A Large Scale Evaluation of Distributional Semantic Models: Parameters, Interactions and Model Selection

Gabriella Lapesa[2,1]
[1]Universität Osnabrück
Institut für Kognitionswissenschaft
Albrechtstr. 28, Osnabrück, Germany
gabriella.lapesa@fau.de

Stefan Evert[2]
[2]FAU Erlangen-Nürnberg
Professur für Korpuslinguistik
Bismarckstr. 6, Erlangen, Germany
stefan.evert@fau.de

Abstract

This paper presents the results of a large-scale evaluation study of window-based Distributional Semantic Models on a wide variety of tasks. Our study combines a broad coverage of model parameters with a model selection methodology that is robust to overfitting and able to capture parameter interactions. We show that our strategy allows us to identify parameter configurations that achieve good performance across different datasets and tasks[1].

1 Introduction

Distributional Semantic Models (DSMs) are employed to produce semantic representations of words from co-occurrence patterns in texts or documents (Sahlgren, 2006; Turney and Pantel, 2010). Building on the Distributional Hypothesis (Harris, 1954), DSMs quantify the amount of meaning shared by words as the degree of overlap of the sets of contexts in which they occur.

A widely used approach operationalizes the set of contexts as co-occurrences with other words within a certain window (e.g., 5 words). A window-based DSM can be represented as a co-occurrence matrix in which rows correspond to target words, columns correspond to context words, and cells store the co-occurrence frequencies of target words and context words. The co-occurrence information is usually weighted by some scoring function and the rows of the matrix are normalized. Since the co-occurrence matrix tends to be very large and sparsely populated, dimensionality reduction techniques are often used to obtain a more compact representation. Landauer and Dumais (1997) claim that dimensionality reduction also improves the semantic representation encoded in the co-occurrence matrix. Finally, distances between the row vectors of the matrix are computed and – according to the Distributional Hypothesis – interpreted as a correlate of the semantic similarities between the corresponding target words. The construction and use of a DSM involves many design choices, such as: selection of a source corpus, size of the co-occurrence window; choice of a suitable scoring function, possibly combined with an additional transformation; whether to apply dimensionality reduction, and the number of reduced dimensions; metric for measuring distances between vectors. Different design choices – technically, the DSM parameters – can result in quite different similarities for the same words (Sahlgren, 2006).

DSMs have already proven successful in modeling lexical meaning: they have been applied in Natural Language Processing (Schütze, 1998; Lin, 1998), Information Retrieval (Salton et al., 1975), and Cognitive Modeling (Landauer and Dumais, 1997; Lund and Burgess, 1996; Padó and Lapata, 2007; Baroni and Lenci, 2010). Recently, the field of Distributional Semantics has moved towards new challenges, such as predicting brain activation (Mitchell et al., 2008; Murphy et al., 2012; Bullinaria and Levy, 2013) and modeling meaning composition (Baroni et al., 2014, and references therein).

Despite such progress, a full understanding of the different parameters governing a DSM and their influence on model performance has not been achieved yet. The present paper is a contribution towards this

[1] The analysis presented in this paper is complemented by supplementary materials, which are available for download at http://www.linguistik.fau.de/dsmeval/. This page will also be kept up to date with the results of follow-up experiments.

Transactions of the Association for Computational Linguistics, vol. 2, pp. 531–545, 2014. Action Editor: Janyce Wiebe.
Submission batch: 3/2014; Revision batch 9/2014; Published 12/2014.

goal: it presents the results of a large-scale evaluation of window-based DSMs on a wide variety of semantic tasks. More complex tasks building on distributional representations (e.g., vector composition or relational analogies) will also benefit from our findings, allowing them to choose optimal parameters for the underlying word-level DSMs.

At the level of parameter coverage, this work evaluates most of the relevant parameters considered in comparable state-of-the-art studies (Bullinaria and Levy, 2007; Bullinaria and Levy, 2012); it also introduces an additional one, which has received little attention in the literature: the *index of distributional relatedness*, which connects distances in the DSM space to semantic similarity. We compare direct use of distance measures to *neighbor rank*. Neighbor rank has already been successfully used to model priming effects with DSMs (Hare et al., 2009; Lapesa and Evert, 2013); the present study extends its evaluation to standard tasks. We show that neighbor rank consistently improves the performance of DSMs compared to distance, but the degree of this improvement varies from task to task.

At the level of task coverage, the present study includes most of the standard datasets used in comparative studies (Bullinaria and Levy, 2007; Baroni and Lenci, 2010; Bullinaria and Levy, 2012). We consider three types of evaluation tasks: multiple choice (TOEFL test), correlation to human similarity ratings, and semantic clustering.

At the level of methodology, our work adopts the approach to model selection proposed by Lapesa and Evert (2013), which is described in detail in section 4. Our results show that parameter interactions play a crucial role in determining model performance.

This paper is structured as follows. Section 2 briefly reviews state-of-the-art studies on DSM evaluation. Section 3 describes the experimental setting in terms of tasks and evaluated parameters. Section 4 outlines our methodology for model selection. In section 5 we report the results of our evaluation study. Finally, section 6 summarizes the main findings and sketches ongoing and future work.

2 Previous work

In this section we summarize the results of previous evaluation studies of Distributional Semantic Models. Among the existing work on DSM evaluation, we can identify two main types of approaches.

One possibility is to evaluate a distributional model with certain new features on a range of tasks, applying little or no parameter tuning, and to compare it to competing models; examples are Pado and Lapata's (2007) Dependency Vectors as well as Baroni and Lenci's (2010) Distributional Memory. Since both studies focus on testing a single new model with fixed parameters (or a small number of new models), we will not go into further detail concerning them.

Alternatively, the evaluation may be conducted via *incremental tuning of parameters*, which are tested sequentially to identify their best performing values on a number of tasks, as has been done by Bullinaria and Levy (2007; 2012), Polajnar and Clark (2014), and Kiela and Clark (2014).

Bullinaria and Levy (2007) report on a systematic study of the impact of a number of parameters (shape and size of the co-occurrence window, distance metric, association score for co-occurrence counts) on a number of tasks (including the TOEFL synonym task, which is also evaluated in our study). Evaluated models were based on the British National Corpus. Bullinaria and Levy (2007) found that vectors scored with Pointwise Mutual Information, built from very small context windows with as many context dimensions as possible, and using cosine distance ensured the best performance across all tasks at issue.

Bullinaria and Levy (2012) extend the evaluation reported in Bullinaria and Levy (2007). Starting from the optimal configuration identified in the first study, they test the impact of three further parameters: application of stop-word lists, stemming, and dimensionality reduction using Singular Value Decomposition. DSMs were built from the ukWaC corpus, and evaluated on a number of tasks (including TOEFL and noun clustering on the dataset of Mitchell et al. (2008), also evaluated in our study). Neither stemming nor the application of stop-word lists resulted in a significant improvement of DSM performance. Positive results were achieved by performing SVD dimensionality reduction and discarding the initial components of the reduced matrix.

Polajnar and Clark (2014) evaluate the impact of context selection (for each target, only the most relevant context words are selected, and the remaining vector entries are set to zero) and vector normalization (used to vary model sparsity and the range of values of the DSM vectors) in standard tasks related to word and phrase similarity. Context selection and normalization improved DSM performance on word similarity and compositional tasks, both with and without SVD.

