Evaluating Neighbor Rank and Distance Measures as Predictors of Semantic Priming

Gabriella Lapesa
Universität Osnabrück
Institut für Kognitionswissenschaft
Albrechtstr. 28, 49069 Osnabrück
glapesa@uos.de

Stefan Evert
FAU Erlangen-Nürnberg
Professur für Korpuslinguistik
Bismarckstr. 6, 91054 Erlangen
severt@fau.de

Abstract

This paper summarizes the results of a large-scale evaluation study of bag-of-words distributional models on behavioral data from three semantic priming experiments. The tasks at issue are (i) identification of consistent primes based on their semantic relatedness to the target and (ii) correlation of semantic relatedness with latency times. We also provide an evaluation of the impact of specific model parameters on the prediction of priming. To the best of our knowledge, this is the first systematic evaluation of a wide range of DSM parameters in all possible combinations. An important result of the study is that neighbor rank performs better than distance measures in predicting semantic priming.

1 Introduction

Language production and understanding make extensive and immediate use of world knowledge information that concerns prototypical events. Plenty of experimental evidence has been gathered to support this claim (see McRae and Matzuki, 2009, for an overview). Specifically, a number of priming studies have been conducted to demonstrate that event knowledge is responsible for facilitation of processing of words that denote events and their participants (Ferretti et al., 2001; McRae et al., 2005; Hare et al., 2009). The aim of our research is to investigate to which extent such event knowledge surfaces in linguistic distribution and can thus be captured by Distributional Semantic Models (henceforth, DSMs). In particular, we test the capabilities of bag-of-words DSMs in simulating priming data from the three aforementioned studies.

DSMs have already proven successful in simulating priming effects (Padó and Lapata, 2007; Herdağdelen et al., 2009; McDonald and Brew, 2004). Therefore, in this work, we aim at a more specific contribution to the study of distributional modeling of priming: to identify the *indexes of distributional relatedness* that produce the best performance in simulating priming data and to assess the impact of specific model parameters on such performance. In addition to *distance in the semantic space*, traditionally used as an index of distributional relatedness in DSMs, we also introduce *neighbor rank* as a predictor of priming effects. Distance and a number of rank-based measures are compared with respect to their performance in two tasks: the identification of congruent primes on the basis of distributional relatedness to the targets (we measure accuracy in picking up the congruent prime) and the prediction of latency times (we measure correlation between distributional relatedness and reaction times). The results of our experiments show that neighbor rank is a better predictor than distance for priming data.

Our approach to DSM evaluation constitutes a methodological contribution of this study: we use linear models with performance (accuracy or correlation) as a dependent variable and various model parameters as independent variables, instead of looking for optimal parameter combinations. This approach is robust to overfitting and allows to analyze the influence of individual parameters as well as their interactions.

The paper is structured as follows. Section 2 provides an overview of the modeled datasets. Section 3 introduces model parameters and indexes of distributional relatedness evaluated in this paper, describes the experimental tasks and outlines our statistical approach to DSM evaluation. Section 4 presents results for the accuracy and correlation tasks and evaluates the impact of model parameters on performance. We conclude in section 5 by sketching ongoing work and future developments of our research.

Dataset	Relation	N	Prime$_c$	Prime$_i$	Target	Fac
V-N	AGENT	28	Pay	Govern	Customer	27*
	PATIENT	18	Invite	Arrest	Guest	32*
	PATIENT FEATURE	20	Comfort	Hire	Upset	33*
	INSTRUMENT	26	Cut	Dust	Rag	32*
	LOCATION	24	Confess	Dance	Court	- 5
N-V	AGENT	30	Reporter	Carpenter	Interview	18*
	PATIENT	30	Bottle	Ball	Recycle	22*
	INSTRUMENT	32	Chainsaw	Detergent	Cut	16*
	LOCATION	24	Beach	Pub	Tan	18*
N-N	EVENT-PEOPLE	18	Trial	War	Judge	32*
	EVENT-THING	26	War	Gun	Banquet	33*
	LOCATION-LIVING	24	Church	Gym	Athlete	37*
	LOCATION-THING	30	Pool	Garage	Car	29*
	PEOPLE-INSTRUMENT	24	Hiker	Barber	Compass	45*
	INSTRUMENT-PEOPLE	24	Razor	Compass	Barber	-10
	INSTRUMENT-THING	24	Hair	Scissors	Oven	58*

Table 1: Overview of datasets: thematic relations, number of triples, example stimuli, facilitation effects

2 Data

This section introduces the priming datasets which are the object of the present study. All the experiments we aim to model were conducted to provide evidence for the immediate effect of event knowledge in language processing.

The first dataset comes from Ferretti et al. (2001), who found that verbs facilitate the processing of nouns denoting prototypical participants in the depicted event and of adjectives denoting features of prototypical participants. In what follows, the dataset from this study will be referred to as **V-N dataset**.

