The Direct Route: Mediated Priming in Semantic Space

Will Lowe
wlowe02@tufts.edu
Center for Cognitive Studies
Tufts University; 11 Miner Hall
Medford MA 02155 USA

Scott McDonald
scottm@cogsci.ed.ac.uk
Institute of Adaptive and Neural Computation
University of Edinburgh; 2 Buccleuch Place
Edinburgh EH8 9LW UK

Abstract

McKoon and Ratcliff (1992) presented a theory of mediated priming where the priming effect is due to a direct but weak relatedness between prime and target. They also introduced a quantitative measure of word relatedness based on pointwise mutual information (Church and Hanks, 1990), and showed that stimuli chosen with the measure produced graded priming effects as predicted by their theory. Using stimuli from Balota and Lorch (1986), Livesay and Burgess (1998a,b) replicated the mediated priming effect in humans, but found that in HAL, a corpus-derived semantic space (Lund et al., 1995), mediated primes were in fact further from their targets than unrelated words. They concluded from this that mediated priming is not due to direct but weak relatedness. In this paper we present an alternative semantic space model based on earlier work (McDonald and Lowe, 1998). We show how this space allows a) a detailed replication of Ratcliff and McKoon's experimental results using their stimuli and b) a replication of Livesay and Burgess's human experimental results showing mediated priming. We discuss the implications for theories of mediated priming.

Mediated Priming

Mediated priming is an important test for theories of semantic memory (Neely, 1991). According to spreading activation theory (e.g. Anderson, 1983), when a word is presented it activates its representation in a network structure in which semantically related words are directly connected; more generally, the semantic similarity of two words depends on the number of links that must be traversed to reach one to the other. The level of activation controls the amount of facilitation received by the corresponding word. Although ultimately every word can be reached from any location in the network, activation decays during memory access so only a few of the most related words are facilitated. Spreading activation theories predict that a prime word should facilitate pronunciation or lexical decision on a target word directly, for example when "tiger" facilitates "stripes". Spreading activation theory also predicts that "lion" will facilitate "stripes" when activation spreads from the representation of "lion" to that of "stripes", via the related concept of tiger (de Groot, 1983; Neely, 1991).

Small but reliable mediated priming effects have been demonstrated for pronunciation tasks though they are less reliable for lexical decision (Balota and Lorch, 1986). Spreading activation theory explains the size of the priming effect by arguing that "lion" and "stripes" are only indirectly related in semantic memory so that activation has decayed significantly by the time activation from "lion" reaches "stripes".

Theories that do not assume the existence of activation or a network structure in semantic memory, e.g. compound cue theory (Ratcliff and McKoon, 1988; McKoon and Ratcliff, 1998), cannot take advantage of either of the priming explanations above. In compound cue theory, direct priming is explained roughly as follows: the prime and target are joined in a compound cue that is compared to representations in long-term memory. The comparison process generates a 'familiarity' value which controls the size of the priming effect. The essential feature of this explanation is that, unlike spreading activation theory, there is no mention of the intermediate representation "tiger" when explaining how "lion" facilitates "stripes". But is less clear how compound cue theory should explain mediated priming.

In response to this difficulty, McKoon and Ratcliff (1992) have argued that the mediated priming effects are not due to activation spreading through an intervening representation, but are the result of direct but weak relatedness between the prime and target words. To address the issue of priming effect magnitude they provided a quantitative method for generating prime target pairs with various degrees of relatedness. The method is based on pointwise mutual information (Church and Hanks, 1990) computed over a corpus. McKoon and Ratcliff's (1992) Experiment 3 showed that their method produced stimuli that reliably generated a range of priming effect sizes, and that the effect sizes could be controlled. They then argued that mediated priming is simply a special case of graded priming.

