Improving Biomedical Analogical Retrieval with Embedding of Structural Dependencies

Amandalynne Paullada*, **Bethany Percha**†, **Trevor Cohen**‡
*Department of Linguistics, University of Washington, Seattle, WA, USA
† Dept. of Medicine and Dept. of Genetics & Genomic Sciences,
Icahn School of Medicine at Mount Sinai, New York, NY, USA
‡Dept. of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA
paullada@uw.edu, bethany.percha@mssm.edu, cohenta@uw.edu

Abstract

Inferring the nature of the relationships between biomedical entities from text is an important problem due to the difficulty of maintaining human-curated knowledge bases in rapidly evolving fields. Neural word embeddings have earned attention for an apparent ability to encode relational information. However, word embedding models that disregard syntax during training are limited in their ability to encode the structural relationships fundamental to cognitive theories of analogy. In this paper, we demonstrate the utility of encoding dependency structure in word embeddings in a model we call Embedding of Structural Dependencies (ESD) as a way to represent biomedical relationships in two analogical retrieval tasks: a relationship retrieval (RR) task, and a literature-based discovery (LBD) task meant to hypothesize plausible relationships between pairs of entities unseen in training. We compare our model to skip-gram with negative sampling (SGNS), using 19 databases of biomedical relationships as our evaluation data, with improvements in performance on 17 (LBD) and 18 (RR) of these sets. These results suggest embeddings encoding dependency path information are of value for biomedical analogy retrieval.

1 Introduction

Distributed vector space models of language have been shown to be useful as representations of relatedness and can be applied to information retrieval and knowledge base augmentation, including within the biomedical domain (Cohen and Widdows, 2009). A vast amount of knowledge on biomedical relationships of interest, such as therapeutic relationships, drug-drug interactions, and adverse drug events, exists in largely human-curated knowledge bases (Zhu et al., 2019). However, the rate at which new papers are published means new relationships are being discovered faster than human curators can manually update the knowledge bases. Furthermore, it is appealing to automatically generate hypotheses about novel relationships given the information in scientific literature (Swanson, 1986), a process also known as 'literature-based discovery.' A trustworthy model should also be able to reliably represent known relationships that are validated by existing literature.

Neural word embedding techniques such as `word2vec`[1] and `fastText`[2] are a widely-used and effective approach to the generation of vector representations of words (Mikolov et al., 2013a) and biomedical concepts (De Vine et al., 2014). An appealing feature of these models is their capacity to solve proportional analogy problems using simple geometric operators over vectors (Mikolov et al., 2013b). In this way, it is possible to find analogical relationships between words and concepts without the need to specify the relationship type explicitly, a capacity that has recently been used to identify therapeutically-important drug/gene relationships for precision oncology (Fathiamini et al., 2019). However, neural embeddings are trained to predict co-occurrence events without consideration of syntax, limiting their ability to encode information about relational structure, which is an essential component of cognitive theories of analogical reasoning (Gentner and Markman, 1997). Additionally, recent work (Peters et al., 2018) has found that contextualized word embeddings from language models such as ELMo, when evaluated on analogy tasks, perform worse on semantic relation tasks than static embedding models.

The present work explores the utility of encoding syntactic structure in the form of dependency paths into neural word embeddings for analogical

[1] https://github.com/tmikolov/word2vec
[2] https://fasttext.cc/

Proceedings of the BioNLP 2020 workshop, pages 38–48

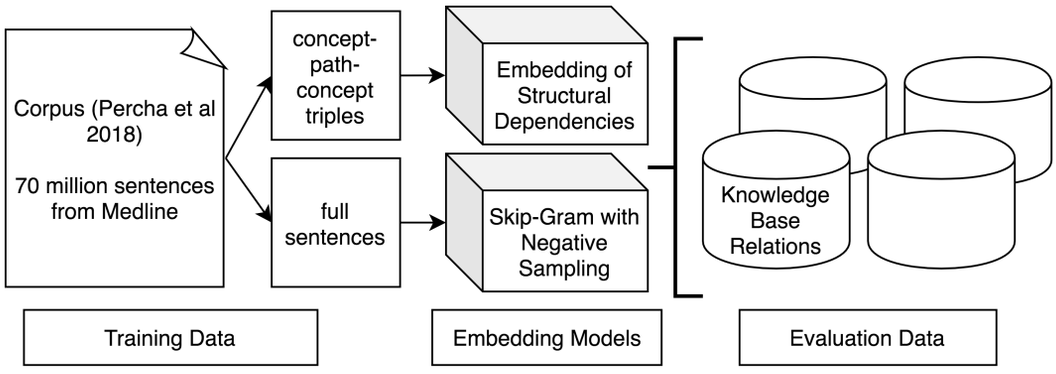


Figure 1: Overview of training and evaluation pipeline. Two embedding models, Embedding of Structural Dependencies (ESD) and Skip-gram with Negative Sampling (SGNS), are trained on data from a corpus of ≈70 million sentences from Medline. The resulting representations are then evaluated on data collected from biomedical knowledge bases.

retrieval of biomedical relations. To this end, we build and evaluate vector space models for representing biomedical relationships, using a corpus of dependency-parsed sentences from biomedical literature as a source of grammatical representations of relationships between concepts.

