What Analogies Reveal about Word Vectors and their Compositionality

Gregory P. Finley
EMR.AI*
San Francisco, CA
gregpfinley@gmail.com

Stephanie Farmer
Department of Linguistics
Macalester College
Saint Paul, MN
sfarmer@macalester.edu

Serguei V.S. Pakhomov
College of Pharmacy
University of Minnesota
Minneapolis, MN
pakh0002@umn.edu

Abstract

Analogy completion via vector arithmetic has become a common means of demonstrating the compositionality of word embeddings. Previous work have shown that this strategy works more reliably for certain types of analogical word relationships than for others, but these studies have not offered a convincing account for why this is the case. We arrive at such an account through an experiment that targets a wide variety of analogy questions and defines a baseline condition to more accurately measure the efficacy of our system. We find that the most reliably solvable analogy categories involve either 1) the application of a morpheme with clear syntactic effects, 2) male–female alternations, or 3) named entities. These broader types do not pattern cleanly along a syntactic–semantic divide. We suggest instead that their commonality is distributional, in that the difference between the distributions of two words in any given pair encompasses a relatively small number of word types. Our study offers a needed explanation for why analogy tests succeed and fail where they do and provides nuanced insight into the relationship between word distributions and the theoretical linguistic domains of syntax and semantics.

1 Introduction

In recent years, low-dimensional vectors have proven an efficient and fruitful means of representing words for numerous computational applications, from calculating semantic similarity to serving as an early layer in deep learning architectures (Baroni et al., 2014; Schnabel et al., 2015; LeCun et al., 2015). Despite these advances, however, strategies for representing meaning compositionally with a vector model remain limited. Given the difficulties in training representations of composed meaning (for example, most possible phrases will be rare or unattested in training data), achieving an accurate means of building complex lexical or phrasal representations from lower-order ones would be a decisive coup in computational semantics.

Another promising avenue of compositional semantics is the representation of concepts that do not map easily to lexemes. A simple averaging of two vectors may yield a concept that is semantically akin to both, and the arithmetic difference between word vectors has been said to represent the relationship between two terms. The ability to model knowledge unbounded by linguistic labels is an exciting prospect for natural language processing and artificial intelligence more broadly.

A common test of the compositional properties of word vectors is complete-the-analogy questions. Word vector arithmetic has achieved surprisingly high accuracy on this type of task. A flurry of recent studies have applied this test under various conditions, but there has been limited focus on defining precisely what types of relations vectors can capture, and less still on explaining these differences. As such, there remains a major gap in our understanding of distributional semantics. Our original experimental work improves upon prior methods by 1) targeting a wide variety of analogy questions drawn from several available resources and 2) defining a baseline condition to control for differences in "difficulty" between questions. These considerations enable an analysis that constitutes a major step towards a comprehensive, theoretically grounded account for the

* This work was done while the first author was a postdoctoral research associate at the University of Minnesota.

*Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 1–11,
Vancouver, Canada, August 3-4, 2017.

observed phenomena. To begin, however, we present a brief review of the analogy problem as usually posed.

2 Background

Several computational approaches have been proposed for representing the meaning of words (and holistic phrases) in terms of their co-occurrence with other words in large text corpora. Some of these, such as latent semantic analysis (Landauer and Dumais, 1997), focus on developing semantic representations based on theories of human cognition, whereas others, such as random indexing (Kanerva, 2009) and word embeddings (Bengio et al., 2003; Mikolov et al., 2013a) focus more on computational efficiency. Despite differences in purpose and implementation, all current distributional semantic approaches rely on the same basic principle of using similarity between co-occurrence frequency distributions as a way to infer the strength of association between words. For many practical purposes, such as information indexing and retrieval and semantic clustering, these approaches work remarkably well.

There is no obvious best way to compose these types of representations into larger arbitrary linguistic units, although it does seem that certain regularities exist between terms that surface through vector subtraction (Mikolov et al., 2013c; Levy et al., 2014). Why should this be the case? Consider the relationships between a difference vector $w_b - w_a$ and other words in the vocabulary: $w_b - w_a$ will be orthogonal to words that co-occur equally frequently with w_a and w_b, highly similar to words that co-occur only with w_b, and dissimilar (negative) to words that co-occur only with w_a.[1] If a word's context is a fair representation of its meaning, as is the key tenet of the distributional hypothesis, then this vector difference should isolate crucial differences in meaning.