Kiela and Clark (2014) evaluate window-based and dependency-based DSMs on a variety of tasks related to word and phrase similarity. A wide range of parameters are involved in this study: source corpus, window size, number of context dimensions, use of stemming, lemmatization and stopwords, similarity metric, score for feature weighting. Best results were obtained with large corpora and small window sizes, around 50000 context dimensions, stemming, Positive Mutual Information, and a mean-adjusted version of cosine distance.

Even though we adopt a different approach than these incremental tuning studies, there is considerable overlap in the evaluated parameters and tasks, which will be pointed out in section 3.

An alternative to incremental tuning is the methodology proposed by Lapesa and Evert (2013) and Lapesa et al. (2014). They systematically test a large number of parameter combinations and use linear regression to determine the importance of individual parameters and their interactions. As their evaluation methodology is adopted in the present work and described in more detail in section 4, we will not discuss it here and instead focus on the main results. DSMs are evaluated in the task of modeling semantic priming. This task, albeit not standard in DSM evaluation, is of great interest as priming experiments provide a window into the structure of the mental lexicon. Both studies showed that *neighbor rank* outperforms *distance* in capturing priming effects. They also found that the scoring function has a crucial influence on model performance and interacts strongly with an additional logarithmic transformation. Lapesa et al. (2014) focused on a comparison of syntagmatic and paradigmatic relations. They found that discarding the initial SVD dimensions is only benefical for certain relations, suggesting that these dimensions may encode syntagmatic information if larger context windows are used. Concerning the scope of the evaluation, both studies consider a wide range of parameters[2] but target only a very specific task. Our study aims at extending their parameter set and evaluation methodology to standard tasks.

3 Experimental setting

3.1 Tasks

The evaluation of DSMs has been conducted on three standard types of semantic tasks.

The first task is a **multiple choice** setting: distributional relatedness between a target word and two or more other words is used to select the best, i.e. most similar candidate. Performance in this task is quantified by the decision accuracy. The evaluated dataset is the well-known **TOEFL multiple-choice synonym test** (Landauer and Dumais, 1997), which was also included in the studies of Bullinaria and Levy (2007; 2012) and Kiela and Clark (2014).

In the second task, we measure the **correlation** between distributional relatedness and native speaker judgments of semantic similarity or relatedness. Following previous studies (Baroni and Lenci, 2010; Padó and Lapata, 2007), performance in this task is quantified in terms of Pearson correlation.[3] Evaluated datasets are the **Rubenstein and Goodenough dataset** (RG65) of 65 noun pairs (Rubenstein and Goodenough, 1965), also evaluated by Kiela and Clark (2014), and the **WordSim-353 dataset** (WS353) of 353 noun pairs (Finkelstein et al., 2002), included in the study of Polajnar and Clark (2014).

The third evaluation task is **noun clustering**: distributional similarity between words is used to assign them to a pre-defined number of semantic classes. Performance in this task is quantified in terms of cluster purity. Clustering is performed with an algorithm based on partitioning around medoids (Kaufman and Rousseeuw, 1990, Ch. 2), using the

[2] The parameter set of Lapesa et al. (2014) fully corresponds to the one used in the present study.

[3] Some other evaluation studies adopt Spearman's rank correlation ρ, which is more appropriate if there is a non-linear relation between distributional relatedness and the human judgements. We computed both coefficients in our experiments and decided to report Pearson's r for three reasons: (i) Baroni and Lenci (2010) already list r scores for a wide range of DSMs in this task; (ii) in most experimental runs, ρ and r values were quite similar, with a tendency for ρ to be slightly lower then r (difference of means RG65: 0.001; WS353: 0.02); (iii) linear regression analyses for ρ and r showed the same trends and patterns for all DSM parameters.

R function pam with standard settings.[4] Evaluated datasets for the clustering task are the **Almuhareb-Poesio set** (henceforth, AP) containing 402 nouns grouped into 21 classes (Almuhareb, 2006); the **Battig set**, containing 83 concrete nouns grouped into 10 classes (Van Overschelde et al., 2004); the **ESSLLI 2008 set**, containing 44 concrete nouns grouped into 6 classes;[5] and the **Mitchell set**, containing 60 nouns grouped into 12 classes (Mitchell et al., 2008), also employed by Bullinaria and Levy (2012).

3.2 Parameters

DSMs evaluated in this paper belong to the class of window-based models. All models use the same large vocabulary of target words (27522 lemma types), which is based on the vocabulary of Distributional Memory (Baroni and Lenci, 2010) and has been extended to cover all items in our datasets. Distributional models were built using the UCS toolkit[6] and the wordspace package for R (Evert, 2014). The following parameters have been evaluated:[7]

- **Source Corpus** (abbreviated in the plots as *corpus*): the corpora from which we compiled our DSMs differ in both size and quality, and they represent standard choices in DSM evaluation. Evaluated corpora in this study are: British National Corpus[8]; ukWaC; WaCkypedia_EN[9];
- **Context window**:
  - **Direction*** (*win.direction*): we collected co-occurrence counts both using a *directed window* (i.e., separate co-occurrence counts for context words to the left and to the right of the target) and an *undirected window* (no distinction between left and right context);
  - **Size** (*win.size*)*†: we expect this parameter to be crucial as it determines the amount of shared context involved in the computation of similarity. We tested windows of 1, 2, 4, 8, and 16 words to the left and right of the target, limited by sentence boundaries;
- **Context selection**: Context words are filtered by part-of-speech (nouns, verbs, adjectives, and adverbs). From the full co-occurrence matrix, we further select dimensions (i.e., columns, corresponding to context words) according to the following two parameters:
  - **Criterion for context selection** (*criterion*): marginal frequency; number of nonzero co-occurrence counts;
  - **Threshold for context selection** (*context.dim*)*†: from the context dimensions ranked according to this criterion, we select the top 5000, 10000, 20000, 50000 or 100000 dimensions;
- **Score for feature weighting** (*score*)*†: we compare plain co-occurrence frequency to tf.idf and to the following association measures: Dice coefficient; simple log-likelihood; Mutual Information (MI); t-score; z-score;[10]
- **Feature transformation** (*transformation*): to reduce the skewness of feature scores, it is possible to apply a transformation function. We evaluate square root, sigmoid (tanh) and logarithmic transformation vs. no transformation.

[4] Other clustering studies have often been carried out using the CLUTO toolkit (Karypis, 2003) with standard settings, which corresponds to spectral clustering of the distributional vectors. Unlike pam, which operates on a pre-computed dissimilarity matrix, CLUTO cannot be used to test different distance measures or neighbor rank. Comparative clustering experiments showed no substantial differences for cosine similarity; in the rank-based setting, pam consistently outperformed CLUTO clustering.

[5] http://wordspace.collocations.de/doku.php/data:esslli2008:concrete_nouns_categorization

[6] http://www.collocations.de/software.html

[7] Parameters also evaluated by Bullinaria and Levy (2007; 2012), albeit with a different range of values, are marked with an asterisk (*); those evaluated by Kiela and Clark (2014) and/or Polajnar and Clark (2014) are marked with a dagger (†).

[8] http://www.natcorp.ox.ac.uk/

[9] Both ukWaC and WaCkypedia_EN are available from http://wacky.sslmit.unibo.it/doku.php?id=corpora.