We compare two methods for learning biomedical concept embeddings, the skip-gram with negative sampling (SGNS) algorithm (Mikolov et al., 2013a) and Embedding of Semantic Predications (ESP) (Cohen and Widdows, 2017), which adapts SGNS to encode concept-predicate-concept triples. In the current work, we adapt ESP to encode dependency paths, an approach we call Embedding of Structural Dependencies (ESD). We train ESD and SGNS on a corpus of approximately 70 million sentences from biomedical research paper abstracts from Medline, and evaluate each model's ability to solve analogical retrieval problems derived from various biomedical knowledge bases. We train ESD on concept-path-concept triples extracted from these sentences, and SGNS on full sentences that have been minimally preprocessed with named entities (see §3). Figure 1 shows the pipeline from training to evaluation.

From an applications perspective, we aim to evaluate the utility of these representations of relationships for two tasks. The first involves correctly identifying a concept that is related in a particular way to another concept, when this relationship has already been described explicitly in the biomedical literature. This task is related to the NLP task of relationship extraction, but rather than considering one sentence at a time, distributional models represent information from across all of the instances in which this pair have co-occurred, as well as information about relationships between similar concepts. We refer to this task as *relationship retrieval (RR)*. The second task involves identifying concepts that are related in a particular way to one another, where this relationship has not been described in the literature previously. We refer to this task as *literature-based discovery (LBD)*, as identifying such implicit knowledge is the main goal of this field (Swanson, 1986).

We evaluate on four kinds of biomedical relationships, characterized by the semantic types of the entity pairs involved, namely *chemical-gene*, *chemical-disease*, *gene-gene*, and *gene-disease* relationships.

The following paper is structured as follows. §2 describes vector space models of language as they are evaluated for their ability to solve proportional analogy problems, as well as prior work in encoding dependency paths for downstream applications in relation extraction. §3 presents the dependency path corpus from Percha and Altman (2018). §4 summarizes the knowledge bases from which we develop our evaluation data sets. §5 describes the training details for each vector space model. §6 and §7 describe the methods and results for the *RR* and *LBD* evaluation paradigms. §8 and §9 offer discussion and conclude the paper. Code and evaluation data will be made available at `https://github.com/amandalynne/ESD`.

2 Background

We look to prior work in using proportional analogies as a test of relationship representation in the general domain with existing studies on vector space models trained on generic English. While our biomedical data is largely in English, we constrain

our evaluation to specific biomedical concepts and relationships as we apply and extend established methods.

Vector space models of language and analogical reasoning

Vector space models of semantics have been applied in information retrieval, cognitive science and computational linguistics for decades (Turney and Pantel, 2010), with a resurgence of interest in recent years. Mikolov et al. (2013a) and Mikolov et al. (2013b) introduce the skip-gram architecture. This work demonstrated the use of a continuous vector space model of language that could be used for analogical reasoning when vector offset methods are applied, providing the following canonical example: if x_i is the vector corresponding to word i, x_{king} - x_{man} + x_{woman} yields a vector that is close in proximity to x_{queen}. This result suggests that the model has learned something about semantic gender. They identified some other linguistic patterns recoverable from the vector space model, such as pluralization: x_{apple} - $x_{\text{apples}} \approx x_{\text{car}}$ - x_{cars}, and developed evaluation sets of proportional analogy problems that have since been widely used as benchmarks for distributional models (see for example (Levy et al., 2015)).

However, work soon followed that pointed out some of the shortcomings of attributing these results to the models' analogical reasoning capacity. For example, Linzen (2016) showed that the vector for 'queen' is itself one of the nearest neighbors to the vector for 'woman,' and so it can be argued that the model does not actually learn relational information that can be applied to analogical reasoning, but rather, can rely on the direct similarity between the target terms in the analogy to produce desirable results.

Furthermore, Gladkova et al. (2016) introduce the Better Analogy Test Set (BATS) to provide an evaluation set for analogical reasoning that includes a broader set of semantic and syntactic relationships between words. This set proved far more challenging for embedding-based approaches. Newman-Griffis et al. (2017) provide results of vector offset methods applied to a dataset of biomedical analogies derived from UMLS triples, showing that certain biomedical relationships are more difficult to learn with analogical reasoning than others.

Because the aim of this project is to robustly learn a handful of biomedical relationships, we are less concerned about the linguistic generalizability of these particular representations, but future work will examine the application of these vector space models to analogies in the general domain.

