mikolov et al. (2013b), where Word2Vec models were first introduced, the phrase "Linguistic Regularities" was used. While it was not made explicit what the word regularities referred to, analogy completion was exclusively used for validation, leaving open the possibility that regularities referred to some pattern of structure allowing for lexical relationships to be expressed. If the regularities relevant to analogy completion are not formed from examples of the lexical relationship contained in the analogy, then how are they formed and how was the accuracy of the completion mostly retained in our experiments in the absence of examples? Instead, it may leverage the robustness of regularities, or features, learned about individual words to lay the structural foundation for inferences to then be made about a lexical relation. Removing co-occurrences of capitals and countries, for example, would not completely remove the concept of capitals and countries from the corpus. The embedding of "Madrid" would likely still encode features associated with a busy city, government buildings, culture, and European regionality. This is also related to work that showed relations and relevant information from relations can be captured from word embeddings (Jadhav et al., 2020), even if the relation cannot be retrieved explicitly from linear transformations (Drozd et al., 2016; Bouraoui et al., 2018). Interestingly, however, our results indicate that the frequency of an individual word in a corpus is only weakly related to the robustness of features leveraged for successful analogy completion.

Finally, even though co-occurrences of pairs from a specific relation are not necessary to learn the necessary features, word pairs still may play a critical role in regularity development. Most

[12] In the appendix we also include the same visualizations for the relations with the second largest performance drop (*city-in-state*) and the smallest performance drop (*nationality-adjective*), which largely share the same conclusions.

word embedding models (including the Skip-Gram model of Word2Vec) are trained in pairwise fashion after-all, making predictions and calculating loss based on each pair of input and context words.

6.2 Cognitive perspective

Neural embeddings come from a cognitive perspective on semantic representation. They stem from a hypothesized architecture of the mind called Connectionism (Feldman and Ballard, 1982) in which emergent concepts (Hopfield, 1982; Hinton, 1986) are learned as distributed representations across the embedding space (Hinton et al., 1986). If neural word embeddings are a candidate model of a component of human cognition, then our results suggest that the faculties of the mind that understand relational concepts (e.g., male and female) may establish these concepts primarily through induction and observations of behavior. For example, this would mean that we learn features of male and female separately, rather than through explicit declaration of representative pairs (i.e., explicit co-occurrences). It is perhaps a separate faculty of the mind that queries this conceptual representation framework for inferences to be made about relationships between new elements. These inferences, conducted by way of analogy, may indeed be key to innovation (Hope et al., 2017) and a possible component of human creativity (Holyoak et al., 1996).

7 Conclusion and Future Work

In this paper we have presented a large-scale analysis on the role of co-occurring relational word pairs in completing analogies. In the analyses we have measured to what extent the loss of co-occurrence information within relation types affects analogy completion using neural word embeddings. Perhaps surprisingly, this effect is quite small, to the point that word embeddings can complete analogies of a relationship in the vector space even if the co-occurrence information from the reference corpus is totally removed.

In order to complement this analysis, for future work it would be interesting to analyze to what extent the conclusions of this analysis apply to purely distributional models, e.g., PMI-based, as they have shown to share similarity properties with word embeddings (Levy et al., 2015), to the point of Skip-Gram being viewed as an implicit co-occurrence matrix factorization (Levy and Goldberg, 2014b). Moreover, the analysis could be extended to other types of relations, not only semantic. Further investigation could then focus on how the main sources of concepts and linguistic regularities in word embeddings are learned, and how they can be leveraged to improve unsupervised relation models, e.g., (Jameel et al., 2018; Joshi et al., 2019). Finally, as a follow-up to recent work aiming at understanding how language models and contextualized embeddings capture relations (Petroni et al., 2019; Bouraoui et al., 2020; Jiang et al., 2020), further research could be devoted to analyze the performance of such models with and without pairwise co-occurrence information.

References

Carl Allen and Timothy Hospedales. 2019. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, pages 223–231.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5(1):135–146.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *Proceedings of AAAI*.

Zied Bouraoui, Shoaib Jameel, and Steven Schockaert. 2018. Relation induction in word embeddings revisited. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1627–1637, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jose Camacho-Collados, Luis Espinosa-Anke, Shoaib Jameel, and Steven Schockaert. 2019. A latent variable model for learning distributional relation vectors. In *Proceedings of IJCAI*.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 3519–3530.

Luis Espinosa-Anke and Steven Schockaert. 2018. SeVeN: Augmenting word embeddings with unsupervised relation vectors. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2653–2665, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262.

Jerome A. Feldman and Dana H. Ballard. 1982. Connectionist models and their properties. *Cognitive Science*, 6(3):205–254.

John R. Firth. 1957. A synopsis of linguistic theory. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32.

Alex Gittens, Dimitris Achlioptas, and Michael W Mahoney. 2017. Skip-gram- zipf+ uniform= vector additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–76.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the Student Research Workshop at NAACL*, pages 8–15.

