Analogy-based Detection of Morphological and Semantic Relations With Word Embeddings: What Works and What Doesn't.

Anna Gladkova
Department of Language and Information Sciences
The University of Tokyo
Tokyo, Japan
gladkova@phiz.c.u-tokyo.ac.jp

Aleksandr Drozd
Global Scientific Information and Computing Center
Tokyo Institute of Technology
Tokyo, Japan
alex@smg.is.titech.ac.jp

Satoshi Matsuoka
Global Scientific Information and Computing Center
Tokyo Institute of Technology
Tokyo, Japan
matsu@is.titech.ac.jp

Abstract

Following up on numerous reports of analogy-based identification of "linguistic regularities" in word embeddings, this study applies the widely used vector offset method to 4 types of linguistic relations: inflectional and derivational morphology, and lexicographic and encyclopedic semantics. We present a balanced test set with 99,200 questions in 40 categories, and we systematically examine how accuracy for different categories is affected by window size and dimensionality of the SVD-based word embeddings. We also show that GloVe and SVD yield similar patterns of results for different categories, offering further evidence for conceptual similarity between count-based and neural-net based models.

1 Introduction

The recent boom of research on analogies with word embedding models is largely due to the striking demonstration of "linguistic regularities" (Mikolov et al., 2013b). In the so-called Google analogy test set (Mikolov et al., 2013a) the task is to solve analogies with vector offsets (a frequently cited example is *king - man + woman = queen*). This test is a popular benchmark for word embeddings, some achieving 80% accuracy (Pennington et al., 2014).

Analogical reasoning is a promising line of research, since it can be used for morphological analysis (Lavallée and Langlais, 2010), word sense disambiguation (Federici et al., 1997), and even for broad-range detection of both morphological and semantic features (Lepage and Goh, 2009). However, it remains to be seen to what extent word embeddings capture the "linguistic regularities". The Google analogy test set includes only 15 relations, and Köper et al. (2015) showed that lexicographic relations such as synonymy are not reliably discovered in the same way.

This study systematically examines how well various kinds of linguistic relations can be detected with the vector offset method, and how this process is affected by window size and dimensionality of count-based word embeddings. We develop a new, more balanced test set (BATS) which includes 99,200 questions in 40 morphological and semantic categories. The results of this study are of practical use in real-world applications of analogical reasoning, and also provide a more accurate estimate of the degree to which word embeddings capture linguistic relations.

2 Related work

Current research on analogical reasoning in word embeddings focuses on the so-called "proportional analogies" of the *a:b::c:d* kind. The task is to detect whether two pairs of words have the same relation. A recent term is "linguistic regularity" (Mikolov et al., 2013b), used to refer to any "similarities between pairs of words" (Levy et al., 2014). Analogies have been successfully used for detecting different semantic relations, such as synonymy and antonymy (Turney, 2008), ConceptNet relations and selectional preferences (Herdadelen and Baroni, 2009), and also for inducing morphological categories from unparsed data (Soricut and Och, 2015).

The fact that analogies are so versatile means that to make any claims about a model being good at

analogical reasoning, we need to show what types of analogies it can handle. This can only be determined with a comprehensive test set. However, the current sets tend to only include a certain type of relations (semantic-only: SAT (Turney et al., 2003), SemEval2012-Task2 (Jurgens et al., 2012), morphology-only: MSR (Mikolov et al., 2013b)). The Google analogy test (Mikolov et al., 2013a) contains 9 morphological and 5 semantic categories, with 20-70 unique word pairs per category which are combined in all possible ways to yield 8,869 semantic and 10,675 syntactic questions.[1]

None of the existing tests is balanced across different types of relations (word-formation getting particularly little attention). With unbalanced sets, and potentially high variation in performance for different relations, it is important to evaluate results on all relations, and not only the average.

Unfortunately, this is not common practice. Despite the popularity of the Google test set, the only study we have found that provides data for individual categories is (Levy et al., 2014). In their experiments, accuracy varied between 10.53% and 99.41%, and much success in the semantic part was due to the fact that the two categories explore the same *capital:country* relation and together constitute 56.72% of all semantic questions. This shows that a model may be more successful with some relations but not others, and more comprehensive tests are needed to show what it can and cannot do.