[10] See Evert (2008) for a thorough description of the association measures and details on their calculation (Fig. 58.4 on p. 1225 and Fig. 58.9 on p. 1235). We selected these measures because they have widely been used in previous work on DSMs (tf.idf, MI and log-likelihood) or are popular choices for the identification of multiword expressions. Based on statistical hypothesis tests, log-likelihood, t-score and z-score measure the significance of association between a target and feature term; MI shows how much more frequently they co-occur than expected by chance; and Dice captures the mutual predictability of target and feature term. Note that we compute sparse versions of the association measures with negative values clamped to zero in order to preserve the sparseness of the co-occurrence matrix. For example, our MI measure corresponds to Positive MI in the other evaluation studies.

- **Distance metric** (*metric*)*†: cosine distance (i.e., angle between vectors); Manhattan distance[11];
- **Dimensionality reduction**: we optionally apply Singular Value Decomposition to 1000 dimensions, using randomized SVD (Halko et al., 2009) for performance reasons. For the SVD-based models, there are two additional parameters:
  - **Number of latent dimensions** (*red.dim*): out of the 1000 SVD dimensions, we select the first 100, 300, 500, 700, 900 dimensions (i.e. those with the largest singular values);
  - **Number of skipped dimensions** (*dim.skip*): when selecting the reduced dimensions, we exclude the first 0, 50 or 100 dimensions. This parameter has already been evaluated by Bullinaria and Levy (2012), who achieved best performance by discarding the initial components of the reduced matrix, i.e., those with the highest variance.
- **Index of distributional relatedness** (*rel.index*). Given two words *a* and *b* represented in a DSM, we consider two alternative ways of quantifying the degree of relatedness between *a* and *b*. The first option (and standard in DSM modeling) is to compute the *distance* (cosine or Manhattan) between the vectors of *a* and *b*. The alternative choice, proposed in this work, is based on *neighbor rank*. Neighbor rank has already been successfully used for capturing priming effects (Hare et al., 2009; Lapesa and Evert, 2013; Lapesa et al., 2014) and for quantifying the semantic relatedness between derivationally related words (Zeller et al., 2014); however, its performance on standard tasks has not been tested yet. For the TOEFL task, we compute *rank* as the position of the target among the nearest neighbors of each synonym candidate.[12] For the correlation and clustering tasks, we compute a symmetric *rank* measure as the average of $\log \text{rank}(a,b)$ and $\log \text{rank}(b,a)$. An exploration of the effects of directionality on the prediction of similarity ratings and its use in clustering tasks (i.e., experiments involving $\text{rank}(a,b)$ and $\text{rank}(b,a)$ as indexes of relatedness) is left for future work.

4 Model selection

As has already been pointed out in the introductory section, one of the main open issues in DSM evaluation is the need for a systematic investigation of the interactions between DSM parameters. Another issue that large-scale evaluation studies face is overfitting: if a large number of models (i.e. parameter combinations) is evaluated, it makes little sense to look at the best model (i.e. the best parameter combination), which will be subject to heavy overfitting, especially on small datasets such as TOEFL. The methodology for model selection applied in this work successfully addresses both issues.

In our evaluation study, we tested all possible combinations of the parameters described in section 3.2. This resulted in a total of 537600 model runs (33600 in the unreduced setting, 504000 in the dimensionality-reduced setting). The models were generated and evaluated on a large HPC cluster within approximately 5 weeks.

Following Lapesa and Evert (2013), DSM parameters are considered predictors of model performance: we analyze the influence of individual parameters and their interactions using general linear models with performance (accuracy, correlation, purity) as a dependent variable and the model parameters as independent variables, including all two-way interactions. More complex interactions are beyond the scope of this paper and are left for future work. Analysis of variance – which is straightforward for our full factorial design – is used to quantify the importance of each parameter or interaction. Robust optimal parameter settings are identified with the help of effect displays (Fox, 2003), which show the partial effect of one or two parameters by marginalizing over all other parameters. Unlike coefficient estimates, they allow an intuitive interpretation of the effect sizes of categorical variables irrespective of the dummy coding scheme used.

[11] In this study, the range of evaluated metrics is restricted to cosine vs. manhattan for a number of reasons: (i) cosine is considered a standard choice in DSM modeling and is adopted by most evaluation studies (Bullinaria and Levy, 2007; Bullinaria and Levy, 2012; Polajnar and Clark, 2014); (ii) for our normalized vectors, Euclidean distance is fully equivalent to cosine; (iii) preliminary experiments with the maximum distance measure resulted in very low performance.

[12] Note that using the positions of the synonym candidates among the neighbors of the target would have been equivalent to direct use of the distance measure, since the transformation from distance to rank is monotonic in this case.

5 Results

This section reports the results of the modeling experiments outlined in section 3. Table 1 summarizes the evaluation results: for each dataset, we report minimum, maximum and mean performance, comparing unreduced and reduced runs. The column *Difference of Means* shows the average difference in performance between an unreduced model and its reduced counterpart (with dimensionality reduction parameters set to the values of the general best setting identified in section 5.5) and the p-value[13] of a Wilcoxon signed rank test with continuity correction.It is evident that dimensionality reduction improves model performances for all datasets[14].

Dataset	Unreduced			Reduced			Difference of Means
	Min	Max	Mean	Min	Max	Mean	
TOEFL	25.0	87.5	63.9	18.7	98.7	64.4	−4.626***
RG65	0.01	0.88	0.59	0.00	0.89	0.63	−0.073***
WS353	0.00	0.73	0.39	0.00	0.73	0.43	−0.074***
AP	0.15	0.73	0.56	0.13	0.76	0.54	0.004n.s.
BATTIG	0.28	0.99	0.77	0.23	0.99	0.78	−0.037***
ESSLLI	0.32	0.93	0.72	0.32	0.98	0.72	−0.003*
MITCH.	0.26	0.97	0.68	0.27	0.97	0.69	−0.031***

Table 1: Summary of performance

While the improvements are only minimal in some cases, dimensionality reduction never has a detrimental effect while offering practical advantages in memory usage and computation speed. Therefore, in our analysis, we focus on the runs involving dimensionality reduction. In the following subsections, we present detailed results for each of the three tasks. In each case, we first discuss the impact of DSM parameters on performance, and then describe the optimal parameter values.

5.1 TOEFL

In the TOEFL task, the linear model achieves an adjusted R^2 of 89%, showing that it explains the influence of model parameters on TOEFL accuracy very well. Figure 1 displays the ranking of the evaluated parameters according to their importance in a *feature ablation* setting. The R^2 values in the plots refer to the proportion of variance explained by the respective parameter together with all its interactions, corresponding to the reduction in adjusted R^2 if this parameter is left out. We do not rely on significance values for model selection because, given the large number of measurements, virtually all parameters have a highly significant effect.

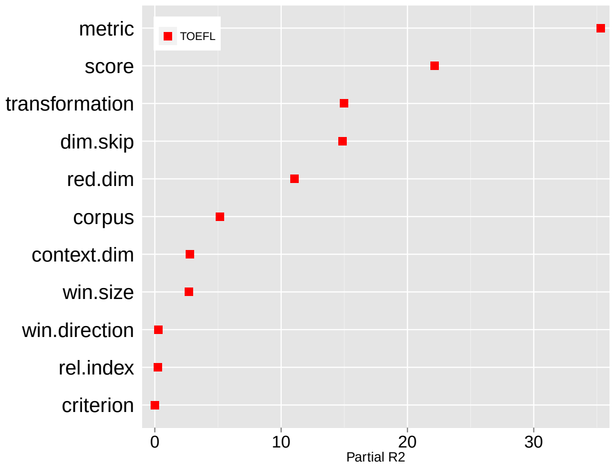

Figure 1: TOEFL, parameters and feature ablation

Table 2 reports all parameter interactions for the TOEFL task that explain more than 0.5% of the total variance (i.e. $R^2 \geq 0.5\%$), as well as the corresponding degrees of freedom (df) and R^2.