Dependency embeddings

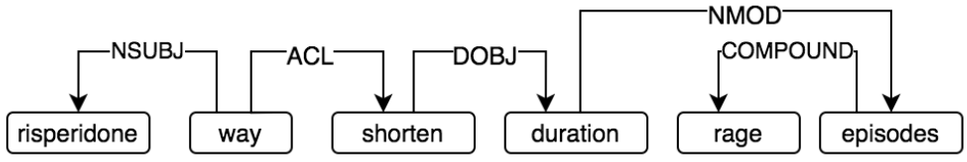
Levy and Goldberg (2014a) adapt the SGNS model to encode direct dependency relationships, rather than dependency paths. In this approach, a dependency-type/relative pair is treated as a target for prediction when the head of a phrase is observed (e.g. $P(scientist/nsubj|discovers)$). The dependency-based skipgram embeddings were shown to better reflect the functional roles of words than those trained on narrative text, which tended to emphasize topical associations. Recent work (Zhang et al. (2018), Zhou et al. (2018), Li et al. (2019)) has also integrated dependency path representations in neural architectures for biomedical relation extraction, framing it as a classification task rather than an analogical reasoning task. The work of Washio and Kato (2018) is perhaps the most closely related to our approach, in that neural embeddings are trained on word-path-word triples. Aside from our application of domain-specific Named Entity Recognition (NER), a key methodological difference between this work and the current work is that their approach represents word pairs as a linear transformation of the concatenation of their embeddings, while we use XOR as a binding operator (following the approach of Kanerva (1996)), which was first used to model biomedical analogical retrieval with semantic predications extracted from the literature by Cohen et al. (2011)[3]. On account of the use of a binding operator, individual entities, pairs of entities and dependency paths are all represented in a common vector space.

3 Text Data

We train both the ESD and SGNS models on data released by Percha and Altman (2018). This corpus[4] consists of about 70 million sentences from a subset of MEDLINE (approximately 16.5 million abstracts) which have `PubTator` (Wei et al., 2013) annotations applied to identify phrases that denote names of *chemicals* (including drugs and other chemicals of interest), *genes* (and the proteins they code for), and *diseases* (including side effects

[3] For related work, see Widdows and Cohen (2014)

[4] Version 7 of the corpus retrieved at `https://zenodo.org/record/3459420`



"Liquid **risperidone** may be a safe and effective way to shorten the duration of **rage** episodes."

`way_nsubj_START_ENTITY way_acl_shorten shorten_dobj_duration duration_nmod_episodes episodes_compound_END_ENTITY`

Figure 2: Example of a path of dependencies between two entities of interest. The full parse is not shown, but rather, the minimum path of dependency relations between the two entities given the sentence.

and other phenotypes). Throughout this paper, we use these shorthand names for each of these categories, following the convention established in Wei et al. (2013) and followed by Percha and Altman (2018).

The following example sentence from an article processed by `PubTator` shows how multi-word phrases that denote biomedical entities of interest, in this case *atypical depression* and *seasonal affective disorder*, are concatenated by underscores to constitute single tokens:

> Chromium has a beneficial effect on eating-related atypical symptoms of depression, and may be a valuable agent in treating atypical_depression and seasonal_affective_disorder.

Percha and Altman (2018) also provide pruned Stanford dependency (De Marneffe and Manning, 2008) parses for the sentences in the corpus, consisting, for each sentence, of the minimal path of dependency relations connecting pairs of biomedical named entities identified by `PubTator`. Specifically, they extract dependency paths that connect chemicals to genes, chemicals to diseases, genes to diseases, and genes to genes. Figure 2 shows an example of a dependency path of relations between two terms, *risperidone* and *rage*. We use these dependency paths as representations for predicates that denote biomedical relationships of interest by concatenating the string representations of each path element, which are shown below the sentence in Figure 2. Following Percha and Altman (2018), we exclude paths that denote a coordinating conjunction between elements and paths that denote an appositive construction, both of which are highly common in the set. In this corpus of 70 million sentences, there are about 44 million unique dependency paths that connect concepts of interest, the vast majority (around 40 million) of which appear just once in the corpus. 540,011 of these paths appear at least 5 times in the corpus.

4 Knowledge Bases

We construct our evaluation data sets with exemplars from knowledge bases for four primary kinds of biomedical relationships, characterized by the interactions between pairs of entities of the following types: *chemical-gene*, *chemical-disease*, *gene-disease*, and *gene-gene*.

We evaluate on pairs of entities from the following knowledge bases: DrugBank (Wishart et al., 2018), Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2005), PharmGKB (PGKB) (Whirl-Carrillo et al., 2012), Reactome (Fabregat et al., 2016), Side Effect Resource (SIDER) (Kuhn et al., 2016), and Therapeutic Target Database (TTD) Wang et al. (2020).

Each knowledge base consists of pairs of entities that relate in a specific way. For example, SIDER Side Effects consists of *chemical-disease*-typed pairs such that the chemical is known to have the disease as a side effect, e.g. (*sertraline*, *insomnia*). Meanwhile, another *chemical-disease* pair from a different database, Therapeutic Target Database (TTD) indications, is such that the chemical is indicated as a treatment for the disease, e.g. (*carphenazine*, *schizophrenia*). In constructing our evaluation sets, we process all terms such that they are lower-cased, and multi-word terms are concatenated by underscores. Furthermore, we eliminate from our evaluation sets any knowledge base terms that do not appear in the training corpus described in §3 at least 5 times. It should be noted that across these sets, a single biomedical entity may appear with numerous spellings and naming conventions.