Model parameters can also have a major impact on performance (Levy et al., 2015; Lai et al., 2015). So far they have been studied in the context of semantic priming (Lapesa and Evert, 2014), semantic similarity tasks (Kiela and Clark, 2014), and across groups of tasks (Bullinaria and Levy, 2012). However, these results are not necessarily transferable to different tasks; e.g. dependency-based word embeddings perform better on similarity task, but worse on analogies (Levy and Goldberg, 2014a). Some studies report effects of changing model parameters on general accuracy on Google analogy test (Levy et al., 2015; Lai et al., 2015), but, to our knowledge, this is the first study to address the effect of model parameters on individual linguistic relations in the context of analogical reasoning task.

[1]For semantic relations there are also generic resources such as EVALution (Santus et al., 2015), and semantic similarity sets such as BLESS and WordSim353 (Baroni and Lenci, 2011), which are sometimes used as sources for compiling analogy tests. For example, (Vylomova et al., 2015) presents a compilation with 18 relations in total (58 to 3163 word pairs per relation): 10 semantic, 4 morphological, 2 affix-derived word relations, animal collective nouns, and verb-object pairs.

3 The Bigger Analogy Test Set (BATS)

We introduce BATS - the Bigger Analogy Test Set. It covers 40 linguistic relations that are listed in table 1. Each relation is represented with 50 unique word pairs, which yields 2480 questions (99,200 in all set). BATS is balanced across 4 types of relations: inflectional and derivational morphology, and lexicographic and encyclopedic semantics.

A major feature of BATS that is not present in MSR and Google test sets is that morphological categories are sampled to reduce homonymy. For example, for verb present tense the Google set includes pairs like *walk:walks*, which could be both verbs and nouns. It is impossible to completely eliminate homonymy, as a big corpus will have some creative uses for almost any word, but we reduce it by excluding words attributed to more than one part-of-speech in WordNet (Miller and Fellbaum, 1998). After generating lists of such pairs, we select 50 pairs by top frequency in our corpus (section 4.2).

The semantic part of BATS does include homonyms, since semantic categories are overall smaller than morphological categories, and it is the more frequently used words that tend to have multiple functions. For example, both *dog* and *cat* are also listed in WordNet as verbs, and *aardvark* is not; an homonym-free list of animals would mostly contain low-frequency words, which in itself decreases performance. However, we did our best to avoid clearly ambiguous words; e.g. prophet Muhammad was not included in the E05 *name:occupations* section, because many people have the same name.

The lexicographic part of BATS is based on SemEval2012-Task2, extended by the authors with words similar to those included in SemEval set. About 15% of extra words came from BLESS and EVALution. The encyclopedic section was compiled on the basis of word lists in Wikipedia and other internet resources[2]. Categories E01 and E10

[2]E06-08: https://en.wikipedia.org/wiki/List_of_animal_names
E02: http://www.infoplease.com/ipa/A0855611.html

	Subcategory	Analogy structure and examples
Inflections	Nouns	I01: regular plurals *(student:students)*
		I02: plurals - orthographic changes *(wife:wives)*
	Adjectives	I03: comparative degree *(strong:stronger)*
		I04: superlative degree *(strong:strongest)*
	Verbs	I05: infinitive: 3Ps.Sg *(follow:follows)*
		I06: infinitive: participle *(follow:following)*
		I07: infinitive: past *(follow:followed)*
		I08: participle: 3Ps.Sg *(following:follows)*
		I09: participle: past *(following:followed)*
		I10: 3Ps.Sg : past *(follows:followed)*
Derivation	No stem change	D01: noun+less *(life:lifeless)*
		D02: un+adj. *(able:unable)*
		D03: adj.+ly *(usual:usually)*
		D04: over+adj./Ved *(used:overused)*
		D05: adj.+ness *(same:sameness)*
		D06: re+verb *(create:recreate)*
		D07: verb+able *(allow:allowable)*
	Stem change	D08: verb+er *(provide:provider)*
		D09: verb+ation *(continue:continuation)*
		D10: verb+ment *(argue:argument)*
Lexicography	Hypernyms	L01: animals *(cat:feline)*
		L02: miscellaneous *(plum:fruit, shirt:clothes)*
	Hyponyms	L03: miscellaneous *(bag:pouch, color:white)*
	Meronyms	L04: substance *(sea:water)*
		L05: member *(player:team)*
		L06: part-whole *(car:engine)*
	Synonyms	L07: intensity *(cry:scream)*
		L08: exact *(sofa:couch)*
	Antonyms	L09: gradable *(clean:dirty)*
		L10: binary *(up:down)*
Encyclopedia	Geography	E01: capitals *(Athens:Greece)*
		E02: country:language *(Bolivia:Spanish)*
		E03: UK city:county *York:Yorkshire*
	People	E04: nationalities *(Lincoln:American)*
		E05: occupation *(Lincoln:president)*
	Animals	E06: the young *(cat:kitten)*
		E07: sounds *(dog:bark)*
		E08: shelter *(fox:den)*
	Other	E09: thing:color *(blood:red)*
		E10: male:female *(actor:actress)*

