Cumulative Contextual Facilitation in Word Activation and Processing: Evidence from Distributional Modelling

Diego Frassinelli (diego.frassinelli@ims.uni-stuttgart.de)
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B, 70569 Stuttgart, Germany

Frank Keller (keller@inf.ed.ac.uk)
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

Information provided by the linguistic context has been shown to have a strong facilitatory effect on the activation and processing of upcoming words. The studies described in this paper aim to model the relation between context and target words using a distributional semantic model. We report three modelling studies in which we show that this model can successfully capture context effects in human-generated data (reading times and association scores).

Keywords: word processing; contextual effects; feature overlap; distributional semantics.

Introduction

Previous work has shown that linguistic context facilitates the activation and processing of upcoming words (e.g., Federmeier & Kutas, 1999). This facilitatory effect can be described in terms of *feature overlap:* the linguistic context activates a set of semantic features, restricting the set of possible upcoming words. The words that match these features are pre-activated and thus processed more quickly when encountered. The feature overlap account predicts that contextual facilitation is *cumulative:* the more biasing words are present in the context, the faster the target word should be processed, as overlapping features accumulate.

Frassinelli, Keller, and Scheepers (2013) studied this facilitation effect by manipulating the number of contextual words that are highly related to a target word occurring at the end of the sentence. They performed a self-paced reading and a visual world paradigm study in which they analysed linguistic contextual constraints on the processing of word meaning.

The aim of the modelling studies in this paper is to capture the effect of context reported in Frassinelli et al. (2013) using a distributional semantic model (DSM, Turney & Pantel, 2010). The "strong version" of the Distributional Hypothesis posits the cognitive validity of those models: DSMs provide an insight into the internal representation and structure of the lexicon in the brain (Lenci, 2008). In recent decades, DSMs have become very popular in psycholinguistics where they are used to successfully model different aspects of human language acquisition and processing such as: vocabulary acquisition (Landauer & Dumais, 1997), category-related deficits (Vigliocco, Vinson, Lewis, & Garrett, 2004), and semantic priming (Lapesa & Evert, 2013).

In this paper we present three studies in which we model several aspects of contextual processing using the bag-of-words distributional semantic model developed by Mitchell (2011). DSMs represent the meaning of a word as a multidimensional vector: each dimension of the vector corresponds to a word co-occurring with the target word in a corpus. Similar to semantic properties, the vector dimensions describe specific aspects of a word that contribute to the meaning of that word. In this framework, two words are similar if they appear in similar contexts, and, consequently, if their vector dimensions overlap. Traditionally, the overlap between two words has been computed in terms of the geometrical distance between the word vectors. We will show that it is possible to describe the relation between low and high biasing words and the target word in terms of vector similarity scores.

Previous Work

In order to manipulate the effect of context on word processing in a self-paced reading study, Frassinelli et al. (2013) constructed linguistic materials that vary the context words that bias the processing of a target word. Their stimuli had the following structure:

(1) *location* – *actor* – verb – *object* – **target** - spill-over area

For each of the 24 target words, they identified three context words highly related to it (high-biasing words, HB) and three context words unrelated to it (low-biasing words, LB). Four possible combinations of HB and LB context words were then used to construct the sentential context, as illustrated by the following examples:

(2) `Zero HB words` (*None*): On the *path*, the *man* was holding a *box* full of **mushrooms** carefully.

(3) `One HB word` (*One*): In the *forest*, the *man* was holding a *box* full of **mushrooms** carefully.

(4) `Two HB words` (*Two*): In the *forest*, the *picker* was

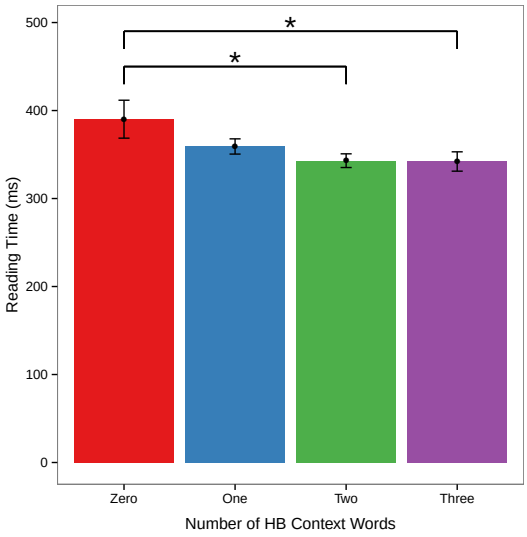

Figure 1: Plot of the reading times (with SE) averaged by the number of HB context words (from Frassinelli et al., 2013).

holding a *box* full of **mushrooms** carefully.

