What just happened? Evaluating retrofitted distributional word vectors

Dmetri Hayes
International Computer Science Institute
University of California, Berkeley
Berkeley, CA, 94720, USA
dmetri@berkeley.edu

Abstract

Recent work has attempted to enhance vector space representations using information from structured semantic resources. This process, dubbed retrofitting Faruqui et al. (2015), has yielded improvements in word similarity performance. Research has largely focused on the retrofitting algorithm, or on the kind of structured semantic resources used, but little research has explored why some resources perform better than others. We conducted a fine-grained analysis of the original retrofitting process, and found that the utility of different lexical resources for retrofitting depends on two factors: the coverage of the resource and the evaluation metric. Our assessment suggests that the common practice of using correlation measures to evaluate increases in performance against full word similarity benchmarks 1) obscures the benefits offered by smaller resources, and 2) overlooks incremental gains in word similarity performance. We propose root-mean-square error (RMSE) as an alternative evaluation metric, and demonstrate that correlation measures and RMSE sometimes yield opposite conclusions concerning the efficacy of retrofitting. This point is illustrated by word vectors retrofitted with novel treatments of the FrameNet data (Fillmore and Baker, 2010).

1 Introduction

One of the most challenging tasks in the field of Natural Language Processing (NLP) is accurately encoding meaning into a computational system. Currently, the predominant approach is to represent the meanings of linguistic units, such as words or phrases, as vectors in a high-dimensional space. Vector embeddings are trained over large text corpora using machine-learning techniques, and have proven useful for a wide range of applications, such as named entity recognition (Turian et al., 2010), semantic role labeling (Collobert et al., 2011), sentiment analysis (Socher et al., 2013), and machine translation (Zou et al., 2013).

Word vectors are typically trained solely on the distributional information from text corpora. Recent work has attempted to improve word vectors by infusing them with information from semantic resources in a post-processing step. This technique, referred to as *retrofitting*, was introduced by Faruqui et al. (2015). They adjusted pre-trained embeddings based on lexical relations in WordNet (Miller, 1995), FrameNet (Fillmore and Baker, 2010), and the Paraphrase Database (Ganitkevitch et al., 2013). In some cases, this method yielded gains in word similarity performance.

Retrofitting has been extended in a variety of ways. Briefly, these include 1) adding word-to-word relations to encompass more than just similarity relations, such as by directly introducing antonymy relations (Mrkšić et al., 2016), or by explicitly modeling the pairwise relations between items (Lengerich et al., 2017); 2) increasing the size of the output vocabulary (Speer et al., 2017), or extending the process to affect the word vectors of words outside of the semantic resource (Glavaš and Vulić, 2018); and 3) constructing sense-specific word vectors using a word sense ontology (Jauhar et al., 2015), or word sense information learned from parallel text corpora (Ettinger et al., 2016).

However, while Faruqui et al. (2015) has certainly spawned a productive line of research into improving pre-trained word vectors, the original study contained a puzzling finding: retrofitting with certain semantic resources actually appeared to harm the quality of the word embeddings. This seems counter-intuitive. In principle, if semantic resources contain information that is not already captured by the word vectors, then retrofitting should always improve them.

Proceedings of NAACL-HLT 2019, pages 1062–1072
Minneapolis, Minnesota, June 2 - June 7, 2019.

In order to understand why some semantic resources appear better suited for retrofitting word vectors, we conducted a fine-grained analysis of Faruqui et al.'s original technique. Given their popularity, we focused on word similarity evaluations. We observe that the perceived usefulness of a semantic resource depends on its coverage of the words in the evaluation benchmark. Furthermore, we report that the choice of evaluation metric can lead to different conclusions. We note that some gains in performance are not captured by correlation measures, and propose that root-mean-square error (RMSE) is more appropriate for measuring changes in word similarity performance.

2 Methods

2.1 Retrofitting

The original retrofitting algorithm from Faruqui et al. (2015) is described below. The process essentially moves the word vectors of related words closer together. A semantic resource can be regarded as a graph which covers a vocabulary $V = \{w_1, ..., w_n\}$ and denotes relations between them as edges $(w_i, w_j) \in E$. Given a set of pre-trained distributional vectors $\vec{\hat{W}} = \{\vec{\hat{w}}_1, ..., \vec{\hat{w}}_d\}$ and a semantic resource with edges E, the goal is to learn a new set of vectors $\vec{W} = \{\vec{w}_1, ..., \vec{w}_d\}$. Here $\vec{w}_i$ is the word vector corresponding to vocabulary item w_i. The objective function to be minimized is the following:

$$\sum_{w_i}^{V} \left(\alpha_i ||\vec{w}_i - \vec{\hat{w}}_i||^2 + \sum_{(w_i,w_j)}^{E} \beta_{ij} ||\vec{w}_i - \vec{w}_j||^2 \right) \quad (1)$$