Table 1: The Bigger Analogy Test Set: categories and examples

are based on the Google test, and category E09 - on the color dataset (Bruni et al., 2012). In most cases we did not rely on one source completely, as they did not make the necessary distinctions, included clearly ambiguous or low-frequency words, and/or were sometimes inconsistent[3] (e.g. *sheep:flock* in EVALution is a better example of *member:collection* relation than *jury:court*).

Another new feature in BATS, as compared to the Google test set and SemEval, is that it contains several acceptable answers (sourced from WordNet), where applicable. For example, both *mammal* and *canine* are hypernyms of *dog*.

E03: http://whitefiles.org/b4_g/5_towns_to_counties_index/

L02: https://www.vocabulary.com/lists/189583#view=notes

L07: http://justenglish.me/2012/10/17/character-feelings

[3]No claims are made about our own work being free from inconsistencies, as no dictionary will ever be so.

4 Testing the test

4.1 The vector offset method

As mentioned above, Mikolov et al. (2013a) suggested to capture the relations between words as the offset of their vector embeddings. The answer to the question "a is to b as c is to $?d$" is represented by hidden vector d, calculated as $argmax_{d \in V}(sim(d, c - a + b))$. Here V is the vocabulary excluding words a, b and c and sim is a similarity measure, for which Mikolov and many other researchers use angular distance: $sim(u, v) = cos(u, v) = \frac{u \cdot v}{||u||||v||}$.

Levy and Goldberg (2014b) propose an alternative optimization objective: $argmax_{d \in V}(cos(d - c, b - a))$ They report that this method produces more accurate results for some categories. Essentially it accounts for $d - c$ and $b - a$ to share the same direction and discards lengths of these vectors.

We supply the BATS test set with a Python evaluation script that implements both methods.[4] We report results calculated by the Mikolov's method for the sake of consistency, but some authors choose the best result for each category from each method (Levy and Goldberg, 2014b).

4.2 Corpus and models

One of the current topics in research on word embeddings is the (de)merits of count-based models as compared to the neural-net-based models. While some researchers find that the latter outperform the former (Baroni et al., 2014), others show that these approaches are mathematically similar (Levy and Goldberg, 2014b). We compare models of both types as a contribution to the ongoing dispute.

Our count-based model is built with Pointwise Mutual Information (PMI)frequency weighting. In the dimensionality reduction step we used the Singular Value Decomposition (SVD), raising Σ matrix element-wise to the power of a where $0 < a \leq 1$ to give a boost to dimensions with smaller variance Caron (2001). In this study, unless mentined otherwise, $a = 1$. The co-occurrence extraction was performed with the kernel developed by Drozd et al. (2015).