Livesay and Burgess (1998a,b) replicated the mediated priming effect in human subjects using a pronunciation task, but had less success with lexical decision (the same situation that was reported in Balota and Lorch's original paper). In an attempt to understand the nature of the priming mechanism they found that mediated primes from the Balota and Lorch stimuli could be divided heuristically into contextually appropriate and contextually inappropriate word pairs. Subsequent analysis revealed that only contextually appropriate pairs were responsible for generated a priming effect.

They then compared distances between each type of prime (direct or mediated) and their targets in HAL, a semantic space model (Lund et al., 1995). Burgess and colleagues have argued that distances in HAL reflect semantic relatedness; shorter distances reflect greater semantic relatedness (Burgess et al., 1998). Directly related primes were on average closer to their targets than the corresponding unrelated primes, so HAL successfully replicated the direct priming effect. However, both contextually appropriate and contextu-

ally inappropriate mediated primes were *further* from their targets than unrelated controls. Thus distances in HAL predict that the mediated primes should slow responses to their targets, relative to an unrelated word baseline. Subsequent analysis showed that even for contextually consistent primes, greater distance correlated 0.6 with larger priming effects.

Livesay and Burgess concluded that mediated priming could not be due to direct but weak relatedness between mediated primes and their targets on the grounds that HAL predicted the wrong effect. They then explored the possibility, suggested in McKoon and Ratcliff's paper, that mediated priming is determined by raw co-occurrence frequencies between prime words and their targets, but found no significant correlations.

Below we present replications of two priming experiments using a semantic space model. In Experiment 1 we replicate human performance on the stimuli generated by McKoon and Ratcliff using pointwise mutual information. We will refer to these stimuli as the mutual information stimuli. These results demonstrate that McKoon and Ratcliff's direct theory of mediated priming is consistent with explanations of priming based on semantic space. In Experiment 2 we tackle mediated priming directly by replicating the results of Livesay and Burgess's mediated priming experiment. From these two experiments we argue that our semantic space constitutes a model of mediated priming that is 'direct' in the way that McKoon and Ratcliff suggested.

Experiment 1

Materials

In this experiment we use materials from McKoon and Ratcliff's Experiment 3. Each target (e.g "grass") has a prime taken from association norms ("green"), a high-t prime ("acres") and a low-t prime ("plane"). High and low-t primes were chosen by first calculating a measure of lexical association based on the T-statistic between each target word and a large number of candidate primes (Church and Hanks, 1990, see Appendix A for details). McKoon and Ratcliff divided the candidate primes for each target into those with high values of the T-statistic (high-t primes) and low values (low-t primes). Unrelated primes were related primes from another target.

Methods

We constructed a semantic space from 100 million words of the British National Corpus, a balanced corpus of British English (Burnage and Dunlop, 1992). Word vectors were generated by passing a moving window through the corpus and collecting co-occurrence frequencies for 536 of the most reliable context words within a 10 word window either side of each stimulus item. Appendix B describes the method of choosing reliable context words. We used positive log odds-ratios to measure the amount of lexical association between each context word and each of the experimental stimuli.

A brief justification of the positive log odds-ratio as a measure of lexical association is appropriate at this point: Table 1 describes the true co-occurrence probabilities for a stimulus word t and context word c. $p(c,\neg t)$ is the probability of seeing c with a word other than t. The odds of seeing t rather than some other word when c is present are $p(c,t)/p(c,\neg t)$ and the odds of seeing t in the absence of c are $p(\neg c,t)/p(\neg c,\neg t)$, so if the presence of c increases the probability of seeing t then

Table 1: The true probabilities of seeing combinations of words t and c in text. $p(c,t)$ is the probability of seeing words c and t together in a window. $p(c,\neg t)$ is the probability of seeing c together with a word that it *not* t.