Interaction	df	R^2
score:transf	18	7.42
metric:dim.skip	2	4.44
score:metric	6	1.77
metric:context.dim	4	0.98
win.size:transf	12	0.91
corpus:score	12	0.84
score:context.dim	24	0.64
metric:red.dim	4	0.63

Table 2: TOEFL task: interactions, R^2

On the basis of their influence in determining model performance, we can identify three parameters that are crucial for the TOEFL task, and which will also turn out to be very influential in the other tasks at issue: *distance metric*, *feature score* and *feature transformation*.

The best *distance metric* is *cosine distance*: this is one of the consistent findings of our evaluation study and it is in accordance with Bullinaria and Levy (2007) and, to a lesser extent, Kiela and Clark (2014).[15] *Score* and *transformation* always have a fundamental impact on model perfor-

[13] $* = p < 0.05$; $*** = p < 0.001$; n.s. = not significant.

[14] Difference of means and Wilcoxon p-value on Spearman's *rho* for ratings datasets: RG65, −0.061***; WS353, −0.091***.

[15] In Kiela and Clark (2014), *cosine* is reported to be the best similarity metric, together with the *correlation similarity metric*

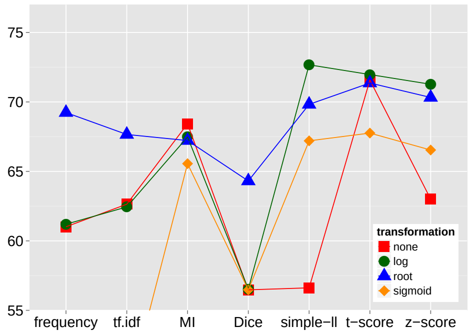

Figure 2: TOEFL, score / transformation

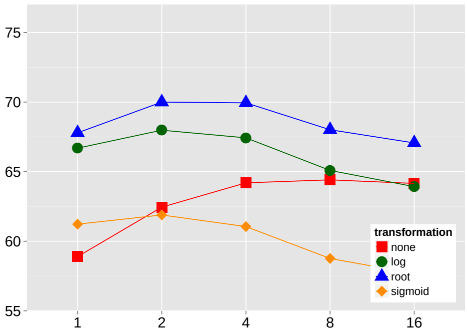

Figure 3: TOEFL, window size / transformation

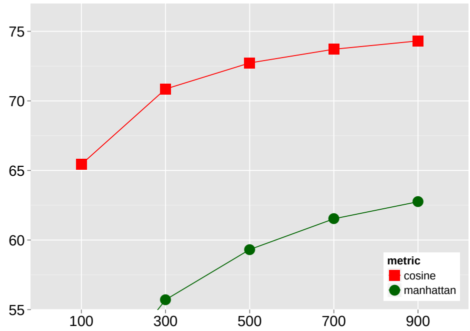

Figure 4: TOEFL, metric / n. of latent dim.

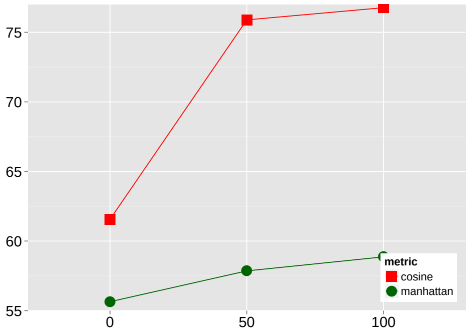

Figure 5: TOEFL, metric / n. of skipped dim.

mance: these parameters affect the distributional space independently of tasks and datasets. We will show that they are systematically involved in a strong interaction and that it is possible to identify a score/transformation combination with robust performance across all tasks. The interaction between *score* and *transformation* is displayed in figure 2. The best results are achieved by association measures based on significance tests (*simple-ll, t-score, z-score*), followed by *MI*. This result is in line with previous studies (Bullinaria and Levy, 2012; Kiela and Clark, 2014), which found Pointwise MI or Positive MI to be the best feature scores. The best choice, *simple-log likelihood*, exhibits a strong variation in performance across different transformations. For all three significance measures, the best *feature transformation* is consistently a *logarithmic* transformation. Raw co-occurrence *frequency*, *tf.idf* and *Dice* only perform well in combination with a square *root* transformation.

The best *window size*, as shown in figure 3, is a 2-word window for all evaluated transformations.

The SVD parameters (*number of latent dimensions* and *number of skipped dimensions*) play a significant role in determining model performance. They are particularly important for the TOEFL task, but we will see that their explanatory power is also quite strong in the other tasks. Interestingly, they show a tendency to participate in interactions with other parameters, but do not interact among themselves. We display the interaction between *metric* and *number of latent dimensions* in figure 4: the steep performance increase for both metrics shows that the widely-used choice of 300 latent dimensions (Landauer and Dumais, 1997) is suboptimal for the TOEFL task. The best value in our experiment is *900 latent dimensions*, and additional dimensions would probably lead to a further improvement. The interaction between *metric* and *number of skipped dimensions* is displayed in figure 5. While *manhattan* performs poorly no matter how many dimensions are skipped, *cosine* is positively affected by skipping 100 and (to a lesser extent) 50 dimensions. The latter trend has already been discussed by Bullinaria and Levy (2012).

Inspection of the remaining interaction plots, not shown here for reasons of space, reveals that the best

(a mean-adjusted version of cosine similarity). The latter, however, turned out to be more robust across different corpora and weighting schemes.

DSM performance in the TOEFL task is achieved by selecting *ukwac* as *corpus* and *10000 original dimensions*. The *index of distributional relatedness* has a very low explanatory power in the TOEFL task, with *neighbor rank* being the best choice (see plots 16 and 17 in section 5.4).

Given the minimal explanatory power of the *direction of the context window* and the *criterion for context selection* in all three tasks, we will not further consider these parameters in our analysis. We recommend to set them to an "unmarked" option: *undirected* and *frequency*.

The best setting identified by inspecting all effects is shown in table 5, together with its performance and with the performance of the (over-trained) best model in this task. Parameters of the latter are reported in appendix A.

5.2 Ratings

Figure 6 displays the importance of the evaluated parameters in the task of predicting similarity ratings. Parameters are ranked according to the average feature ablation R^2 values across both datasets (adj. R^2 of the full linear model: RG65: 86%; WS353: 90%).

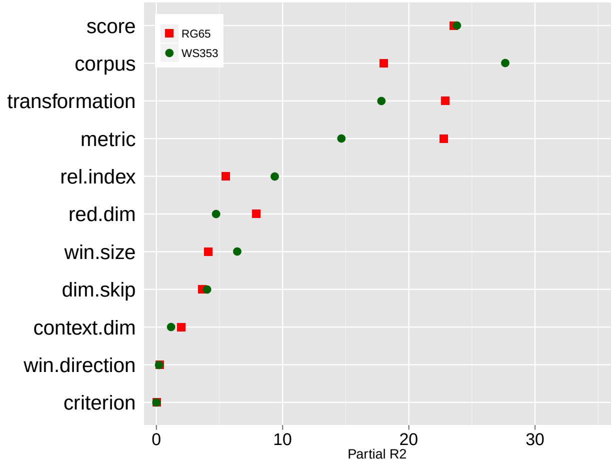

Figure 6: Ratings, parameters and feature ablation

Table 3 reports all interactions that explain more than 0.5% of the total variance in both datasets. For reasons of space, we only discuss the interactions and best parameter values on RG65; the corresponding plots for WS353 are shown only if there are substantial differences.