Table 2 shows the corresponding relationship type for each of the knowledge bases we use, as well as the number of pairs from each that are used in our evaluation data. The relationship retrieval data consists of knowledge base pairs that appear in our training corpus connected by a dependency

path at least once, while the literature-based discovery targets are those knowledge base pairs that do not appear connected by a dependency path in the corpus.

5 Training Details

SGNS With SGNS, a shallow neural network is trained to estimate the probability of encountering a context term, t_c, within a sliding window centered on an observed term, t_o. The training objective involves maximizing this probability for true context terms $P(t_c|t_o)$, and minimizing it for randomly drawn counterexamples $t_{\neg c}$, $P(t_{\neg c}|t_o)$, with probability estimated as the sigmoid function of the scalar product between the input weight vector for the observed term and the output weight vector of the context term, $\sigma(\overrightarrow{t_o}.\overrightarrow{t_{c|\neg c}})$. We used the `Semantic Vectors`[5] implementation of SGNS (which performs similarly to the `fastText` implementation across a range of analogical retrieval benchmarks (Cohen and Widdows, 2018)) to train 250-dimensional embeddings, with a sliding window radius of two, on the complete set of full sentences from the corpus described in §3 as the training corpus. As previously mentioned, multi-word phrases corresponding to named entities recognized by the `PubTator` system in these sentences are concatenated by underscores, and consequently receive a single vector representation.

ESD With ESD, a shallow neural network is trained to estimate the probability of encountering the object, o, of a subject-predicate-object triple sPo. The training objective involves maximizing this probability for true objects $P(o|s,P)$ and minimizing it for randomly drawn counterexamples, $\neg o$, $P(\neg o|s,P)$. We adapted the `Semantic Vectors`[5] implementation of ESP to encode dependency paths, with binary vectors as representational basis (Widdows and Cohen, 2012) and the non-negative normalized Hamming distance ($NNHD$) to estimate the similarity between them.

$$\text{NNHD} = \max\left(0, 1 - \frac{2 \times \text{Hamming distance}}{\text{dimensionality}}\right)$$

With this representational paradigm, probability can be estimated as $NNHD(o, s \otimes P)$, where $\otimes$ represents the use of pairwise exclusive OR as a *binding operator*, in accordance with the Binary Spatter Code (Kanerva, 1996). While ESP was originally developed to encode knowledge extracted from the literature using a small set of predefined predicates (e.g. TREATS), we adapt it here to encode a large variety (n=546,085) of dependency paths. For training, we concatenate the dependency relations (the underscored parts in Figure 2) into a single predicate token for which a vector is learned. Some examples of path tokens (concatenated dependency relations) can be seen in Table 1. Unlike the original ESP implementation where predicate vectors were held constant, we permit dependency path vectors to evolve during training[6]. Further details on ESP can be found in (Cohen and Widdows, 2017). For the current work, we set the dimensionality at 8000 bits (as this is equivalent in representational capacity to 250-dimensional single precision real vectors). For ESD, Table 1 shows the nearest neighboring dependency path vectors to the bound product $I(metformin) \otimes O(diabetes)$, illustrating paths that indicate the relationship between these terms, and ESD's capability to learn similar representations for paths with similar meaning.

Both SGNS and ESD were trained over five epochs, with a subsampling threshold of 10^{-5}, a minimum term frequency threshold of 5 (which includes concatenated dependency paths for ESD), and a maximum frequency threshold of 10^6.

6 Evaluation Methods

We use a proportional analogy ranked retrieval task for both the RR and LBD tasks, following prior work as described in §2. Figure 3 visualizes this process. From a set of (X, Y) entity pairs from a knowledge base, given a term C and all terms D such that (C, D) is a pair in the set, we select n random (A, B) cue pairs from a disjoint set of pairs. We refer to (C, D) pairs as 'target pairs,' correct D completions as 'targets,' and (A, B) pairs as 'cues.' The vectors for the cue terms (A, B) and the term C are summed in the following fashion to produce the resulting vector v. Given an analogical pair *A:B::C:D*, where *A* and *C*, *B* and *D* are of the same semantic type, respectively, we develop cue vectors for the target *D* in each model as follows:

$$SGNS: \overrightarrow{v} = \overrightarrow{B} - \overrightarrow{A} + \overrightarrow{C}$$
$$ESD: \overrightarrow{v} = \overrightarrow{I(A)} \otimes \overrightarrow{O(B)} \otimes \overrightarrow{I(C)}$$

[5] https://github.com/semanticvectors/semanticvectors

[6] This capability has been used to to predict drug interactions, with performance exceeding that of models with orders of magnitude more parameters (Burkhardt et al., 2019).

SCORE	PATH
0.974	controlled_nmod_start_entity_end_entity_amod_controlled
0.935	add-on_nmod_start_entity_end_entity_amod_add-on
0.565	reduces_nsubj_start_entity_reduces_dobj_requirement_requirement_nmod_end_entity
0.537	associated_compound_start_entity_end_entity_nsubj_associated
0.516	start_entity_conj_efficacy_efficacy_acl_treating_treating_dobj_end_entity
0.438	treatment_amod_start_entity_treatment_nmod_end_entity

Table 1: Nearest neighboring dependency path embeddings to $I(metformin) \otimes O(diabetes)$ where I and O indicate input and output weight vectors respectively.

