Contrasting Syntagmatic and Paradigmatic Relations: Insights from Distributional Semantic Models

Gabriella Lapesa[3,1]
[1]Universität Osnabrück
Institut für
Kognitionswissenschaft
glapesa@uos.de

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

Sabine Schulte im Walde[3]
[3]Universität Stuttgart
Institut für Maschinelle
Sprachverarbeitung
schulte@ims.uni-stuttgart.de

Abstract

This paper presents a large-scale evaluation of bag-of-words distributional models on two datasets from priming experiments involving syntagmatic and paradigmatic relations. We interpret the variation in performance achieved by different settings of the model parameters as an indication of which aspects of distributional patterns characterize these types of relations. Contrary to what has been argued in the literature (Rapp, 2002; Sahlgren, 2006) – that bag-of-words models based on second-order statistics mainly capture paradigmatic relations and that syntagmatic relations need to be gathered from first-order models – we show that second-order models perform well on both paradigmatic and syntagmatic relations if their parameters are properly tuned. In particular, our results show that size of the context window and dimensionality reduction play a key role in differentiating DSM performance on paradigmatic vs. syntagmatic relations.

1 Introduction

Distributional takes on the representation and acquisition of word meaning rely on the assumption that words with similar meaning tend to occur in similar contexts: this assumption, known as *distributional hypothesis*, has been first proposed by Harris (1954). Distributional Semantic Models (henceforth, DSMs) are computational models that operationalize the distributional hypothesis; they produce semantic representations for words in the form of distributional vectors recording patterns of co-occurrence in large samples of language data (Sahlgren, 2006; Baroni and Lenci, 2010; Turney and Pantel, 2010). Comparison between distributional vectors allows the identification of shared contexts as an empirical correlate of the semantic similarity between the target words. As noted in Sahlgren (2008), the notion of semantic similarity applied in distributional approaches to meaning is an easy target of criticism, as it is employed to capture a wide range of semantic relations, such as synonymy, antonymy, hypernymy, up to topical relatedness.

The study presented in this paper contributes to the debate concerning the nature of the semantic representations built by DSMs, and it does so by comparing the performance of several DSMs in a classification task conducted on priming data and involving paradigmatic and syntagmatic relations. Paradigmatic relations hold between words that occur in similar contexts; they are also called relations *in absentia* (Sahlgren, 2006) because paradigmatically related words do not co-occur. Examples of paradigmatic relations are *synonyms* (e.g., *frigid–cold*) and *antonyms* (e.g., *cold–hot*). Syntagmatic relations hold between words that co-occur (relations *in praesentia*) and therefore exhibit a similar distribution across contexts. Typical examples of syntagmatic relations are phrasal associates (e.g., *help–wanted*) and syntactic collocations (e.g., *dog–bark*).

Distributional modeling has already tackled the issue of paradigmatic and syntagmatic relations (Sahlgren, 2006; Rapp, 2002). Key contributions of the present work are the scope of its evaluation (in terms of semantic relations and model parameters) and the new perspective on paradigmatic vs. syntagmatic models provided by our results.

Concerning the scope of the evaluation, this is the first study in which the comparison involves such a wide range of semantic relations (*paradigmatic:* synonyms, antonyms and co-hyponyms; *syntagmatic:* syntactic collocations, backward and forward phrasal associates). Moreover, our evaluation covers a large number of DSM parameters: source corpus, size and direction of the context window, criteria for feature selection, feature

*Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 160–170,
Dublin, Ireland, August 23-24 2014.

weighting, dimensionality reduction and index of distributional relatedness. We consider the variation in performance achieved by different parameter settings as a cue towards characteristic aspects of specific relations (or groups of relations).

Our work also differs from previous studies (Sahlgren, 2006; Rapp, 2002) in its focus on second-order models. We aim to show that they are able to capture both paradigmatic and syntagmatic relations with appropriate parameter settings. In addition, this focus provides a uniform experimental design for the evaluation. For example, parameters like window size and directionality apply to bag-of-words DSMs and collocation lists but not to term-context models; dimensionality reduction, whose effect has not yet been explored systematically in the context of syntagmatic and paradigmatic relations, is not applicable to collocation lists.

This paper is structured as follows. Section 2 summarizes previous work. Section 3 describes the experimental setup, in terms of task, datasets and evaluated parameters. Section 4 introduces our model selection methodology. Section 5 presents the results of our evaluation study. Section 6 summarizes main findings and sketches ongoing and future work.

2 Previous Work

In this section we discuss previous work relevant to the distributional modeling of paradigmatic and syntagmatic relations. For space constraints, we focus only on two studies (Rapp, 2002; Sahlgren, 2006) in which the two classes of relations are compared at a global level, and not on studies that are concerned with specific semantic relations, e.g., *synonymy* (Edmonds and Hirst, 2002; Curran, 2003), *hypernymy* (Weeds et al., 2004; Lenci and Benotto, 2012) or syntagmatic predicate preferences (McCarthy and Carroll, 2003; Erk et al., 2010), etc.

