Insights into Analogy Completion from the Biomedical Domain

Denis Newman-Griffis[†,‡] and **Albert M. Lai**♣ and **Eric Fosler-Lussier**[†]
[†]The Ohio State University, Columbus, OH
[‡]National Institutes of Health, Clinical Center, Bethesda, MD
♣Washington University in St. Louis, St Louis, MO
`newman-griffis.1@osu.edu, amlai@wustl.edu,`
`fosler@cse.ohio-state.edu`

Abstract

Analogy completion has been a popular task in recent years for evaluating the semantic properties of word embeddings, but the standard methodology makes a number of assumptions about analogies that do not always hold, either in recent benchmark datasets or when expanding into other domains. Through an analysis of analogies in the biomedical domain, we identify three assumptions: that of a *Single Answer* for any given analogy, that the pairs involved describe the *Same Relationship*, and that each pair is *Informative* with respect to the other. We propose modifying the standard methodology to relax these assumptions by allowing for multiple correct answers, reporting MAP and MRR in addition to accuracy, and using multiple example pairs. We further present BMASS, a novel dataset for evaluating linguistic regularities in biomedical embeddings, and demonstrate that the relationships described in the dataset pose significant semantic challenges to current word embedding methods.

1 Introduction

Analogical reasoning has long been a staple of computational semantics research, as it allows for evaluating how well implicit semantic relations between pairs of terms are represented in a semantic model. In particular, the recent boom of research on learning vector space models (VSMs) for text (Turney and Pantel, 2010) has leveraged analogy completion as a standalone method for evaluating VSMs without using a full NLP system. This is due largely to the observations of "linguistic regularities" as linear offsets in context-based semantic models (Mikolov et al., 2013c; Levy and Goldberg, 2014; Pennington et al., 2014).

In the analogy completion task, a system is presented with an example term pair and a query, e.g., *London:England::Paris:_____*, and the task is to correctly fill in the blank. Recent methods consider the vector difference between related terms as representative of the relationship between them, and use this to find the closest vocabulary term for a target analogy, e.g., *England - London + Paris ≈ France*. However, recent analyses reveal weaknesses of such offset-based methods, including that the use of cosine similarity often reduces to just reflecting nearest neighbor structure (Linzen, 2016), and that there is significant variance in performance between different kinds of relations (Köper et al., 2015; Gladkova et al., 2016; Drozd et al., 2016).

We identify three key assumptions encoded in the standard offset-based methodology for analogy completion: that a given analogy has only one correct answer, that all relationships between the example pair and the query-target pair are the same, and that the example pair is sufficiently informative with respect to the query-target pair. We demonstrate that these assumptions are violated in real-world data, including in existing analogy datasets. We then propose several modifications to the standard methodology to relax these assumptions, including allowing for multiple correct answers, making use of multiple examples when available, and reporting mean average precision (MAP) and mean reciprocal rank (MRR) to give a more complete picture of the implicit ranking used in finding the best candidate for completing a given analogy.

Furthermore, we present the BioMedical Analogic Similarity Set (BMASS), a novel dataset for

19

analogical reasoning in the biomedical domain. This new resource presents real-world examples of semantic relations of interest for biomedical natural language processing research, and we hope it will support further research into biomedical VSMs (Chiu et al., 2016; Choi et al., 2016).[1]

2 Related work

Analogical reasoning has been studied both on its own and as a component of downstream tasks, using a range of systems. Early work used rule-based systems for world knowledge (Reitman, 1965) and syntactic (Federici and Pirelli, 1997) relationships. Supervised models were used for SAT (Scholastic Aptitude Test) analogies (Veale, 2004), and later for synonymy, antonymy, and some world knowledge (Turney, 2008; Herdağdelen and Baroni, 2009). Analogical reasoning has also been used in support of downstream tasks, including word sense disambiguation (Federici et al., 1997) and morphological analysis (Lepage and Goh, 2009; Lavallée and Langlais, 2010; Soricut and Och, 2015).

Recent work on analogies has largely focused on their use as an intrinsic evaluation of the properties of a VSM. The analogy dataset of Mikolov et al. (2013a), often referred to as the Google dataset, has become a standard evaluation for general-domain word embedding models (Pennington et al., 2014; Levy and Goldberg, 2014; Schnabel et al., 2015; Faruqui et al., 2015), and includes both world knowledge and morphosyntactic relations. Other datasets include the MSR analogies (Mikolov et al., 2013c), which describe morphological relations only; and BATS (Gladkova et al., 2016), which includes both morphological and semantic relations. The semantic relations from SemEval-2012 Task 2 (Jurgens et al., 2012) have also been used to derive analogies; however, as with the lexical Sem-Para dataset of Köper et al. (2015), the semantic relationships tend to be significantly more challenging for embedding-based methods (Drozd et al., 2016). Additionally, Levy et al. (2015b) demonstrate that even for some lexical relations where embeddings appear to perform well, they are actually learning prototypicality as opposed to relatedness.