Analogy tasks have been used to test how well vector differences capture consistent semantic differences. Four-word proportional analogies, typically written as w_1:w_2::w_3:w_4, feature two pairs of words such that the relationship between w_1 and w_2 is the same as between w_3 and w_4. If these words are represented with vectors, then, it is assumed that the differences between each pair are roughly equal:

$$w_2 - w_1 \approx w_4 - w_3 \tag{1}$$

In the most popular version of this task, a system is given the first three words in the analogy and asked guess the best candidate for w_4. Solving for w_4,

$$w_4 \approx w_3 + w_2 - w_1 \tag{2}$$

and thus a system selects its hypothesis w_{hyp} from the vocabulary V—typically excluding w_1, w_2 and w_3—by finding the word with maximum angular (cosine) similarity to the hypothesis vector (expressed as vector dot product, assuming all word vectors are unit length):

$$w_{hyp} = \arg\max_{w \in V} \left(w \cdot (w_3 + w_2 - w_1)\right) \tag{3}$$

We call this algorithm 3CosAdd following Levy *et al.* (2014). Levy and Goldberg (2014) note that this strategy is equivalent to finding the word in the lexicon that is the best match for w_3 and w_2 while also being most distant from w_1. This reframing suggests that it may not be necessary at all to represent ineffable concepts through intermediate stages of vector composition; 3CosAdd could be solving analogies simply through term similarity. Indeed, words in a pair sharing some relation tend to be similar to each other; when they are extremely similar, the difference between w_2 and w_1 is negligible, and the task becomes trivial.

Linzen (2016) makes this observation as well and goes on to demonstrate that accuracy falls to near zero across the board when not excluding w_1, w_2, and w_3 from contention in the hypothesis space, which shows how strongly dependent 3CosAdd is upon vector similarity. We agree wholeheartedly with that paper's claim that it is important to measure the consistency of vector differences in a way that is mindful of the typically high similarity between paired terms.

2.1 Analogy Test Sets

Several categorized sets of semantic and syntactic analogies are publicly available. One of the earliest was published by Microsoft Research (Mikolov et al., 2013c) and consists of 16 categories of inflectional morphological relations for English nouns, verbs, and adjectives. The most commonly reported test set, which we refer to as the Google set, is included with the distribution of

[1] These assertions are supported by the distributivity of a dot product, which is the standard calculation for similarity, over addition: $w_x \cdot (w_b - w_a) = w_x \cdot w_b - w_x \cdot w_a$.

the word2vec tool (Mikolov et al., 2013a). The Google set comprises 14 categories, mostly involving inflectional or geographical relationships between terms. Categories are grouped into a "semantic" and a "syntactic" subset, and results are often reported averaged over each rather than by category. This practice is rather problematic in our view, as the syntactic/semantic division is quite coarse and even questionable in some cases. We explore the relationship between syntax, semantics, and morphology in detail later on.

The "Better Analogy Test Set" (BATS) is a large set developed to contain a balanced sampling of a wide range of categories (Gladkova et al., 2016). BATS features 40 categories of 50 word pairs each, covering inflectional and derivational morphology as well as several semantic relations.

The relational similarity task in SemEval-2012 featured relations between word pairs targeting a massive range of lexical semantic relationships (Jurgens et al., 2012). By drawing on the aggregated results of the task's participants, we have extracted highly representative pairs for each relation to build an analogy set.

2.2 Accounting for Analogy Performance

In addition to those already cited, numerous other recent papers have evaluated word embeddings by benchmarking on analogy questions (Mikolov et al., 2013b; Garten et al., 2015; Lofi et al., 2016). There is some consensus regarding performance across question types: systems do well on questions of inflectional morphology (especially so for English (Nicolai et al., 2015)), but far less reliably so for various non-geographical semantic questions—although some gains in performance are possible by adjusting the embedding algorithms used or their hyperparameters (Levy et al., 2015), or by training further on subproblems (Drozd et al., 2016).

Amongst all of these findings, however, we found lacking a cohesive, thorough, and satisfying account of why vector arithmetic works where it does to solve analogies. To that end, we conducted an experiment to arrive at such an explanation, with some notable departures from previously used methods. We included a wide range of available test data, which is key because individual sets usually feature some bias towards one type or a few types of question, and benchmarkers often report nothing more than accuracy averaged over an entire set (Schnabel et al., 2015). Additionally, we define a baseline, which is critical not only to gauge effectiveness, but also to understand the mechanism behind solving analogies using compositional methods.

In the following sections we present the design of the experiment, baseline condition, and question sets; a discussion of how performance on analogy questions breaks down by broad category; and finally, a theoretical accounting for the observed patterns and the implications for distributional semantics.