In previous studies, the comparison of syntagmatic and paradigmatic relations has been implemented in terms of an opposition between different classes of corpus-based models: term-context models (words as targets, documents or context regions as features) vs. bag-of-words models (words as targets and features) in Sahlgren (2006); collocation lists vs. bag-of-words models in Rapp (2002). Given the high terminological variation in the literature, in this paper we will adopt the labels *syntagmatic* and *paradigmatic* to characterize different types of semantic relations, and we will use the labels *first-order* and *second-order* to characterize corpus-based models with respect to the kind of co-occurrence information they encode. We will refer to collocation lists and term-document DSMs as *first-order models*, and to bag-of-words DSMs as *second-order models*[1].

Rapp (2002) integrates first-order (co-occurrence lists) and second-order (bag-of-words DSMs) information to distinguish syntagmatic and paradigmatic relations. Under the assumption that paradigmatically related words will be found among the closest neighbors of a target word in the DSM space and that paradigmatically and syntagmatically related words will be intermingled in the list of collocates of the target word, Rapp proposes to exploit a comparison of the most salient collocates and the nearest DSM neighbors to distinguish between the two types of relations.

Sahlgren (2006) compares term-context and bag-of-words DSMs in a number of tasks involving syntagmatic and paradigmatic relations. First, a comparison between the thesaurus entries for target words (containing both paradigmatically and syntagmatically related words) and neighbors in the distributional spaces is conducted. It shows that, while term-context DSMs produce both syntagmatically and paradigmatically related words, the nearest neighbors in a bag-of-words DSM mainly provide paradigmatic information. Bag-of-words models also performed better than term-context models in predicting association norms, in the TOEFL multiple-choice synonymy task and in the prediction of antonyms (although the difference in performance was less significant here). Last, word neighborhoods are analysed in terms of their part-of-speech distribution. Sahlgren (2006) observes that bag-of-words spaces contain more neighbors with the same part of speech as the target than term-context spaces. He concludes that bag-of-words spaces privilege paradigmatic relations, based on the assumption that paradigmatically related word pairs belong to the same part of speech, while this is not necessarily the case for syntagmatically related word pairs.

[1] Term-document models encode first-order information because dot products between row vectors are related to co-occurrence counts of the corresponding words (within documents). More precisely, for a binary term-document matrix, cosine similarity is identical to the square root of the MI^2 association measure. Please note that our terminology differs from that of Schütze (1998) and Peirsman et al. (2008).

Summing up, in both Rapp (2002) and Sahlgren (2006) it is claimed that second-order models perform poorly in predicting syntagmatic relations. However, neither of those studies involves datasets containing *exclusively syntagmatic relations*, as the evaluation focuses either on paradigmatic relations (TOEFL multiple choice test, antonymy test) or on resources containing both types of relations (thesauri, association norms).

3 Experimental Setting

3.1 Evaluation Task and Data

In this study, bag-of-words DSMs are evaluated on two datasets containing experimental items from two priming studies. Each item is a word triple (target, consistent prime, inconsistent prime) with a particular semantic relation between target and consistent prime. Following previous work on modeling priming effects as a comparison between prime-target pairs (McDonald and Brew, 2004; Padó and Lapata, 2007; Herdağdelen et al., 2009), we evaluate our models in a classification task. The goal is to identify the consistent prime on the basis of its distributional relatedness to the target: if a particular DSM (i.e., a certain parameter combination) is sensitive to a specific relation (or group of relations), we expect the consistent primes to be closer to the target in semantic space than the inconsistent ones.

The first dataset is derived from the **Semantic Priming Project** (SPP) (Hutchison et al., 2013). To the best of our knowledge, our study represents the first evaluation of bag-of-words DSMs on items from this dataset. The original data consist of 1661 word triples (target, consistent prime, inconsistent prime) collected within a large-scale project aiming at characterizing English words in terms of a set of lexical and associative/semantic characteristics, along with behavioral data from visual lexical decision and naming studies[2]. We manually discarded all triples containing proper nouns, adverbs or inflected words. We then selected five subsets involving different semantic relations, namely: **synonyms** (SYN), 436 triples (example of a consistent prime and target: *frigid–cold*); **antonyms** (ANT): 135 triples (e.g., *hot–cold*); **cohyponyms** (COH): 159 triples (e.g., *table–chair*); **forward phrasal associates** (FPA): 144 triples (e.g., *help–wanted*); **backward phrasal associates** (BPA): 89 triples (e.g., *wanted–help*).