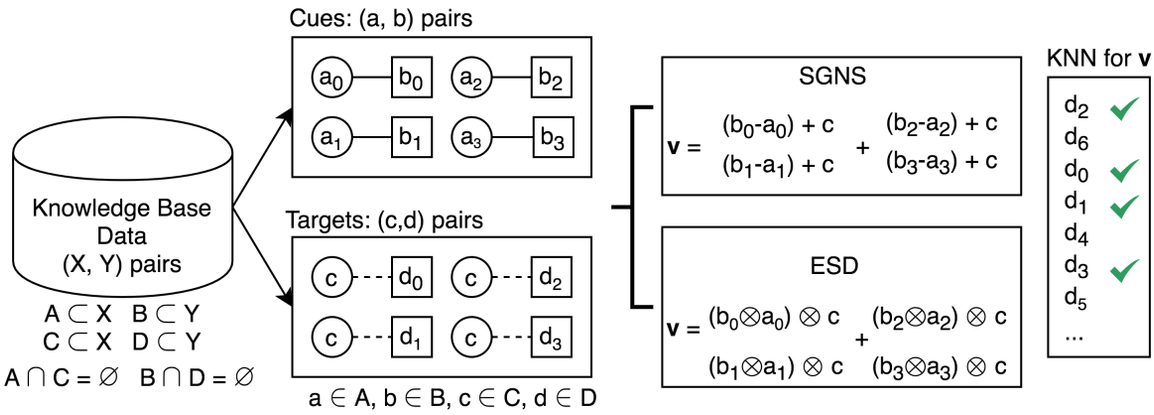


Figure 3: Overview of analogical ranked retrieval paradigm.

where I and O represent the input and output weight vectors of the ESD model, respectively. The SGNS method is the same as the 3COSADD method as described in Levy and Goldberg (2014b).

A K-nearest neighbor search is performed for **v** (using cosine distance for SGNS, NNHD for ESD) over the search space, and we record the ranks for each correct D target. The search space is constrained such that it consists of those terms from our training corpus that have a vector in both ESD and SGNS, a total of about 300,000 terms overall. For ESD, this space consists of the output weight vectors for each concept. For the proportional analogy task using K-nearest neighbors to rank completions to the analogy, the desired outcome is for the correct targets to be highly similar to the analogy cue vector v, such that the highest ranks are assigned to the correct target terms D in a search over the entire vector space. In this fashion, we perform this KNN search for every (X, Y) pair in the knowledge base and record the ranks for correct targets. We then compare the ranks of terms D across both vector spaces; the higher the ranks, the better the model is at capturing relational similarity.

Table 2 shows, for each knowledge base, how many total unique X terms and total (X, Y) pairs are used for each task. Additionally, we show the average number of correct Y terms per X and the maximum number of correct Y terms per X. For the relationship retrieval task, we consider those (X, Y) pairs which are connected by at least one dependency path in our corpus. Meanwhile, (X, Y) pairs for the LBD task must *not* be connected by a dependency path in the corpus (we treat these held-out pairs as a proxy for estimating the quality of novel hypotheses). We know from the (X, Y) pair's presence in the knowledge base that it is a gold standard pair for the given relationship type, but from the models' perspective this information is not available from the text alone. Thus, we believe it is a good test of the models' ability to generate plausible hypotheses. To reiterate, the methodology for both the relationship retrieval and literature-based discovery evaluations is the same; the only difference is in which pairs of terms from each knowledge base are used for evaluation data.

We examine the role of increasing the number of cues in improving retrieval. For example, for a given (C, D) target pair, we can combine vectors

		Relationship Retrieval				Literature-based Discovery			
		Total X	Total Pairs	Mean Y / X	Max Y / X	Total X	Total Pairs	Mean Y / X	Max Y / X
Chem-Gene	Gene Targets (DrugBank)	1626	6290	4	107	3569	37162	10	420
	PGKB	535	2089	4	48	1563	28053	18	144
	Agonists (TTD)	148	172	1	3	307	462	2	7
	Antagonists (TTD)	188	200	1	2	508	620	1	5
	Gene Targets (TTD)	1179	1436	1	7	4088	6430	2	15
	Inhibitors (TTD)	522	669	1	7	1273	2082	2	15
Chem-Disease	Side Effects (SIDER)	334	1289	4	31	892	6591	7	46
	Drug Indication (SIDER)	1077	2737	3	22	2160	8356	4	45
	Biomarker-Disease (TTD)	298	417	1	11	253	321	1	6
	Drug Indication (TTD)	1749	1958	1	6	2664	2999	1	10
	Disease Targets (TTD)	710	1502	2	22	1085	3088	3	27
Gene-Disease	OMIM	2197	2870	1	9	3461	5545	2	11
	PGKB	600	1693	3	34	1609	12605	8	73
Gene-Gene	Enzymes (DrugBank)	966	3622	4	33	1781	16242	9	71
	Carriers (DrugBank)	203	345	2	27	444	1174	3	18
	Transporters (DrugBank)	510	2357	5	44	1140	13889	12	94
	PGKB	497	2595	5	50	940	14142	15	89
	Complex (Reactome)	1757	3061	2	9	2550	6593	3	31
	Reaction (Reactome)	579	1031	2	9	1274	4024	3	29