The second dataset comes from McRae et al. (2005). In this experiment, nouns were found to facilitate the processing of verbs denoting events in which they are prototypical participants. In this paper, this dataset is referred to as **N-V dataset**.

The third dataset comes from Hare et al. (2009), who found a facilitation effect from nouns to nouns denoting events or their participants. We will refer to this dataset as **N-N dataset**.

Experimental items and behavioral data from these three experiments have been pooled together in a global dataset that contains 404 word **triples** (Target, Congruent Prime, Incongruent Prime). For every triple, the dataset contains mean reaction times for the congruent and incongruent conditions, and a label for the thematic relation involved. Table 1 provides a summary of the experimental data. It specifies the number of triples for every relation in the datasets (*N*) and gives an example triple ($Prime_{congruent}$, $Prime_{incongruent}$, *Target*). Facilitation effects and stars marking significance by participants and items reported in the original studies are also specified for every relation (*Fac*). Relations for which the experiments showed no priming effect are highlighted in bold.

3 Method

3.1 Models

Building on the Distributional Hypothesis (Harris, 1954), DSMs are employed to produce semantic representations of words from patterns of co-occurrence in texts or documents (Sahlgren, 2006; Turney and Pantel, 2010). Semantic representations in the form of distributional vectors are compared to quantify the amount of shared contexts as an empirical correlate of semantic similarity. For the purposes of this study, similarity is understood in terms of topical relatedness (words connected to a particular situation) rather than attributional similarity (synonyms and near-synonyms).

DSMs evaluated in this study belong to the class of bag-of-words models: the distributional vector of a target word consists of co-occurrence counts with other words, resulting in a word-word co-occurrence matrix. The models cover a large vocabulary of target words (27668 words in the untagged version; 31713 words in the part-of-speech tagged version). It contains the stimuli from the datasets described in section 2 and further target words from state-of-the-art evaluation studies (Baroni and Lenci, 2010; Baroni and Lenci, 2011; Mitchell and Lapata, 2008). Contexts are filtered by part-of-speech (nouns, verbs, adjectives, and adverbs) and by frequency thresholds. Neither syntax nor word order were taken into account when gathering co-occurrence information. Distributional models were built using the UCS

toolkit[1] and the `wordspace` package for R[2]. The evaluated parameters are:

- **Corpus**: British National Corpus[3]; ukWaC[4]; WaCkypedia_EN[5]; WP500[6]; and a concatenation of BNC, ukWaC, and WaCkypedia_EN (called the *joint corpus*);
- **Window size**: 2, 5, or 15 words to the left and to the right of the target;
- **Part of speech**: no part of speech tags; part of speech tags for targets; part of speech tags for targets and contexts;
- **Scoring measure**: frequency; Dice coefficient; simple log-likelihood; Mutual Information; t-score; z-score;[7]
- **Vector transformation**: no transformation; square root, sigmoid or logarithmic transformation;
- **Dimensionality reduction**: no dimensionality reduction; Singular Value Decomposition to 300 dimensions using randomized SVD (Halko et al., 2009); Random Indexing (Sahlgren, 2005) to 1000 dimensions;
- **Distance measure**: cosine, euclidean or manhattan distance.

3.2 Indexes of Distributional Relatedess

3.2.1 Distance and Rank

The indexes of distributional relatedness described in this section represent alternative perspectives on the semantic representation inferred by DSMs from co-occurrence data.

Given a *target*, a *prime*, and a matrix of distances produced by a distributional model, we test the following indexes of relatedness between *target* and *prime*:

- **Distance**: distance between the vectors of *target* and *prime* in the semantic space;
- **Backward association**: rank of *prime* among the neighbors of *target*, as in Hare et al. (2009);[8]
- **Forward association**: rank of *target* in the neighbors of *prime*;
- **Average rank**: average of backward and forward association.

Indexes of distributional relatedness were considered as an additional parameter in the evaluation, labeled **relatedness index** below. Every combination of the parameters described in section 3.1 with each value of the *relatedness index* parameter defines a DSM. The total number of models evaluated in our study amounts to 38880.

3.2.2 Motivation for Rank

This section provides some motivation for the use of neighbor rank as a predictor of priming effects in DSMs, on the basis of general cognitive principles and of previous modeling experiments.

In distributional semantic modeling, similarity between words is calculated according to Euclidean geometry: the more similar two words are, the closer they are in the semantic space. One of the axioms of spatial models is *symmetry* (Tversky, 1977): the distance between point *a* and point *b* is equal to the distance between point *b* and point *a*. Cognitive processes, however, often violate the symmetry axiom. For example, asymmetric associations are often found in word association norms (Griffiths et al., 2007).