The second priming dataset is the **Generalized Event Knowledge** dataset (henceforth GEK), already evaluated in Lapesa and Evert (2013): a collection of 402 triples (target, consistent prime, inconsistent prime) from three priming studies conducted to demonstrate that event knowledge is responsible for facilitation of the processing of words that denote events and their participants. The first study was conducted by 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. The study covered five thematic relations: agent (e.g., *pay–customer*), patient, feature of the patient, instrument, location. The second study (McRae et al., 2005) focussed on priming from nouns to verbs. It involved four relations: agent (e.g., *reporter–interview*), patient, instrument, location. The third study (Hare et al., 2009) investigated priming from nouns to nouns, referring to participants of the same event or the event itself. The dataset involves seven relations: event-people (e.g., *trial–judge*), event-thing, location-living, location-thing, people-instrument, instrument-people, instrument-thing.

In the presentation of our results we group synonyms with antonyms and cohyponyms from SPP as paradigmatic relations, and the entire GEK dataset with backward and forward phrasal associates from SPP as syntagmatic relations.

3.2 Evaluated Parameters

DSMs evaluated in this paper belong to the class of bag-of-words models. We defined a large vocabulary of target words (27522 lemma types) containing all the items from the evaluated datasets as well as items from other state-of-the-art evaluation studies (Baroni and Lenci, 2010; Baroni and Lenci, 2011). Context words were filtered by part-of-speech (nouns, verbs, adjectives, and adverbs). Distributional models were built using the UCS toolkit[3] and the `wordspace` package for R[4]. The following parameters have been evaluated:

- **Source corpus** (abbreviated as *corpus* in plots 1-4): We compiled DSMs from three corpora often used in DSM evaluation studies and that

[2]The dataset is available at http://spp.montana.edu/

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

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

differ in both size and quality: British National Corpus[5], ukWaC, and WaCkypedia_EN[6].

- **Size of the context window** (*win.size*): As this parameter quantifies the amount of shared context involved in the computation of similarity, we expect it to be crucial in determining whether syntagmatic or paradigmatic relations are captured. We therefore use a finer granularity for window size than Lapesa and Evert (2013): 1, 2, 4, 8 and 16 words.
- **Directionality of the context window** (*win.direction*): When collecting co-occurrence information from the source corpora, we use either a directed window (i.e., separate frequency counts for co-occurrences of a context term to the left and to the right of the target term) or an undirected window (i.e., no distinction between left and right context when collecting co-occurrence counts).
- **Context selection**: From the full co-occurrence matrix collected as described above, we select dimensions (columns) according to the following parameters:
  - **Criterion for context selection** (*criterion*): We select the top-ranked dimensions either according to marginal frequency (i.e., we use the most frequent words as context terms) or number of nonzero co-occurrence counts (i.e., we use the context terms that co-occur with the highest number of targets).
  - **Number of context dimensions** (*context.dim*): We select the top-ranked 5000, 10000, 20000, 50000 or 100000 dimensions, according to the criterion above.
- **Feature scoring** (*score*): Co-occurrence counts are weighted using one of the following association measures: frequency, Dice coefficient, simple log-likelihood, Mutual Information, t-score, z-score or tf.idf.[7]
- **Feature transformation** (*transformation*): A transformation function may be applied to reduce the skewness of feature scores. Possible transformations are: none, square root, logarithmic and sigmoid.
- **Distance metric** (*metric*): We apply cosine distance (i.e., angle between vectors) or Manhattan distance.
- **Dimensionality reduction**: We apply singular value decomposition in order to project distributional vectors to a relatively small number of latent dimensions and compare the results to the unreduced runs[8]. For the SVD-based models, there are two additional parameters:
  - **Number of latent dimensions** (*red.dim*): Whether to use the first 100, 300, 500, 700 or 900 latent dimensions from the SVD analysis.
  - **Number of skipped dimensions** (*dim.skip*): When selecting latent dimensions, we optionally skip the first 50 or 100 SVD components. This parameter was inspired by Bullinaria and Levy (2012), who found that discarding the initial components of the reduced matrix, i.e. the SVD components with highest variance, improves evaluation results.
- **Index of distributional relatedness** (*rel.index*): We propose two alternative ways of quantifying the degree of relatedness between two words *a* and *b* represented in a DSM. The first option (and standard in distributional modeling) is to compute the *distance* (cosine or Manhattan) between the vectors of *a* and *b*. The second option, proposed in this work, is based on *neighbor rank*, i.e. we determine the rank of the target among the nearest neighbors of each prime. We expect that the target will occur in a higher position among the neighbors of the consistent prime than among those of the inconsistent prime. Since this corresponds to a lower numeric rank value for the consistent prime, we can treat neighbor rank as a measure of dissimilarity. Neighbor rank is particularly interesting as an index of relatedness because, unlike a distance metric, it can capture asymmetry effects[9].

4 Methodology

In our evaluation study, we tested all the possible combinations of the parameters listed in section

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

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

[7] See Evert (2008) for a description of these measures and details on the calculation of association scores. Note that we compute "sparse" versions of the association measures (where negative values are clamped to zero) in order to preserve the sparseness of the co-occurrence matrix.

[8] For efficiency reasons, we use randomized SVD (Halko et al., 2009) with a sufficiently high oversampling factor to ensure a good approximation.