(5) `Three HB words` (*Three*): In the *forest*, the *picker* was holding a *basket* full of **mushrooms** carefully.

The plausibility of the resulting sentences and the biasing effect exerted by the contexts were carefully tested in a series of norming studies.

Frassinelli et al. (2013) then investigated the effect of HB context words on the reading time (RT) at the target word. Figure 1 reports RTs averaged by condition with standard errors (SE). Significant differences occur only between the conditions *None* and *Two* ($\beta_{\text{Two}} = -.10$, $p < .05$) and the conditions *None* and *Three* ($\beta_{\text{Three}} = -.09$, $p < .05$). Overall, Frassinelli et al. (2013) results show that the time required to read a target word decreases with increasing amount of biasing context.

The Model

For the studies reported in the next sections, we use a re-implementation (by Blacoe & Lapata, 2012) of the bag-of-words distributional model developed by Mitchell (2011).

We trained this model on the lemmatised and part-of-speech tagged version of ukWaC, an English corpus of two billion tokens extracted from the Web (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009). The use of this corpus provided full coverage of the experimental items.

Mitchell presented an evaluation of four association measures to weight vector components: conditional probabilities, point-wise mutual information, ratios of probabilities, and positive point-wise mutual information. We compared these four measures on our data and overall positive point-wise mutual information (posPmi) obtained the best results. It is defined as follows:

$$posPmi = \max(0, \log(\frac{freq_{ct} freq_{total}}{freq_c freq_t})) \quad (1)$$

where $freq_{ct}$ is the frequency of the target t in the context c; $freq_t$ is the overall frequency of t; $freq_c$ is the overall frequency of c; and $freq_{total}$ is the total frequency of all the words. The substitution of negative values with zeros in the posPmi model makes this association measure more suitable for dealing with sparsity and low frequency words, as they occur in our dataset (Mitchell, 2011, p. 44).

Moreover, we experimented with changing vector dimensionality of the model (from 1,000 up to 50,000 dimensions). We found that model performance stabilizes at 30,000 dimensions and shows no further improvement at higher dimensions. In this work, we therefore report only the results obtained with vectors of 30,000 dimensions.

Study 1: Predicting Reading Times

Aim In this study we test the bag-of-words DSM of Mitchell (2011) by using it to predict the reading times collected by Frassinelli et al. (2013). The authors showed that the amount of time required to read the target word decreases (though not linearly) with an increase of the number of biasing words in the context. In modelling terms, this means that the averaged distance between the context vectors and the vector of the target word should decrease when we increase the number of high biasing words in the context. Conversely, when we increase the number of low biasing words in the context, the distance should increase.

Method In the self-paced reading experiment, the authors analysed the effect of context on word meaning averaging the RTs based on the number of HB words available (from zero up to three) (see Previous Work and Figure 1). In this study, we compute the distance between each context vector and the vector of the target word (e.g., for condition *Three*: forest-mushroom, picker-mushroom, basket-mushroom), we average the three resulting cosines and we average again by condition as for the RTs.

Results Figure 2 presents the average cosine distance per condition. It shows that an increasing amount of biasing context produces a reduction in the distance between the context and the target word vectors. Table 1 reports the coefficients of a linear mixed effects (LME, version 1.1-7) model analysis of these data (Baayen, Davidson, & Bates, 2008). The model has the *cosine distance* as the dependent variable, the *contextual condition* as main factor (contrast coded with the *Zero* condition as the reference level) and random slope and intercept under *Item*. The model shows a significant difference between condition *Zero* and all the other conditions. We also performed a Tukey post-hoc test to compare the conditions pairwise. The analysis shows a significant difference between all the conditions ($p < .001$).