Our study also contains a case of asymmetry. In particular, the results from Hare et al. (2009), which constitute our N-N dataset, show priming from PEOPLE to INSTRUMENTs, but not from INSTRUMENTs to PEOPLE. This asymmetry cannot be captured by distance measures for reasons stated above. However, the use of rank-based indexes allows to overcome the limitation of symmetrical distance measures by introducing directionality (in our case, target $\rightarrow$ prime vs. prime $\rightarrow$ target), and this without discarding the established and proven measures.

Rank has already proven successful in modeling priming effects with DSMs. Hare et al. (2009) conducted a simulation on the N-N dataset using LSA (Landauer and Dumais, 1997) and BEAGLE (Jones and Mewhort, 2007) trained on the TASA corpus. Asymmetric priming was correctly predicted by the context-only version of BEAGLE using rank (namely, rank of prime among neighbors of target, cf. backward rank in section 3.2.1).

Our study extends the approach of Hare et al. (2009) in a number of directions. First, we introduce and evaluate several different rank-based measures (section 3.2.1). Second, we evaluate rank in connection with specific parameters and on

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

[2] http://r-forge.r-project.org/projects/wordspace/

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

[4] http://wacky.sslmit.unibo.it/doku.php?id=corpora

[5] http://wacky.sslmit.unibo.it/doku.php?id=corpora

[6] A subset of WaCkypedia_EN containing the initial 500 words of each article, which amounts to 230 million tokens.

[7] See Evert (2004) for a description of these measures and details on the calculation of association scores.

[8] This type of association is labeled as "backward" because it goes from targets to primes, while in the experimental setting targets are shown after primes.

larger corpora. Third, we extend the use of rank-based measures to the distributional simulation of two other experiments on event knowledge (Ferretti et al., 2001; McRae et al., 2005). Note that our simulation differs from the one by Hare et al. (2009) with respect to tasks (they test for a significant difference of mean distances between target and related vs. unrelated prime) and the class of DSMs (we use term-term models, rather than LSA; our models are not sensitive to word order, unlike BEAGLE).

3.3 Tasks and Analysis of Results

The aim of this section is to introduce the experimental tasks whose results will be discussed in section 4 and to describe the main features of the analysis we applied to interpret these results.

Two experiments have been carried out:

- A **classification** experiment: given a target and two primes, distributional information is used to identify the congruent prime. Performance in this task is measured by classification accuracy (section 4.1).
- A **prediction** experiment: the information concerning distributional relatedness between targets and congruent primes is tested as a predictor for latency times. Performance in this task is quantified by Pearson correlation (section 4.2).

Concerning the interpretation of the evaluation results, it would hardly be meaningful to look at the best parameter combination or the average across all models. The best model is likely to be overfitted tremendously (after testing 38880 parameter settings over a dataset of 404 data points). Mean performance is largely determined by the proportions of "good" and "bad" parameter settings among the evaluation runs, which include many non-optimal parameter values that were only included for completeness.

Instead, we analyze the influence of individual DSM parameters and their interactions using linear models with performance (accuracy or correlation) as a dependent variable and the various model parameters as independent variables. This approach allows us to identify parameters that have a significant effect on model performance and to test for interactions between the parameters. Based on the partial effects of each parameter (and significant interactions) we can select a best model in a robust way.

This statistical analysis contains some elements of novelty with respect to the state-of-the-art DSM evaluation. Broadly speaking, approaches to DSM evaluation described in the literature fall into two classes. The first one can be labeled as *best model first*, as it implies the identification of the optimal configuration of parameters on an initial task, considered more basic; the best performing model on the general task is therefore evaluated on other tasks of interest. This is the approach adopted, for example, by Padó and Lapata (2007). In the second approach, described in Bullinaria and Levy (2007; 2012), evaluation is conducted via *incremental tuning of parameters*: parameters are evaluated sequentially to identify the best performing value on a number of tasks. Such approaches to DSM evaluation have specific limitations. The former approach does not assess which parameters are crucial in determining model performance, since its goal is the evaluation of performance of the same model on different tasks. The latter approach does not allow for parameter interactions, considering parameters individually. Both limitations are avoided in the analysis used here.

4 Results

4.1 Identification of Congruent Prime

This section presents the results from the first task evaluated in our study. We used the DSMs to identify which of the two primes is the congruent one based on their distributional relatedness to the target. For every triple in the dataset, the different indexes of distributional relatedness (parameter *relatedness index*) were used to compare the association between the target and the congruent prime with the association between the target and the incongruent prime. Accuracy of DSMs in picking up the congruent prime was calculated on the global dataset and separately for each subset.[9]

Figure 1 displays the distribution of the accuracy scores of all tested models in the task, on the global dataset. All accuracy values are specified as percentages. Minimum, maximum, mean and standard deviation of the accuracy values for the global dataset and for the three subsets are displayed in table 2. The respective best models are reported in table 6 in the Appendix.