[9] Note that our use of neighbor rank is fully consistent with the experimental design (primes are shown before targets). See Lapesa and Evert (2013) for an analysis of the performance of neighbor rank as a predictor of priming and discussion of the implications of using rank in cognitive modeling.

3.2, resulting in a total of 537600 different model runs (33600 in the setting without dimensionality reduction, 504000 in the dimensionality-reduced setting). The models were generated and evaluated on a large HPC cluster within approx. 4 weeks.

Our methodology for model selection follows the proposal of Lapesa and Evert (2013), who consider DSM parameters as predictors of model performance. We analyze the influence of individual parameters and their interactions using general linear models with performance (percent accuracy) as a dependent variable and the model parameters as independent variables, including all two-way interactions. 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 marginalize over all the parameters not shown in a plot and thus allow an intuitive interpretation of the effect sizes of categorical variables irrespective of the dummy coding scheme.

For each dataset, a separate linear model was fitted. The results are reported and compared in section 5. Table 1 lists the global goodness-of-fit (R^2) on each dataset, for the reduced and unreduced runs. Despite some variability across relations and between unreduced and reduced runs, the R^2 values are always high ($\geq 75\%$), showing that the linear model explains a large part of the observed performance differences. It is therefore justified to base our analysis on the linear models.

Relation	Dataset	Unreduced	Reduced
Syntagmatic	GEK	93%	87%
Syntagmatic	FPA	90%	79%
Syntagmatic	BPA	88%	77%
Paradigmatic	SYN	92%	85%
Paradigmatic	COH	89%	75%
Paradigmatic	ANT	89%	76%

Table 1: Evaluation, Global R^2

5 Results

In this section, we present the results of our study. We begin by looking at the distribution of accuracy for different datasets, and by comparing reduced and unreduced experimental runs in terms of minimum, maximum and mean performance.

The results displayed in table 2 show that dimensionality reduction with SVD improves the performance of the models for all datasets but GEK. We conclude that the information lost by applying SVD reduction (namely, meaningful distributional features, which are replaced by the generalization encoded in the reduced dimensions) is irrelevant to other tasks, but crucial for modeling the relations in the GEK dataset. This interpretation is consistent with the detrimental effect of SVD in tasks involving vector composition reported in the literature (Baroni and Zamparelli, 2010).

Relation	Dataset	Unreduced			Reduced		
		Min	Max	Mean	Min	Max	Mean
Syntagmatic	GEK	54.8	98.4	86.6	48.0	97.0	80.8
Syntagmatic	FPA	41.0	98.0	82.3	43.0	98.6	82.1
Syntagmatic	BPA	49.4	97.7	83.8	41.6	98.9	83.9
Paradigmatic	SYN	54.8	98.4	86.6	57.3	99.0	88.2
Paradigmatic	COH	49.0	100.0	92.6	54.3	100.0	94.0
Paradigmatic	ANT	69.6	100.0	94.2	57.8	100.0	94.3

Table 2: Distribution of Accuracy

5.1 Importance of Parameters

To obtain further insights into DSM performance we explore the effect of specific model parameters, comparing syntagmatic vs. paradigmatic relations and reduced vs. unreduced runs.

In order to establish a ranking of the parameters according to their importance wrt. model performance, we use a feature ablation approach. The ablation value for a given parameter is the proportion of variance (R^2) explained by this parameter together with all its interactions, corresponding to the reduction in adjusted R^2 of the linear model fit if the parameter were left out. In other words, it allows us to find out whether a certain parameter has a substantial effect on model performance (on top of all other parameters). Figures 1 to 4 display the feature ablation values of all the evaluated parameters in the unreduced and reduced setting, for paradigmatic and syntagmatic relations. Parameters are ranked according to their average feature ablation values in each setting.

Two parameters, namely **feature score** and **feature transformation**, are consistently crucial in determining DSM performance, both in reduced and unreduced runs, and for both paradigmatic and syntagmatic relations. In the next section we will show that it is possible to identify optimal (or nearly optimal) values for those parameters that are constant across relations.

A comparison of figures 1 and 2 with figures 3 and 4 allows us to identify parameters that lose or gain explanatory power when SVD comes into play. Feature ablation shows that the effect of the **index of distributional relatedness** is substantially smaller in the SVD-reduced runs, but this parameter still plays an important role. On the other hand, two parameters gain explanatory power in a