The first term of the inner sum ensures that the vectors do not stray too far away from their original representations (controlled by α), while the second term compels the vectors to move closer to their neighbors in the semantic resource (controlled by β). In Faruqui et al.'s experiments, all $\alpha_i = 1$, and all $\beta_{ij} = \frac{1}{degree(w_i)}$, where $degree(w_i)$ refers to the number of neighbors w_i had in the resource. This is equivalent to specifying that half of the new retrofitted vector will come from the distributional data while the other half will be an average of its neighbors' word vectors. They allowed the process to run for 10 iterations. We retained these settings in our experiments.

Resource	Terms	Groupings
WordNet+	147,306	117,659
PPDB	84,467	102,899
FrameNet	8,483	1,074
FrameNet-Anno	37,855	7,146

Table 1: Number of word forms and word groupings per semantic resource.

2.2 Semantic resources

We employed three semantic resources in our analyses. Table 1 shows the number of terms and groupings in each resource after removing terms containing numbers or punctuation.[1]

WordNet. WordNet (Miller, 1995) is a large lexical database of English words. The resource is composed of synsets, groupings of synonyms. Synsets are linked together through a small number of semantic relations. We follow Faruqui et al. (2015) and link each word form to its synonyms, hypernyms, and hyponyms (WN+). For instance, the word *dog* is linked to *canine* (synonym), *corgi* (hyponym) and *domestic_animal* (hypernym). In order to faithfully replicate Faruqui et al., we collapsed part of speech and sense distinctions, meaning that a word form was linked to all of its related words through all of its synsets. For instance, *dog*'s neighbors include *corgi* through the noun *dog* (e.g. "Sam pet the dog.") and *track* through the verb *to dog* (e.g. "The task dogged me.") Although the word vectors and evaluations used in this study are insensitive to part of speech and sense distinctions, the number and order of groupings affects the retrofitting procedure. In particular, as noted by Speer and Chin (2016), the results depend on the order in which the groupings are iterated over. Though we attempted to group words by their synsets, this appeared to lead to poorer performance and we do not report those results here.

PPDB. The paraphrase database (Ganitkevitch et al., 2013) contains millions of English paraphrases automatically extracted from bilingual parallel corpora. The core idea is that if a non-English phrase translates to two distinct English strings, then these may be considered paraphrases of each other. For instance, since German *festgenommen* translates to both "thrown into jail"

[1]The number of groupings for PPDB is approximate, taken as the number of unique sets of words in Faruqui et al.'s pre-processed lexicon file.

and "imprisoned", the latter two are listed as paraphrases. Faruqui et al. (2015) used the XL lexical pack from PPDB 1.0. Since this version is no longer publically available, we used their preprocessed file (PPDB).

FrameNet. FrameNet (Fillmore and Baker, 2010) is a highly-interconnected lexical database of English containing sense-annotated sentences. The basic units of FrameNet are semantic frames, which specify the conceptual structure necessary to understand sets of lexical units (LUs). For instance, the frame Attack contains LUs such as *attack.v*, *attack.n* and *offensive.a*, which can be understood in light of the frame elements (FEs) *Assailant* and *Victim*. We performed two experiments with the FrameNet data. In the first, we grouped words together if they shared a frame (FN). Note that this differs from the treatment of WordNet because the frame groupings retain part of speech and sense distinctions. Although this method follows Faruqui et al. (2015), we located a bug in their code which led to a loss of about 1/3 of the data: the original code did correctly handle polysemy, which is widespread in FrameNet.

For our second experiment, we grouped words together based on the FEs that they filled (FN-ANNO). All of the FrameNet FEs were used in this task (i.e. both core and non-core FEs). Since FEs are defined with respect to their frames, each semantic role is frame-specific. The rationale is that words which can occupy the same semantic role should be more similar. We created groupings from the last nouns which appeared in the FE fillers in the annotation data. To illustrate, since the annotation data linked to the FE *Assailant* of the Attack frame included the nouns *enemy*, *troop*, *terrorist* and *forces*, their corresponding word vectors were moved closer together. Note that all of our retrofitting analyses ignored the frequency of a word's neighbor: even if *enemy* filled the FE *Assailant* 100 times, its effect on its neighbors would be identical to if it had only filled the FE once.

We recognize that the last noun heuristic is simplistic. However, we estimate that around 73% of the syntactic heads of FE fillers are nouns. Of these, 68% contain only one noun, and 18% contain only two nouns. Taken together, this implies that a more sophisticated approach is unlikely to alter the results. In addition to the last noun heuristic, we considered grouping the first nouns in the FE fillers, all of the nouns in the FE fillers, and the nouns from FE fillers which contained only one noun. All of these experiments yielded similar results, so we only report the last noun condition here. Nouns were identified using the default NLTK (Bird and Loper, 2004) English part-of-speech tagger.