[9]The small number of triples for which no prediction could be made because of missing words in the DSMs were considered mistakes. The coverage of the models ranges from 97.8% to 100% of the triples, with a mean of 99%.

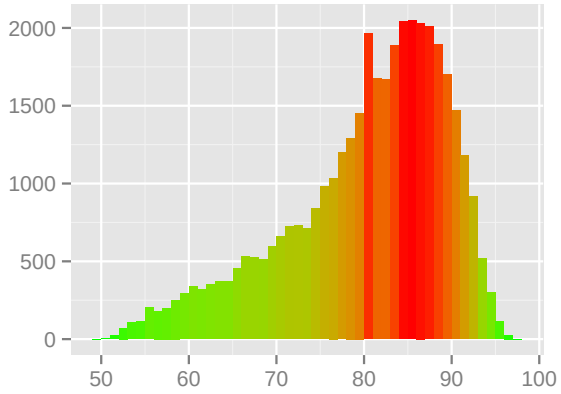

Figure 1: Identification of congruent prime: distribution of accuracy (%) for global dataset

Dataset	Min	Max	Mean	σ
Global	50.2	96.5	80.2	9.2
V-N	45.8	95.8	80.0	8.4
N-V	49.1	99.1	82.7	9.7
N-N	47.6	97.6	78.7	10.0

Table 2: Identification of congruent prime: mean and range for global dataset and subsets

The mean performance on N-N is lower than on N-V and slightly lower than on V-N. This effect may be interpreted as being due to mediated priming, as no verb is explicitly involved in the N-N relationship. Yet, the relatively high accuracy on N-N and its relatively small difference from N-V and V-N does not speak in favor of a different underlying mechanism responsible for this effect. Indeed, McKoon and Ratcliff (1992) suggested that effects traditionally considered as instances of mediated priming are not due to activation spreading through a mediating node, but result from a direct but weaker relatedness between prime and target words. This hypothesis found computational support in McDonald and Lowe (2000).[10]

4.1.1 Model Parameters and Accuracy

The aim of this section is to assess which parameters have the most significant impact on the performance of DSMs in the task of identification of the congruent prime.

We trained a linear model with the eight DSM parameters as independent variables ($R^2 = 0.70$) and a second model that also includes all two-way interactions ($R^2 = 0.89$). Given the improvement in R^2 as a consequence of the inclusion of two-way interactions in the linear model, we will focus on the results from the model with interactions. Table 3 shows results from the analysis of variance for the model with interactions. For every parameter (and interaction of parameters) we report degrees of freedom (*df*), percentage of explained variance (R^2), and a significance code (*signif*). We only list significant interactions that explain at least 1% of the variance. Even though all parameters and many interactions are highly significant due to the large number of DSMs that were tested, an analysis of their predictive power in terms of explained variance allows us to make distinctions between parameters.

Parameter	df	R^2	signif
corpus	4	7.44	***
window	2	4.39	***
pos	2	0.92	***
score	5	7.39	***
transformation	3	3.79	***
distance	2	22.20	***
dimensionality reduction	2	10.52	***
relatedness index	3	13.67	***
score:transformation	15	4.53	***
distance:relatedness index	12	2.24	***
distance:dim.reduction	4	2.16	***
window:dim.reduction	4	1.73	***

Table 3: Accuracy: Parameters and interactions

Results in table 3 indicate that *distance*, *dimensionality reduction* and *relatedness index* are the parameters with the strongest explanatory power, followed by *corpus* and *score*. *Window* and *transformation* have a weaker explanatory power, while *pos* falls below the 1% threshold. There is a strong interaction between *score* and *transformation*, which has more influence than one of the individual parameters, namely *transformation*.

Figures 2 to 7 display the partial effects of different model parameters (*pos* was excluded because of its low explanatory power). One of the main research questions behind this work was whether neighbor rank performs better than distance in predicting priming data. The partial effect of *relatedness index* in Figure 6 confirms our hypothesis: forward rank achieves the best performance, distance the worst.[11]

Accuracy improves for models trained on bigger corpora (parameter *corpus*, figure 2; corpora are ordered by size) and larger context windows (parameter *window*, figure 3). Cosine is the best performing *distance measure* (figure 4). Interestingly, dimensionality reduction is found to negatively affect model performance: as shown in

[10] The interpretation of the N-N results in terms of spreading activation is also rejected by Hare et al. (2009, 163).

[11] Backward rank is equivalent to distance in this task.