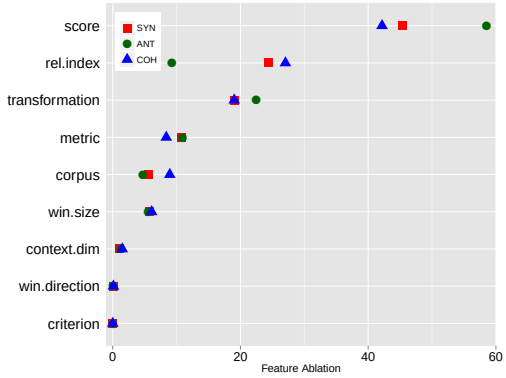

Figure 1: Paradigmatic, unreduced

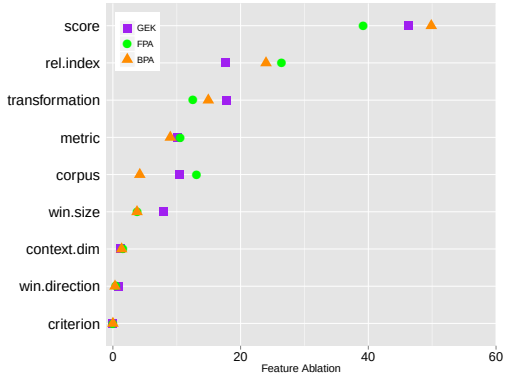

Figure 2: Syntagmatic, unreduced

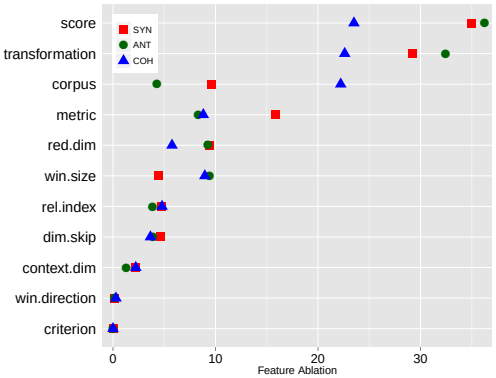

Figure 3: Paradigmatic, reduced

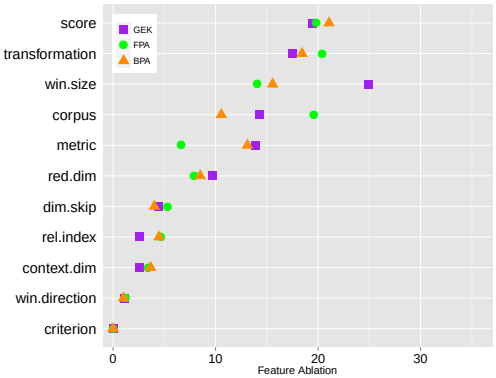

Figure 4: Syntagmatic, reduced

SVD-reduced setting: the **size of the context window** and the **source corpus**. Optimal values are discussed in section 5.2.

Three parameters consistently have little or no explanatory power: **directionality of the context window**, **criterion for context selection** and **number of context dimensions**.

We conclude this section by comparing relations within groups. Within paradigmatic relations, we note a significant drop in explanatory power for the **relatedness index** when it comes to antonyms. Within syntagmatic relations, the **size of the context window** appears to be more crucial on the GEK dataset than it is for FPA and BPA: in the next section, the analysis of the best choices for this parameter will provide a clue for the interpretation of this opposition.

5.2 Best Parameter Values

In this section, we identify the best parameter values for syntagmatic and paradigmatic relations by inspecting partial effects plots[10]. Our discussion starts from the parameters that contribute to the leading topic of this paper, namely the comparison between syntagmatic and paradigmatic relations: window size, parameters related to dimensionality reduction, and relatedness index.

As already anticipated in the feature ablation analysis, the **size of the context window** plays a crucial role in contrasting syntagmatic and paradigmatic relations, as well as different relations within those general groups. The plots in figures 5 and 6 display its partial effect for paradigmatic relations in the unreduced and reduced settings, respectively. The plots in figures 7 and 8 display its partial effect for syntagmatic relations. When no dimensionality reduction is involved, a very small context window (i.e., one word) is sufficient for all paradigmatic relations, and DSM performance decreases as soon as we enlarge the context window. The picture changes when applying dimensionality reduction: a 4-word window is a robust choice for all paradigmatic relations (although ANT show a further increase in performance with an 8-word window), even in the SYN task that is traditionally associated with very small windows of 1 or 2 words (cf. Sahlgren (2006)).

A significant interaction between window size and number of skipped dimensions (not shown for reasons of space) sheds further light on this matter. Without skipping SVD dimensions, the reduced models achieve optimal performance for a 2-word window and degrade more (COH) or less (ANT)

[10]The partial effect plots in figures 5 to 12 display parameter values on the x-axis and their effect size in terms of predicted accuracy on the y-axis (see section 4 for more details concerning the calculation of effect size).

