Improving sparse word similarity models with asymmetric measures

Jean Mark Gawron
San Diego State University
gawron@mail.sdsu.edu

Abstract

We show that asymmetric models based on Tversky (1977) improve correlations with human similarity judgments and nearest neighbor discovery for both frequent and middle-rank words. In accord with Tversky's discovery that asymmetric similarity judgments arise when comparing sparse and rich representations, improvement on our two tasks can be traced to heavily weighting the feature bias toward the rarer word when comparing high- and mid-frequency words.

1 Introduction

A key assumption of most models of similarity is that a similarity relation is symmetric. This assumption is foundational for some conceptions, such as the idea of a similarity space, in which similarity is the inverse of distance; and it is deeply embedded into many of the algorithms that build on a similarity relation among objects, such as clustering algorithms. The symmetry assumption is not, however, universal, and it is not essential to all applications of similarity, especially when it comes to modeling human similarity judgments. Citing a number of empirical studies, Tversky (1977) calls symmetry directly into question, and proposes two general models that abandon symmetry. The one most directly related to a large body of word similarity work that followed is what he calls the **ratio model**, which defines $\mathrm{sim}(\mathrm{a},\mathrm{b})$ as:

$$\frac{f(\mathrm{A}\cap\mathrm{B})}{f(\mathrm{A}\cap\mathrm{B})+\alpha f(\mathrm{A}\backslash\mathrm{B})+\beta f(\mathrm{B}\backslash\mathrm{A})} \quad (1)$$

Here A and B represent feature sets for the objects a and b respectively; the term in the numerator is a function of the set of shared features, a measure of similarity, and the last two terms in the denominator measure dissimilarity: α and β are real-number weights; when $\alpha \neq \beta$, symmetry is abandoned.

To motivate such a measure, Tversky presents experimental data with asymmetric similarity results, including similarity comparisons of countries, line drawings of faces, and letters. Tversky shows that many similarity judgment tasks have an inherent asymmetry; but he also argues, following Rosch (1975), that certain kinds of stimuli are more naturally used as foci or standards than others. Goldstone (in press) summarizes the results succinctly: "Asymmetrical similarity occurs when an object with many features is judged as less similar to a sparser object than vice versa; for example, North Korea is judged to be more like China than China is [like] North Korea." Thus, one source of asymmetry is the comparison of sparse and dense representations.

The relevance of such considerations to word similarity becomes clear when we consider that for many applications, word similarity measures need to be well-defined when comparing very frequent words with infrequent words. To make this concrete, let us consider a word representation in the word-as-vector paradigm (Lee, 1997; Lin, 1998), using a dependency-based model. Suppose we want to measure the semantic similarity of *boat*, rank 682 among the nouns in the BNC corpus studied below, which has 1057 nonzero dependency features based on 50 million words of data, with *dinghy*, rank 6200, which has only 113 nonzero features. At the level of the vector representations we are using, these are events of very different dimensionality; that is, there are ten times as many features in the representation of *boat* as there are in the representation of *dinghy*. If in Tversky/Rosch terms, the more frequent word is also a more likely focus, then this is exactly the kind of situation in which asymmetric similarity judgments will arise. Below we show that an

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 296–301,
Baltimore, Maryland, USA, June 23-25 2014.

asymmetric measure, using α and β biased in favor of the less frequent word, greatly improves the performance of a dependency-based vector model in capturing human similarity judgments.

Before presenting these results, it will be helpful to slightly reformulate and slightly generalize Tversky's ratio model. The reformulation will allow us to directly draw the connection between the ratio model and a set of similarity measures that have played key roles in the similarity literature. First, since Tversky has primarily additive f in mind, we can reformulate $\mathrm{f}(\mathrm{A} \cap \mathrm{B})$ as follows

$$\mathrm{f}(\mathrm{A} \cap \mathrm{B}) = \sum_{f \in \mathrm{A} \cap \mathrm{B}} \mathrm{wght}(f) \tag{2}$$

Next, since we are interested in generalizing from sets of features, to real-valued vectors of features, w_1, w_2, we define

$$\sigma_{\mathrm{SI}}(w_1, w_2) = \textstyle\sum_{f \in w_1 \cap w_2} \mathrm{SI}(w_1[f], w_2[f]). \tag{3}$$