[4]http://vsm.blackbird.pw/bats

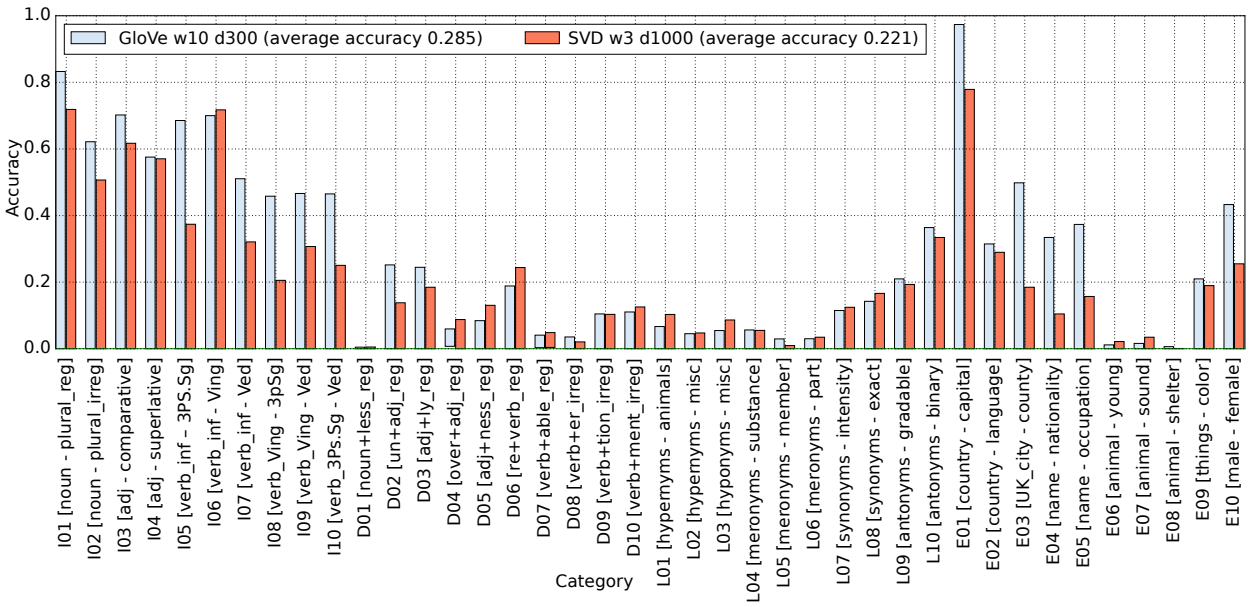

Figure 1: GloVe and SVD: accuracy on different types of relations

As a representative of implicit models we chose GloVe (Pennington et al., 2014) that achieved the highest performance on the Google test set to this date. Our source corpus combines the English Wikipedia snapshot from July 2015 (1.8B tokens), Araneum Anglicum Maius (1.2B) (Benko, 2014) and ukWaC (2B) (Baroni et al., 2009). We discarded words occurring less than 100 times, resulting in vocabulary of 301,949 words (uncased).

To check the validity of our models we evaluate it with the Google test set for which there are numerous reported results. For GloVe we used the parameters from the original study (Pennington et al., 2014): 300 dimensions, window 10, 100 iterations, x_{max}= 100, a = 3/4, sentence borders ignored. For comparison we also built an SVD model with 300 dimensions and window size 10. On our 5 B corpus GloVe achieved 80.4% average accuracy (versus 71.7% on 6 B corpus in the original study). The comparable SVD model achieved 49.9%, as opposed to with 52.6% result reported by Levy et al. (2015) for 500 dimensions, window size 10 on 1.5 B Wikipedia corpus.

To evaluate effects of window size and dimensionality we built 19 SVD-based models for windows 2-8 at 1000 dimensions, and for dimensions 100-1200 for window size 5.

5 Results and discussion

5.1 Word category effect

Figure 1 presents the results of BATS test on the GloVe model (built with the parameters from the original study (Pennington et al., 2014)), and the best performing SVD model, which was the model with window size 3 at 1000 dimensions. The model built with the same parameters as GloVe achieved only 15.9% accuracy on BATS, and is not shown.

While GloVe outperforms the SVD-based model on most categories, neither of them achieves even 30% accuracy, suggesting that BATS is much more difficult than the Google test set. Many categories are either not captured well by the embedding, or cannot be reliably retrieved with vector offset, or both. The overall pattern of easier and more difficult categories is the same for GloVe and SVD, which supports the conclusion of Levy and Goldberg (2014b) about conceptual similarity of explicit and implicit models. The overall performance of both models could perhaps be improved by parameters that we did not consider, but the point is that the current state-of-the-art in analogical reasoning with word embeddings handles well only certain types of linguistic relations, and there are directions for improvement that have not been considered so far.