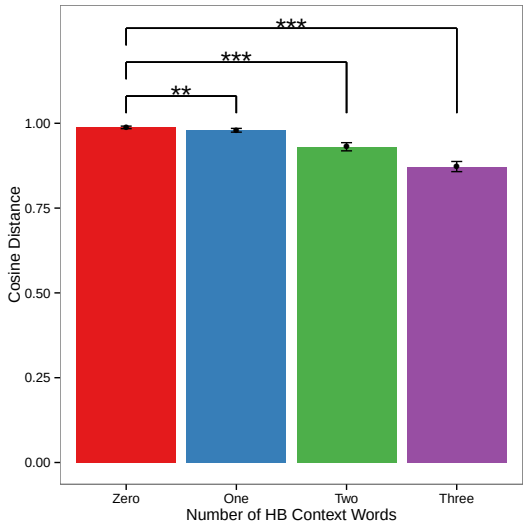

Figure 2: Study 1: Plot of the cosine distances (with standard errors) between context vector and target vector, averaged across all items with the same number of HB words.

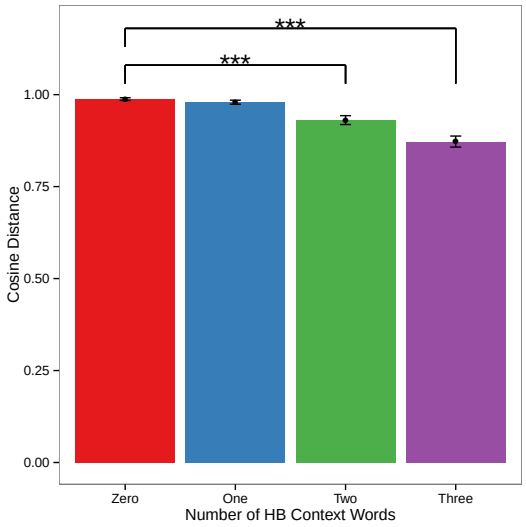

Figure 3: Study 2: Plot of the cosine distances (with standard errors) between pairs of context words, averaged across all items with the same number of HB words.

Predictor	β
(Intercept)	0.90^{***}
One	-0.03^{**}
Two	-0.08^{***}
Three	-0.11^{***}

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 1: Study 1: LME coefficients for data in Figure 2.

Predictor	β
(Intercept)	0.95^{***}
One	-0.01
Two	-0.06^{***}
Three	-0.09^{***}

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 3: Study 2: LME coefficients for data in Figure 3.

Discussion Figure 1 and Figure 2 allow a graphical comparison of the trends in the reading time experiment and in the cosine similarity study. The modelling study shows higher differences between conditions than those in the reading times produced by humans. Similar results have been described in the semantic priming literature, where it was found that Latent Semantic Analysis (LSA) predicts stronger effects than those observed in humans (Hare, Jordan, Thomson, Kelly, & McRae, 2009; Jones, Kintsch, & Mewhort, 2006). Overall, however, the modelling study captures the RT results well: cosine distance (as the reading time) decreases with increasing contextual bias.

So far we considered only the semantic relation between target and context words. We did not include in our analysis the relation between context words, but HB context words are likely to also be related to each other. The relation between context words alone (without a need for the target) could produce a similar effect on the cosine distance as we see in this study. In Study 2 we test this assumption.

Study 2: Predicting the Relation between Context Words

Aim The aim of this study is to model the interaction between context words without taking into account the effect of the target word. We also compare the outcome of this study with Study 1 in order to understand the relationship occurring between context words in isolation and between those words and the target.

Method We computed the cosine distance between each pair of context words (e.g., for condition *Three*: forest-picker, picker-basket, forest-basket) and we averaged them. The resulting cosines were averaged again by condition.

Results Figure 3 shows the average cosine distance per condition. Similarly to the previous results, the increasing amount of biasing context produces a reduction in the distance between the context vectors. Table 3 reports the LME coefficients for these data. The model structure is the same as in the previous study. The difference between condition

Predictor	β Model 1	β Model 2	β Model 3	β Model 4
(Intercept)	2.7^{***}	2.5^{***}	2.6^{***}	2.7^{***}
AvgTarget	-15.5^{***}		-10.7^{***}	-8.5^{***}
AvgContext		-13.4^{***}	-7.7^{***}	-10.1^{***}
AvgTarget:AvgContext				-81.6^{***}

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 2: LME coefficients for the model comparison study.