Here SI is some numerical operation on real-number feature values (SI stands for **shared information**). If the operation is MIN and $w_1[f]$ and $w_2[f]$ both contain the feature weights for f, then

$$\begin{aligned} \sum_{f \in \mathrm{A} \cap \mathrm{B}} \mathrm{wght}(f) &= \sigma_{\mathrm{MIN}}(w_1, w_2) \\ &= \textstyle\sum_{f \in w_1 \cap w_2} \mathrm{MIN}(w_1[f], w_2[f]), \end{aligned}$$

so with SI set to MIN, Equation (3) includes Equation (2) as a special case. Similarly, $\sigma(w_1, w_1)$ represents the summed feature weights of w_1, and therefore,

$$f(w_1 \backslash w_2) = \sigma(w_1, w_1) - \sigma(w_1, w_2)$$

In this generalized form, then, (1) becomes

$$\begin{aligned} &\frac{\sigma(w_1,w_2)}{\sigma(w_1,w_2)+\alpha[\sigma(w_1,w_1)-\sigma(w_1,w_2)]+\beta[\sigma(w_2,w_2)-\sigma(w_1,w_2)]} \\ &= \frac{\sigma(w_1,w_2)}{\alpha\sigma(w_1,w_1)+\beta\sigma(w_2,w_2)+\sigma(w_1,w_2)-(\alpha+\beta)\sigma(w_1,w_2)} \end{aligned} \tag{4}$$

Thus, if $\alpha + \beta = 1$, Tversky's ratio model becomes simply:

$$\mathrm{sim}(w_1, w_2) = \frac{\sigma(w_1,w_2)}{\alpha\sigma(w_1,w_1)+(1-\alpha)\sigma(w_2,w_2)} \tag{5}$$

The computational advantage of this reformulation is that the core similarity operation $\sigma(w_1, w_2)$ is done on what is generally only a small number of shared features, and the $\sigma(w_i, w_i)$ calculations (which we will call self-similarities), can be computed in advance. Note that $\mathrm{sim}(w_1, w_2)$ is symmetric if and only if $\alpha = 0.5$. When $\alpha > 0.5$, $\mathrm{sim}(w_1, w2)$ is biased in favor of w_1 as the referent; When $\alpha < 0.5$, $\mathrm{sim}(w_1, w2)$ is biased in favor of w_2.

Consider four similarity functions that have played important roles in the literature on similarity:

$$\begin{aligned} \text{DICE PROD}(w_1, w_2) &= \frac{2 * w_1 \cdot w_2}{\|w_1\|^2 + \|w_2\|^2} \\ \text{DICE}^{\dagger}(w_1, w_2) &= \frac{2 * \sum_{f \in w_1 \cap w_2} \min(w_1[f], w_2[f])}{\sum w_1[f] + \sum w_2[f]} \\ \text{LIN}(w_1, w_2) &= \frac{\sum_{f \in w_1 \cap w_2} w_1[f] + w_2[f]}{\sum w_1[f] + \sum w_2[f]} \\ \text{COS}(w_1, w_2) &= \text{DICE PROD applied to unit vectors} \end{aligned} \tag{6}$$

The function DICE PROD is not well known in the word similarity literature, but in the data mining literature it is often just called Dice coefficient, because it generalized the set comparison function of Dice (1945). Observe that cosine is a special case of DICE PROD. DICE† was introduced in Curran (2004) and was the most successful function in his evaluation. Since LIN was introduced in Lin (1998); several different functions have born that name. The version used here is the one used in Curran (2004).

The three distinct functions in Equation 6 have a similar form. In fact, all can be defined in terms of σ functions differing only in their SI operation.

Let σ_{SI} be a shared feature sum for operation SI, as defined in Equation (3). We define the Tversky-normalized version of σ_{SI}, written T_{SI}, as:[1]

$$\mathrm{T}_{\mathrm{SI}}(w_1, w_2) = \frac{2 \cdot \sigma_{\mathrm{SI}}(w_1, w_2)}{\sigma_{\mathrm{SI}}(w_1, w_1) + \sigma_{\mathrm{SI}}(w_2, w_2)} \tag{7}$$

Note that T_{SI} is just the special case of Tversky's ratio model (5) in which $\alpha = 0.5$ and the similarity measure is symmetric.

We define three SI operations σ_{PROD}[2], σ_{MIN}, and σ_{AVG} as follows:

SI	$\sigma_{\mathrm{SI}}(w_1, w_2)$
PROD	$\sum_{f \in w_1 \cap w_2} w_1[f] * w_2[f]$
AVG	$\sum_{f \in w_1 \cap w_2} \frac{w_1[f]+w_2[f]}{2}$
MIN	$\sum_{f \in w_1 \cap w_2} \mathrm{MIN}(w_1[f], w_2[f])$

[1] Paralleling (7) is Jaccard-family normalization:

$$\sigma_{\mathrm{JACC}}(w_1, w_2) = \frac{\sigma(w_1, w_2)}{\sigma(w_1, w_1) + \sigma(w_2, w_2) - \sigma(w_1, w_2)}$$

It is easy to generalize the result from van Rijsbergen (1979) for the original set-specific versions of Dice and Jaccard, and show that all of the Tversky family functions discussed above are monotonic in Jaccard.