	Target	Non-target
Context	$p(c,t)$	$p(c,\neg t)$
Non-context	$p(\neg c,t)$	$p(\neg c,\neg t)$

the odds ratio

$$\theta(c,t) = \frac{p(c,t)/p(c,\neg t)}{p(\neg c,t)/p(\neg c,\neg t)} = \frac{p(c,t)\ p(\neg c,\neg t)}{p(c,\neg t)\ p(\neg c,t)}$$

is greater than 1. When $\theta > 1$ c and t are said to be positively associated. In contrast, if the presence of c makes it *less* likely that t will occur then $\theta < 1$ and c and t are negatively associated. Finally, when the presence of c makes no difference to the probability of seeing t then $\theta = 1$ and we can conclude that c and t are distributionally independent.

An important advantage of the odds ratio for measuring lexical association is that takes into account differing marginal word frequencies. For example, consider two target words t_1 and t_2 that have baseline occurrence probabilities $p(t_1)$ and $p(t_2)$. For simplicity we assume that co-occurrences are counted in a window extending exactly one word to one side of stimulus. When neither word is related to a context word c then all three words will distributionally independent. Under distributional independence the expected values of co-occurrence counts $f(c,t_1)$ and $f(c,t_2)$ depend only on their occurrence probabilities:

$$\begin{aligned} E[f(c,t_1)] &= p(c)\ p(t_1)\ N \\ E[f(c,t_2)] &= p(c)\ p(t_2)\ N \end{aligned}$$

where N is the number of words in the corpus[1]. If $p(t_1)$ is much larger than $p(t_2)$ then the expected co-occurrence counts may differ substantially, despite the fact that c has no relation to t_1 or t_2. In other words if raw co-occurrence counts are used to measure lexical association then a more frequent target word will be judged more strongly associated with c than a less frequent target word, whether or not they are actually related. Also, the fact that vector elements for two target words with different frequencies will be tend to have different magnitudes will bias the Euclidean distance measure to treat target words from different frequency bands as further away from each other than those in the same band. This occurs because the measure depends on squared differences between vector elements.

The odds ratio is well-known to be a measure of association that takes chance co-occurrence into account (Agresti, 1990). When t_1 and c are distributionally independent then $p(t_1,c) =$

[1]Strictly speaking N is the number of bigrams in the corpus, which is one less than the number of words.

$p(t_1)p(c)$. The odds ratio is

$$\theta(c,t_1) = \frac{p(c)p(t_1)\ p(\neg c)p(\neg t_1)}{p(c)p(\neg t_1)\ p(\neg c)p(t_1)} = 1,$$

and it is clear that the value of $\theta(c,t_1)$ does not depend on target and context word frequencies.

$\theta(c,t_1)$ is estimated from a corpus by setting the elements of Table 1 to their Maximum Likelihood values. The odds ratio estimate can then be computed using only occurrence and co-occurrence frequencies (see e.g. Agresti, 1990)

$$\hat{\theta}(c,t) \quad = \quad \frac{f(c,t)\ f(\neg c,\neg t)}{f(c,\neg t)\ f(\neg c,t)}.$$

We log the odds ratio to make the measure symmetric around 0 (denoting distributionally independent words) and set all negative odds-ratios to zero. This reflects our belief that information about the whether a word occurs with another *more* often than chance is psychologically salient, whereas the knowledge that a word tends *not* to occur with some other word (one of, say, 60,000 others in the lexicon) is not psychologically salient and need not be represented in the model. Empirical studies show that neither logging nor truncation of the basic odds-ratio measure make much difference to the results presented below. The most important step seems to be taking into account chance when using co-occurrence to quantify lexical association. The g-score (Dunning, 1993) is another useful measure for this purpose (McDonald and Lowe, 1998).

We created vectors for each of the experimental stimuli by calculating lexical association values between it and each context word. Unrelated primes were primes from the previous target word[2]. We use the cosine of the angle between word vectors as a similarity measure corresponding to semantic relatedness (McDonald and Lowe, 1998).

When modeling priming experiments, the cosine between a prime and its target should be inversely proportional to the corresponding reaction time. The size of a priming effect is calculated by subtracting the cosine between the unrelated prime and target from the cosine between the related prime and target. Cosines are entered directly into analyses of variance.