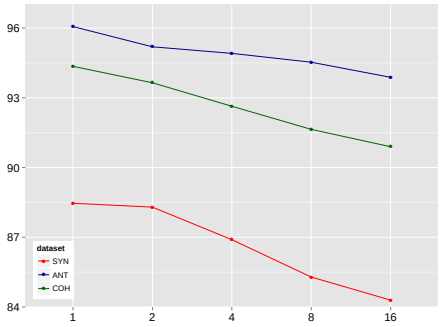

Figure 5: Window, paradigmatic, unreduced

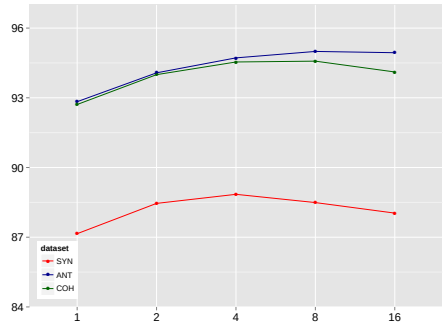

Figure 6: Window, paradigmatic, reduced

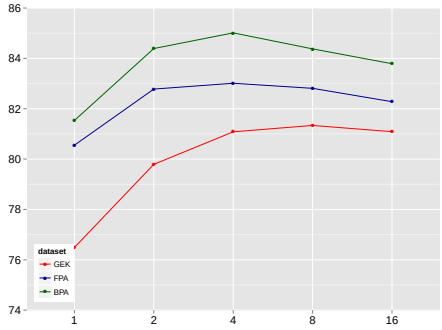

Figure 7: Window, syntagmatic, unreduced

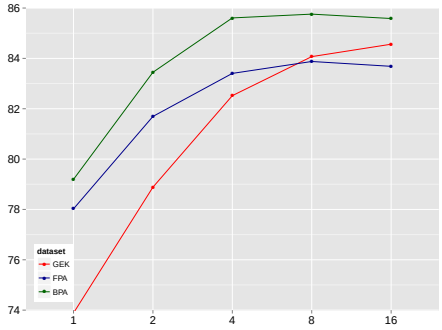

Figure 8: Window, syntagmatic, reduced

quickly for larger windows. With 50 or 100 dimensions skipped, performance improves up to a 4- or 8-word window. Our interpretation is that the first SVD dimensions capture general domain and topic information dominating the co-occurrence data; removing these dimensions reveals paradigmatic semantic relations even for larger windows. For syntagmatic relations without dimensionality reduction, a larger context window of 4 words is needed for FPA and BPA; a further increase of the window is detrimental. For the GEK dataset, performance peaks at 8 words, and decreases only minimally for even larger windows. Again, dimensionality reduction improves performance for large co-occurrence windows. For FPA and BPA, the optimum seems to be achieved with a window of 4–8 words; performance on GEK continues to increase up to 16 words, the largest window size considered in our experiments. Such patterns reflect differences in the nature of the semantic relations involved: smaller windows provide better contextual representations for paradigmatic relations while larger windows are needed to capture syntagmatic relations with bag-of-words DSMs (because co-occurring words then share a large portion of their context windows). Intermediate window sizes are sufficient for phrasal collocates (which are usually adjacent), while event-based relatedness (GEK) requires larger windows. Returning briefly to the slight preference shown by ANT for a larger window, we notice that ANT seems to be more similar to the syntagmatic relations than SYN and COH. This is in line with the observations of Justeson and Katz (1992) concerning the tendency of antonyms to co-occur (e.g., in coordinations such as *short and long*). Like synonyms, antonyms are interchangeable *in absentia*; but they also enter into syntagmatic patterns that are uncommon for synonyms.

We now focus on the parameters related to dimensionality reduction, namely the **number of latent dimensions** (figures 9 and 10) and the **number of skipped dimensions** (figures 11 and 12). These parameters represent an extension of the experiments conducted on the GEK dataset by Lapesa and Evert (2013). They have already been applied by Bullinaria and Levy (2012) to a different set of tasks, including the TOEFL multiple-choice synonymy task. In particular, Bullinaria and Levy found that discarding the initial SVD dimensions (with highest variance) leads to substantial improvements, especially in the TOEFL task. In our experiments, we found no difference between syntagmatic and paradigmatic relations wrt. the *number of latent dimensions*: the more, the better in both cases (900 dimensions). The *number of skipped dimensions*, however, shows some variability across the different relations. The results for SYN are in agreement with the findings of Bullinaria and Levy (2012) on TOEFL: skipping 50 or 100 initial dimensions improves performance. Skipping dimensions makes minimal difference