3 Method

3.1 Word Embeddings

We used word embeddings trained on the plain text of all articles from Wikipedia as of September 2015, processed to remove all punctuation and case distinctions. We tested the word2vec and GloVe (Pennington et al., 2014) training algorithms. Results were qualitatively very similar between the two, although word2vec scored slightly higher on our metrics. Due to space considerations, we discuss only the word2vec results.

Hyperparameters were set as recommended for analogy tasks by the developers: 200-dimensional vectors, continuous bag-of-words sampling, 8-word window size. (We also tested a skip-gram model in word2vec and saw only slight and occasional differences—more subtle even than those seen between word2vec and GloVe.)

3.2 Test Set

We used a pooled set of analogy questions comprising the Google, Microsoft, SemEval 2012, and BATS test sets. At test time, any analogies that featured a word absent from our lexicon were discarded. (Note that the Microsoft categories testing the English possessive enclitic *'s* were not tested, as preprocessing for our vector training corpus removed all punctuation.) The sizes of each set following the removal of out-of-vocabulary analogies are given in Table 1.

Note that the BATS and SemEval data sets feature a number of word pairs in each category but not four-word analogy questions. We simply took every possible pair of pairs from the same category, so long as this did not result in an analogy in which w_1 and w_2 were the same word or in which w_4 was not unique. Some pairs in BATS have more than one correct answer; for uniformity

SOURCE	CATEGORIES	ANALOGIES
Microsoft Research	14	7,000
Google (word2vec)	14	19,544
SemEval2012	79	30,082
BATS	40	95,625
Total	147	152,251

Table 1: Summary of test data sources.

with other test sets, we use only the first answer provided for each of these pairs.

For SemEval, we used the "platinum standard" data distribution, which includes rankings of word pairs in each category based on how well they represent the relationship as defined. We took only the best half of pairs from that ranking to generate the test set. This was necessary because pairs lower down the list tend to poorly represent the relationship, or even to represent its opposite.

3.3 Measures

Virtually all existing studies of automated analogy solving report accuracy as the main measure. Accuracy is indeed a relevant measure when the goal is to simulate human performance on a particular task. Our purpose, however, is to understand the nature of semantic representations and account for when vector arithmetic does and does not function well as a model of relationships.

For every analogy question, we calculate the ranking of the correct w_4 in the hypothesis space—that is, the ordering of all words in the lexicon in descending order of the result of the 3COSADD hypothesis function (3). A "correct" answer would correspond to a ranking of 1.

Accuracy is a coarse measure in that it is insensitive to any ranking other than 1. Rather than accuracy, we borrow a measure from information retrieval (Voorhees, 1999)—the reciprocal of rank (RR) averaged across analogy questions in each category, which is always a positive fraction in the range:

$$\frac{1}{||V||} \leq RR \leq 1 \tag{4}$$

Numerically, RR acts as a "softer" version of accuracy, with rankings other than 1 contributing somewhat to the average.

Besides being coarse, accuracy is also an uncontrolled measure in that it is insensitive to differences in analogy "difficulty," by which we mean the prior degree of similarity between single word vectors. An example: nominal plural analogies, such as *dog*:*dogs*::*horse*:*horses*, often achieve high accuracy, but this may follow naturally from the high similarity between most singular nouns and their plural forms—indeed, for both of these pairs, the singular and plural forms are the closest terms to each other in our trained vector space.

To measure the efficacy of vector arithmetic in a manner controlled for variances in prior vector similarity, we propose a baseline, defined for each analogy as the best ranking between the word most similar to w_2 and the word most similar to w_3:

$$rank_{base} = \min(rank(\arg\max_{w \in V}(w \cdot w_2)), rank(\arg\max_{w \in V}(w \cdot w_3))) \tag{5}$$

For the above example, as *dog* is the most similar word to *dogs*, there is no improvement to be made upon baseline. Likewise, for the analogy *banana*:*yellow*::*sky*:*blue*, baseline would likely be high because *yellow* and *blue* are very similar.

Consistent with reporting RR for 3COSADD, we report baseline reciprocal rank (BRR). We suspect that using RR will be especially illustrative for baseline, where there may be many "near misses" that are informative but would all be reduced to zero if measuring only accuracy.