---

3 Analogy completion task

3.1 Standard methodology

Given an analogy $a{:}b{::}c{:}d$, the evaluation task is to guess d out of the vocabulary, given a, b, c as evidence. Recent methods for this involve using the vector difference between embedded representations of the related pairs to rank all terms in the vocabulary by how well they complete the analogy, and choosing the best fit. The vector difference is most commonly used in one of three ways, where cos is cosine similarity:

$$argmax_{d \in V} \big(cos(d, b - a + c)\big) \qquad (1)$$

$$argmax_{d \in V} \big(cos(d - c, b - a)\big) \qquad (2)$$

$$argmax_{d \in V} \frac{cos(d, b)cos(d, c)}{cos(d, a) + \epsilon} \qquad (3)$$

Following the terminology of Levy and Goldberg (2014), we refer to Equation 1 as 3COSADD, Equation 2 as PAIRWISEDISTANCE, and Equation 3 (which is equivalent to 3COSADD with log cosine similarities) as 3COSMUL.

In order to generate analogy data for this task, recent datasets have followed a similar process (Mikolov et al., 2013a,c; Köper et al., 2015; Gladkova et al., 2016). First, relations of interest were manually selected for the target domains: syntactic/morphological, lexical (e.g., hypernymy, synonymy), or semantic (e.g., *CapitalOf*). Then, for each relation, example word pairs were manually selected or automatically generated from existing resources (e.g., WordNet). The final analogies were then generated by exhaustively combining the sets of word pairs within each relation.

3.2 Assumptions

Several key assumptions are inherent in this standard methodology that are not reflected in recent benchmark analogy datasets. The first we refer to as the *Single-Target* assumption: namely, that there is a single correct answer for any given analogy. Since the target d is chosen via argmax, if we consider the following two analogies:

flu:nausea::fever:?cough
flu:nausea::fever:?light-headedness

we must necessarily get at least one answer wrong. Gladkova et al. (2016) convert these analogies into a single case:

flu:nausea::fever:?[cough, light-headedness]

Pair	Relations
brother:sister	**FemaleCounterpart** *SiblingOf*
husband:wife	**FemaleCounterpart** *MarriedTo*

Table 1: Binary semantic relations in "brother is to sister as husband is to wife." The target common relation is shown in bold.

where either *cough* or *lightheadedness* is a correct guess. However, this still misses our desire to get both correct answers, if possible. Relations with multiple correct targets are present in all of Google, BATS, and Sem-Para.

The second key assumption is that all the information relating a to b also relates c to d. While the pairs are chosen based on a single common relationship, each pair may actually pertain to multiple relationships. An example from the Google dataset is *brother:sister::husband:wife*; Table 1 shows the semantic relations involved in this analogy. While the target relation *FemaleCounterpart* is present in both pairs, by comparing the offsets $sister - brother$ and $wife - husband$, we assume that either all ways in which each pair is related are present in both, or that *FemaleCounterpart* dominates the offset. We refer to this as the *Same-Relationship* assumption.

Finally, it is not sufficient for two pairs to share a common relationship label; that relationship must be both representative and informative for analogies to make sense (the *Informativity* assumption). Relation labels may be sufficiently broad as to be meaningless, as we encountered when drawing unfiltered binary relations from the Unified Medical Language System (UMLS) Metathesaurus. One sample analogy from the *RO:Null* relation (indicating "related in some way") was *socks:stockings::Finns:Finnish language*. While both pairs are of related terms, they are in no way related to one another.

Furthermore, even when two pairs are examples of the same kind of clearly-defined relation, they may still be relatively uninformative. For example, in the Sem-Para *Meronym* analogy *apricot:stone::trumpet:mouthpiece* the meronymic relationship between *apricot* and *stone* could plausibly identify a number of parts of a trumpet: *mouthpiece, valves, slide*, etc.[2] The extremely

---

[2]While this is similar to the Single-Target assumption, it

---

high-level nature of several of the Sem-Para relations (hypernymy, antonymy, and synonymy) suggests that some of the difficulty observed by Köper et al. (2015) is due to violations of Informativity.

4 BMASS

We present BMASS (the BioMedical Analogic Similarity Set), a dataset of biomedical analogies, generated using the expert-curated knowledge in the Unified Medical Language System (UMLS)[3] (Bodenreider, 2004) in order to identify medical term pairs sharing the same relationships. We followed the standard process for dataset generation outlined in Section 3.1, with some adjustments for the assumptions in Section 3.2.