The high variation we observe in this experiment

is consistent with evidence from systems competing at SemEval2012-Task2, where not a single system was able to achieve superior performance on all subcategories. Fried and Duh (2015) also showed a similar pattern in 7 different word embeddings.

As expected, inflectional morphology is overall easier than semantics, as shown even by the Google test results (see Skip-Gram (Mikolov et al., 2013a; Lai et al., 2015), GloVe (Pennington et al., 2014), and K-Net (Cui et al., 2014), among others). But it is surprising that derivational morphology is significantly more difficult to detect than inflectional: only 3 categories out of ten yield even 20% accuracy.

The low accuracy on the lexicographic part of BATS is consistent with the findings of Köper et al. (2015). It is not clear why lexicographic relations are so difficult to detect with the vector offset method, despite numerous successful word similarity tests on much the same relations, and the fact that BATS make the task easier by accepting several correct answers. The easiest category is binary antonyms of the *up:down* kind - the category for which the choice should be the most obvious in the semantic space.

A typical mistake that our SVD models make in semantic questions is suggesting a morphological form of one of the source words in the *a:b::c:d* analogy: *cherry:red* :: *potato:?potatoes* instead of *potato:brown*. It would thus be beneficial to exclude from the set of possible answers not only the words a, b and c, but also their morphological forms.

5.2 Window size effect

Evaluating two count-based models on semantic and syntactic parts of the Google test set, Lebret and Collobert (2015) shows that the former benefit from larger windows while the latter do not. Our experiments with SVD models using different window sizes only partly concur with this finding.

Table 2 presents the accuracy for all categories of BATS using a 1000-dimension SVD model with window size varying between 2 and 8. The codes and examples for each category are listed in table 1. All categories are best detected between window sizes 2-4, although 9 of them yield equally good performance in larger windows. This indicates that there is not a one-on-one correspondence between "semantics" and "larger windows" or "morphology" and "smaller windows". Also, different categories benefit from changing window size in different ways: for noun plurals the difference between the best and the worse choice is 13%, but for categories where accuracy is lower overall there is not much gain from altering the window size.

	2	3	4	5	6	7	8		2	3	4	5	6	7	8
I01	62	**71**	70	68	67	65	58	L01	**11**	10	9	8	7	6	6
I02	41	**50**	47	44	42	40	34	L02	**5**	4	4	4	4	**5**	4
I03	57	**61**	58	52	47	41	32	L03	**10**	8	8	8	7	6	4
I04	49	**57**	51	45	40	35	25	L04	5	**5**	**5**	**5**	**5**	**5**	4
I05	27	37	**39**	36	34	32	29	L05	**2**	0	1	1	1	1	1
I06	62	**71**	67	63	60	58	53	L06	3	3	**4**	3	3	3	3
I07	26	**32**	**36**	**36**	**36**	**36**	34	L07	**13**	12	9	7	6	5	4
I08	**21**	20	19	18	18	18	16	L08	**19**	16	13	12	10	9	6
I09	23	30	34	35	**36**	**36**	35	L09	15	**19**	17	14	12	11	9
I10	**25**	**25**	23	21	19	19	17	L10	32	**33**	30	28	27	25	24
D01	0	0	0	0	0	0	0	E01	69	77	**79**	77	74	71	69
D02	12	**13**	12	12	11	10	9	E02	**29**	**28**	24	22	21	20	17
D03	10	18	**20**	**20**	**20**	**20**	19	E03	11	**18**	**18**	**18**	**18**	**18**	17
D04	**12**	8	6	5	4	3	2	E04	**19**	10	3	3	3	3	4
D05	7	**13**	**13**	11	9	8	5	E05	**20**	15	15	14	14	13	13
D06	15	**24**	18	13	10	8	5	E06	**2**	**2**	1	1	1	1	1
D07	**4**	**4**	3	2	2	1	1	E07	2	**3**	**3**	2	2	1	1
D08	1	2	**2**	2	1	1	1	E08	0	0	0	0	0	0	0
D09	6	10	**11**	**11**	**11**	**11**	10	E09	**19**	18	**19**	18	18	**19**	18
D10	3	**12**	**12**	10	10	9	9	E10	20	**25**	**25**	**25**	24	23	21

Table 2: Accuracy of SVD-based model on 40 BATS categories, window sizes 2-8, 1000 dimensions

Our results are overall consistent with the evaluation of an SVD-based model on the Google set by Levy et al. (2015). This study reports 59.1% average accuracy for window size 2 yields, 56.9% for window size 5, and 56.2% for window size 10. However, using window sizes 3-4 clearly merits further investigation. Another question is whether changing window size has different effect on different models, as the data of Levy et al. (2015) suggest that GloVe actually benefits from larger windows.