Table 2: Total unique X terms, total (X, Y) pairs, average number of correct Y terms per X, and maximum number of correct Y terms per X for each knowledge base.

for multiple (A, B) pairs with the C term vector to produce a final cue vector that is closer to the target D. When multiple cues are used, we superpose the cue vector for each of the cues, and normalize the resulting vector, with normalization of real vectors to unit length in SGNS, and normalization of binary vectors using the majority rule with ties split at random with ESD. Cues are always selected from the subset of knowledge base pairs that co-occur in our training corpus. We ensure that none of the (A, B) cue terms overlap with each other, nor with the (C, D) target terms, to assure that self-similarity does not inflate performance. We produced results for a range of 1, 5, 10, 25, and 50 cues, finding that the best results come from using 25 cues; we only report these resulting scores in §7.

As a baseline inspired partly by Linzen (2016), we compute the similarity of vectors for *B* and *D* terms and *C* and *D* terms compared directly to each other, omitting the analogical task. The intuition here is that *C* and *D* terms are potentially close together in the vector space merely due to frequent co-occurrence in the corpus, and any analogical reasoning performance is merely relying on that fact. Meanwhile, terms *B* and *D* can be close together in the vector space simply because they are the same semantic type, and thus occur in similar contexts. In this case, relational analogy might not explain the performance, but mere distributional similarity. In the *B:D* comparison setting, cues *B* are added together to create a single cue vector with which to perform the KNN ranking over terms in which to find the target term *D*. These cue terms *B* are extracted from the same *A, B* cue pairs as those used for the full analogy setting to ensure a reasonable comparison across methods. In the *C:D* comparison setting, no cues are aggregated.

7 Results

We present qualitative and quantitative results for each vector space model's ability to represent and retrieve relational information.

Qualitative Results Table 3 shows a side-by-side comparison of the top 10 retrieved terms given the vector for the term *risperidone* composed with 25 randomly selected (drug, indication) cues from SIDER. The goal is to complete the proportional analogy corresponding to the treatment relationship. Of the top 10 terms retrieved in the ESD vector space, 4 are correct completions to the analogy, while 3 more are plausible completions based on literature. 'Tardive oromandibular dystonia,' while of the correct semantic type targeted by this analogy, is actually a side effect of risperidone. A majority of the retrieved results, however, are known or plausible treatment targets. Meanwhile, most of the top 10 terms retrieved by SGNS are names of other drugs that are similar to risperidone. Additionally, 'psychiatric and visual disturbances' and 'tardive dyskinesia' are side effects of risperidone, not treatment targets. Notably, all of the results retrieved with ESD are of the correct semantic type, i.e., they are disorders, while SGNS retrieves a mix of drugs and side effects.

Quantitative Results For each C term in each evaluation set, we record the ranks of all D tar-

rank	ESD (ours)	SGNS
1	separation anxiety	risperidone ×
2	**schizophrenia**	olanzapine ×
3	depressed state	quetiapine ×
4	**bipolar mania**	aripiprazole ×
5	tardive oromanibular dystonia	clozapine ×
6	treatment of trichotillomania ∗	psychiatric and visual disturbances
7	pervasive developmental disorder (NOS) ∗	ziprasidone ×
8	borderline personality disorder	amisulpride ×
9	**psychotic disorders**	paliperidone ×
10	**mania**	tardive dyskinesia

Table 3: Top 10 results for a K-nearest neighbor search over terms for treatment targets for the drug risperidone (an antipsychotic drug), using 25 (drug, indication) pairs from SIDER as cues. **Bolded** terms are correct targets, i.e., they are listed as treatment targets for risperidone in SIDER. ∗: a disorder that risperidone treats or might treat, based on external literature or a synonym for a target from SIDER; ×: a chemical, i.e., something that could not be a treatment target for a drug.

get terms resulting from the K-nearest neighbor search. For ease of comparison, we normalize all raw ranks by the length of the full search space (324363 terms in total), and then subtract this value from 1 so that lower ranks (i.e., better results) are displayed as higher numbers, for ease of interpretation. For a baseline score, we ran a simulation in which the entire search space was shuffled randomly 100 times, and recorded the median ranks of multiple target D terms, given some C. We find that the median rank for D terms in a randomly shuffled space tended toward the middle of the ranked list. Thus, the baseline score is established as 0.5; any score lower than this means the model performed worse than a random shuffle at retrieving target terms. In Table 4, 1 is the highest possible score, and 0 is the lowest.