The UMLS Metathesaurus is centered around normalized *concepts*, represented by Concept Unique Identifiers (CUIs). Each concept can be represented in textual form by one or more *terms* (e.g., C0009443 → "Common cold", "acute rhinitis"). These terms may be multi-word expressions (MWEs); in fact, many concepts in the UMLS have no unigram terms.

The Metathesaurus also contains ⟨*subject, relation, object*⟩ triples describing binary relationships between concepts. These relationships are specified at two levels: relationship types (RELs), such as *broader-than* and *qualified-by*, and specific relationships (RELAs) within each type, e.g., *tradename-of* and *has-finding-site*. For this work, we used the 721 unique REL/RELA pairings as our source relationships, and treated the ⟨*subject, object*⟩ pairs linked within each of these relationships as candidates for generating analogies.

To enable a word embedding–based evaluation, we first identified terms that appeared at least 25 times in the 2016 PubMed baseline collection of biomedical abstracts,[4] and removed all ⟨*subject, object*⟩ pairs involving concepts that did not correspond to these frequent terms. Most relationships in the Metathesaurus are many-to-many (i.e., each subject can be paired with multiple objects and

---

bears separate consideration in that Single-Target refers to multiple valid objects of a specific relationship, while this is an issue of multiple valid relationships being described.

[3]We use the 2016AA release of the UMLS.

[4]We chose 25 as our minimum frequency to ensure that each term appeared often enough to learn reasonable embeddings for its component words. To determine term frequency, we first lowercased and stripped punctuation from both the PubMed corpus and the term list extracted from UMLS, then searched the corpus for exact term matches.

vice versa), and thus may challenge Single-Target and Informativity assumptions; we therefore next identified relations that had at least 50 1:1 instances, i.e., a subject and object that are only paired with one another within a specific relationship. Since 1:1 instances are not sufficient to guarantee Informativity, we then manually reviewed the remaining relations to identify those those that we deemed to satisfy Informativity constraints. For example, the *is-a* relationship between *tongue muscles* and *head muscle* is not specific enough to suggest that *carbon monoxide* should elicit *gasotransmitters* as its corresponding answer. However, for *associated-with*, sampled pairs such as *leg injuries : leg* and *histamine release : histamine* were sufficiently consistent that we deemed it Informative. This gave us a final set of 25 binary relations, listed in Table 2.[5]

We follow Gladkova et al. (2016) in generating a balanced dataset, to enable a more robust comparative analysis between relations. We randomly sampled 50 ⟨*subject, object*⟩ pairs from each relation, again restricting to concepts with strings appearing frequently in PubMed. For each subject concept that we sampled, we collected all valid object concepts and bundled them as a single ⟨*subject, objects*⟩ pair. We then exhaustively combined each concept pair with the others in its relation to create 2,450 analogies, giving us a total dataset size of 61,250 analogies. Finally, for each concept, we chose a single frequent term to represent it, giving us both CUI and string representations of each analogy.

5 Evaluation

We assess how well biomedical word embeddings can perform on our dataset, and explore modifications to the standard evaluation methodology to relax the assumptions described in Section 3.2. We use the skip-gram embeddings trained by Chiu et al. (2016) on the PubMed citation database, one set using a window size of 2 (PM-2) and another set with window size 30 (PM-30). All other word2vec hyperparameters were tuned by Chiu et al. on a combination of similarity and relatedness and named entity recognition tasks.

Additionally, we use the hyperparameters they identified (minimum frequency=5, vector dimension=200, negative samples=10, sample=1e-4,

ID	Name	Amb
Lab/Rx		
L1	form-of	1.0
L2	has-lab-number	1.1
L3	has-tradename	1.5
L4	tradename-of	1.3
L5	associated-substance	1.6
L6	has-free-acid-or-base-form	1.0
L7	has-salt-form	1.1
L8	measured-component-of	1.3
Hierarchical		
H1	refers-to	1.0
H2	same-type	10.4
Morphological		
M1	adjectival-form-of	1.1
M2	noun-form-of	1.0
Clinical		
C1	associated-with-malfunction-of-gene-product	2.6
C2	gene-product-malfunction-associated-with-disease	1.5
C3	causative-agent-of	4.6
C4	has-causative-agent	2.0
C5	has-finding-site	1.9
C6	associated-with	1.2
Anatomy		
A1	anatomic-structure-is-part-of	1.6
A2	anatomic-structure-has-part	5.4
A3	is-located-in	1.4
Biology		
B1	regulated-by	1.0
B2	regulates	1.0
B3	gene-encodes-product	1.1
B4	gene-product-encoded-by	2.4

Table 2: List of the relations kept after manual filtering; Amb is the average ambiguity, i.e., the average number of correct answers per analogy.