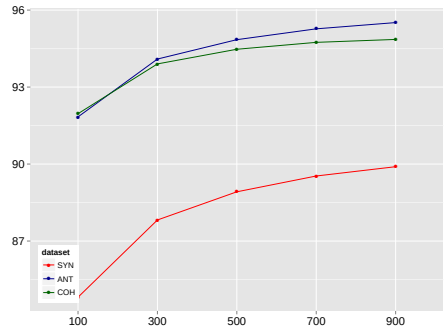

Figure 9: Latent dimensions, paradigmatic

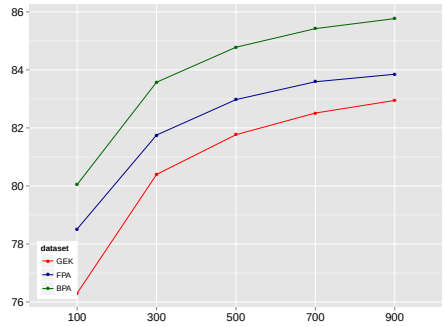

Figure 10: Latent dimensions, syntagmatic

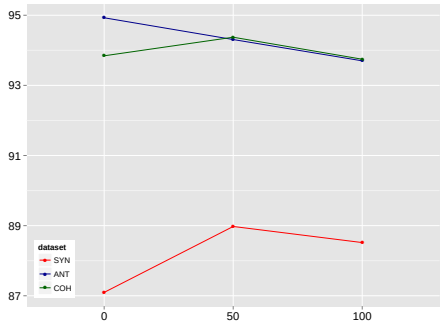

Figure 11: Skipped dimensions, paradigmatic

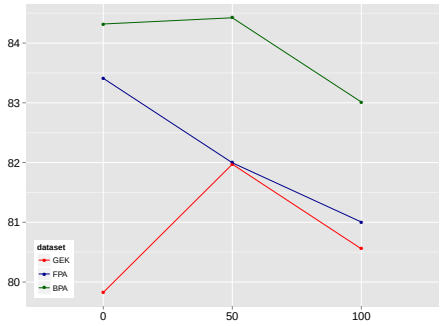

Figure 12: Skipped dimensions, syntagmatic

for COH (best choice is 50 dimensions), while the full range of reduced dimensions is necessary for ANT. Within syntagmatic relations, the full range of latent dimensions ensures good performance on phrasal associates (even if skipping 50 dimensions is not detrimental for BPA). GEK shows a pattern similar to SYN, with 50 skipped dimensions leading to a considerable improvement.

We now inspect the best values for the **relatedness index**. As shown in figure 13 for the unreduced runs and in figure 14 for the reduced runs, *neighbor rank* is consistently better than *distance* on all datasets. This is not surprising because, as discussed in section 3.2, our use of neighbor rank captures asymmetry and mirrors the experimental setting, in which targets are shown after primes. A further observation may be made relating to the degree of asymmetry of different relations. The unreduced setting in particular shows that syntagmatic relations are subject to stronger asymmetry effects than the paradigmatic ones, presumably due to the directional nature of the relations involved (phrasal associates and syntactic collocations). Among paradigmatic relations, antonyms appear to be the least asymmetric ones (because using neighbor rank instead of distance makes a comparatively small difference).

We conclude by briefly summarizing the optimal choices for the remaining parameters. The corresponding partial effects plots are not shown because of space constraints.

A very strong interaction between **score** and **transformation** characterizes all four settings (paradigmatic or syntagmatic datasets, reduced or unreduced experimental runs). Association measures outperform raw co-occurrence frequency. Measures based on significance tests (simple-ll, t-score, z-score) are better than Dice, and to a lesser extent, MI. Simple-ll is the best choice in combination with a logarithmic transformation for paradigmatic relations, z-score appears to be the best measure for syntagmatic relations in combination with a square root transformation. The difference is small, however, and *simple-ll with log transformation* works well across all datasets. Ongoing experiments with standard tasks show a similar pattern, suggesting that this combination of score and transformation parameters is appropriate for DSMs, regardless of the task involved.

The optimal **distance metric** is the *cosine distance*, consistently outperforming *Manhattan*. Concerning **source corpus**, *BNC* consistently yields the worst results, while *WaCkypedia* and *ukWaC* appear to be almost equivalent in the unreduced runs. The trade-off between quality and quantity appears to be strongly biased towards sheer corpus size in the case of distributional models. For syntagmatic relations and SVD-reduced models, *ukWaC* is clearly the best choice. This suggests that syntagmatic relations are better captured by features from a larger lexical inventory, combined with the abstraction performed by SVD.

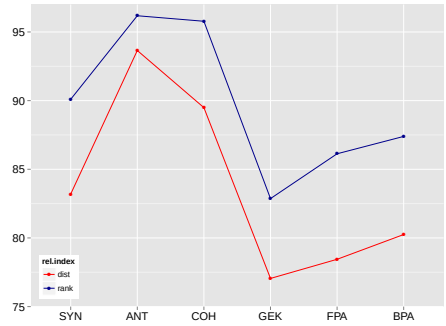

Figure 13: Relatedness index, unreduced

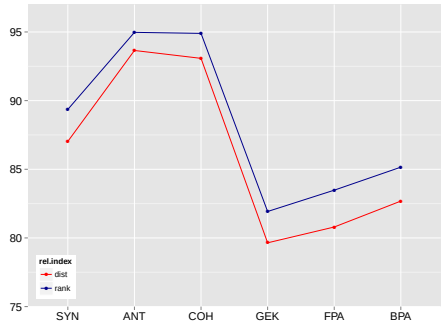

Figure 14: Relatedness index, reduced

Concerning minimally explanatory parameters, inspection of partial effect plots supported the choice of "unmarked" default values for **directionality of the context window** (i.e., *undirected*) and **criterion for context selection** (i.e., *frequency*), as well as an intermediate **number of context dimensions** (i.e., *50000* dimensions).