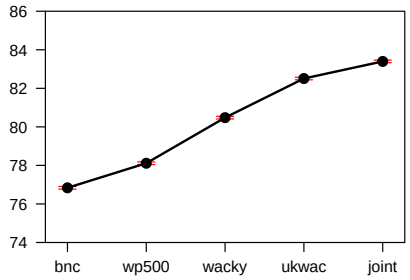

Figure 2: Corpus

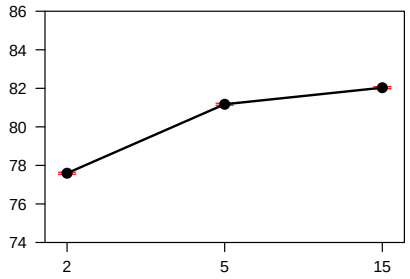

Figure 3: Window

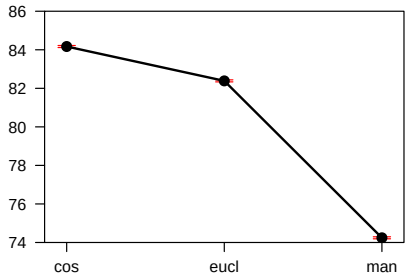

Figure 4: Distance

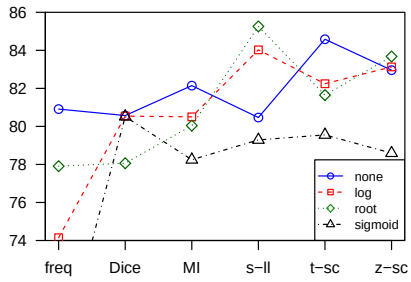

Figure 5: Score + Transformation

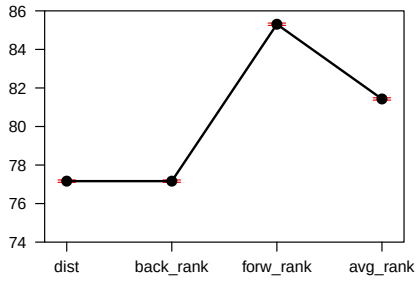

Figure 6: Rel. Index

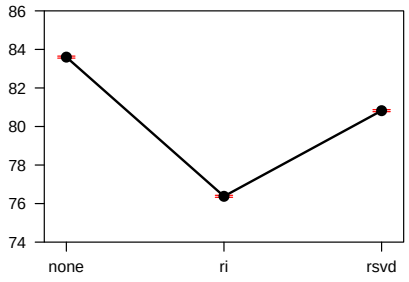

Figure 7: Dim. Reduction

figure 7, both random indexing (ri) and singular value decomposition (rsvd) cause a decrease in predicted accuracy.

Because of the strong interaction between *score* and *transformation*, only their combined effect is shown (figure 5). Among the scoring measures, stochastic association measures perform better than frequency: in particular log-likelihood (simple-ll), z-score and t-score are the best measures. We can identify a general tendency of *transformation* to lower accuracy. This is true for all scores except log-likelihood: square root and (to a lesser extent) logarithmic transformation result in an improvement for this measure.

Figure 8 displays the interaction between the parameters *distance* and *dimensionality reduction*. Despite a general tendency for *dimensionality reduction* to lower accuracy, we found an interaction between cosine distance and singular value decomposition: in this combination, accuracy remains stable and is even minimally higher compared to no dimensionality reduction.

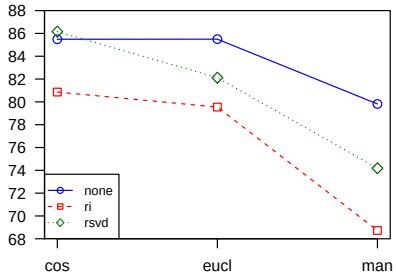

Figure 8: Distance + Dimensionality Reduction

4.2 Correlation to Reaction Times

The results reported in section 4.1 demonstrate that forward rank is the best index for identifying which of the two primes is the congruent one. The aim of this section is to find out whether rank is also a good predictor of latency times. We check correlation between distributional relatedness and reaction times and evaluate the impact of model parameters on this task. Figure 9 displays the distribution of Pearson correlation coefficient achieved by the different DSMs on the global dataset.

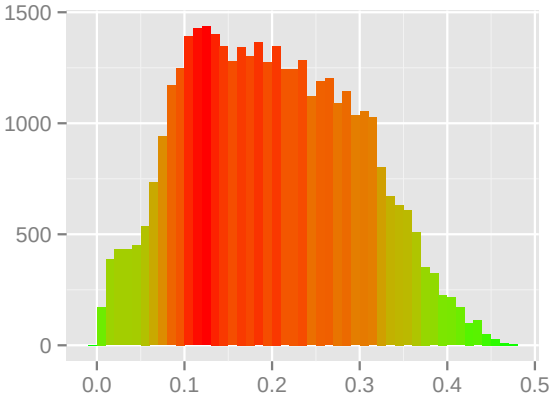

Figure 9: Distribution of Pearson correlation between relatedness and RT in the global dataset

Figure 9 shows that the majority of the models perform rather poorly, and that only few models achieve moderate correlation with RT. DSM performance in the correlation task appears to be less

robust to non-optimal parameter settings than in the accuracy task (cf. figure 1). Minimum, maximum, mean and standard deviation correlation for the global dataset and for the three evaluation subsets are shown in table 4. In all the cases, absolute correlation values are used so as not to distinguish between positive and negative correlation. Best models for the global dataset and for the three subsets are reported in table 7 in the Appendix.