α=0.05, window size=2) to train our own embeddings on a subset of the 2016 PubMed Baseline (14.7 million documents, 2.7 billion tokens). We train word2vec (Mikolov et al., 2013a) samples with the continuous bag-of-words (CBOW) and skip-gram (SGNS) models, trained for 10 iterations, and GloVe (Pennington et al., 2014) samples, trained for 50 iterations.

We performed our evaluation with each of 3COSADD, PAIRWISEDISTANCE, and 3COS-MUL as the scoring function over the vocabulary. In contrast to the prior findings of Levy and Gold-

---

[5]Examples of each relation, along with their mappings to UMLS REL/RELA values, are available online.

berg (2014) on the Google dataset, performance on BMASS is roughly equivalent among the three methods, often differing by only one or two correct answers. We therefore only report results with 3COSADD, since it is the most familiar method.

5.1 Modifications to the standard method

We consider 3COSADD under three settings of the analogies in our dataset. For a given analogy *a:b::c:?d*, we refer to $\langle a, b \rangle$ as the exemplar pair and $\langle c, d \rangle$ as the query pair; *?d* signifies the target answer.

Single-Answer puts analogies in *a:b::c:d* format, with a single example object b and a single correct object d, by taking the first object listed for each term pair. This enforces the Single-Answer assumption.

Multi-Answer takes the first object listed for the exemplar term pair, but keeps all valid answers, i.e. *a:b::c:[d_1,d_2,...]*; this is similar to the approach of Gladkova et al. (2016). There are approximately 16k analogies in our dataset with multiple valid answers.

All-Info keeps all valid objects for both the exemplar and query pairs. The exemplar offset is then calculated over $B = [b_1, b_2, \dots]$ as

$$a - B = \frac{1}{|B|} \sum_i a - b_i$$

Though this is superficially similar to 3COSAVG (Drozd et al., 2016), we average over objects for a specific subject, as opposed to averaging over all subject-object pairs.

We report a relaxed accuracy (denoted Acc_R), in which the guess is correct if it is in the set of correct answers. (In the Single-Answer case, this reduces to standard accuracy.) Acc_R, as with standard accuracy, necessitates ignoring a, b, or c if they are the top results (Linzen, 2016).

In order to capture information about all correct answers, we also report Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) over the set of correct answers in the vocabulary, as ranked by Equation 1. Since MAP and MRR do not have a cutoff in terms of searching for the correct answer in the ranked vocabulary, they can be used without the adjustment of ignoring a, b, and c; thus, they can give a more accurate picture of how close the correct terms are to the calculated guesses.

5.2 MWEs and candidate answers

As noted in Section 4, the terms in our analogy dataset may be multi-word expressions (MWEs). We follow the common baseline approach of representing an MWE as the average of its component words (Mikolov et al., 2013b; Chen et al., 2013; Wieting et al., 2016). For phrasal terms containing one or more words that are out of our embedding vocabulary, we only consider the in-vocabulary words: thus, if "parathyroid" is not in the vocabulary, then the embedding of *parathyroid hypertensive factor* will be

$$\frac{hypertensive + factor}{2}$$

For any individual analogy *a:b::c:?d*, the vocabulary of candidate phrases to complete the analogy is derived by calculating averaged word embeddings for each UMLS term appearing in PubMed abstracts at least 25 times. Terms for which none of the component words are in vocabulary are discarded. This yields a candidate set of 229,898 phrases for the PM-2 and PM-30, and 263,316 for our CBOW, SGNS, and GloVe samples.

Since prior work on analogies has primarily been concerned with unigram data, we also identified a subset of our data for which we could find single-word string realizations for all concepts in an analogy, using the full vocabulary of our trained embeddings. Even in the All-Info setting, we could only identify 606 such analogies; Table 3 shows MAP results for PM-2 and CBOW embeddings on the three relations with at least 100 unigram analogies. The unigram analogies are slightly better captured than the full MWE data for *has-lab-number* (L2) and *has-tradename* (L3); however, lower performance on the unigram subset in *tradename-of* (L4) shows that unigram analogies are not always easier. We see a small effect from the much larger set of candidate answers in the unigram case (>1m unigrams), as shown by the slightly higher MAP numbers in the Uni_M case. In general, it is clear that the difficulty of some of the relations in our dataset is not due solely to using MWEs in the analogies.