Results

McKoon and Ratcliff's subjects responded fastest to target words preceded by an associated prime, next fastest to a high-t prime, slower to a low-t prime and slowest of all to an unrelated prime (see Table 2, line 1.) Priming effects were reliable in all except the low-t condition.

The cosine similarity measure shows similar results (see Table 2, line 2). The following analyses are for items only since there are no subjects. The prime conditions were significantly different, F(3,156)=33.32, p<.001 so we performed pairwise analyses of variance to examine the differences more closely, correcting for multiple comparisons according to the Bonferroni method. There was a reliable associative priming effect: associated pairs were significantly more related than non-associated pairs (0.412 vs. 0.078), F(1,78)=80.645 p<.001 and high-t pairs were significantly more related than unrelated pairs (0.216 vs. 0.078), F(1,78)=19.727 p<.001. The mean value for low-t pairs was higher than the unrelated baseline (0.139 vs. 0.078), but this was not significant F(1,78)=5.268 p=.024.

[2]Since the stimuli have no inherent ordering, this will not produce any spurious effects. Other methods of choosing primes have been tested and give equivalent results.

Table 2: Mean reaction times in msec. (line 1) and cosines on (line 2) for the mutual information stimuli (from McKoon and Ratcliff, 1992)

	Related	High-t	Low-t	Unrelated
M&R	500	528	532	549
Space	0.412	0.216	0.139	0.078

Discussion

Experiment 1 shows a close fit to human reaction time data. The experiment also demonstrates that semantic space models are capable of representing the kind of weak but direct relatedness that McKoon and Ratcliff argue underlies mediated priming. If we can also account for mediated priming data, we will not only have uncovered additional evidence that direct but weak relatedness is sufficient to explain mediated priming, but also have found a 'direct' alternative explanation for the apparent mediation process. We address mediated priming in Experiment 2.

Experiment 2

Materials

Materials for Experiment 2 are taken from Balota and Lorch's (1986) paper. Each target (e.g. "stripes") has a directly related prime ("tiger") and a mediated prime ("lion"). One target had to be discarded because it had a prime with very low frequency in the corpus. A randomly chosen prime target combination was discarded from each of the other two prime conditions to maintain balance.

Method

The semantic space was the same as in Experiment 1.

Results

In the pronunciation task both Balota and Lorch and Livesay and Burgess's subjects showed direct and mediated priming (see Table 3, lines 1 and 2). The semantic space measure for related, mediated and unrelated pairs is shown in Table 3, line 3. The prime conditions were significantly different F(2,132)=12.065 p<.001 and we performed pairwise analyses of variance to examine the differences in more detail. There was a reliable direct priming effect (0.212 vs. 0.085), F(1,88)=24.105 p<.001 and also a reliable mediated priming effect of smaller magnitude (cosines 0.164 vs. 0.084), F(1,88)=13.107 p<.001.

Discussion

The results of Experiment 2 show that it is possible to model mediated priming using a semantic space. The experiment also demonstrates the plausibility of McKoon and Ratcliff's theory that direct but weak relatedness underlies mediated priming phenomena. There is no mediation mechanism in

Table 3: Mean reaction times in for the pronunciation experiments of Balota and Lorch (B&L, line 1) and Livesay and Burgess (L&B, line 2) in msec. Similarity measures for the same materials are on line 3.

	Related	Mediated	Unrelated
B&L Pron.	549	558	575
L&B Pron.	576	588	604
Space	0.212	0.164	0.084

the space, so the most parsimonious explanation of mediated priming is that it is due to direct relatedness.

On the other hand, Livesay and Burgess's distinction between contextually consistent and contextually inconsistent prime target pairs suggests an alternative view. Perhaps only some of the mediated priming stimuli are causing priming, and the rest are unnecessary.