Our baseline is similar to the so-called ONLY-B baseline tested by Linzen (2016), except that the latter considers only w_3. We include w_2 because this term has just as much effect on the 3COSADD hypothesis as w_3. Note that our baseline would not itself be implementable as a solving strategy because it presumes access to w_4 to select between w_2 and w_3; nevertheless, we contend that it is helpful to define the baseline as we have done to account for those categories in the test data where all w_2 and w_4 are drawn from a small semantic cluster—most notably, the color example in the previous paragraph. (Overall, 16–18% of analogies across our test sets show similarity to w_2 as a better baseline than to w_3.)

Improvement is defined as the difference between 3COSADD RR and baseline RR, a measure we will refer to as reciprocal rank gain (RRG). RRG is more sensitive to shifts in rank that might not result in perfect accuracy. Analogies that show improvement from a very poor rank to first place will show a gain of nearly 1, whereas moving from second to first place is only 0.5 (and moving from

poor rank to second is nearly 0.5). If 3COSADD yields a worse hypothesis, this will be reflected as a negative RRG.

We also tested other solving methods suggested by Levy and Goldberg (Levy et al., 2014), 3COSMUL and PAIRDIRECTION, although we do not report them here—results with the former were virtually indistinguishable from 3COSADD, and poorer overall with the latter.

The raw results of our similarity experiments, as well as source code to replicate all steps of the experiments and analysis, can be downloaded at `https://github.com/gpfinley/analogies`.

4 Results

Most broadly, we confirm prior findings that vector arithmetic can be used to solve analogy questions, with a mean RRG of .165 across all questions in all categories ($t = 187$, $p \ll .01$). For a more nuanced analysis, we sorted analogy tests into four broad supercategories of analogical relationship: 30 categories of inflectional morphology, 12 of derivational morphology, 10 of named entities, and 95 of semantics of non-named entities (79 of which are from SemEval).

The gain in RR from baseline for all categories is presented visually in Figure 1, where they are grouped into our four supercategories for ease of interpretation. (See the appendix for the names of the top performing categories.) Each individual category is represented by a line between its BRR and 3COSADD RR. Within each supercategory, we also consider intermediate groupings of categories, and these are visualized by differences in line stroke in the figure. Note that some patterns are evident between and within supercategories:

- **Inflectional:** Although all inflectional categories show positive RRG, adjectival and verbal inflection shows reliably higher RRG than nominal inflection.
- **Derivational:** Derivational morphemes whose primary function is to shift syntactic class (*-tion*, *-ment*, *-ly*, *-ness*) show on average higher RRG than those with stronger regular semantic consequences (*-less*, *-able*, *over-*, adjectival *un-*, repetitive *re-*, agentive *-er*).
- **Named entities:** All categories—and particularly those dealing with country capitals—show high RRG.
- **Lexical:** Analogy relationships based on gender difference exhibit high RRG, while most other categories have low or even negative RRG.

We performed a linear regression analysis to predict RRG as a function of supercategory ($F = 24600$, $p \ll .01$, $R^2 = .39$). The model is summarized in Table 4. (Note that the model contains no intercept term, so the coefficient for each supercategory is equivalent to its mean RRG.) A positive RRG can be demonstrated with high statistical significance for all supercategories except lexical semantics.

We also investigated possible effects of word frequency on analogy performance. Multicollinearity poses a major challenge here: the frequencies of all four words in an analogy are highly correlated, and frequency can change dramatically across category. A comprehensive analysis of this complex problem is beyond the scope of this study, although we did find that the difference between an analogy's w_4 frequency and the mean w_4 frequency in that category correlates positively with RRG, although this effect is subtle ($r = .016$, $t = 6.28$, $p \ll .01$).

5 Discussion

It is clear from our results that vector arithmetic is a better approach for certain types of analogy questions than for others. Almost as clear is the hierarchy of the four broad types of questions that we have defined: excellent performance for inflection and named entities, with decidedly mixed results for derivational morphology and poorer still for lexical semantics—with the notable exception of male–female analogies. Below, we account for these patterns in the context of two domains of linguistic theory: the interaction between morphology and syntax, and the type-theoretic difference between individuals and sets.