5.3 Metric comparison

Figure 1 shows Acc_R, MAP, and MRR results for each relation in BMASS, using PM-2 embeddings in the Multi-Answer setting. Overall, performance varies widely between relations, with all
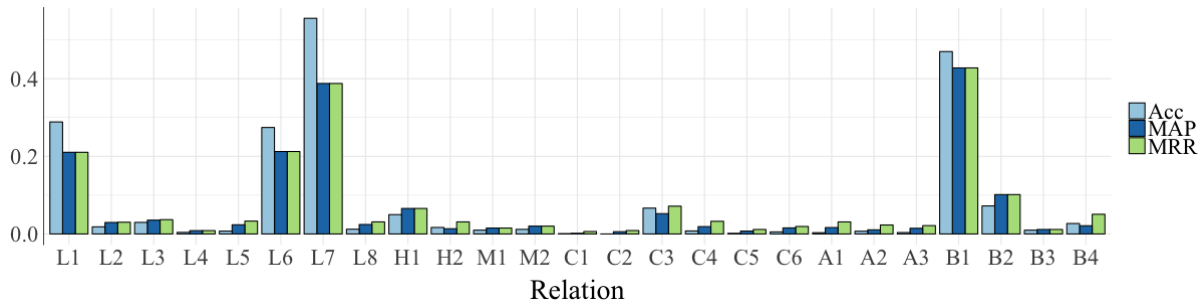
Figure 1: Acc_R, MAP, and MRR for each relation, using PM-2 embeddings under the Multi-Answer setting. Note that MAP is calculated using the position of all correct answers in the ranked list, while MRR reflects only the position of the first correct answer found in the ranked list for each individual query.
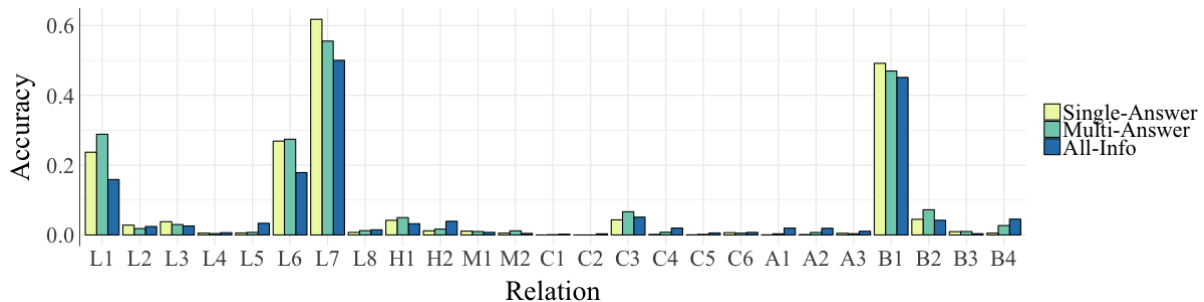

Figure 2: Acc_R per relation for PM-2 on BMASS, under Single-Answer, Multi-Answer, and All-Info settings.
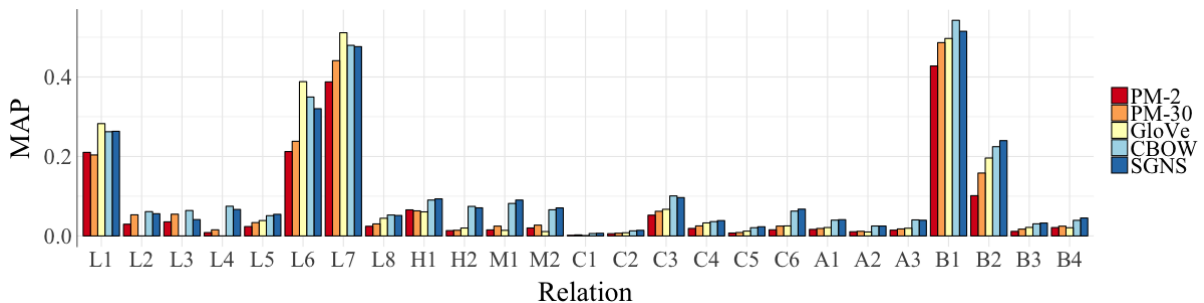

Figure 3: Per-relation MAP for all embeddings under the Multi-Answer setting.

three metrics staying under 0.1 in the majority of cases; this mirrors previous findings on other analogy datasets (Levy and Goldberg, 2014; Gladkova et al., 2016; Drozd et al., 2016).

MAP further fleshes out these differences by reporting performance over all correct answers for a given analogy. This lets us distinguish between relations like *has-salt-form* (L7), where noticeably lower MAP numbers reflect a wider distribution of the multiple correct answers, and relations like *regulates* (B2) or *associated-with* (C6), where a low Acc_R reflects many incorrect answers, but a higher MAP indicates that the correct answers are relatively near the guess.