2.3 Word vectors

Our analyses included two popular pre-trained word vector embeddings.

SG. word2vec (Mikolov et al., 2013) is widely-used to learn vector representations from distributional information. In the continuous skip-gram architecture (SG), the target word is fed into a log-linear classifier to predict surrounding words within a given context window. The available vectors were trained on about 100 billion words from the Google News dataset.

GloVe. Global Vectors for Word Representation (Pennington et al., 2014) is a global log-bilinear regression model which captures both global and local word co-occurrence statistics. We use the 300 dimension vectors trained on 6 billion words from Wikipedia and the English Gigaword corpus.

2.4 Word similarity

Word similarity judgments are the most widely-used method of intrinsic evaluation. We chose four commonly used word similarity datasets comprised of nouns, verbs and adjectives.

MEN3K (Bruni et al., 2012) contains 3,000 pairs of words from a set of labels for an image database. Interestingly, although Bruni et al. claim that their dataset "contains 3,000 pairs of randomly selected words that occur [as labels]", it only contains 751 unique words.[2] Therefore, as an additional evaluation of high-frequency words, we included MTURK-771 (Halawi et al., 2012), a crowd-sourced dataset of 771 word-pairs consisting of 1,113 unique words which we will refer to as **MT771**. The Stanford Rare Words (**RW**) dataset (Luong et al., 2013) is comprised of 2,034 word-pairs formed from 2,951 unique words. **SL999** (Hill et al., 2015) explicitly quantifies semantic similarity between pairs of words. The dataset contains 999 word pairs from 1,028 unique words. The word pairs in SL999 were chosen to cover the full range of concreteness within each part of speech category. We included RW

[2]By our calculations, the expected number of unique words obtained from 3,000 random pairs drawn from 20,515 labels (the number in their image database) is around 5,200.

and SL999 to examine whether the results of our analyses would differ for benchmarks containing common vs. rare words and for those capturing association and relatedness vs. similarity only.

3 Evaluation procedure

The standard approach to evaluate the performance of word vectors on word similarity judgments is to compute the cosine similarity values between each pair of words in the dataset and then calculate the correlation between these values and the similarity scores collected from human raters. A similar technique is used to assess the utility of different semantic resources in retrofitting word vectors: increases in correlation are taken to be indicative that information from the resource has been successfully injected into the word vectors. For both types of evaluations, Spearman correlation has become the preferred correlation measure.

However, there are several reasons that this method may be misleading. The first concerns the issue of the relative coverage of each resource. Simply put, not every resource contains all of the words in the evaluation dataset. If a resource lacks the words for a particular similarity judgment, then the predicted score will be the same for both the baseline and retrofitted vectors. This may have important consequences on the evaluation metric: the fixed scores can throw off the global ranking of the predicted scores, which is measured by the Spearman correlation.

For every word pair in a word similarity dataset, a resource can contain 1) both words, 2) one of the words, or 3) neither of the words. If the goal of the evaluation is to determine whether the knowledge of particular semantic resources can be added to word vectors, then it seems reasonable to only evaluate the resource on the word pairs it covers. In this case, the resource will either group the two words together or place them in separate groups, which can be interpreted as explicitly indicating whether the two words are semantically related or not. Conversely, it is obvious that retrofitting will not improve the vectors for the word pairs for which neither word is in the semantic resource.

The situation where only one word is present is more complicated. For example, imagine that a resource contained the word *view* but not the word *skyline*. Following retrofitting the vector for *view* will move while the vector for *skyline* will stay the same. The relationship between *view* and *skyline* will either become more accurate or less accurate, but this change does not directly stem from the semantic resource. If the goal of the retrofitting evaluation is to assess the usefulness of particular semantic resources, then including these kinds of word pairs is misleading, since the observed changes are incidental and do not reflect the semantic groupings in the resource.

In our analyses, "all pairs" shows the performance of the word vectors using all of the word similarity judgments, and "pairs in resource" shows their performance using only the subset comprised of judgments for which both words were contained in the semantic resource.

Our more radical proposal is to consider an entirely different evaluation metric altogether. Measures of correlation indicate how well word vectors are able to predict the similarity judgments. Spearman correlation specifically measures how well word vectors are able to predict the correct rankings of similarity judgments. For example, according to the MEN3K dataset, *brick* and *construction* should be ranked as less similar than *town* and *village*. Another conceivable way to test the word vectors ability to capture word similarity knowledge would be to directly compare the word vectors' predicted score with the human score. According to MEN3K, the average rated similarity for *town* and *village* was 43 out of 50. Taken literally, after normalizing the original scores the cosine similarity should be exactly 0.86. We operationalized this by evaluating word vectors using root-mean-square error (RMSE). This approach seems particularly appealing for measuring the effects of retrofitting because each similarity judgment contributes independently to the RMSE score.