We report results at 25 (A, B) cues, the setting for which performance was best for both ESD and SGNS. 'Full' in Table 4 refers to evaluation with a full *A:B::C:D* analogy, while 'B:D' refers to the baseline that compares vectors for terms directly, rather than using relational information. We do not report *C:D* comparison results, as they were categorically worse than both Full and *B:D* results.

8 Discussion

The results in Table 4 show that ESD outperforms SGNS on the *RR* task for 18 of 19 databases, and for 17 of 19 databases on the *LBD* task. It is clear that literature-based discovery is harder than relationship retrieval, as the scores are generally lower across the board for this task. We discuss the results for each task separately.

8.1 Relationship retrieval

For a total of 12 out of 19 sets, ESD on full analogies outperforms ESD on direct *B:D* comparisons, suggesting that the model has learned generalizable relationship information for these types of relations rather than relying on distributional term similarity. Because *gene-gene* pairs consist of entities of the same semantic type, it can be argued that *B:D* similarity should be very high, and yet scores are higher for the full analogy over the *B:D* baseline for most of these sets, for both ESD and SGNS. For SIDER side effects, the *B:D* baseline for ESD shows higher scores than the full analogy for both *LBD* and *RR*; one reason for this could be that there is a high degree of side effect overlap between drugs, and so the side effect terms themselves are highly similar to each other.

8.2 Literature-based discovery

The best performance on a majority of the sets comes from the ESD *B:D* model, suggesting that the model relies on term similarity over relational information for performance. Although SGNS doesn't perform the best overall, the full analogy model tends to outperform its B:D counterpart, suggesting that SGNS has managed to extrapolate relational information to the retrieval of held-out targets. As previously mentioned, performance on this task is made difficult due to the lack of normalization of concepts across our datasets. Additionally, as Table 4 shows, several top ranked terms are plausible analogy completions, but do not appear as

		Relationship retrieval				LBD			
		ESD (ours)		SGNS		ESD (ours)		SGNS	
		Full	B:D	Full	B:D	Full	B:D	Full	B:D
Chem-Gene	Gene Targets (DrugBank)	**0.912**	0.897	0.839	0.212	0.715	**0.806**	0.496	0.250
	PGKB	0.969	**0.994**	0.705	0.361	0.737	**0.918**	0.366	0.317
	Agonists (TTD)	0.997	0.907	**0.998**	0.647	0.802	0.781	**0.924**	0.708
	Antagonists (TTD)	**1.000**	0.900	0.999	0.732	0.802	0.703	**0.831**	0.750
	Gene Targets (TTD)	**0.998**	0.867	0.994	0.387	0.746	**0.760**	0.625	0.479
	Inhibitors (TTD)	**0.998**	0.874	0.993	0.415	**0.773**	0.759	0.682	0.392
Chem-Disease	Side Effects (SIDER)	0.997	**0.999**	0.967	0.942	0.952	**0.994**	0.799	0.932
	Drug Indication (SIDER)	**1.000**	0.995	0.949	0.588	0.969	**0.988**	0.663	0.605
	Biomarker-Disease (TTD)	0.996	**0.997**	0.944	0.781	0.932	**0.977**	0.799	0.726
	Drug Indication (TTD)	**1.000**	0.994	0.981	0.675	0.977	**0.992**	0.722	0.661
	Disease Targets (TTD)	0.990	**0.997**	0.900	0.711	0.887	**0.989**	0.663	0.648
Gene-Disease	OMIM	**0.997**	0.911	0.950	0.599	0.668	**0.792**	0.578	0.578
	PGKB	0.982	**0.996**	0.781	0.624	0.836	**0.969**	0.592	0.618
Gene-Gene	Enzymes (DrugBank)	**1.000**	**1.000**	0.987	0.981	0.979	**0.999**	0.900	0.975
	Carriers (DrugBank)	0.987	**1.000**	0.636	0.555	0.841	**0.962**	0.360	0.487
	Transporters (DrugBank)	**1.000**	**1.000**	0.974	0.947	0.996	**0.999**	0.870	0.951
	PGKB	**0.999**	0.995	0.899	0.471	0.907	**0.956**	0.479	0.425
	Complex (Reactome)	**1.000**	0.819	**1.000**	0.206	**0.866**	0.731	0.838	0.399
	Reaction (Reactome)	**1.000**	0.917	0.996	0.273	**0.878**	0.826	0.699	0.366

Table 4: Results for relationship retrieval (RR) and literature-based discovery (LBD) for full analogy (A:B::C:D) and B:D retrieval. Scores are displayed here as the median of scores (1 - normalized rank) for all D terms in a knowledge base evaluation set.

gold-standard targets in the databases. Considering the case of SIDER, which is built from automatically extracted information (not human-curated) the plausible results here are missing from the database but are supported by evidence from published papers (e.g. Oravecz and Štuhec (2014)).

9 Conclusion

We have compared two vector space models of language, Embedding of Structural Dependencies and Skip-gram with Negative Sampling, for their ability to represent biomedical relationships from literature in an analogical retrieval task. Our results suggest that encoding structural information in the form of dependency paths connecting biomedical entities of interest can improve performance on two analogical retrieval tasks, relationship retrieval and literature-based discovery. In future work, we would like to compare our methods with knowledge base completion techniques using contextualized vectors from language models as in Bosselut et al. (2019) as another method applicable to literature-based discovery.