[2] σ_{PROD}, of course, is dot product.

This yields the three similarity functions cited above:

$$\begin{aligned} \text{DICE PROD}(w_1, w_2) &= \text{T}_{\text{PROD}}(w_1, w_2) \\ \text{DICE}^{\dagger}(w_1, w_2) &= \text{T}_{\text{MIN}}(w_1, w_2) \\ \text{LIN}(w_1, w_2) &= \text{T}_{\text{AVG}}(w_1, w_2) \end{aligned} \quad (8)$$

Thus, all three of these functions are special cases of symmetric ratio models. Below, we investigate asymmetric versions of all three, which we write as $\text{T}_{\alpha,\text{SI}}(w_1, w_2)$, defined as:

$$\frac{\sigma_{\text{SI}}(w_1, w_2)}{\alpha \cdot \sigma_{\text{SI}}(w_1, w_1) + (1-\alpha) \cdot \sigma_{\text{SI}}(w_2, w_2)} \quad (9)$$

Following Lee (1997), who investigates a different family of asymmetric similarity functions, we will refer to these as α-skewed measures.

We also will look at a **rank-biased** family of measures:

$$\begin{aligned} &\text{R}_{\alpha,\text{SI}}(w_1, w_2) = \text{T}_{\alpha,\text{SI}}(w_h, w_l) \\ &\text{where } w_l = \arg\min_{w \in \{w_1, w_2\}} \text{Rank}(w) \\ &\qquad w_h = \arg\max_{w \in \{w_1, w_2\}} \text{Rank}(w) \end{aligned} \quad (10)$$

Here, $\text{T}_{\alpha,\text{SI}}(w_h, w_l)$ is as defined in (9), and the α-weighted word is always the less frequent word. For example, consider comparing the 100-feature vector for *dinghy* to the 1000 feature vector for *boat*: if α is high, we give more weight to the proportion of *dinghy*'s features that are shared than we give to the proportion of *boat*'s features that are shared.

In the following sections we present data showing that the performance of a dependency-based similarity system in capturing human similarity judgments can be greatly improved with rank-bias and α-skewing. We will investigate the three asymmetric functions defined above.[3] We argue that the advantages of rank bias are tied to improved similarity estimation when comparing vectors of very different dimensionality. We then turn to the problem of finding a word's nearest semantic neighbors. The nearest neighbor problem is a rather a natural ground in which to try out ideas on asymmetry, since the nearest neighbor relation is itself not symmetrical. We show that α-skewing can be used to improve the quality of nearest neighbors found for both high- and mid-frequency words.

[3]Interestingly, Equation (9) does not yield an asymmetric version of cosine. Plugging unit vectors into the α-skewed version of DICE PROD still leaves us with a symmetric function (COS), whatever the value of α.

2 Systems

1. We parsed the BNC with the Malt Dependency parser (Nivre, 2003) and the Stanford parser (Klein and Manning, 2003), creating two dependency DBs, using basically the design in Lin (1998), with features weighted by PMI (Church and Hanks, 1990).

2. For each of the 3 rank-biased similarity systems ($\text{R}_{\alpha,\text{SI}}$) and cosine, we computed correlations with human judgments for the pairs in 2 standard wordsets: the combined Miller-Charles/Rubenstein-Goodenough word sets (Miller and Charles, 1991; Rubenstein and Goodenough, 1965) and the Wordsim 353 word set (Finkelstein et al., 2002), as well as to a subset of the Wordsim set restricted to reflect semantic similarity judgments, which we will refer to as Wordsim 201.

3. For each of 3 α-skewed similarity systems ($\text{T}_{\alpha,\text{SI}}$) and cosine, we found the nearest neighbor from among BNC nouns (of any rank) for the 10,000 most frequent BNC nouns using the the dependency DB created in step 2.

4. To evaluate of the quality of the nearest neighbors pairs found in Step 4, we scored them using the Wordnet-based Personalized Pagerank system described in Agirre (2009) (UKB), a non distributional WordNet based measure, and the best system in Table 1.

3 Human correlations

Table 1 presents the Spearman's correlation with human judgments for Cosine, UKB, and our 3 α-skewed models using Malt-parser based vectors applied to the combined Miller-Charles/Rubenstein-Goodenough word sets, the Wordsim 353 word set, and the Wordsim 202 word set.