One may wonder whether Pearson correlation, which measures linear association, might serve as a better comparison to RMSE. To address this concern, we employed the harmonic mean of the Pearson and Spearman correlations as our correlation measure. This blends the linear measure (Pearson) with the standardly-employed measure (Spearman). However, we note that the resulting baseline and retrofitted scores were very similar across correlation measures, and so our conclusions regarding the choice of evaluation metric were unaffected by this decision.

In the analysis that follows, we considered the effect of resource coverage and evaluation metric

	NB	GloVe	SG
MT771	**0.80 / 0.35**	0.65 / 0.36	0.66 / 0.39
MEN3K	**0.85** / 0.28	0.75 / 0.27	0.78 / **0.27**
RW	**0.54 / 0.37**	0.35 / 0.55	0.45 / 0.45
SL999	**0.66 / 0.21**	0.38 / 0.26	0.45 / 0.25

Table 2: Baseline word vector similarity performance. Scores are listed in the form "Correlation/RMSE". Bold face indicates the best-performing set of vectors for each similarity dataset for their correlation score and for their RMSE score. Recall that lower RMSE is better.

on the results of retrofitting. There were four conditions: 1) Correlation, all word pairs in the benchmark, 2) Correlation, only those pairs in which both words were in resource, 3) RMSE, all word pairs in the benchmark, and 4) RMSE, only those pairs in which both words were in resource. If one of the words in a word pair was missing from the word vectors, then it was assigned a predicted cosine similarity of zero. (This only occurred with the RW dataset, and was limited to the all pairs conditions.)

4 Results

Table 2 shows the baseline word similarity performance according to the harmonic mean of the correlation measures and RMSE. As a reference, we include the NumberBatch (NB) vectors, which recently demonstrated state-of-the-art word similarity performance (Speer et al., 2017). Correlation and RMSE give similar baseline results among the vector sets and their ability to predict the four similarity benchmarks: NB performs the best. The exception is that SG scores a slightly better RMSE score on the MEN3K dataset.

4.1 All word pairs

Figure 1 shows the measured improvements in correlation due to retrofitting. This mirrors Faruqui et al. (2015)'s original finding that the PPDB offers the most improvements, and that grouping words by FrameNet frames (FN) usually leads to worse performance. Note that this finding is observed after correcting for the issue from Faruqui et al. which omitted data from FrameNet. This plot also suggests that using FrameNet frame elements (FN−ANNO) to group words is very detrimental to word vectors.

As shown in Figure 2, simply switching the evaluation metric to RMSE paints a much different picture. (Since RMSE measures error rather than improvement, the y-axis has been inverted so that improvement is still in the upward direction.) The most obvious difference is that according to RMSE all of the semantic resources appear to help. Compared to Figure 1, there is a noticeable boost in performance for WN+, especially when evaluated against RW. Remarkably, FN−ANNO almost completely flips polarity. The result is especially dramatic against the evaluation sets containing common words (i.e. MT771 and MEN3K): FN−ANNO goes from being the worst-performing resource to one of the best-performing resources.

4.2 Word pairs in resource

Figure 3 shows the measured improvements in correlation when considering only the word pairs in which both words were present in resources. The ranked order of the semantic resources is virtually the same. Note, however, that the measured performance of the relatively low-coverage resource, FrameNet (FN), has jumped considerably: in the RW with GloVe condition, it overtakes PPDB as the resource providing the best improvement.

Figure 4 measures the change of RMSE for the word pairs covered by the resources. FrameNet (FN) appears to yield a substantial gain in performance for the subset of the similarity judgments that it covers, and again emerges as the highest-performance resource when evaluated against RW. A direct comparison of the "all pairs" to "pairs in resource" figures shows that the scores of the other resources change very little. The difference is attenuated because these resources are much larger and therefore cover most of the words in the similarity datasets.