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. UMBC EBIQUITY-CORE: Semantic textual similarity systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1, pages 44–52.

Geoffrey E. Hinton. 1986. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA.

Geoffrey E. Hinton, James L. McClelland, and David E. Rumelhart. 1986. Distributed representations. *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*, pages 77–109.

Keith J Holyoak, Keith James Holyoak, and Paul Thagard. 1996. *Mental leaps: Analogy in creative thought*. MIT press.

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 235–243.

John J. Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558.

Aishwarya Jadhav, Yifat Amir, and Zachary A. Pardos. 2020. Lexical relation mining in neural word embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Shoaib Jameel, Zied Bouraoui, and Steven Schockaert. 2018. Unsupervised learning of distributional relation vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23–33, Melbourne, Australia. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Mandar Joshi, Eunsol Choi, Omer Levy, Daniel Weld, and Luke Zettlemoyer. 2019. pair2vec: Compositional word-pair embeddings for cross-sentence inference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3597–3608, Minneapolis, Minnesota. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014a. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of HLT-NAACL*, pages 746–751.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Denis Paperno and Marco Baroni. 2016. When the whole is less than the sum of its parts: How composition affects pmi values in distributional semantic vectors. *Computational Linguistics*, 42(2):345–350.

Zachary A. Pardos and Andrew J. H. Nam. 2020. A university map of course knowledge. *PLoS ONE*, 15(9).

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

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Natalie Schluter. 2018. The word analogy testing caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 242–246.

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

Koki Washio and Tsuneaki Kato. 2018. Filling missing paths: Modeling co-occurrences of word pairs and dependency paths for recognizing lexical semantic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1123–1133, New Orleans, Louisiana. Association for Computational Linguistics.

Appendix

A Additional Experiments

In this appendix we include some additional results that complement our main experiments.

Skip-gram results. Table 5 and 6 present the results of the Skip-gram model of Word2Vec trained in UMBC and Wikipedia, respectively.

Performance drop. Figure 3 presents the performance drop percentage of the remove setting (Rm) with respect to the default setting in UMBC.

B Visualizations

Figures 4 and 5 present visualizations of the word embedding space for the relations *nationality-adjective* and *city-in-state*, respectively, in both default (Def) and remove (Rm) settings. These figures complement Figure 2 of the main paper.

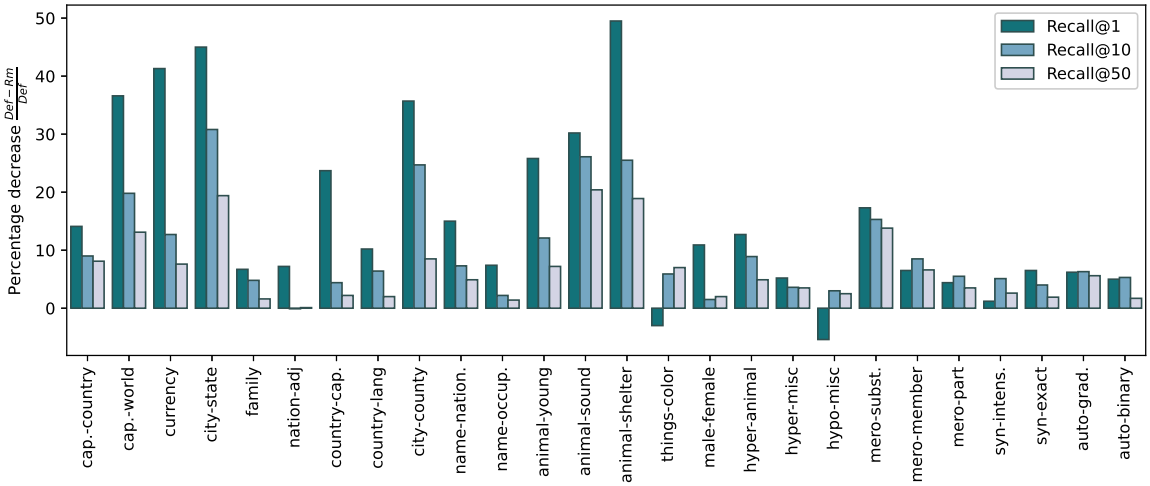

Figure 3: Average decrease in performance (%) of the CBOW model trained on the original UMBC corpus with respect to the removal strategy (Rm) per relation type (Google and BATS datasets).

		Recall@1			Recall@10			Recall@50		
		Def	Rp	Rp+	Def	Rp	Rp+	Def	Rp	Rp+
Google	**cap.-country**	61.7	51.0	48.2	91.7	85.2	86.2	98.4	95.1	93.1
	cap.-world	49.8	34.0	33.5	79.1	66.5	65.3	92.2	81.0	80.4
	currency	12.1	9.5	5.8	35.7	30.6	28.1	61.8	57.7	53.0
	city-state	16.6	10.2	10.7	54.6	37.5	38.4	76.3	62.5	62.8
	family	82.0	67.2	63.2	100.0	90.9	82.0	100.0	94.9	87.6
	nation-adj	92.4	89.2	87.9	98.0	96.6	95.9	99.8	99.3	97.2
	AVERAGE	**52.4**	**43.5**	**41.5**	**76.5**	**67.9**	**66.0**	**88.1**	**81.7**	**79.0**

Table 5: UMBC corpus word analogy results using Skip-gram with three different configurations: Default (Def), Replace (Rp), its more aggressive setting replacing all pairwise co-occurrences, i.e., Rp+.