As already noted for the TOEFL task, *score* and *transformation* have a large explanatory power and they are involved in a strong interaction showing the same tendencies and optimal values already identified for TOEFL. For reasons of space, we do not elaborate on this interaction here.

Interaction	df	RG65	WS353
score:transf	18	10.28	8.66
metric:red.dim	4	2.18	1.42
score:metric	6	1.91	0.59
win.size:transf	12	1.43	1.01
corpus:metric	2	1.83	0.51
metric:context.dim	4	1.08	0.62
corpus:score	12	0.77	0.82
win.size:score	24	0.77	0.69
score:dim.skip	12	0.58	0.85

Table 3: Ratings datasets: interactions, R^2

The analysis of the main effects shows that for both datasets *WaCkypedia* is the best option as a *source corpus*, suggesting that this task benefits from a trade-off between quality and quantity (WaCkypedia being smaller and cleaner than ukWaC, but less balanced than the BNC).

Index of distributional relatedness plays a much more important role than for the TOEFL task, with *neighbor rank* clearly outperforming *distance* (see figures 16 and 17 and the discussion in section 5.4 for more details).

The choice of the optimal *window size* depends on *transformation*: on the RG65 dataset, figure 7 shows that for a *logarithmic* transformation – which we already identified as the best *transformation* in combination with significance association measures – the highest performance is achieved with a *4 word window*. The corresponding effect display for WS353 (figure 8) suggests that a further small improvement may be obtained with an *8 word window* in this case. One possible explanation for this observation is the different composition of the WS353 dataset, which includes examples of semantic relatedness beyond attributional similarity. The 4 word window is a robust choice across both datasets, though.

The *number of latent dimensions* is involved in a strong interaction with the distance *metric* (figure 9). Best results are achieved with the *cosine metric* and at least *300 latent dimensions*, as well as 50 *skipped dimensions*. The interaction plot between *metric* and *number of original dimensions* in figure 10 shows that *50000 context dimensions* are sufficient for good performance, and no further improvement can be expected from even higher-dimensional spaces.

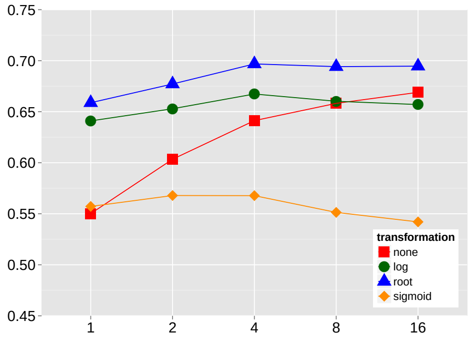

Figure 7: RG65, window size / transformation

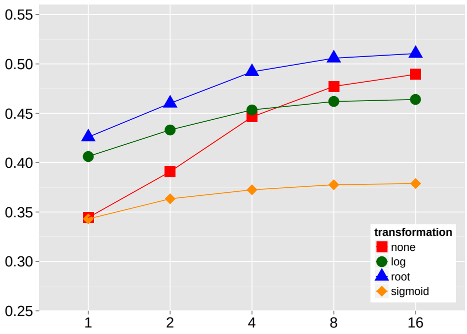

Figure 8: WS353, window size / transformation

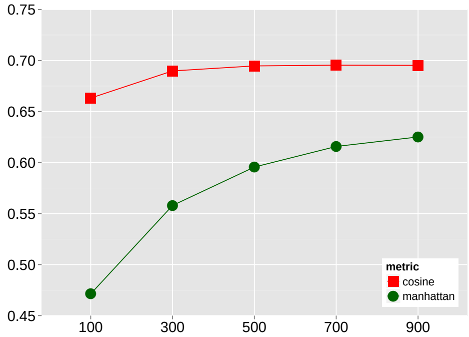

Figure 9: RG65, metric / n. latent dim.

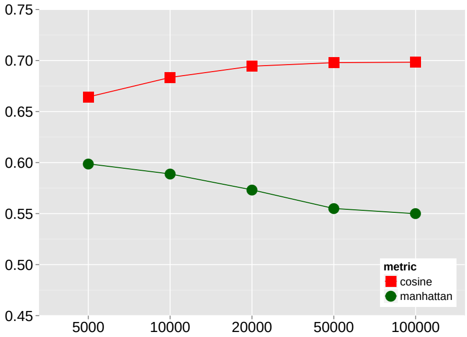

Figure 10: RG65, metric / n. context dimensions

Best settings for both datasets are summarized in table 5. Refer to appendix A for best models.

5.3 Clustering

Figure 11 displays the importance of the evaluated parameters in the clustering task (adj. R^2 of the full linear model: AP: 82%; BATTIG: 77%; ESSLLI: 58%; MITCHELL: 73%). Parameter ranking is determined by the average of the feature ablation R^2 values over all four datasets.

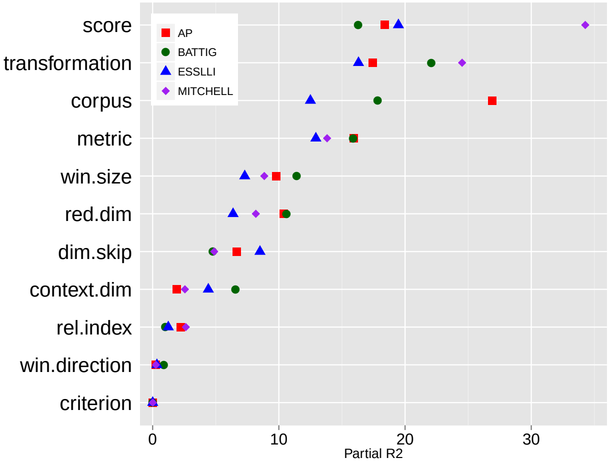

Figure 11: Clustering, parameters and feat. ablation

Interaction	df	AP	BATTIG	ESSLLI	MITCHELL
score:transf	18	7.10	7.95	7.56	11.42
metric:red.dim	4	3.29	3.16	2.03	2.03
win.size:metric	4	2.22	1.26	2.97	2.72
win.size:transf	12	2.00	2.95	0.88	2.66
corpus:metric	2	1.42	2.91	2.79	1.11
metric:dim.skip	2	2.25	1.54	2.77	0.86
corpus:win.size	8	2.36	1.18	1.49	1.23
score:dim.skip	12	0.56	1.15	0.99	1.39
win.size:score	24	0.74	0.77	0.54	0.65

Table 4: Clustering task: interactions, R^2

Table 4 reports all parameter interactions that explain more than 0.5% of the total variance for each of the four datasets.

In the following discussion, we focus on the AP dataset, which is larger and thus more reliable than the other three datasets. We mention remarkable differences between the datasets in terms of best parameter values. For a full overview of the best parameter setting for each dataset, see table 5.