5.3 Vector dimensionality effect

Intuitively, larger vectors capture more information about individual words, and therefore should increase accuracy of detecting linguistic patterns. In our data this was true of 19 BATS categories (I01-02, I04, I06, D02-03, D05-07, E01, E03, E07, E10, L03-04, L07-10): all of them either peaked at 1200 dimensions or did not start decreasing by that point.

However, the other 20 relations show all kinds of patterns. 14 categories peaked between 200 and 1100 dimensions, and then performance started decreasing (I03, I05, I07-10, D01, D04, D09, E02, E05, E09, L1, L6). 2 categories showed negative effect of higher dimensionality (D08, E04). Finally, 2 categories showed no dimensionality effect (E08,

L05), and 3 more - idiosyncratic patterns with several peaks (D10, E02, L06); however, this could be chance variation, as in these categories performance was generally low (under 10%). Figure 2 shows several examples of these different trends[5].

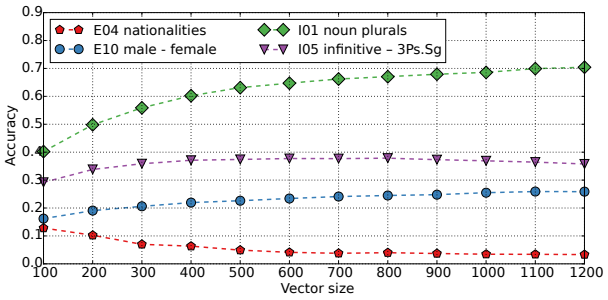

Figure 2: Effect of vector dimensionality: example categories

The main takeaway from this experiment is that, although 47.5% of BATS categories do perform better at higher dimensions (at least for SVD-based models), 40% do not, and, like with window size, there is no correlation between type of the relation (semantic or morphological) and its preference for a higher or low dimensionality. One possible explanation for lower saturation points of some relations is that, once the dimensions corresponding to the core aspects of a particular relation are included in the vectors, adding more dimensions increases noise. For practical purposes this means that choosing model parameters would have to be done to target specific relations rather than relation types.

5.4 Other parameters

In scope of this study we did not investigate all possible parameters, but our pilot experiments show that changing the power a for the Σ matrix of the SVD transformation can boost or decrease the performance on individual categories by 40-50%. Smaller value of a gives more weight to the dimensions which capture less variance in the original data, which can correspond to subtle linguistic nuances. However, as with windows and dimensions, no setting yields the best result for all categories.

A big factor is word frequency, and it deserves more attention than we can provide in scope of this paper. Some categories could perform worse because they contain only low-frequency vocabulary; in our corpus, this could be the case for D01 and D04-06[6]. But other derivational categories still do not yield higher accuracy even if the frequency distribution is comparable with that of an "easier" category (e.g. D8 and E10). Also, SVD was shown to handle low frequencies well (Wartena, 2014).

6 Conclusion

This study follows up on numerous reports of successful detection of linguistic relations with vector offset method in word embeddings. We develop BATS - a balanced analogy test set with 40 morphological and semantic relations (99,200 questions in total). Our experiments show that derivational and lexicographic relations remain a major challenge. Our best-performing SVD-based model and GloVe achieved only 22.1% and 28.5% average accuracy, respectively. The overall pattern of "easy" and "difficult" categories is the same for the two models, offering further evidence in favor of conceptual similarity between explicit and implicit word embeddings. We hope that this study would draw attention of the NLP community to word embeddings and analogical reasoning algorithms in context of lexicographic and derivational relations[7].