Zero and *One* is not significant, while all the other differences are strongly significant ($p < .001$). A post-hoc analysis shows significant differences between condition *Two* and *Three* ($p = .007$) and all the remaining conditions ($p < .001$).

A rank correlation analysis between the cosines for each item described in Study 1 and Study 2 shows an association of medium strength ($\rho = .65$, $p < .05$). In order to test the interaction of the information captured by the two models we performed a series of LME analyses treating the *contextual condition* as the dependent variable, the average cosines from Study 1 (*AvgTarget*) and Study 2 (*AvgContext*) as continuous predictors, random slopes and intercepts under *Item*. To reduce collinearity in our analyses, we centered the predictors and we assessed the collinearity between them estimating the conditional number κ. As suggested in Baayen (2008), $\kappa = 3.3$ indicates the absence of collinearity between the predictors.

Four different LME analyses including different combinations of these predictors were performed. Table 2 reports the coefficients for each model (1–4) and shows significant effects for all the predictors. In Model 1 we included only *AvgTarget* as predictor, while in Model 2 we included only *AvgContext*. In Model 3 we included both the predictors but not their interaction. Model 4 is the most complex one and includes the interaction between *AvgTarget* and *AvgContext*.

A model selection procedure that compares the log likelihoods of the different models was performed in order to determine the model that provides the best fit to the data. Both the comparison of Model 1 and Model 3 and the comparison of Model 2 and Model 3 show that the inclusion of both the predictors (as in Model 3) significantly enhances the accuracy of the model ($p < .001$). On the other hand, the comparison of Model 3 and Model 4 does not show any significant difference (p=.08) and suggests that the inclusion of the interaction between the two predictors does not improve the fit significantly.

Discussion Similarly to what we described in the previous modelling study, we show a strong relationship between HB context words that produces a significant reduction in the average cosine distance when increasing the amount of HB information provided. The model selection procedure indicates that the inclusion of both contextual relations and target-context relations is essential to improve the fit of the model to the data.

The basic assumption behind this study is that HB context words are not only highly related to the target word but also strongly associated to each other. In order to test this hypothesis experimentally, in Experiment 1 we collected association scores between context words. The outcome of this experiment should shed more light on the relation between contextual words. Moreover, in Study 3 we model these scores using the distributional model.

Experiment 1: Association between Context Words

Aim The aim of the study is to test our assumption that highly biasing context words are not only related to the target word, but also to each other, thus explaining why a context-only model (Study 2) was able to predict the reading time data from Frassinelli et al. (2013).

Method This experiment was performed on Amazon Mechanical Turk. Subjects were required to rate how related two words are on a scale from 1 (not at all related) to 5 (very related). Subjects were all native speakers of English with an US account. They were paid $ 0.20 to produce 24 association scores and were allowed to complete only one hit from the same batch.

144 participants took part to the experiment producing a total of 3,456 association scores. Each item was evaluated by twelve subjects. Participants were asked to judge the relation between each combination of LB and HB context words.

Results On average, two high biasing words have the highest association score: 3.77 (SE $\pm$.05) out of 5. On the other hand, two low biasing words obtain the lowest score: 2.81 (SE $\pm$.04). The associations between a high biasing and a low biasing word (*HB − LB* and *LB − HB*) are 2.93 (SE $\pm$.05) and 2.96 (SE $\pm$.05) respectively. A LME analysis was performed. The only significant difference is between *LB−LB* and *HB−HB* words ($\beta_{\text{HighBias}} = 0.94$, $p < 0.001$). We performed a post-hoc analysis to compare pairwise the different conditions. The *HB−HB* condition obtained significantly higher association scores than the other conditions ($p < .001$).

Discussion The aim of this study was to analyse the semantic relationship between pairs of context words. In this experiment we directly evaluated this relation without the influ-

ence of the target word. The *HB−HB* condition obtained the highest scores, while the *LB−LB* condition the lowest ones. The mixed situations (*HB−LB* and *LB−HB*) are negatively biased by the presence of the LB property in the pair. These results highlights the fact that the association of a HB word with a LB word is similar to the association of two LB words even though HB words have more specific meanings than LB words.