MRR, on the other hand, more optimistically reports how close we got to finding any correct answer. Thus, for the *has-causative-agent* (C4) relation, low Acc_R is belied by a noticeably higher MRR, suggesting that even when we guess wrong, the correct answer is close. This contrasts with relations like *refers-to* (H1) or *causative-agent-of* (C3), where MRR is more consistent with Acc_R, indicating that wrong guesses tend to be farther from the truth. Since most of our analogies (45,178 samples, or about 74%) have only a single correct answer, MAP and MRR tend to be highly similar. However, in high-ambiguity relations like *same-type* (H2), higher MRR numbers

24

give a better sense of our best case performance.

5.4 Analogy settings

To compare across the Single-Answer, Multi-Answer, and All-Info settings, we first look at Acc_R for each relation in BMASS, shown for PM-2 embeddings in Figure 2 (the observed patterns are similar with the other embeddings). Unsurprisingly, allowing for multiple answers in Multi-Answer and All-Info slightly raises Acc_R in most cases. What is surprising, however, is that including more sample exemplar objects in the All-Info setting had widely varying results. In some cases, such as *same-type* (H2), *associated-substance* (L5), and *gene-product-encoded-by* (B4), the additional exemplars gave a noticeable improvement in accuracy. In others, accuracy actually went down: *form-of* (L1) and *has-free-acid-or-base-form* (L6) are the most striking examples, with absolute decreases of 4% and 8% respectively from the Multi-Answer case for PM-2 (the decreases are similar with other embeddings). Thus, it seems that multiple examples may help with Informativity in some cases, but confuse it in others. Taken together with the improvements seen in Drozd et al. (2016) from using 3CosAvg, this is another indication that any single subject-object pair may not be sufficiently representative of the target relationship.

5.5 Embedding methods

Averaging over all relations, the five embedding settings we tested behaved roughly the same, with our trained embeddings slightly outperforming the pretrained embeddings of Chiu et al. (2016); summary Acc_R, MAP, and MRR performances are given in Table 4. At the level of individual relations, Figure 3 shows MAP performance in the Multi-Answer setting. The four word2vec samples tend to behave similarly, with some inconsistent variations. Interestingly, CBOW outperforms the other embeddings by a large margin in several relations, including *regulated-by* (B1) and *has-tradename* (L3).

GloVe varies much more widely across the relations, as reflected in the higher standard deviations in Table 4. While GloVe consistently outperforms word2vec embeddings on *has-free-acid-or-base-form* (L6) and *has-salt-form* (L7), it significantly underperforms on the morphological and hierarchical relations, among others. Most notably, while the word2vec embeddings show minor differences in performance between the Multi-Answer and All-Info settings, GloVe Acc_R performance falls drastically on *form-of* (L1) and *has-free-acid-or-base-form* (L6), as shown in Table 5. However, its MAP and MRR numbers stay similar, suggesting that there is only a reshuffling of results closest to the guess.

5.6 Error analysis

Several interesting patterns emerge in reviewing individual $a{:}b{::}c{:}?d$ predictions. A number of errors follow directly from our word averaging approach to MWEs: words that appear in b or c often appear in the predictions, as in *gosorelin:<u>ici</u> 118630::letrozole:*<u>ici</u> 164384*. Prefix substitutions also occurred, as with *mianserin hydrochloride:mianserin::scopolamine <u>hydrobromide</u>:*scopolamine <u>methylbromide</u>*.

Often, the b term(s) would outweigh c, leading to many of the top guesses being variants on b. In one analogy, *sodium acetylsalicyclate:<u>aspirin</u>::intravenous immunoglobulins:?immunoglobulin g*, the top guesses were: **<u>aspirin</u> prophylaxis*, **<u>aspirin</u>*, **<u>aspirin</u> antiplatelet*, and **low-dose <u>aspirin</u>*.

In other cases, related to the nearest neighborhood over-reporting observed by Linzen (2016), we saw guesses very similar to c, regardless of a or b, as with *acute inflammations:acutely inflamed::<u>endoderm</u>:*embryonic endoderm*; other near guesses included **endoderm cell* and *epiblast*.

Finally, we found several analogies where the incorrect guesses made were highly related to the correct answer, despite not matching. One such analogy was *oropharyngeal suctioning:substances::thallium scan:?radioisotopes*; the top guess was **radioactive substances*, and

Rel	PM-2			CBOW		
	Uni	Uni$_M$	MWE	Uni	Uni$_M$	MWE
L2	0.05	0.04	0.03	0.11	0.13	0.06
L3	0.10	0.12	0.04	0.12	0.17	0.06
L4	0.00	0.00	0.01	0.04	0.05	0.07

Table 3: MAP performance on the three BMASS relations with ≥ 100 unigram analogies. Uni is using unigram embeddings on unigram data, Uni$_M$ is using MWE embeddings on unigram data, and MWE is performance with MWE embeddings over the full MWE data.