5.1 Morphology and Syntax

Verbal and adjectival inflection show much more improvement over baseline than nominal inflection. It may simply be that the nominal categories have too high a baseline value to show much evidence of improvement by 3COSADD. It is also possible, however, that the nominal plural has fewer syntactic implications than verbal and adjectival morphology: nouns in non-subject position do not participate in number agreement in

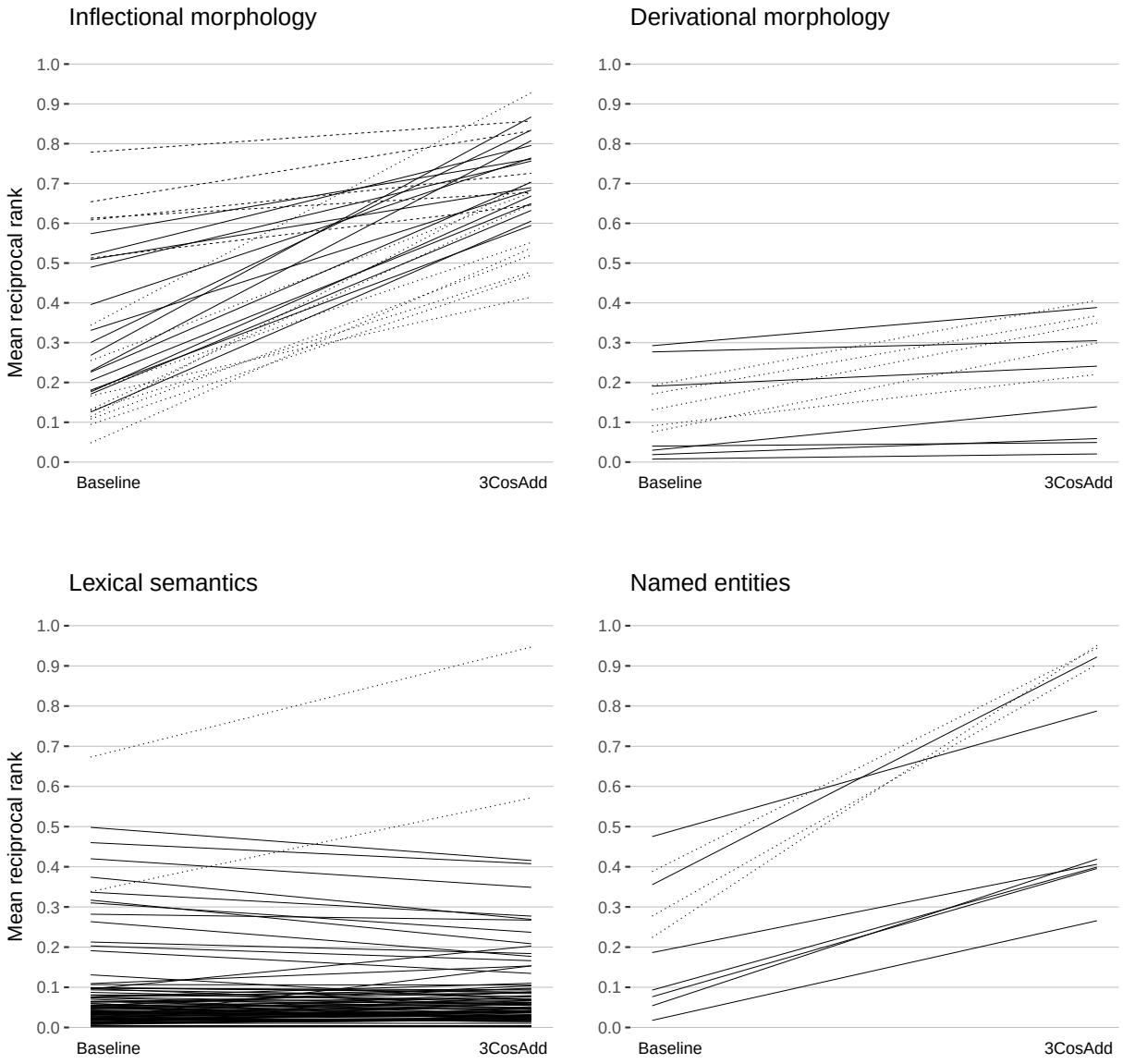

Figure 1: Mean reciprocal rank shifts between baseline and 3COSADD for four supercategories. Each line is a single category of analogy questions ("country - capital" or "male - female," for example). Some lines are differentiated by stroke type (dotted, solid, or dashed), the meaning of which is idiosyncratic to each supercategory: for inflectional, dashed lines are for nouns, dotted lines for adjectives, and solid lines for verbs; for derivational, dotted lines are for morphemes that change syntactic class with minimal semantic impact (e.g., *-ly*, as opposed to *re-*); for named entities, dotted lines are for country capitals; for lexical semantics, dotted lines are for gender relationships. Within supercategory, the difference in RRG between categories of different stroke types is significant in every case ($|t|$ between 14.5 and 58.7, $p \ll .01$).

English, so the plurality of many nouns in a text has little syntactic consequence.