5.3 Best Settings

We conclude by comparing the performance achieved by our robust choice of optimal parameter values ("best setting") from section 5.2 with the performance of the best model for each dataset. For space constraints, the analysis of best settings focuses on the reduced experimental runs. Our best settings, shown in table 3, perform fairly well on the respective datasets[11].

dataset	corpus	win	score	transf	r.dim	d.sk	acc	best
GEK	ukwac	16	s-ll	log	900	50	96.0	97.0
FPA	ukwac	8	z-sc	root	900	0	93.0	98.6
BPA	ukwac	8	z-sc	root	900	0	95.5	98.9
SYN	ukwac	4	s-ll	log	900	50	96.3	99.0
COH	ukwac	4	s-ll	log	900	50	98.7	100
ANT	wacky	8	s-ll	log	900	0	100	100

Table 3: Best settings: datasets, parameter values, accuracy (*acc*), accuracy of the best model (*best*)

best setting	corpus	win	score	transf	r.dim	d.sk
Syntagmatic	ukwac	8	z-sc	root	900	0
Paradigmatic	ukwac	4	s-ll	log	900	50
General	ukwac	4	s-ll	log	900	0

Table 4: General best settings: parameter values

Dataset	Best Synt.	Best Para.	General
GEK	92.5	94.8	91.3
FPA	93.0	90.2	91.7
BPA	95.5	97.7	95.5
SYN	94.4	96.3	96.3
COH	99.3	98.7	98.7
ANT	99.2	99.2	99.2

Table 5: General best settings: accuracy

[11] Abbreviations in tables 3 and 4: win = window size; transf = transformation; z-sc = z-score; s-ll = simple-ll; r.dim = number of latent dimensions; d.sk = number of skipped dimensions. Parameters with fixed values for all datasets: number of context dimensions = 50k; direction = undirected; criterion = frequency; metric = cosine; relatedness index = rank.

As a next step, we identified parameter combinations that work well for all types of syntagmatic and paradigmatic relations, as well as an even more general setting that is suitable for paradigmatic and syntagmatic relations alike. Best settings are shown in table 4, their performance on each dataset is reported in table 5. General models achieve fairly good performance on all relations.

6 Conclusion

We presented a large-scale evaluation study of bag-of-words DSMs on a classification task derived from priming experiments. The leading theme of our study is a comparison between syntagmatic and paradigmatic relations in terms of the aspects of distributional similarity that characterize them. Our results show that second-order DSMs are capable of capturing both syntagmatic and paradigmatic relations, if parameters are properly tuned. Size of the co-occurrence window as well as parameters connected to dimensionality reduction play a key role in adapting DSMs to particular relations. Even if we do not address the more specific task of distinguishing between relations (e.g., synonyms vs. antonyms; see Scheible et al. (2013) and references therein), we believe that such applications may benefit from our detailed analyses on the effects of DSM parameters.

Ongoing and future work is concerned with the expansion of the evaluation setting to other classes of models (first-order models, dependency-based second-order models) and parameters (e.g., dimensionality reduction with Random Indexing).

Acknowledgments

We are grateful to Ken MacRae for providing us the GEK priming data and to the three reviewers. This research was funded by the DFG Collaborative Research Centre SFB 732 (Gabriella Lapesa) and the DFG Heisenberg Fellowship SCHU-2580/1-1 (Sabine Schulte im Walde).

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*, pages 1–10.

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.

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.

James Curran. 2003. *From distributional to semantic similarity*. Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

Philip Edmonds and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

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.

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.

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

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. Hutchison, David A. Balota, James H. Neely, Michael J. Cortese, Emily R. Cohen-Shikora, Chi-Shing Tse, Melvin J. Yap, Jesse J. Bengson, Dale Niemeyer, and Erin Buchanan. 2013. The semantic priming project. *Behavior Research Methods*, 45(4):1099–1114.

John. S. Justeson and Slava M. Katz. 1992. Redefining antonymy: The textual structure of a semantic relation. *Literary and Linguistic Computing*, 7(3):176–184.

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.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1*, pages 75–79.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

Scott McDonald and Chris Brew. 2004. A distributional model of semantic context effects in lexical processing. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 17–24.

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.

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

Yves Peirsman, Kris Heylen, and Dirk Speelman. 2008. Putting things in order. First and second order context models for the calculation of semantic similarity. In *JADT 2008: 9es Journées internationales d'Analyse statistique des Données Textuelles*.

Reinhard Rapp. 2002. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, pages 1–7.

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.

Magnus Sahlgren. 2008. The distributional hypothesis. *Rivista di Linguistica (Italian Journal of Linguistics)*, 20(1):33–53.

Silke Scheible, Sabine Schulte im Walde, and Sylvia Springorum. 2013. Uncovering Distributional Differences between Synonyms and Antonyms in a Word Space Model. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 489—497.

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

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

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference of Computational Linguistics*, pages 1015–1021, Geneva, Switzerland.