Dataset	Min	Max	Mean	σ
Global	-0.26	0.47	0.19	0.10
V-N	-0.34	0.57	0.2	0.12
N-V	-0.35	0.41	0.11	0.06
N-N	-0.29	0.42	0.16	0.09

Table 4: Mean and range of Pearson correlation coefficients on global dataset and subsets

4.2.1 Model Parameters and Correlation

In this section we discuss the impact of different model parameters on correlation with reaction times. We trained a linear model with absolute Pearson correlation on the global dataset as dependent variable and the eight DSM parameters as independent variables ($R^2 = 0.53$), and a second model that includes two-way interactions ($R^2 = 0.77$). Table 5 is based on the model with interactions; it reports the degrees of freedom (*df*), proportion of explained variance (R^2) and a significance code (*signif*) for every parameter and every interaction of parameters (above 1% of explained variance).

Parameter	df	R^2	signif
corpus	4	7.45	***
window	2	0.47	***
pos	2	0.20	***
score	5	3.03	***
transformation	3	3.52	***
distance	2	4.27	***
dimensionality reduction	2	10.57	***
relatedness index	3	23.40	***
dim.reduction:relatedness index	6	5.21	***
distance:dim.reduction	4	4.11	***
distance:relatedness index	6	3.77	***
score:transformation	15	3.22	***
score:relatedness index	15	1.37	***

Table 5: Correlation: Parameters and interactions

Relatedness index is the most important parameter, followed by *dimensionality reduction* and *corpus*. The explanatory power of the other parameters (*score*, *transformation*, *distance*) is lower than for the accuracy task, and two parameters (*window* and *pos*) explain less than 1% of the variance each. By contrast, the explanatory power of interactions is higher in this task. Table 5 shows the five relevant interactions with an overall higher R^2 compared to the accuracy task (cf. table 3).

The partial effect plot for *relatedness index* (figure 15) confirms the findings of the accuracy task: forward rank is the best value for this parameter. The best values for the other parameters, however, show opposite tendencies with respect to the accuracy task. Models trained on smaller corpora (figure 11) perform better than those trained on bigger ones. Cosine is still the best distance measure, but manhattan distance performs equally well in this task (parameter *distance*, figure 13). Singular value decomposition (parameter *dimensionality reduction*, figure 16) weakens the correlation values achieved by the models, but no significant difference is found between random indexing and the unreduced data.

Co-occurrence frequency performs better than statistical association measures and *transformation* improves correlation: figure 14 displays the interaction between these two parameters. *Transformation* has a positive effect for every score, but the optimal transformation differs. Its impact is particularly strong for the Dice coefficient, which reaches the same performance as frequency when combined with a square root transformation.

Let us conclude by discussing the interaction between *distance* and *dimensionality reduction* (figure 10). Based on the partial effects of the individual parameters, any combination of manhattan or cosine distance with random indexing or no dimensionality reduction should be close to optimal. However, the interaction plot reveals that manhattan distance with random indexing is the best combination, outperforming the second best (cosine without dimensionality reduction) by a considerable margin. The positive effect of random indexing is quite surprising and requires further investigation.