Our evaluation of the effect of vector dimensionality on accuracy of analogy detection with SVD-based models shows that roughly half BATS categories are best discovered with over 1000 dimensions, but 40% peak between 200 and 1100. There does not seem to be a correlation between type of linguistic relation and preference for higher or low dimensionality. Likewise, our data does not confirm the intuition about larger windows being more beneficial for semantic relations, and smaller windows - for morphological, as our SVD model performed best on both relation types in windows 2-4. Further research is needed to establish whether other models behave in the same way.

[5]All data for all categories can be found at `http://vsm.blackbird.pw/bats`

[6]Data on frequency distribution of words in BATS categories in our corpus can be found at `http://vsm.blackbird.pw/bats`

[7]BATS was designed for word-level models and does not focus on word phrases, but we included WordNet phrases as possible correct answers, which may be useful for phrase-aware models. Also, morphological categories involving orthographic changes may be of interest for character-based models.

References

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pages 1–10. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Marco Baroni, Georgiana Dinu, and Germn Kruszewski. 2014. Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247.

Vladimír Benko. 2014. Aranea: Yet another family of (comparable) web corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, speech, and dialogue: 17th international conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings*, LNCS 8655, pages 257–264. Springer.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

John A. Bullinaria and Joseph P. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, 44(3):890–907.

John Caron. 2001. Computational information retrieval. chapter Experiments with LSA Scoring: Optimal Rank and Basis, pages 157–169. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Qing Cui, Bin Gao, Jiang Bian, Siyu Qiu, and Tie-Yan Liu. 2014. Learning effective word embedding using morphological word similarity.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2015. Python, performance, and natural language processing. In *Proceedings of the 5th Workshop on Python for High-Performance and Scientific Computing*, PyHPC '15, pages 1:1–1:10, New York, NY, USA. ACM.

Stefano Federici, Simonetta Montemagni, and Vito Pirrelli. 1997. Inferring semantic similarity from distributional evidence: an analogy-based approach to word sense disambiguation. In *Proceedings of the ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 90–97.

Daniel Fried and Kevin Duh. 2015. Incorporating both distributional and relational semantics in word representations.

Ama Herdadelen and Marco Baroni. 2009. BagPack: A general framework to represent semantic relations. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 33–40. Association for Computational Linguistics.

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-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 356–364. Association for Computational Linguistics.

Douwe Kiela and Stephen Clark. 2014. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL*, pages 21–30.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. 2015. Multilingual reliability and semantic structure of continuous word spaces. In *Proceedings of the 11th International Conference on Computational Semantics 2015*, pages 40–45. Association for Computational Linguistics.

Siwei Lai, Kang Liu, Liheng Xu, and Jun Zhao. 2015. How to generate a good word embedding?

Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. 2:531–545.

Jean-Franois Lavallée and Philippe Langlais. 2010. Unsupervised morphological analysis by formal analogy. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pages 617–624. Springer.

Rmi Lebret and Ronan Collobert. 2015. Rehabilitation of count-based models for word vector representations. In *Computational Linguistics and Intelligent Text Processing*, pages 417–429. Springer.

Yves Lepage and Chooi-ling Goh. 2009. Towards automatic acquisition of linguistic features. In *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA 2009), eds., Kristiina Jokinen and Eckard Bick*, pages 118–125.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *ACL (2)*, 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, Ann Arbor, Michigan, June. Association for Computational Linguistics.

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

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

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

Tomas 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 (NAACL-HLT-2013)*, pages 746–751. Association for Computational Linguistics.

George Miller and Christiane Fellbaum. 1998. *Wordnet: An electronic lexical database*. MIT Press: Cambridge.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, volume 12, pages 1532–1543.

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. EVALution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015)*, pages 64–69.

Radu Soricut and Franz Och. 2015. Unsupervised morphology induction using word embeddings. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1627–1637.

Peter Turney, Michael Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems. In *In Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 482–489.

Peter D. Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 905–912.

Ekaterina Vylomova, Laura Rimmel, Trevor Cohn, and Timothy Baldwin. 2015. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning.

Christian Wartena. 2014. On the effect of word frequency on distributional similarity. In *Proceedings of the 12th edition of the KONVENS conference - Hildesheim*, volume 1, pages 1–10.