We interpret the jumps in performance from the "all pairs" to "pairs in resource" condition as evidence that evaluating a resource on word pairs containing a mixture of words within and outside of its vocabulary may obscure its benefits. Of course, low coverage is problematic if the goal is to improve word vectors on a large number of word judgments. The "pairs in resource" assessment is particularly antithetical to the spirit of RW, which is often employed to assess word vector coverage, and we admit that FN only contains 6.3% of the RW word pairs. However, we

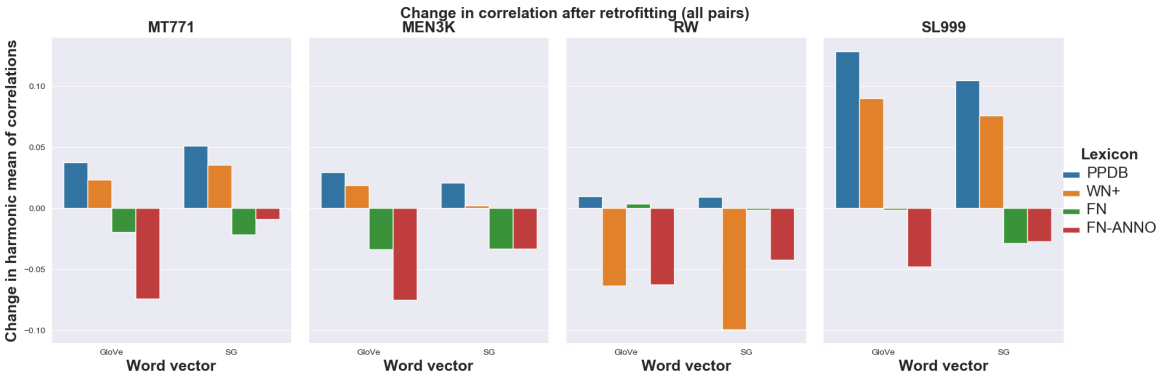

Figure 1: Change in correlation after retrofitting, considering all word pairs

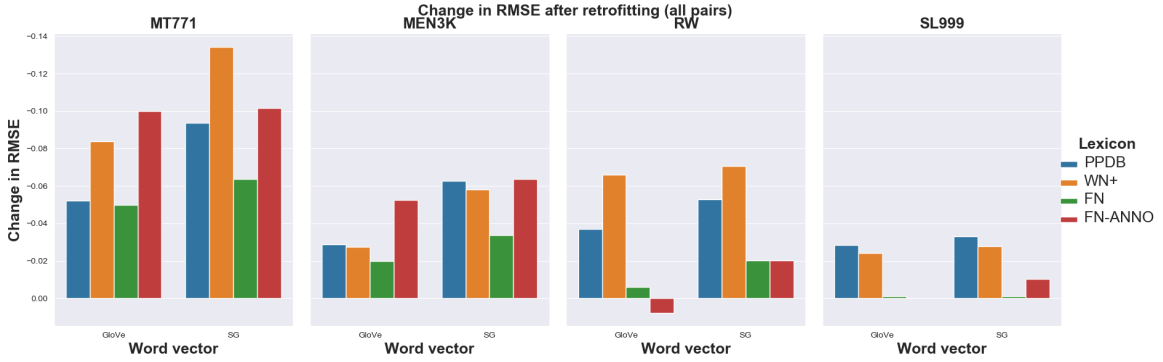

Figure 2: Change in Root-mean-square error after retrofitting, considering all word pairs

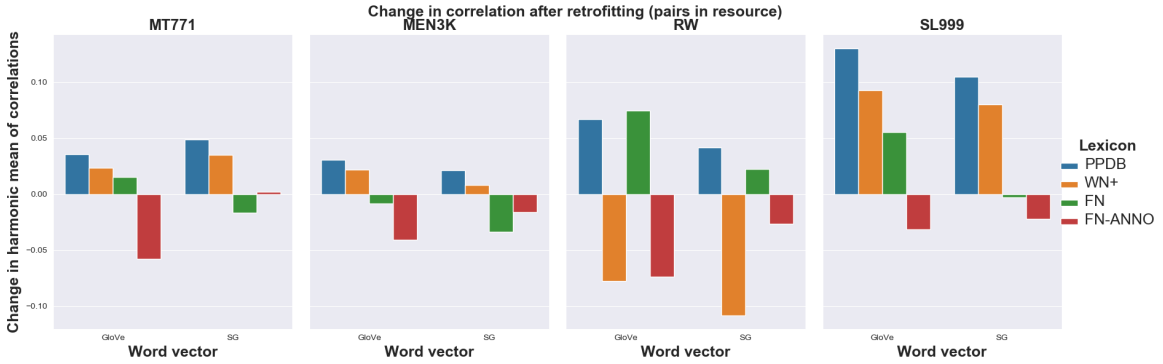

Figure 3: Change in correlation after retrofitting, considering only the word pairs in each resource