		Recall@1			Recall@10			Recall@50		
		Def	Rm	Rm+	Def	Rm	Rm+	Def	Rm	Rm+
Google	**cap.-country**	68.0	52.0	45.9	96.1	86.8	84.6	99.2	97.0	95.9
	cap.-world	70.0	44.6	42.2	93.7	78.6	78.1	97.9	89.1	89.9
	currency	31.2	23.6	19.2	59.9	53.1	44.8	73.2	67.7	59.5
	city-state	19.9	8.9	10.5	56.8	29.9	32.4	75.4	50.4	51.2
	family	71.2	61.5	59.1	94.9	83.8	82.6	95.5	90.7	89.5
	nation-adj	99.7	98.7	96.7	100.0	99.9	99.3	100.0	100.0	99.8
	AVERAGE	**60.0**	**48.2**	**45.6**	**83.6**	**72.0**	**70.3**	**90.2**	**82.5**	**80.9**

Table 6: Wikipedia corpus word analogy results using Skip-gram with three different configurations: Default (Def), Remove (Rm), its more aggressive setting removing all pairwise co-occurrences, i.e., Rm+.

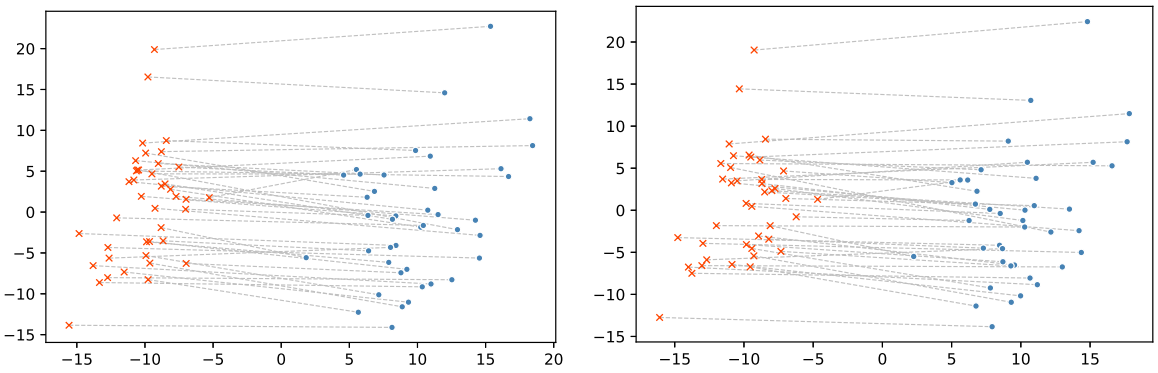

Figure 4: *Nationality-adjective* vectors projected by PCA. All models trained on different subsets of UMBC: default (left) and remove (right). Red marks (x) correspond to *country-adjectives* and blue dots, *countries*.

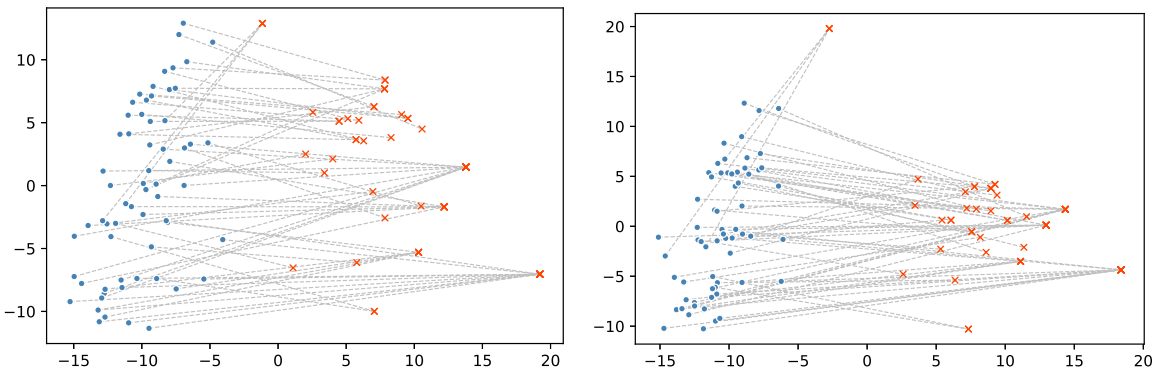

Figure 5: *City-in-state* vectors projected by PCA. All models trained on different subsets of UMBC: default (left) and remove (right). Red marks (x) correspond to *states* and blue dots, *cities*.