Overall, this study shows that the words that are highly related to the target word are also highly related with each other. LB words, in contrast, being more general words, are less strongly associated with each other and also with HB words.

Study 3: Predicting Association between Context Words

Aim Here, we use the model from Study 1 and Study 2 for the task of predicting human generated association scores between two contextual words (rather than RTs as in Study 2). From Experiment 1 it emerged that context words that are highly related to the target are also highly related to each other. We therefore expect the semantic similarity between two high biasing context words to be higher than the similarity between two low biasing context words.

Task In Experiment 1, participants had to evaluate on a scale from 1 (not related) to 5 (completely related) the association between two contextual words (both high and low biasing words). The model has to correctly predict the 3,456 human-generated association scores.

Method We computed the cosine similarity between pairs of context words.

Results A LME analysis was performed. The *association scores* are the dependent variable, the *cosine similarities* (posPmi, dim=30,000) the continuous factor, *subject* and *item* the random slopes and intercepts. The model shows a significant positive relation between word similarity and association scores ($\beta_{\text{Cosine}} = 6.431$, $p < 0.001$).

Discussion This study demonstrated a positive relationship between the human association scores and model similarity scores for context words. This indicates that the model successfully captures the relation between context words we found in the association study (Experiment 1). Words that are highly related to the target word are also strongly associated with each other because they occur in similar contexts.

Taken together, the results from this study and Experiment 1 confirm the assumptions underlying in Study 2. They provide evidence that the facilitation effect found in the RT data of Frassinelli et al. (2013) cannot be attributed solely to the relationship between the context words and the target; the relationship of the context words with each other also plays a role.

General Discussion

The studies reported in this paper aimed to test if a bag-of-words DSM can capture the relation between context and target words and, consequently, describe context effects on word processing in terms of feature overlap. We conducted three studies where we analysed different aspects of the target–context relation.

In Study 1 we showed that the DSM can successfully predict the RTs from Frassinelli et al. (2013). We averaged the cosine distance between each context vector and the target vector and we averaged again the resulting values by condition. When increasing the amount of HB context words the cosine distance significantly decreases.

In Study 2 we modelled the relation between context words without the influence of the target. The outcomes of this model show that the inclusion of more HB information reduces the cosine distance in the different conditions. In order to test the validity of these results we collected association scores between pairs of context words. As shown in Study 3 the model can successfully predict these scores indicating that words that are highly related to the target word are also highly related to each other. The model comparison study highlights the importance of including both these relations in order to have the best fit of our data.

Overall, this study provides support for feature overlap theory by showing that contextual facilitation is cumulative, i.e., it increases with the number of highly biasing context words. This theory however does not account for the context-only effect that we found in Study 2 and Experiment 1.

We demonstrated that the accumulation of semantic features can be modeled as the combination of the distributional vectors of the context words. Distributional semantics therefore provides a computational implementation of feature overlap theory, with semantic features represented as vectors components (i.e., word co-occurrences).

Acknowledgments

We are grateful to William Blacoe for making available the distributional model used in this paper. The support of the European Research Council under award number 203427 “Synchronous Linguistic and Visual Processing” is gratefully acknowledged.

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. New York: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–231.

Blacoe, W., & Lapata, M. (2012). A Comparison of Vector-based Representations for Semantic Composition. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning.* (pp. 546–556).

Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, *41*(4), 469–495.

Frassinelli, D., Keller, F., & Scheepers, C. (2013). The Effect of Incremental Context on Conceptual Processing : Evidence from Visual World and Reading Experiments. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 460–466). Berlin.

Hare, M., Jordan, M. I., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition*, *111*(2), 151–167.

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552.

Landauer, T., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211–240.

Lapesa, G., & Evert, S. (2013). Evaluating Neighbor Rank and Distance Measures as Predictors of Semantic Priming. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 66–74).

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica*, *20*(1), 1–31.

Mitchell, J. (2011). *Composition in distributional models of semantics*. Unpublished doctoral dissertation, University of Edinburgh.

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning : Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*(4), 422–488.