As already discussed for TOEFL and the ratings task, we find *score* and *transformation* at the top of the feature ablation ranking. Table 4 confirms that the two parameters are involved in a strong interaction. The interaction plot (figure 12) shows the behavior we are already familiar with: significance

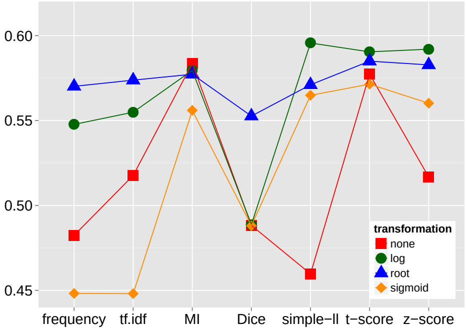

Figure 12: AP, score / transformation

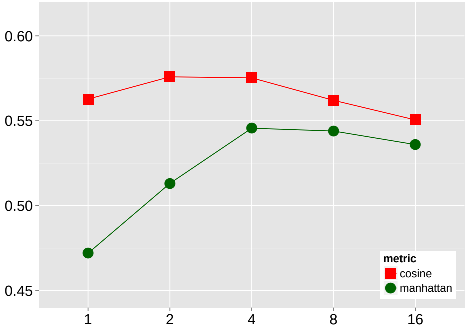

Figure 13: AP, window size / metric

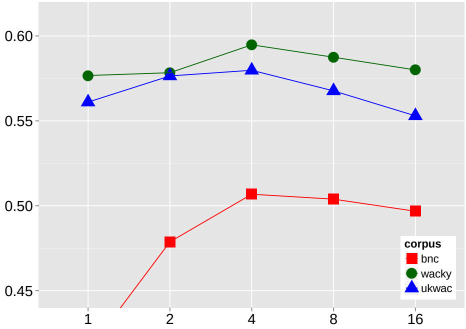

Figure 14: AP, corpus / window size

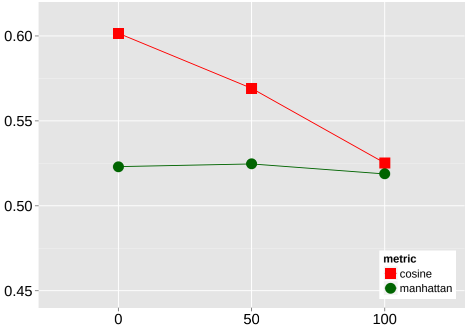

Figure 15: AP, metric / n. of skipped dim.

measures (*simple-ll, t-score and z-score*) reach the best performance in combination with *log transformation*: this combination is a robust choice also for the other datasets, with minor differences that can be observed in table 5 .

The interaction between *window size* and *metric* is displayed in figure 13: best performance is achieved with a *2 or 4 word window* in combination with *cosine distance*. Results on the other datasets suggest a preference for the *4 word window*. This is confirmed by interaction plots with *source corpus* (figure 14), which also reveal that *WaCkypedia* is again the best compromise between size and quality.

A very clear picture concerning the *number of skipped dimensions* emerges from figure 15 and is the same for all datasets: skipping dimensions is not necessary to achieve good performance (even though skipping 50 dimensions turned out at least to be not detrimental for BATTIG and MITCHELL).

Further effect displays, not shown here for reasons of space, suggest that *300 or 500 latent dimensions* – with some variation across the datasets (cf. table 5) – and a medium-sized co-occurrence matrix (*20000 or 50000 dimensions*) are needed to achieve good performance. *Neighbor rank* is the best choice as *index of distributional relatedness* (see section 5.4). See appendix A for best models.

5.4 Relatedness index

A novel contribution of our work is the systematic evaluation of a parameter that has received little attention in DSM research so far, and only in studies limited to a narrow choice of datasets (Lapesa and Evert, 2013; Lapesa et al., 2014; Zeller et al., 2014): the *index of distributional relatedness*.

The aim of this section is to provide a full overview of the impact of this parameter in our experiments. Despite the main focus of the paper on the reduced setting, in this section we also show results from the unreduced setting, for two reasons: first, since this parameter is relatively novel and evaluated here for the first time on standard tasks, we consider it necessary to provide a full picture concerning its behavior; second, *relatedness index* turned out to be much more influential in the unreduced setting than in the reduced one.

Figure 16 and 17 display the partial effect of *relatedness index* for each dataset, in the unreduced and reduced setting respectively. To allow for a comparison between the different measures of perfor-

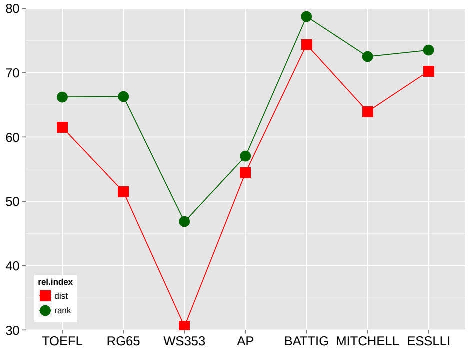

Figure 16: Unreduced Setting

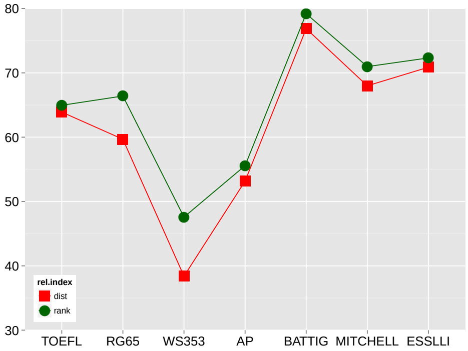

Figure 17: Reduced Setting

mance, correlation and purity values have been converted to percentages. The picture emerging from the two plots is very clear: *neighbor rank* is the best choice for both settings across all seven datasets. The degree of improvement over vector distance, however, shows considerable variation between different datasets. The rating task benefits the most from the use of neighbor rank.

On the other hand, *neighbor rank* has very little effect for the TOEFL task in a reduced setting, where its high computational complexity is clearly not justified; the improvement on the AP clustering dataset is also fairly small. While the TOEFL result seems to contradict the substantial improvement of *neighbor rank* found by Lapesa and Evert (2013) for a multiple-choice task based on stimuli from priming experiments, there were only two choices (consistent and inconsistent prime) in this case rather than four. We do not rule out that a more refined use of the rank information (for example, different strategies for rank combinations) may produce better results on the TOEFL and AP datasets.

As discussed in section 3.2, we have not yet explored the potential of neighbor rank in modeling directionality effects in semantic similarity. Unlike Lapesa and Evert (2013), who adopt four different indexes of distributional relatedness (vector distance; forward rank, i.e., rank of the target in the neighbors of the prime; backward rank, i.e, rank of the prime in the neighbors of the target; average of backward and forward rank), we used only a single rank-based index (cf. section 3.2), mostly for reasons of computational complexity. We consider the results of this study more than encouraging, and expect further improvements from a full exploration of directionality effects in the tasks at issue.

5.5 Best settings

We conclude the result overview by evaluating the best parameter combinations identified for each task and data set, showing how well our approach to model selection works in practice.

Table 5 summarizes the optimal parameter settings identified for each task and compares the performance of this model (*B.set* = best setting) with the over-trained best run in the experiment (*B.run* = best run).[16] In most cases, the result of our robust parameter optimization is close to the best run. The only exception is the ESSLLI dataset, which is smaller than the other datasets and particularly susceptible to over-training (cf. the low R^2 of the regression analysis in section 5.3). Table 5 also reports the current state of the art for each task (*SoA* = *state-of-the-art*), taken from the ACL wiki[17] where available (TOEFL and similarity ratings), from Baroni and Lenci (2010) for the clustering tasks, and from more recent studies of which we are aware. Our results are comparable to the state of the art, even though the latter includes a much broader range of approaches than our window-based DSMs. In one case (BATTIG), our optimized model even improves on the best previous result.