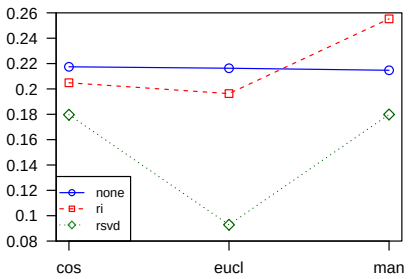

Figure 10: Distance + Dimensionality Reduction

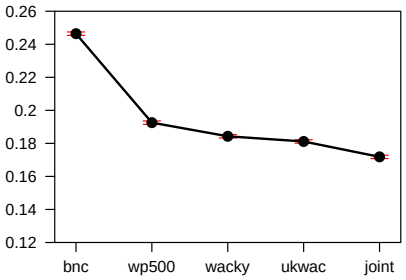

Figure 11: Corpus

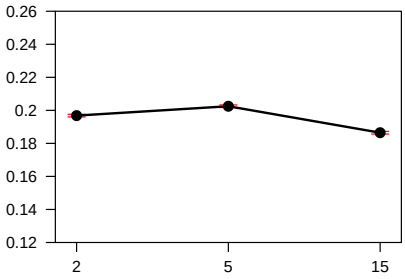

Figure 12: Window

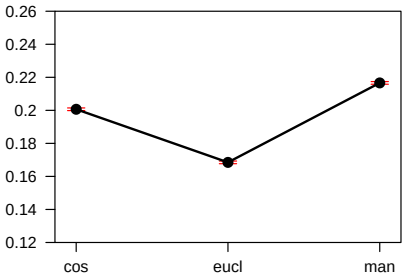

Figure 13: Distance

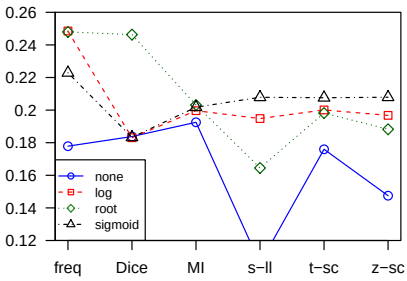

Figure 14: Score + Transformation

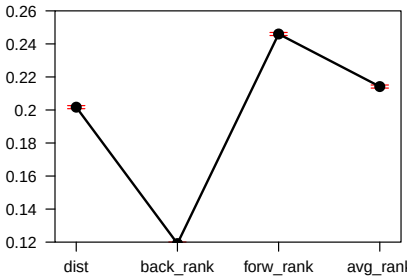

Figure 15: Rel. Index

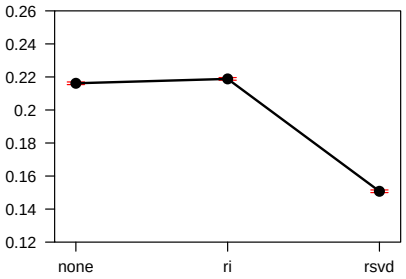

Figure 16: Dim. Reduction

5 Conclusion

In this paper, we presented the results of a large-scale evaluation of distributional models and their parameters on behavioral data from priming experiments. Our study is, to the best of our knowledge, the first systematic evaluation of such a wide range of DSM parameters in all possible combinations. Our study also provides a methodological contribution to the problem of DSM evaluation. We propose to apply linear modeling to determine the impact of different model parameters and their interactions on the performance of the models. We believe that this type of analysis is robust against overfitting. Moreover, effects can be tested for significance and various forms of interactions between model parameters can be captured.

The main findings of our evaluation can be summarized as follows. Forward association (rank of target among the nearest neighbors of the prime) performs better than distance in both tasks at issue: identification of congruent prime and correlation with latency times. This finding confirms and extends the results of previous studies (Hare et al., 2009). The relevance of rank-based measures for cognitive modeling is discussed in section 3.2.2.

Identification of congruent primes on the basis of distributional relatedness between prime and target is improved by employing bigger corpora and by using statistical association measures as scoring functions, while correlation to reaction times is strengthened by smaller corpora and co-occurrence frequency or Dice coefficient. A significant interaction between transformation and scoring function is found in both tasks: considering the interaction between these two parameters turned out to be vital for the identification of optimal parameter values.

Some preliminary analyses of individual thematic relations showed substantial improvements of correlations. Therefore, future work will focus on finer-grained linear models for single relations and on further modeling of reaction times, extending the study by Hutchinson et al. (2008).

Further research steps also include an evaluation of syntax-based models (Baroni and Lenci, 2010; Padó and Lapata, 2007) and term-document models on the tasks tackled in this paper, as well as an evaluation of all models on standard tasks.

Acknowledgments

We are grateful to Ken MacRae for providing us the priming data modeled here and to Alessandro Lenci for his contribution to the development of this study. We would also like to thank the Computational Linguistics group at the University of Osnabrück and the Corpus Linguistics group at the University Erlangen for feedback. Thanks also go to three anonymous reviewers, whose comments helped improve our analysis, and to Sascha Alexeyenko for helpful advice. The first author's PhD project is funded by a Lichtenberg grant from the Ministry of Science and Culture of Lower Saxony.

References

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

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

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.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, IMS, University of Stuttgart.

Todd Ferretti, Ken McRae, and Ann Hatherell. 2001. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516–547.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114:211–244.

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.

Amac Herdağdelen, Marco Baroni, and Katrin Erk. 2009. Measuring semantic relatedness with vector space models and random walks. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 50–53.

Keith A. Hutchinson, David A. Balota, Michael J. Cortese, and Jason M. Watson. 2008. Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, 61(7):1036–1066.

Michael Jones and Douglas Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114:1–37.

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.

Will Lowe and Scott McDonald. 2000. The direct route: mediated priming in semantic space. Technical report, Division of Informatics, University of Edinburgh.

Scott McDonald and Chris Brew. 2004. A distributional model of semantic context effects in lexical processing. In *Proceedings of ACL-04*, pages 17–24.