Acknowledgements

This research was supported by U.S. National Library of Medicine Grant No. R01 LM011563. The authors would like to thank the anonymous reviewers for their feedback.

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Hannah A Burkhardt, Devika Subramanian, Justin Mower, and Trevor Cohen. 2019. Predicting adverse drug-drug interactions with neural embedding of semantic predications. In *AMIA Annual Symposium Proceedings*, volume 2019, page 992. American Medical Informatics Association.

Trevor Cohen and Dominic Widdows. 2009. Empirical distributional semantics: methods and biomedical applications. *Journal of biomedical informatics*, 42(2):390–405.

Trevor Cohen and Dominic Widdows. 2017. Embed-

ding of semantic predications. *Journal of biomedical informatics*, 68:150–166.

Trevor Cohen and Dominic Widdows. 2018. Bringing order to neural word embeddings with embeddings augmented by random permutations (earp). In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 465–475.

Trevor Cohen, Dominic Widdows, Roger Schvaneveldt, and Thomas C Rindflesch. 2011. Finding schizophrenia's prozac emergent relational similarity in predication space. In *International Symposium on Quantum Interaction*, pages 48–59. Springer.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8. Association for Computational Linguistics.

Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. 2014. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1819–1822.

Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, et al. 2016. The reactome pathway knowledgebase. *Nucleic acids research*, 44(D1):D481–D487.

Safa Fathiamini, Amber M Johnson, Jia Zeng, Vijaykumar Holla, Nora S Sanchez, Funda Meric-Bernstam, Elmer V Bernstam, and Trevor Cohen. 2019. Rapamycin- mtor+ braf=? using relational similarity to find therapeutically relevant drug-gene relationships in unstructured text. *Journal of biomedical informatics*, 90:103094.

Dedre Gentner and Arthur B Markman. 1997. Structure mapping in analogy and similarity. *American psychologist*, 52(1):45.

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 NAACL-HLT SRW*, pages 47–54, San Diego, California, June 12-17, 2016. ACL.

Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. 2005. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl_1):D514–D517.

Pentti Kanerva. 1996. Binary spatter-coding of ordered k-tuples. In *International Conference on Artificial Neural Networks*, pages 869–873. Springer.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.

Omer Levy and Yoav Goldberg. 2014b. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

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.

Zhiheng Li, Zhihao Yang, Chen Shen, Jun Xu, Yaoyun Zhang, and Hua Xu. 2019. Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text. *BMC medical informatics and decision making*, 19(1):22.

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, Berlin, Germany. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Denis Newman-Griffis, Albert Lai, and Eric Fosler-Lussier. 2017. Insights into analogy completion from the biomedical domain. In *BioNLP 2017*, pages 19–28, Vancouver, Canada,. Association for Computational Linguistics.

Robert Oravecz and Matej Štuhec. 2014. Trichotillomania successfully treated with risperidone and naltrexone: a geriatric case report. *Journal of the American Medical Directors Association*, 15(4):301–302.

Bethany Percha and Russ B Altman. 2018. A global network of biomedical relationships derived from text. *Bioinformatics*, 34(15):2614–2624.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Don R Swanson. 1986. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18.

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

Yunxia Wang, Song Zhang, Fengcheng Li, Ying Zhou, Ying Zhang, Zhengwen Wang, Runyuan Zhang, Jiang Zhu, Yuxiang Ren, Ying Tan, et al. 2020. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic acids research*, 48(D1):D1031–D1041.

Koki Washio and Tsuneaki Kato. 2018. Filling missing paths: Modeling co-occurrences of word pairs and dependency paths for recognizing lexical semantic relations. *arXiv preprint arXiv:1809.03411*.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522.

Michelle Whirl-Carrillo, Ellen M McDonagh, JM Hebert, Li Gong, K Sangkuhl, CF Thorn, Russ B Altman, and Teri E Klein. 2012. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417.

Dominic Widdows and Trevor Cohen. 2012. Real, complex, and binary semantic vectors. In *International Symposium on Quantum Interaction*, pages 24–35. Springer.

Dominic Widdows and Trevor Cohen. 2014. Reasoning with vectors: A continuous model for fast robust inference. *Logic Journal of the IGPL*, 23(2):141–173.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2018. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, Shaowu Zhang, Yuanyuan Sun, and Liang Yang. 2018. A hybrid model based on neural networks for biomedical relation extraction. *Journal of biomedical informatics*, 81:83–92.

Huiwei Zhou, Shixian Ning, Yunlong Yang, Zhuang Liu, Chengkun Lang, and Yingyu Lin. 2018. Chemical-induced disease relation extraction with dependency information and prior knowledge. *Journal of biomedical informatics*, 84:171–178.

Yongjun Zhu, Olivier Elemento, Jyotishman Pathak, and Fei Wang. 2019. Drug knowledge bases and their applications in biomedical informatics research. *Briefings in bioinformatics*, 20(4):1308–1321.