Setting	Single-Answer			Multi-Answer			All-Info		
	Acc_R	MAP	MRR	Acc_R	MAP	MRR	Acc_R	MAP	MRR
PM-2	.08 (.16)	.07 (.13)	.07 (.13)	.08 (.15)	.07 (.12)	.08 (.11)	.07 (.13)	.07 (.11)	.07 (.11)
PM-30	.08 (.16)	.08 (.13)	.08 (.13)	.09 (.16)	.08 (.13)	.09 (.13)	.08 (.13)	.08 (.12)	.09 (.12)
GloVe	.11 (.22)	.09 (.16)	.09 (.16)	.11 (.22)	.09 (.16)	.10 (.15)	.10 (.18)	.09 (.16)	.10 (.15)
CBOW	.11 (.18)	.11 (.14)	.11 (.14)	.12 (.18)	.12 (.14)	.12 (.14)	.11 (.17)	.12 (.14)	.13 (.14)
SGNS	.11 (.18)	.11 (.14)	.11 (.14)	.11 (.18)	.11 (.14)	.12 (.13)	.11 (.17)	.12 (.14)	.12 (.14)

Table 4: Average performance over all relations in the dataset, for each set of embeddings. Results are reported as "Mean (Standard deviation)" for each metric.

Metric	L1			L6		
	SA	MA	AI	SA	MA	AI
Acc_R	0.49	0.49	0.24	0.62	0.62	0.39
MAP	0.28	0.28	0.28	0.39	0.39	0.39
MRR	0.28	0.28	0.28	0.39	0.39	0.39

Table 5: Acc_R, MAP, and MRR performance variation between Single-Answer (SA), Multi-Answer (MA), and All-Info (AI) settings for GloVe embeddings on *form-of* (L1) and *has-free-acid-or-base-form* (L6)

gallium compounds was two guesses farther down. Showing some mixed effect from the neighborhood of *b*, *performance-enhancing substances* was the next-ranked candidate.

6   Discussion

Relaxing the Single-Answer, Same-Relationship, and Informativity assumptions by including multiple correct answers and multiple exemplar pairs and by reporting MAP and MRR in addition to accuracy paints a more complete picture of how well word embeddings are performing on analogy completion, but leaves a number of questions unanswered. While we can more clearly see the relations where we correctly complete analogies (or come close), and contrast with relations where a vector arithmetic approach completely misses the mark, what distinguishes these cases remains unclear. Some more straightforward relationships, such as *gene-encodes-product* (B3) and its inverse *gene-product-encoded-by* (B4), show surprisingly poor results, while the very broad synonymy of *refers-to* (H1) is captured comparatively well. Additionally, in contrast to prior work with morphological relations, *adjectival-form-of* (M1) and *noun-form-of* (M2) are much more challenging in the biomedical domain, as we see non-morphological related pairs such as *predisposed:disease susceptibility* and *venous lumen:endovenous*, in addition to more normal pairs like *sweating:sweaty* and *muscular:muscle*. Fur-

ther analysis may provide some insight into specific challenges posed by the relations in our dataset, as well as why performance with PAIRWISEDISTANCE and 3COSMUL did not noticeably differ from 3COSADD.

In terms of specific model errors, we did not evaluate the effects of any embedding hyperparameters on performance in BMASS, opting to use hyperparameter settings tuned for general-purpose use in the biomedical domain. Levy et al. (2015a) and Chiu et al. (2016), among others, show significant impact of embedding hyperparameters on downstream performance. Exploring different settings may be one way to get a better sense of exactly what incorrect answers are being highly-ranked, and why those are emerging from the affine organization of the embedding space. Additionally, the higher variance in per-relation performance we observed with GloVe embeddings suggests that there is more to unpack as to what the GloVe model is capturing or failing to capture compared to word2vec approaches.

Finally, while we considered Informativity during the generation of BMASS, and relaxed the Single-Answer assumption in our evaluation, we have not really addressed the Same-Relationship assumption. Using multiple exemplar pairs is one attempt to reduce the impact of confusing extraneous relationships, but in practice this helps some relations and harms others. Drozd et al. (2016) tackle this problem with the LRCos method; however, their findings of mis-applied features and errors due to very slight mis-rankings show that there is still room for improvement. One question is whether this problem can be addressed at all with non-parametric models like the vector offset approaches, to retain the advantages of evaluating directly from the word embedding space, or if a learned model (like LRCos) is necessary to separate out the different aspects of a related term pair.