Unfortunately the distinction between contextually consistent and inconsistent pairs appears to resist characterization in quantitative terms, e.g. in terms of distance in HAL. To investigate the possibility that a subset of primes were carrying the mediated priming effect we examined the distribution of differences between a) cosines between unrelated primes and their targets and b) mediated primes and their targets. The larger these differences are, the larger the mediated priming effect. If only a subset of materials carry the priming effect then we might expect that some targets have larger differences than the rest. However, we found that differences clustered symmetrically around the mean effect size. Ideally we would correlate priming effect size in milliseconds to the cosine measure to identify a subset of relevant primes; this is further work.

In an attempt to understand why HAL does not produce mediated priming, we attempted to replicate its behaviour on the mediated priming stimuli by changing the parameters of our semantic space. First, we used co-occurrence counts for the 536 reliable context words to create vectors for the Balota and Lorch materials and computed Euclidean distances between each prime and target combination. There were no significant differences between conditions, $F(2,132)=0.043$ $p=.958$. We then performed the same analysis with vectors normalized to length 1 to offset the effects of large co-occurrence counts. The conditions were still not reliably different $F(2,132)=1.257$, $p=.288$. However, in this case the model hinted at a direct priming effect and a smaller mediated effect. Finally we constructed vectors from 500 higher frequency context words[3], in case our choice of context words had adversely affected the measure. We used normalized vectors because they had previously given a slightly better match to the priming magnitudes. Again there was no significant difference between the conditions $F(2,132)=0.493$ $p=0.612$, but the model suggested a larger direct than mediated priming effect.

In conclusion, we were not able to replicate HAL's behaviour by changing the parameters of our model, so it is not easy to explain why the cosines in the space replicate human mediated priming effects while distances in HAL do not. It is possible that relevant differences between the space and HAL depend on HAL's method of choosing context words, or its window weighting function for collecting co-occurrence counts. Comparisons between the space and HAL are the subject of ongoing work.

[3]The context words had rank frequencies from 200 to 700. Occurrence frequencies ranged between 61926 to 220 occurrences per million.

Conclusion

In Experiments 1 and 2 we have presented detailed replications of human performance on graded and mediated priming stimuli using a semantic space. Since there is no mediation mechanism in the space we have argued that direct but weak relatedness, as reflected by the cosine measure in our space, is sufficient to yield a mediated semantic priming effect. This result supports McKoon and Ratcliff's contention that weak relatedness, rather than spreading activation, underlies mediated priming effects.

The results presented here stand in marked contrast to HAL's failure to generate mediated priming effects. However, we were not able to replicate HAL's behaviour in our model, so it is presently unclear why the HAL model does not work for this data.

We conclude by noting that graded and mediated priming can now be added to the list of psycholinguistic phenomena which may be accounted for by semantic space models.

Acknowledgments

WL is grateful to the Medical Research Council for funding, and to Daniel Dennett and the Center for Cognitive Studies at Tufts for providing a supportive and stimulating research environment. SM acknowledges the support of NSERC Canada and the ORS Awards Scheme.

References

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley and Sons.

Anderson, J. R. (1983). *The Architecture of Cognition*. Harvard University Press.

Balota, D. A. and Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning Memory and Cognition*, (12):336–345.

Burgess, C., Livesay, K., and Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, (25):211–257.

Burnage, G. and Dunlop, D. (1992). Encoding the British National Corpus. In *Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora*.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, (16):22–29.

de Groot, A. M. B. (1983). The range of automatic spreading activation in word priming. *Journal of Verbal Learning and Verbal Behavior*, pages 417–436.

Dunning, T. (1993). Accurate methods for the statistics for surprise and coincidence. *Computational Linguistics*, (19):61–74.

Finch, S. (1993). *Finding Structure in Language*. PhD thesis, Centre for Cognitive Science, University of Edinburgh.

Livesay, K. and Burgess, C. (1998a). Mediated priming does not rely on weak semantic relatedness or local co-occurrence. In *Proceedings of the Cognitive Science Society*, pages 609–614.