The first of each of the column pairs is a symmetric system, and the second a rank-biased variant, based on Equation (10). In all cases, the biased system improves on the performance of its symmetric counterpart; in the case of $\text{DICE}^{\dagger}$and DICE PROD, that improvement is enough for the biased system to outperform cosine, the best of the symmetric distributionally based systems. The value .97 was chosen for α because it produced the best α-system on the MC/RG corpus. That value

		MC/RG		Wdsm201		Wdsm353	
		$\alpha = .5$	$\alpha = .97$	$\alpha = .5$	$\alpha = .97$	$\alpha = .5$	$\alpha = .97$
Dice	DICE PROD	.59	.71	.50	.60	.35	.44
	LIN	.48	.62	.42	.54	.29	.39
	DICE†	.58	.67	.49	.58	.34	.43
Euc	Cosine	.65	NA	.56	NA	.41	NA
WN	UKB WN	.80	NA	.75	NA	.68	NA

Table 1: System/Human correlations. Above the line: MALT Parser-based systems

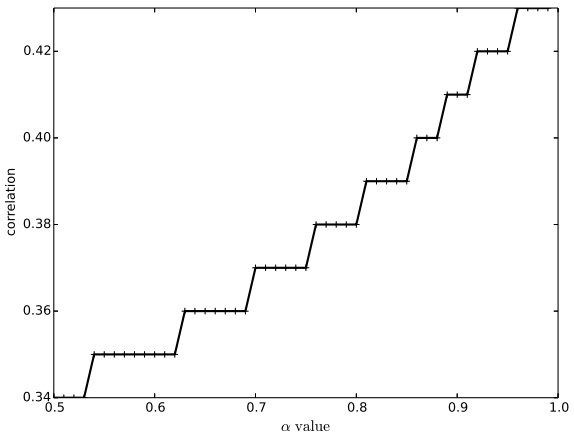

Figure 1: Scores monotonically increase with α

is probably probably an overtrained optimum. The point is that α-skewing always helps: For all three systems, the improvement shown in raising α from .5 to whatever the optimum is is monotonic. This is shown in Figure 1. Table 2 shows very similar results using the Stanford parser, demonstrating the pattern is not limited to a single parsing model.

In Table 3, we list the pairs whose reranking on the MC/RG dataset contributed most to the improvement of the $\alpha = .9$ system over the default $\alpha = .5$ system. In the last column an approximation of the amount of correlation improvement provided by that pair (δ):[4] Note the 3 of the 5 items contributing the most improvement this system were pairs with a large difference in rank. Choosing $\alpha = .9$, weights recall toward the rarer word. We conjecture that the reason this helps is Tversky's principle: It is natural to use the sparser representation as the focus in the comparison.

[4]The approximation is based on the formula for computing Spearman's R with no ties. If n is the number of items, then the improvement on that item is:

$$\frac{6 * [(\text{baseline} - \text{gold})^2 - (\text{test} - \text{gold})^2]}{n * (n^2 - 1)}$$

Word 1	Rank	Word 2	Rank	δ
automobile	7411	car	100	0.030
asylum	3540	madhouse	14703	0.020
coast	708	hill	949	0.018
mound	3089	stove	2885	0.017
autograph	10136	signature	2743	0.009

Table 3: Pairs contributing the biggest improvement, MC/RG word set

4 Nearest neighbors

Figure 2 gives the results of our nearest neighbor study on the BNC for the case of DICE PROD. The graphs for the other two α-skewed systems are nearly identical, and are not shown due to space limitations. The target word, the word whose nearest neighbor is being found, always receives the weight $1 - \alpha$. The x-axis shows target word rank; the y-axis shows the average UKB similarity scores assigned to nearest neighbors every 50 ranks. All the systems show degraded nearest neighbor quality as target words grow rare, but at lower ranks, the $\alpha = .04$ nearest neighbor system fares considerably better than the symmetric $\alpha = .50$ system; the line across the bottom tracks the score of a system with randomly generated nearest neighbors. The symmetric DICE PROD system is as an excellent nearest neighbor system at high ranks but drops below the $\alpha = .04$ system at around rank 3500. We see that the $\alpha = .8$ system is even better than the symmetric system at high ranks, but degrades much more quickly.

We explain these results on the basis of the principle developed for the human correlation data: To reflect natural judgments of similarity for comparisons of representations of differing sparseness, α should be tipped toward the sparser representation.