A side-by-side inspection of the main effects and interaction plots for different data sets allowed us to identify parameter settings that are robust *across datasets* and even *across tasks*. Table 6 shows recommended settings for each task (independent of the

[16] Abbreviations in the table: *win* = window size; *c.dim* = number of context dimensions; *tr* = transformation; *red.dim* = number of latent dimensions; *d.sk*= number of skipped dimensions; *r.ind* = relatedness index; Parameter values: *s-ll* = simple-ll; *t-sc* = t-score; *cos* = cosine; *man* = manhattan.

[17] http://aclweb.org/aclwiki

Dataset	corpus	win	c.dim	score	tr	metric	r.ind	red.dim	d.sk	B.set	B.run	SoA	Reference
TOEFL	ukwac	2	10k	s-ll	log	cos	rank	900	100	92.5	98.7	100.0	Bullinaria and Levy (2012)
RG65	wacky	4	50k	s-ll	log	cos	rank	500	50	0.87	0.89	0.86	Hassan and Mihalcea (2011)[18]
WS353	wacky	8	50k	s-ll	log	cos	rank	300	50	0.68	0.73	0.81	Halawi et al. (2012)[19]
AP	wacky	4	20k	s-ll	log	cos	rank	300	0	0.69	0.76	0.79	Rotenhäusler and Schütze (2009)
BATTIG	wacky	8	50k	s-ll	log	cos	rank	500	0	0.98	0.99	0.96	Baroni and Lenci (2010)
ESSLLI	wacky	2	20k	t-sc	log	cos	rank	300	0	0.77	0.98	0.91	Katrenko, ESSLLI workshop[20]
MITCHELL	wacky	4	50k	s-ll	log	cos	rank	500	0	0.88	0.97	0.94	Bullinaria and Levy (2012)

common for all datasets: window direction = undirected; criterion for context selection = frequency

Table 5: Best Settings

particular dataset) and a more general setting that achieves good performance in all three tasks. Evaluation results for these settings on each dataset are reported in table 7. In most cases, the general model is close to the performance of the task- and dataset-specific settings. Our robust evaluation methodology has enabled us to find a good trade-off between portability and performance.

Task	corpus	win	c.dim	score	tr	metric	r.ind	red.dim	d.sk
TOEFL	ukwac	2	10k	s-ll	log	cos	rank	900	100
Rating	wacky	4	50k	s-ll	log	cos	rank	300	50
Clustering	wacky	4	50k	s-ll	log	cos	rank	500	0
General	**wacky**	**4**	**50k**	**s-ll**	**log**	**cos**	**rank**	**500**	**50**

Table 6: General Best Settings

Dataset	TOEFL	RATINGS	CLUSTERING	GENERAL
TOEFL	92.5	85.0	75.0	90.0
RG65	0.84	0.86	0.84	0.87
WS353	0.62	0.67	0.64	0.68
AP	0.62	0.66	0.67	0.67
BATTIG	0.87	0.91	0.98	0.90
ESSLLI	0.66	0.77	0.80	0.77
MITCHELL	0.75	0.83	0.88	0.83

Table 7: General best Settings – Performance

6 Conclusion

In this paper, we reported the results of a large-scale evaluation of window-based Distributional Semantic Models, involving a wide range of parameters and tasks. Our model selection methodology is robust to overfitting and sensitive to parameter interactions. It allowed us to identify parameter configurations that perform well across different datasets within the same task, and even across different tasks. We recommend the setting highlighted in bold font in table 5 as a general-purpose DSM for future research. We believe that many applications of DSMs (e.g. vector compositon) will benefit from using such a parameter combination that achieves robust performance in a variety of semantic tasks. Moreover, an extensive evaluation based on a robust methodology like the one presented here is the first necessary step for further comparisons of bag-of-words DSMs to different techniques for modeling word meaning, such as neural embeddings (Mikolov et al., 2013). Let us now summarize our main findings.

- Our experiments show that a cluster of three parameters, namely *score, transformation* and *distance metric*, plays a consistently crucial role in determining DSM performance. These parameters also show a homogeneous behavior across tasks and datasets with respect to best parameter values: *simple-ll, log transformation* and *cosine distance*. These tendencies confirm the results in Polajnar and Clark (2014) and Kiela and Clark (2014). In particular, the finding that sparse association measures (with negative values clamped to zero) achieve the best performance can be connected to the positive impact of context selection highlighted by Polajnar and Clark (2014): ongoing work targets a more specific analysis of their "thinning" effect on distributional vectors.
- Another group of parameters (*corpus, window size, dimensionality reduction parameters*) is also influential in all tasks, but shows more variation wrt. the best parameter values. Except for the TOEFL task, best results are obtained with the *WaCkypedia* corpus, confirming the observation of Sridharan and Murphy (2012) that corpus qual-

[18] The ACL wiki lists the hybrid model of Yih and Qazvinian (2012) as the best model on RG65 with $\rho = 0.89$, but does not specify its Pearson correlation r. In our comparison table, we show the best Pearson correlation, achieved by Hassan and Mihalcea (2011), which is also the best corpus-based model.

[19] Halawi et al. (2012) report Spearman's ρ. The ρ values for our best setting are: RG65: 0.85, WS353: 0.70; best setting for the ratings task: RG65: 0.82, WS353: 0.67; best general setting: RG65: 0.87, WS353: 0.70.

[20] http://wordspace.collocations.de/

ity compensates for size to some extent. *Window size* and *dimensionality reduction* show a more task-specific behavior, even though it is possible to find a good compromise in a *4 word window*, a reduced space of *500 dimensions* and *skipping of the first 50* dimensions. The latter result confirms the findings of Bullinaria and Levy (2007; 2012) in their clustering experiments.

- The *number of context dimensions* turned out to be less crucial. While very high-dimensional spaces usually result in better performance, the increase beyond *20000 or 50000 dimensions* is rarely sufficient to justify the increased processing cost.
- A novel contribution of our work is the systematic evaluation of a parameter that has been given little attention in DSM research so far: the *index of distributional relatedness*. Our results show that, even if the parameter is not among the most influential ones, *neighbor rank* consistently outperforms *distance*. Without SVD dimensionality reduction, the difference is more pronounced: this result is particularly interesting for compositionality tasks, where SVD has been reported to be detrimental (Baroni and Zamparelli, 2010). In such cases, the benefits of using *neighbor rank* clearly outweigh the increased (but manageable) computational complexity.

Ongoing work focuses on the extension of the evaluation setting to further parameters (e.g., new distance metrics and association scores, Caron's (2001) exponent *p*) and tasks (e.g., compositionality tasks, meaning in context), as well as the evaluation of dependency-based models. We are also working on a refined model selection methodology involving a systematic analysis of three-way interactions and the exclusion of inferior parameter values (such as Manhattan distance, sigmoid transformation and Dice score), which may have a confounding effect on some of the effect displays.