Livesay, K. and Burgess, C. (1998b). Mediated priming in high-dimensional meaning space: What is mediated in mediated priming? In *Proceedings of the Cognitive Science Society*, pages 436–441.

Lund, K., Burgess, C., and Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665. Mahwah, NJ: Lawrence Erlbaum Associates.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

McDonald, S. and Lowe, W. (1998). Modelling functional priming and the associative boost. In Gernsbacher, M. A. and Derry, S. D., editors, *Proceedings of the 20th Annual Meeting of the Cognitive Science Society*, pages 675–680, New Jersey. Lawrence Erlbaum Associates.

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

McKoon, G. and Ratcliff, R. (1998). Memory-based language processing: Psycholinguistic research in the 1990s. *Annual Review of Psychology*, (49):25–42.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In Besner, D. and Humphreys, G. W., editors, *Basic processes in reading: Visual word recognition*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Ratcliff, R. and McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, (95):385–408.

Appendix A

The pointwise mutual information or *association ratio* between a target word and candidate prime is

$$AR \quad = \quad \log_2 \frac{p(\text{prime and target})}{p(\text{prime})p(\text{target})}.$$

The numerator is estimated by normalizing the number of co-occurrences between prime and target words over the corpus. The denominator is estimated from the occurrence frequencies of the prime and target words separately. When prime and target words are distributionally independent *AR* should, like the log odds-ratio, take the value zero. When the prime word is occurs with the target more than would be expected by chance *AR* is positive with greater magnitude for greater levels of association. The T-statistic may used to determine whether the ratio is significantly different than 0, although Church and Hanks (1990) use the the value of the statistic itself as a lexical association measure. The *AR* measure is called pointwise mutual information in analogy to mutual information, a information theoretic measure which is the expectation of *AR* with respect the distribution $p(\text{prime and target})$. Manning and Schütze 1999 discuss uses and shortcomings of pointwise mutual information as an association measure.

Appendix B

We assume that the ease that two words can be substituted for one another in text reflects their semantic similarity. Substitutability in context, defined over word pairs or *targets*, is the underlying continuous quantity that a semantic space model needs to capture (Finch, 1993). Measuring substitutability in context entails holding linguistic context constant and swapping in targets. This is equivalent to holding targets constant and examining possible surrounding linguistic contexts because targets that are easily substitutable are those that occur in similar contexts.

Any large balanced corpus, such as the BNC, realizes a subset of the possible linguistic contexts that can surround a target. Given sufficient target instances the subset will be representative because the number of times a context surrounds a target is proportional to how meaningful the resulting sentence is. We represent contexts using finite set of *context words*. The linguistic contexts that surround a target are represented by the number of times each context word occurs within a 10 word window surrounding the target. These co-occurrence counts and the marginal frequencies of each context word and the target are used to create vectors of positive log odds ratios. To represent linguistic context adequately context words should be *reliable*.

To quantify reliability we treat context words like human raters and use standard ANOVA methods to assess their reliability: First, we choose several thousand candidate context words from the high frequency portion of the BNC (excluding stop words). Second, we pick randomly another set of words called meta-context words, and compute log odds ratios as described above for each context and meta-context word combination over k disjoint sections of the corpus. The resulting k matrices can be seen either as sets of column vectors describing the positions of the meta-context words in a space defined by the candidate context words, or as a set of row vectors describing the positions of the candidate context words in a space given by the meta-context words. The meta-context words are so-called because they are context words for the candidate context words. Each candidate context word is then associated with k vectors. We consider the vectors to be the results of k rating tasks and use a within subjects ANOVA to

test whether each context word generates significant variation in vector elements between the k tests. Context words that are reliable have k vectors with similar values so their rating do no vary significantly across corpus sections. Context words for which we cannot reject the null hypothesis of no variation between corpus sections are retained.

In these experiments we chose k=4 sections from the BNC, each containing 10M words, and used the rather conservative critical significance level 0.1. The procedure generated 536 context words.