Gain McKoon and Roger Ratcliff. 1992. Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18:1155–1172.

Ken McRae and Kazunaga Matzuki. 2009. People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass*, 3(6):1417–1429.

Ken McRae, Mary Hare, Jeffrey L. Elman, and Todd Ferretti. 2005. A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, 33(7):1174–1184.

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

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

Magnus Sahlgren. 2005. An introduction to random indexing. In *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*.

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.

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

Amos Tversky. 1977. Features of similarity. *Psychological Review*, 84:327–352.

Appendix

The tables[12] in this appendix report best models for the global dataset and the three subsets. In particular, they include the following information for

[12]For the parameter *Pos* (*part of speech*), abbreviated values read as follows: *no* – no part of speech information; *t* –

the two tasks: (a) the best model or one of the best models, as well as a specification of how many tight models achieved the best performance (row *Model*, tights in parentheses); (b) the best parameter setting according to the linear model analysis (row *Setting*)[13]; (c) the model which performed best in the respective other task (row *Correlation* in table 6, row *Accuracy* in table 7).

Full Dataset									
Best	**Corpus**	**Win**	**Pos**	**Score**	**Trans**	**Dist**	**D.Red**	**R.Ind**	**Acc**
Model (2)	joint	15	t	s-ll	none	cos	none	fw_r	96.5
Setting	joint	15	t	s-ll	root	cos	none	fw_r	94.3
Correlation	bnc	5	no	freq	log	cos	none	fw_r	83.9
N-V									
Best	**Corpus**	**Win**	**Pos**	**Score**	**Trans**	**Dist**	**D.Red**	**R.Ind**	**Acc**
Model (57)	ukwac	5	no	freq	none	cos	rsvd	fw_r	99.1
Setting	joint	15	t	s-ll	root	cos	none	fw_r	97.4
Correlation	bnc	2	no	freq	sigm	man	rsvd	fw_r	74.1
V-N									
Best	**Corpus**	**Win**	**Pos**	**Score**	**Trans**	**Dist**	**D.Red**	**R.Ind**	**Acc**
Model (2)	wacky	15	no	z-sc	none	cos	none	fw_r	95.8
Setting	joint	15	t	s-ll	root	cos	none	fw_r	93.2
Correlation	joint	5	no	MI	log	man	ri	fw_r	85.6
N-N									
Best	**Corpus**	**Win**	**Pos**	**Score**	**Trans**	**Dist**	**D.Red**	**R.Ind**	**Acc**
Model (27)	ukwac	15	t	z-sc	none	cos	rsvd	av_r	97.6
Setting	joint	15	t	s-ll	root	cos	none	fw_r	92.9
Correlation	bnc	5	t+f	freq	log	cos	ri	av_r	69.4

Table 6: Best Models: Identification of congruent prime, full dataset and subsets

Full Dataset									
Best	**Corpus**	**Win**	**Pos**	**Score**	**Trans**	**Dist**	**D.Red**	**R.Ind**	**Cor**
Model (6)	bnc	5	no	freq	log	cos	none	fw_r	0.47
Setting	bnc	5	no	freq	none	man	ri	fw_r	0.42
Accuracy	joint	15	t	s-ll	none	cos	none	fw_r	0.16
N-V									
Best	**Corpus**	**Win**	**Pos**	**Score**	**Trans**	**Dist**	**D.Red**	**R.Ind**	**Cor**
Model (1)	bnc	2	no	freq	sigm	man	rsvd	fw_r	0.41
Setting	bnc	5	no	freq	none	man	ri	fw_r	0.24
Accuracy	ukwac	5	no	freq	none	cos	none	fw_r	-.01
V-N									
Best	**Corpus**	**Win**	**Pos**	**Score**	**Trans**	**Dist**	**D.Red**	**R.Ind**	**Cor**
Model (7)	joint	5	no	MI	log	man	ri	fw_r	0.57
Setting	bnc	5	no	freq	none	man	ri	fw_r	0.45
Accuracy	wacky	15	no	z-sc	none	cos	none	fw_r	0.31
N-N									
Best	**Corpus**	**Win**	**Pos**	**Score**	**Trans**	**Dist**	**D.Red**	**R.Ind**	**Cor**
Model (1)	bnc	5	t+f	freq	log	cos	ri	av_r	0.42
Setting	bnc	5	no	freq	none	man	ri	fw_r	0.27
Accuracy	ukwac	15	t	z-sc	none	cos	rsvd	av_r	0.06

Table 7: Best Models: Pearson correlation to reaction times, full dataset and subsets

part of speech on targets; *tf* – part of speech on targets and contexts. For the parameter *R.Ind* (*relatedness index*), *fw_r* reads as *forward rank* and *av_r* as *average rank*.

[13]For the three subsets we report the performance of the best setting on the global dataset.