Appendix A: Best models

This appendix reports the best runs for every dataset.[21]

corpus	w	dir	e	c.dim	score	tr	m	r.ind	red.dim	d.sk	acc
ukwac	2	undir	f	5000	MI	none	cos	rank	900	100	98.75
ukwac	4	dir	f	50000	t-score	log	cos	rank	900	100	98.75
ukwac	4	undir	f	50000	t-score	root	cos	dist	900	100	98.75
ukwac	4	dir	f	5000	simple-ll	log	cos	dist	900	100	98.75

Table 8: TOEFL dataset – 23 models tied for best result (4 hand-picked examples shown)

corpus	w	dir	e	c.dim	score	tr	m	r.ind	red.dim	d.sk	cor
ukwac	16	undir	nz	20000	MI	none	cos	rank	700	100	0.89
ukwac	8	dir	f	20000	MI	none	cos	rank	700	100	0.89
wacky	4	dir	nz	50000	simple-ll	log	cos	rank	700	50	0.89
wacky	4	undir	f	100000	z-score	log	cos	rank	900	50	0.89

Table 9: Ratings, RG65 dataset – 19 models tied for best result (4 hand-picked examples shown)

corpus	w	dir	e	c.dim	score	tr	m	r.ind	red.dim	d.sk	cor
wacky	16	dir	f	5000	MI	none	man	rank	900	50	0.73
wacky	16	undir	f	5000	MI	none	man	rank	900	50	0.72
wacky	16	undir	f	5000	z-score	log	man	rank	900	50	0.72
wacky	16	dir	f	10000	z-score	root	man	rank	900	50	0.72

Table 10: Ratings, WordSim353 dataset – best model (3 additional hand-picked models with similar performance are shown)

corpus	w	dir	e	c.dim	score	tr	m	r.ind	red.dim	d.sk	pur
ukwac	4	dir	nz	10000	t-score	log	man	rank	900	50	0.76
wacky	1	dir	nz	10000	z-score	log	man	rank	900	50	0.75
wacky	1	undir	f	20000	simple-ll	log	man	rank	900	50	0.75
wacky	2	dir	f	100000	z-score	log	cos	rank	500	0	0.75

Table 11: Clustering, Almuhareb-Poesio dataset – best model (plus 3 additional hand-picked models)

corpus	w	dir	e	c.dim	score	tr	m	r.ind	red.dim	d.sk	pur
ukwac	1	undir	f	20000	Dice	root	man	rank	300	100	0.99
ukwac	2	undir	f	100000	freq	log	cos	dist	300	50	0.99
wacky	16	undir	f	50000	z-score	log	man	dist	500	50	0.99
wacky	8	undir	f	10000	Dice	root	man	rank	500	0	0.99

Table 12: Clustering, Battig dataset – 1037 models tied for best result (4 hand-picked examples shown)

corpus	w	dir	e	c.dim	score	tr	m	r.ind	red.dim	d.sk	pur
wacky	16	dir	nz	50000	z-score	none	man	dist	900	0	0.98
ukwac	1	dir	nz	100000	simple-ll	log	cos	dist	100	50	0.95
ukwac	2	undir	f	50000	tf.idf	none	man	dist	700	0	0.95
wacky	8	undir	f	100000	tf.idf	root	man	rank	500	0	0.95

Table 13: Clustering, ESSLLI dataset – best model (plus 3 additional hand-picked models)

corpus	w	dir	e	c.dim	score	tr	m	r.ind	red.dim	d.sk	pur
bnc	2	undir	nz	100000	simple-ll	log	cos	rank	900	0	0.97
bnc	2	undir	f	50000	simple-ll	log	cos	rank	700	0	0.97
bnc	2	undir	nz	50000	simple-ll	log	cos	rank	900	0	0.97

Table 14: Clustering, Mitchell dataset – 3 models tied for best result

[21] Some abbreviations are different from tables 5 and 6. Parameters: *w* = window; *dir* = direction; *e* = exclusion criterion for context selection; *m* = metric. Performance: *acc* = accuracy; *cor* = correlation; *pur* = purity. Parameter values: *dir* = directed; *undir* = undirected; *f* = frequency; *nz* = non-zero.

Acknowledgments

We are grateful to the editor and the anonymous reviewers, whose comments helped us improve the paper. We would like to thank Sabine Schulte im Walde and the SemRel group at the IMS Stuttgart, the Corpus Linguistics group at the University of Erlangen-Nürnberg and the Computational Linguistics group at the IKW Osnabrück for their feedback on our work.

Gabriella Lapesa's PhD research was funded by the DFG Collaborative Research Centre SFB 732 at IMS Stuttgart, where she conducted a large part of the work presented in this paper.

References

Abdulrahman Almuhareb. 2006. *Attributes in Lexical Acquisition*. Ph.D. thesis, University of Essex.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):1–49.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, MIT, Massachusetts, USA.

Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology (LiLT)*, 9(6):5–109.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

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

John A. Bullinaria and Joseph P. Levy. 2013. Limiting factors for mapping corpus-based semantic representations to brain activity. *PLoS ONE*, 8(3):1–12.

John Caron. 2001. Experiments with LSA scoring: Optimal rank and basis. In Michael W. Berry, editor, *Computational Information Retrieval*, pages 157–169. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, chapter 58. Mouton de Gruyter, Berlin, New York.

Stefan Evert. 2014. Distributional semantics in R with the wordspace package. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 110–114, Dublin, Ireland.

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.

John Fox. 2003. Effect displays in R for generalised linear models. *Journal of Statistical Software*, 8(15):1–27.

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, New York, NY, USA.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. 2009. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. Technical Report 2009-05, ACM, California Institute of Technology.

Mary Hare, Michael Jones, Caroline Thomson, Sarah Kelly, and Ken McRae. 2009. Activating event knowledge. *Cognition*, 111(2):151–167.

Zelig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Samer Hassan and Rada Mihalcea. 2011. Semantic relatedness using salient semantic analysis. In *Proceedings of the Twenty-fifth AAAI Conference on Artificial Intelligence*, pages 884 – 889, San Francisco, California.

George Karypis. 2003. CLUTO: A clustering toolkit (release 2.1.1). Technical Report 02-017, Minneapolis: University of Minnesota, Department of Computer Science.

Leonard Kaufman and Peter J. Rousseeuw. 1990. *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons.

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

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Gabriella Lapesa and Stefan Evert. 2013. Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2013)*, pages 66–74, Sofia, Bulgaria.

Gabriella Lapesa, Stefan Evert, and Sabine Schulte im Walde. 2014. Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 160–170, Dublin, Ireland.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, pages 768–774, Montreal, Quebec, Canada.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation and Computers*, 28:203–208.

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

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio.

Tom Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - SemEval '12*, pages 114–123.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Tamara Polajnar and Stephen Clark. 2014. Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 230–238, Gothenburg, Sweden.

Klaus Rothenhäusler and Hinrich Schütze. 2009. Unsupervised classification with dependency based word spaces. In *Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics*, pages 17–24, Athens, Greece.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627—633.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, University of Stockolm.

Gerard Salton, Andrew Wong, and ChungShu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 27(1):97–123.

Seshadri Sridharan and Brian Murphy. 2012. Modeling word meaning: Distributional semantics and the corpus quality-quantity trade-off. In *Proceedings of the 3rd workshop on Cognitive Aspects of the Lexicon (CogAlex-III)*, pages 53–68, Mumbai, India.

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

James Van Overschelde, Katherine Rawson, and John Dunlosky. 2004. Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, 50:289–335.

Wen-tau Yih and Vahed Qazvinian. 2012. Measuring word relatedness using heterogeneous vector space models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12)*, pages 616–620, Montreal, Canada.

Britta Zeller, Sebastian Padó, and Jan Šnajder. 2014. Towards semantic validation of a derivational lexicon. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1728–1739, Dublin, Ireland.