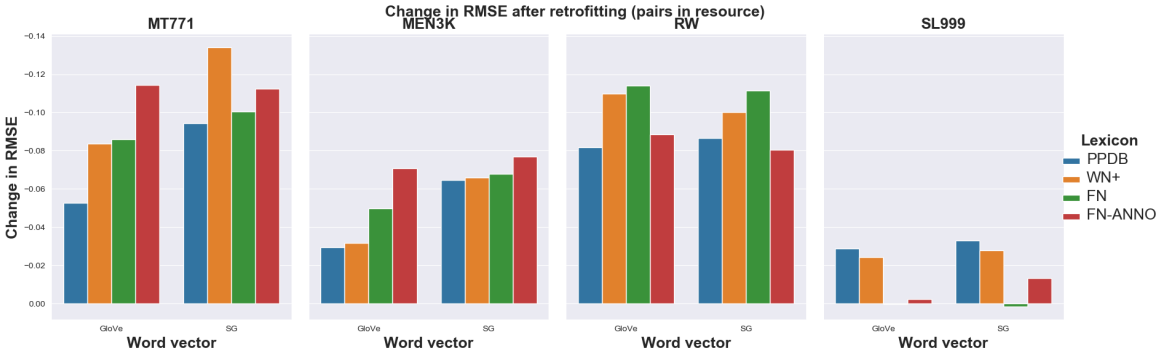

Figure 4: Change in Root-mean-square error after retrofitting, considering only the word pairs in each resource

would argue that there is an important difference between concluding that a semantic resource does not yield gains in retrofitting vs. concluding that the resource improves the quality of the vectors it covers.

4.3 SimLex

We note that our four conditions yield similar conclusions according to the SL999 evaluation set. PPDB and WN+ consistently offer strong improvements, in contrast to FN and FN−ANNO. This is not surprising, and follows from the design principles underlying each resource: while PPDB and WordNet specifically group synonyms, FrameNet groups words which evoke the same semantic frame. In particular, some frames intentionally contain antonyms. As discussed above, the FrameNet groupings still appear useful in improving against MT771, MEN3K and RW, which have been argued to conflate association and similarity (Hill et al., 2015).

4.4 Further analysis

Our most striking finding is that correlation measures and RMSE occasionally yield opposite conclusions regarding the utility of semantic resources. How can the retrofitted data simultaneously show a drop in correlation and a gain in RMSE? To examine this further, we plotted the effects of retrofitting GloVe with FN−ANNO against the MT771 benchmark (Figure 5). Vector cosine similarity (x-axis) is plotted against the human similarity judgments (y-axis). The left and right panels compare the vector performance before and after retrofitting. Each point represents a single word pair in the MT771 dataset. The dashed line corresponds to a model which perfectly predicts the gold standard. Points are color-coded with respect to this line: green points mark word pairs whose computed cosine similarity moved closer to the human judgments, while red points indicate word pairs who moved in the opposite direction. A small number of blue points indicate predictions which were unaffected by retrofitting because the word pairs were not present in the resource.

The color-coding in Figure 5 helps illustrate how both Spearman correlation (a measure of goodness) and RMSE (a measure of error) decrease. Most of the points are green, which means that from the perspective of individual word pairs, the predictions from the retrofitted vectors are more in line with the gold standard. This is directly reflected in RMSE. However, while most of the mass moves closer to the dashed line, retrofitting increases the scatter of the points, resulting in a worse association between the vector cosine similarity human similarity judgments.

Three points are labeled in Figure 5 to show the effect of retrofitting on individual word similarity predictions. The diamond marks the word pair *find & occurrence*, which yields the most improvement according to MT771, with its absolute residual (i.e. distance from the human judgment) dropping 0.25. In comparison, the worst-performing word pair is *occasion & second*, marked with an X, whose residual increases by 0.14. This point is part of a noticeable band of red points located near the dotted line. Interestingly, for these points the predicted scores for the baseline word vectors were nearly correct, and retrofitting pushed them to overpredict similarity. The square marks *film & movie*, whose residual drops an almost imperceptible 0.003.

The reason that retrofitting may lead to a worse correlation but a better RMSE score stems from how these measures are computed from the data. Each word pair contributes independently to the RMSE score. Whether a word pair improves or decreases in performance, it is simply tallied onto the running RMSE score. In this case, it is irrelevant whether retrofitting leads to a large increase in scatter. In contrast, correlation measures are anchored to the sample means of the two variables. After retrofitting, there may be an increase in the scatter in the predicted cosine similarity values. Since on average the word pairs will be further away from the sample mean, there will be a drop in correlation. Put another way, a word pair's contribution to the correlation score depends on the positions of the all of the other word pairs.

The particularly large drop in correlation for FN−ANNO likely stems from the unusual heterogeneity of its groupings. For example, the word *film* occurs in the annotation data of 108 distinct FEs in FrameNet, and is grouped with dozens of varied words, such as *book*, *movie*, but also *DNA* and *meeting*. Each of the 108 retrofitting adjustments introduces some scatter. In contrast, the neighbors in other resources can be straightforwardly interpreted as related words, and each word will appear in a small number of groupings.