7 Conclusions

We identified three key assumptions in the standard methodology for analogy-based evaluations of word embeddings: Single-Answer (that there is a single correct answer for an analogy), Same-Relationship (that the exemplar and query pairs are related in the same way), and Informativity (that the exemplar pair is informative with respect to the query pair). We showed that these assumptions do not hold in recent benchmark datasets or in biomedical data. Therefore, to relax these assumptions, we modified analogy evaluation to allow for multiple correct answers and multiple exemplar pairs, and reported Mean Average Precision and Mean Reciprocal Recall over the ranked vocabulary, in addition to accuracy of the highest-ranked choice.

We also presented the BioMedical Analogic Similarity Set (BMASS), a novel analogy completion dataset for the biomedical domain. In contrast to existing datasets, BMASS was automatically generated from a large-scale database of ⟨subject, relation, object⟩ triples in the UMLS Metathesaurus, and represents a number of challenging real-world relationships. Similar to prior results, we find wide variation in word embedding performance on this dataset, with accuracies above 50% on some relationships such as *has-salt-form* and *regulated-by*, and numbers below 5% on others, e.g., *anatomic-structure-is-part-of* and *measured-component-of*.

Finally, we are able to address the Single-Answer assumption by modifying the analogy evaluation to accommodate multiple correct answers, and we consider Informativity in generating our dataset and using multiple example pairs. However, the Same-Relationship assumption remains a challenge, as does a more automated approach to either evaluating or relaxing Informativity. These offer promising directions for future work in analogy-based evaluations.

Acknowledgments

References

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32(90001):D267–D270.

Danqi Chen, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2013. Learning New Facts From Knowledge Bases With Neural Tensor Networks and Semantic Word Vectors. *arXiv preprint arXiv:1301.3618* pages 1–4.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to Train Good Word Embeddings for Biomedical NLP. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing* pages 166–174.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer Representation Learning for Medical Concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco, California, USA, KDD '16, pages 1495–1504.

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*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 3519–3530.

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

Stefano Federici, Simonetta Montemagni, and Vito Pirelli. 1997. Inferring Semantic Similarity from Distributional Evidence: An Analogy-Based Approach to Word Sense Disambiguation. *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* pages 90–97.

Stefano Federici and Vito Pirelli. 1997. Analogy, computation, and linguistic theory. In *New Methods in Language Processing*, UCL Press, pages 16–34.

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. *Proceedings of the NAACL Student Research Workshop* pages 8–15.

Amaç Herdağdelen and Marco Baroni. 2009. Bag-Pack: A General Framework to Represent Semantic Relations. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*. Association for Computational Linguistics, Athens, Greece, pages 33–40.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 Task 2: Measuring Degrees of Relational Similarity. In **SEM 2012, Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Association for Computational Linguistics, Montréal, Canada, pages 356–364.

Maximilian Köper, Christian Scheible, and Sabine im Walde. 2015. Multilingual Reliability and "Semantic" Structure of Continuous Word Spaces. In *Proceedings of the 11th International Conference on Computational Semantics*. Association for Computational Linguistics, London, UK, pages 40–45.

Jean-François Lavallée and Philippe Langlais. 2010. *Unsupervised Morphological Analysis by Formal Analogy*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 617–624.

Yves Lepage and Chooi Ling Goh. 2009. Towards automatic acquisition of linguistic features. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*. Northern European Association for Language Technology (NEALT), Odense, Denmark, pages 118–125.

Omer Levy and Yoav Goldberg. 2014. Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 171–180.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015a. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics* 3:211–225.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015b. Do Supervised Distributional Methods Really Learn Lexical Inference Relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 970–976.

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. *arXiv preprint arXiv:1301.3781* pages 1–12.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., NIPS '13, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. 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*. Association for Computational Linguistics, Atlanta, Georgia, pages 746–751.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Doha, Qatar, pages 1532–1543.

Walter R Reitman. 1965. *Cognition and Thought: An Information Processing Approach*. John Wiley and Sons, New York, NY.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 298–307.

Radu Soricut and Franz Och. 2015. Unsupervised Morphology Induction Using Word Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1627–1637.

Peter D Turney. 2008. A Uniform Approach to Analogies, Synonyms, Antonyms, and Associations. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '08, pages 905–912.

Peter D Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37:141–188.

Tony Veale. 2004. WordNet Sits the S.A.T. A Knowledge-based Approach to Lexical Analogy. In *Proceedings of the 16th European Conference on Artificial Intelligence*. IOS Press, Amsterdam, The Netherlands, The Netherlands, ECAI'04, pages 606–610.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards Universal Paraphrastic Sentence Embeddings. In *Proceedings of the 4th International Conference on Learning Representations*.