Derivational morphology might be expected to perform worse than inflectional morphology for a number of reasons. Even for highly productive morphemes, derivation tends to have more idiosyncratic meaning (Haspelmath and Sims, 2010, 100). For example, although 'recruitment' refers to the act of recruiting, 'government' refers to a governing body rather than the act of governing; similarly, the adverb 'sadly' can be used as a sentential adverb (expressing the speaker's attitude

SUPERCATEGORY	ESTIMATE	STD ERROR	t
Inflectional	.345	.0015	228 ***
Derivational	.106	.0018	57.7 ***
Lexical semantics	$-.000$	.0012	-0.293
Named entities	.420	.0020	207 ***

Table 2: Summary of regression model for reciprocal rank gain as a function of analogy supercategory. All starred levels are highly significant ($p \ll .01$).

about the statement) as well as a manner adverb, whereas 'angrily' cannot. These semantic characteristics introduce lexically dependent variance that is far less pronounced for inflection.

From our results with derivational sets, there is evidence of a trend in which morphemes with predominantly syntactic consequences are better handled than those with stronger semantic consequences (see dotted/solid lines in Figure 1). Significant further experimental work is needed to quantify the syntactic versus semantic effects of derivational morphemes.

We predict that such work would support the notion of a continuum between morphemes with only syntactic effects and those with only (lexically) semantic effects. Those towards the syntactic end of the continuum will tend to be better captured by vector offsets in distributional representations. There would be a partial overlap between this continuum and the inflectional–derivational continuum in that derivational morphology tends to have more idiosyncratic meanings and is less relevant to syntax. There would be differences as well, especially as regards the property that word class-changing morphology is more derivational: the repetitive *re-* in English, for example, may be considered less derivational than the deverbal nominalizer *-ment* because it does not change word class, but *re-* has virtually no syntactic consequences for the verb to which it affixes.

5.2 Semantics: Named Entities as Individuals

Our results show that analogy sets containing named entities are more readily solvable than those that contain other lexical categories (common nouns, verbs, etc.).

A possible explanation for this is that named entities have a single real-world referent—there is, for instance, only one Amsterdam—while there is a large set of real-world referents that correspond to a common noun like 'dog'. We would expect the co-occurrences of 'dog', then, to be more diverse than those of a named entity like 'Amsterdam'.

The distinction drawn here between named entities and other parts of speech is analogous to the distinction between words of type e ("individuals") and words of type $\langle e, t\rangle$ in Montagovian set-theoretic semantics (Montague, 1973). According to Montague, proper names (arguments of type e) denote *individuals*, while verbs and common nouns (predicates of type $\langle e, t\rangle$) denote *sets* of individuals. Thus, 'Amsterdam' denotes an individual, while 'dog' denotes the set of dogs.

To better appreciate how this distinction might lead to "fuzzier" representations for some words, consider that training a vector on separate references to numerous members of a set of individuals is akin to a massive case of pseudo-polysemy—the vector can only capture the average of all referents rather than a single, clear referent. Polysemy is a well-known problem in training word vectors (Reisinger and Mooney, 2010), although this case of multiple referents has not been considered before to our knowledge.

Overall, named entity categories show very good RRG results, especially when both terms in a pair are named entities (as opposed to 'name - occupation', say). Country capitals show excellent performance in particular. In the broader history of this line of reserach, it is worth noting that the composition of the Google test set plays to this strength: country capital questions constitute over a quarter of its analogies (and over half of those in the "semantic" set, as noted by Gladkova *et al.* (2016)). As our experiments and others have demonstrated, however, the vector arithmetic approach struggles for most semantic questions.

Given the enormous influence of word2vec, it is worth asking whether prevailing knowledge in this field has been influenced by a selective focus on easier tasks. As further illustration of this point, note that the classic go-to example, *king*:*queen*::*man*:*woman*, is drawn from the sole

category in lexical semantics with any clear positive result in our experiments.

As a matter of fact, we should address the exceptional performance on analogies in male–female categories; why, of all lexical semantic sets, do we see such high performance here? We suspect these categories does well for the same reason that inflectional analogies do well: English features gender agreement with some personal pronouns—and, of course, with coreferential gendered terms—so there are concrete and regular distributional consequences of a noun's semantic gender.

5.3 A Unified Account

A recurrent thread in our accounting for all categories is that 3COSADD does well with relationships that have predictable effects on distribution—i.e., nearby terms and their morphology and syntax (although all morphology is effectively suppletive for these embeddings). This is especially evident with inflectional morphology, and true as well for certain types of derivational morphology as well as classes that participate in agreement, such as gender.