We note that while it may be instructive to track the performance of individual word pairs, it is dif-

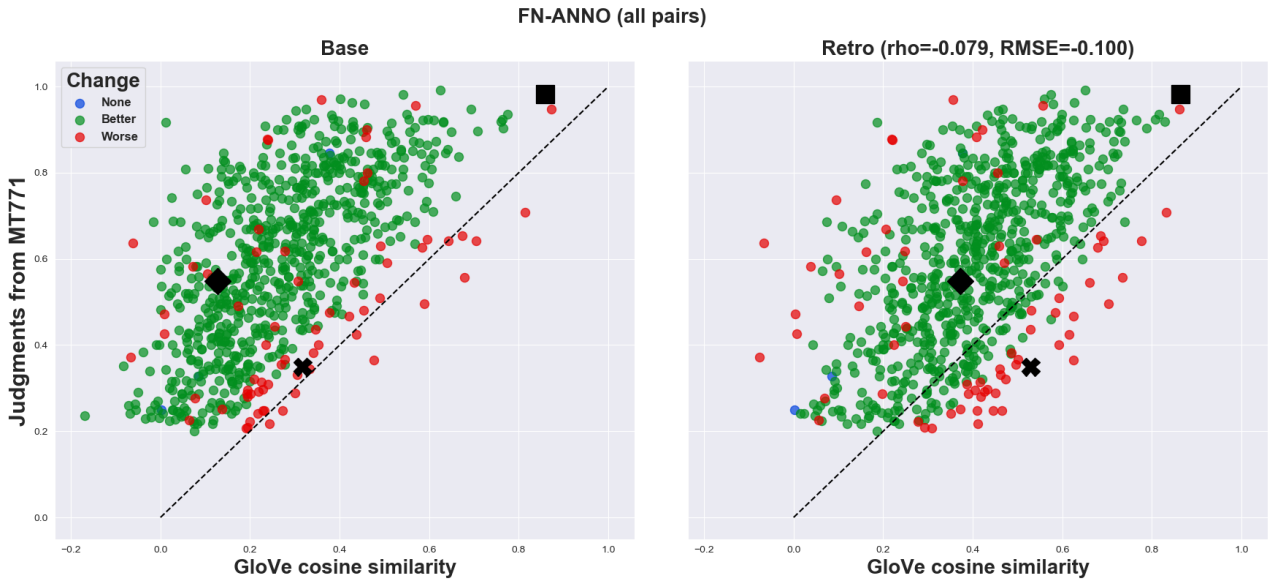

Figure 5: Effects of retrofitting GloVe by grouping nouns filling the same frame element in the FrameNet annotation data, considering all word pairs. Vector-computed similarity is plotted against the MT771 gold standard judgments using the original word vectors (left panel) and the retrofitted vectors (right panel). The dashed line illustrates a model which exactly predicts the human judgments. Predicted scores which moved closer to that line are colored green, while points which moved away from the line are colored red. Blue points represent word pairs which were not present in the resource, and so were unaffected by retrofitting. The changes in Spearman correlation and RMSE are shown above the right panel. The symbols are discussed in the text.

ficult to pinpoint the exact source of the change. For instance, in Figure 5 the words corresponding to the square (little change) and the X (worse change) are paired together, while the word pair linked to the diamond (best change) are not.

5 Related work

Faruqui et al. (2015) attributed FrameNet's comparatively poor performance to the fact that it groups words according to abstract concepts, noting that *push* and *grow* are in the same frame. Such an argument might explain why FrameNet does not yield gains in performance against SL999, which was designed to capture true similarity judgments. However, we have shown that conclusions on the other similarity benchmarks rest on the evaluation metric and on the types of word pairs considered. In the RMSE and "pairs in resource" condition, grouping words by FrameNet frames appears at least as useful as PPDB and WordNet. Alternatively, FrameNet can be interpreted as a useful resource for retrofitting the vectors of the words it contains as lexical units.

Our novel treatment of FrameNet groups nouns using its collection of sense-annotated sentences. Although all of the frame elements in these sentences were annotated by hand, the words filling the FEs are not, adding a component of randomness. Especially with more semantically general frames, frame elements can be realized by a large number of words. This contrasts with FrameNet frames, in which the placement of word senses are painstakingly deliberated, and a particular sense can only be put into one frame.

PropBank (Bonial et al., 2014) is a large semantically-annotated corpus. The semantic roles ("rolesets") in PropBank are defined with respect to individual verb and noun word senses. The types of words that fill these roles are presumably less varied than those that fill the semantically broader FrameNet frame elements. Additionally, PropBank is considerably larger than FrameNet. Consequently, we might predict that retrofitting word vectors to PropBank would yield stronger gains in word similarity judgment than to the FrameNet annotation data. We leave this task for future research.