Relations between named entities are not governed by syntactic differences as inflectional relationships are, but there is a certain *distributional* parallel between the two: terms with a single referent will generally exhibit a less blurred co-occurrence profile than those with multiple referents; similarly, the difference between two realizations of the same root (e.g. 'hot' and 'hotter') will be highly non-orthogonal primarily with words of syntactic relevance, which is also a small set. The common theme is clear: the smaller the set of unique word types that co-occur with either word 1 or word 2 but not both (i.e.,the symmetric difference), the more cleanly the relationship between word 1 and word 2 can be captured.

Recall that our results also suggest that analogy questions containing frequent words are easier to solve with vector arithmetic than those containing less frequent words. We suspect that this is because the distributional representations of frequent words are more robust and less noisy. We believe, however, that more targeted investigation into the effects of frequency might qualify this generalization. For instance, it is reasonable to assume that a word's frequency correlates with the diversity of its co-occurrence, and that this diversity could signal distinct word senses, which are notoriously tricky for distributional representations. This is a ripe topic for further study.

5.4 Challenges

One challenge in interpreting our results is that categories with seemingly identical relations can show marked discrepancies in performance: note the differences between Google 'comparative' and Microsoft 'JJ_JJR', which examine the same inflectional relationship but show rather different levels of performance. Similarly, note the extreme difference in baseline rank for Google 'gender' (called 'family' in the original set) and BATS 'male - female' categories. Clearly, lexical choices make a significant difference and can even overshadow the inter-category differences that we are trying to measure. Note that in both of the above examples, the version of the category featuring more unique word types showed lower baseline *and* lower gain.

The explanations we put forward here may need to be extended to address other types of relationships that we did not evaluate. One particular interesting example might be Linzen *et al.*'s (2016) tests of analogies between quantifiers across domains—e.g., *everything*:*nothing*::*always*:*never*—which show intriguingly mixed results.

6 Conclusion

We evaluated syntactic and semantic analogy questions from a large and highly diverse test set using metrics more controlled and more sensitive than accuracy. Inspecting the results across categories, we were able to account for the differences in performance we observed across types of word relationships in terms that are consistent with the distributional training objectives of word embeddings.

Vector arithmetic with word embeddings is most effective when co-occurrence are limited to a small number of words, either by syntactic regularities or ease of semantic representation. It is possible to account for both of these by considering distributional phenomena directly.

Still, questions remain—do our negative results reflect the failure of word vectors to model semantic nuances, or the failure of vector arithmetic to capture them, or is the semantic data simply too noisy for current methods? Further experiments

with special attention paid to smoothing lexical semantic representations will be key to solving this problem.

Acknowledgments

This work was partially supported by a University of Minnesota Academic Health Center Faculty Development Award and by the National Institute of General Medical Sciences (GM102282).

References

Marco Baroni, Georgiana Dinu, and German Kruszewski. 2014. Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the ACL*. pages 238–247.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3(Feb):1137–1155.

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*. pages 3519–3530.

Justin Garten, Kenji Sagae, Volkan Ustun, and Morteza Dehghani. 2015. Combining distributed vector representations for words. In *Proceedings of NAACL-HLT*. pages 95–101.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesnt. In *Proceedings of NAACL-HLT*. pages 8–15.

Martin Haspelmath and Andrea Sims. 2010. *Understanding Morphology*. Routledge.

David A Jurgens, Peter D Turney, Saif M Mohammad, and Keith J Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*. pages 356–364.

Pentti Kanerva. 2009. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation* 1(2):139–159.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104(2):211.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521(7553):436–444.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., 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.

Omer Levy, Yoav Goldberg, and Israel Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of CoNLL*. pages 171–180.

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

Tal Linzen, Emmanuel Dupoux, and Benjamin Spector. 2016. Quantificational features in distributional word representations. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM)*. pages 1–11.

Christoph Lofi, Athiq Ahamed, Pratima Kulkarni, and Ravi Thakkar. 2016. Benchmarking semantic capabilities of analogy querying algorithms. In *Database Systems for Advanced Applications*. Springer, pages 463–478.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 3111–3119.

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

Richard Montague. 1973. The proper treatment of quantification in ordinary English. In K. J. J. Hintikka, M. E. Moravcsik, and P. Suppes, editors, *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, Dordrecht.

Garrett Nicolai, Colin Cherry, and Grzegorz Kondrak. 2015. Morpho-syntactic regularities in continuous word representations: A multilingual study. In *Proceedings of NAACL*. pages 129–134.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*. volume 14, pages 1532–1543.

Joseph Reisinger and Raymond J Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Proceedings of NAACL*. pages 109–117.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of EMNLP*.

Ellen M Voorhees. 1999. The trec-8 question answering track report. In *NIST Special Publication 500-246: The Eighth Text REtrieval Conference*. pages 77–82.

Appendix: Mean Rank by Category

CATEGORY	RR	CATEGORY	BRR	CATEGORY	RRG
G:capital	.950	G:plural	.711	**G:capital**	.750
G:capital-all	.945	noun - plural_reg	.674	**country - capital**	.659
G:gender	.933	***G:gender***	.618	verb_inf - 3pSg	.604
G:nationality-adj	.917	noun - plural_irreg	.603	G:superlative	.600
country - capital	.909	NN_NNS	.596	**G:capital-all**	.584
G:comparative	.896	G:pres-participle	.566	VBZ_VB	.580
verb_inf - 3pSg	.843	***X is opp. dir. from Y***	.535	G:comparative	.578
G:plural	.841	verb_inf - Ving	.496	VB_VBZ	.573
noun - plural_reg	.835	**G:city-in-state**	.486	**G:nationality-adj**	.548
VB_VBZ	.818	NNS_NN	.484	JJS_JJR	.536
verb_inf - Ving	.783	verb_Ving - Ved	.478	JJR_JJS	.496
VBZ_VB	.781	G:past-tense	.463	VBD_VB	.470
G:city-in-state	.774	***X, Y same category***	.462	VBD_VBZ	.469
G:pres-participle	.755	***antonyms - binary***	.436	VBZ_VBD	.465
G:plural-verbs	.752	G:plural-verbs	.371	verb_inf - Ved	.465
G:past-tense	.739	**G:nationality-adj**	.369	VB_VBD	.443
G:superlative	.713	**G:capital-all**	.361	JJ_JJR	.443
NN_NNS	.710	***things - color***	.340	JJ_JJS	.426
VBD_VB	.677	verb_Ving - 3pSg	.336	adj - comparative	.422
verb_Ving - Ved	.670	***can't X and Y at same time***	.320	verb_3pSg - Ved	.400
noun - plural_irreg	.662	G:comparative	.317	G:plural-verbs	.381
verb_Ving - 3pSg	.661	***male - female***	.317	adj - superlative	.373
JJ_JJR	.659	***antonyms - gradable***	.306	**name - occupation**	.340
JJS_JJR	.653	*G:opposite*	.292	verb_Ving - 3pSg	.325
NNS_NN	.626	***X, Y two kinds in category***	.283	JJR_JJ	.321
VBD_VBZ	.623	***X and Y are contrary***	.279	***G:gender***	.316
VB_VBD	.621	*un+adj_reg*	.268	**name - nationality**	.312
verb_inf - Ved	.604	**country - capital**	.250	**G:city-in-state**	.288
VBZ_VBD	.571	***X, Y similar type of thing***	.245	verb_inf - Ving	.287
adj - comparative	.570	VB_VBZ	.245	**country - language**	.278
male - female	.557	verb_inf - 3pSg	.239	G:past-tense	.276
verb_3pSg - Ved	.553	***X will become Y***	.239	**G:currency**	.246
JJR_JJS	.543	JJ_JJR	.217	***male - female***	.240
JJ_JJS	.520	*G:adj-to-adverb*	.208	*verb+tion_irreg*	.240
adj - superlative	.468	VBD_VB	.207	*verb+ment_irreg*	.231
JJR_JJ	.437	*re+verb_reg*	.207	JJS_JJ	.228
X is opp. dir. from Y	.421	VBZ_VB	.201	**UK_city - county**	.219
G:adj-to-adverb	.402	**G:capital**	.200	*G:adj-to-adverb*	.195
name - occupation	.389	***synonyms - exact***	.199	*adj+ly_reg*	.192
JJS_JJ	.376	VB_VBD	.178	verb_Ving - Ved	.192
⋮		⋮		⋮	

Table 3: The top 40 categories for reciprocal rank using 3COSADD (RR), baseline reciprocal rank (BRR), and reciprocal rank gain (RRG = RR − BRR) as calculated from embeddings trained on Wikipedia text using word2vec. Categories based on inflectional morphology are in plain text, derivational morphology in *italics*, named entity semantics in **bold**, and lexical in ***bold italic***. Sources for analogy questions can be identified from category names: those starting with 'G:' are from the Google set; in all capital letters, the Microsoft set; with reference to 'X' and 'Y', the SemEval set; all others, BATS. Some category names are abbreviated from their original names.