Grouping nouns using the FrameNet annotation data led to large drops in correlation against word similarity benchmarks. However, these same data yielded large gains in RMSE performance. It might be inferred that semantic resources which

have a similar stochastic component may result lower correlation. The PPDB is automatically generated, introducing a similar element of randomness, but this is curtailed by its conservative criteria: paraphrases must be attested as translation equivalents.

BabelNet (Navigli and Ponzetto, 2012) and ConceptNet (Speer et al., 2017) are knowledge resources derived from a number of collaboratively-constructed sources, such as Wikipedia and Wiktionary. Though their collaborative nature likely makes them less accurate than hand-curated resources such as WordNet, they have potential in improving the quality of word vectors (e.g. Speer and Chin, 2016). As we observed with FN−ANNO, RMSE may be a more informative measure of comparison than correlation in future retrofitting experiments involving heterogeneous resources.

More generally, there does not seem to be a strong theoretical reason to prefer correlation-based measures over residual-based ones. Although the current practice is to report the Spearman's rank correlation coefficient between the vector cosine similarities and human word similarity judgments, for over a decade the standard was to report Pearson product-moment correlation coefficient. When Resnik (1995) pioneered the technique of comparing computed measures of similarity with human similarity ratings, he used (Pearson) correlation as "one reasonable way to judge [computational measures of semantic similarity]".

The switch to Spearman correlation appears to have occurred in Gabrilovich and Markovitch (2007), who employed it without comment. Agirre et al. (2009) did provide a justification, saying, "In our belief Pearson is less informative, as the Pearson correlation suffers much when the scores of two systems are not linearly correlated, something which happens often due to the different nature of the techniques applied." Unfortunately, Agirre et al. (2009) mischaracterized the popularity of Spearman correlation by claiming that all researchers have used Spearman in evaluating WordSim-353 dataset (Finkelstein et al., 2002). This likely stems from a misinterpretation of Gabrilovich and Markovitch's Table 4, which compares their methodology with earlier studies using Spearman correlation. The latter authors apparently recomputed word relatedness with the associated algorithms, as the cited studies report Pearson correlation values.

Willmott (1981; 1982) specifically argues that Pearson correlation should not be used to evaluate model performance, and that RMSE is superior at comparing observed and simulated data.[3] However, as far as we know, no previous work has seriously considered evaluating the performance of computed word similarity scores using RMSE. Reliance on Spearman correlation may lead to incorrect conclusions regarding the quality of word vectors.

6 Conclusion

Retrofitting distributional word vectors using relational information in semantic resources can yield improvements in word similarity performance. Our fine-grained analysis of the original retrofitting process shows that 1) the evaluation metric matters: root-mean-square error (RMSE) is more sensitive to gains in performance than correlation measures; and 2) coverage matters: improvements offered by resources are highly dependent on their coverage of the evaluation benchmark. Future attempts to improve word vectors can only succeed if gains in word vector performance are inspected carefully.

Acknowledgments

This research was supported in part by the Defense Threat Reduction Agency (DTRA). Disclaimer: The project or effort depicted was or is sponsored by the Department of the Defense, Defense Threat Reduction Agency. The content of the information does not necessarily reflect the position or the policy of the federal government, and no official endorsement should be inferred.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL*

[3] Willmott and Matsuura (2005) list several advantages of a related metric, mean absolute error (MAE), over RMSE. In our experiments, there were no qualitative differences between MAE and RMSE.

2004 on Interactive poster and demonstration sessions, page 31. Association for Computational Linguistics.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D Hwang, and Martha Palmer. 2014. Propbank: Semantics of new predicate types. In *LREC*, pages 3013–3019.

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.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Allyson Ettinger, Philip Resnik, and Marine Carpuat. 2016. Retrofitting sense-specific word vectors using parallel text. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1378–1383.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*, pages 1606–1615.

Charles J Fillmore and Collin Baker. 2010. A frames approach to semantic analysis. In *The Oxford handbook of linguistic analysis*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1):116–131.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, volume 7, pages 1606–1611.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Goran Glavaš and Ivan Vulić. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 34–45.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414. ACM.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 683–693.

Benjamin J Lengerich, Andrew L Maas, and Christopher Potts. 2017. Retrofitting distributional embeddings to knowledge graphs with functional relations. *arXiv preprint arXiv:1708.00112*.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, volume 1, pages 448–453.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Robert Speer and Joshua Chin. 2016. An ensemble method to produce high-quality word embeddings. *arXiv preprint arXiv:1604.01692*.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451. Association for the Advancement of Artificial Intelligence.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Cort J Willmott. 1981. On the validation of models. *Physical geography*, 2(2):184–194.

Cort J Willmott. 1982. Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society*, 63(11):1309–1313.

Cort J Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.