Thus, $\alpha = .80$ works best for high rank target words, because most nearest neighbor candi-

	MC/RG			Wdsm201			Wdsm353		
	$\alpha = .5$	opt	opt α	$\alpha = .5$	opt	opt α	$\alpha = .5$	opt	opt α
DICE PROD	.65	.70	.86	.42	.57	.99	.36	.44	.98
LIN	.58	.68	.90	.41	.56	.94	.30	.41	.99
DICE†	.60	.71	.91	.43	.53	.99	.32	.43	.99

Table 2: System/Human correlations for Stanford parser systems

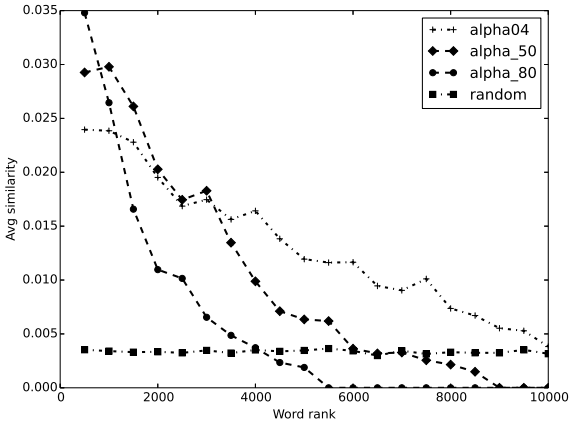

Figure 2: UKB evaluation scores for nearest neighbor pairs across word ranks, sampled every 50 ranks.

dates are less frequent, and $\alpha = .8$ tips the balance toward the nontarget words. On the other hand, when the target word is a low ranking word, a high α weight means it never receives the highest weight, and this is disastrous, since most good candidates are higher ranking. Conversely, $\alpha = .04$ works better.

5 Previous work

The debt owed to Tversky (1977) has been made clear in the introduction. Less clear is the debt owed to Jimenez et al. (2012), which also proposes an asymmetric similarity framework based on Tversky's insights. Jimenez et al. showed the continued relevance of Tversky's work.

Motivated by the problem of measuring how well the distribution of one word w_1 captures the distribution of another w_2, Weeds and Weir (2005) also explore asymmetric models, expressing similarity calculations as weighted combinations of several variants of what they call precision and recall. Some of their models are also Tverskyan ratio models. To see this, we divide (9) everywhere by $\sigma(w_1, w_2)$:

$$\mathrm{T}_{\mathrm{SI}}(w_1, w_2) = \frac{1}{\frac{\alpha \cdot \sigma(w_1,w_1)}{\sigma(w_1,w_2)} + \frac{(1-\alpha)\cdot\sigma(w_2,w_2)}{\sigma(w_1,w_2)}}$$

If the SI is MIN, then the two terms in the denominator are the inverses of what W&W call difference-weighted precision and recall:

$$\mathrm{PREC}(w_1, w_2) = \frac{\sigma_{\mathrm{MIN}}(w_1,w_2)}{\sigma_{\mathrm{MIN}}(w_1,w_1)}$$
$$\mathrm{REC}(w_1, w_2) = \frac{\sigma_{\mathrm{MIN}}(w_1,w_2)}{\sigma_{\mathrm{MIN}}(w_2,w_2)},$$

So for T_{MIN}, (9) can be rewritten:

$$\frac{1}{\frac{\alpha}{\mathrm{PREC}(w_1,w_2)} + \frac{1-\alpha}{\mathrm{REC}(w_1,w_2)}}$$

That is, $\mathrm{T}_{\mathrm{MIN}}$ is a weighted harmonic mean of precision and recall, the so-called weighted F-measure (Manning and Schütze, 1999). W&W's additive precision/recall models appear not to be Tversky models, since they compute separate sums for precision and recall from the $f \in w_1 \cap w_2$, one using $w_1[f]$, and one using $w_2[f]$.

Long before Weed and Weir, Lee (1999) proposed an asymmetric similarity measure as well. Like Weeds and Weir, her perspective was to calculate the effectiveness of using one distribution as a proxy for the other, a fundamentally asymmetric problem. For distributions q and r, Lee's α-skew divergence takes the KL-divergence of a mixture of q and r from q, using the α parameter to define the proportions in the mixture.

6 Conclusion

We have shown that Tversky's asymmetric ratio models can improve performance in capturing human judgments and produce better nearest neighbors. To validate these very preliminary results, we need to explore applications compatible with asymmetry, such as the TOEFL-like synonym discovery task in Freitag et al. (2005), and the PP-attachment task in Dagan et al. (1999).

Acknowledgments

This work reported here was supported by NSF CDI grant # 1028177.

References

E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of NAACL-HLT 09*, Boulder, Co.

K.W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

J.R. Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, University of Edinburgh. College of Science and Engineering. School of Informatics.

I. Dagan, L. Lee, and F.C.N. Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1):43–69.

L.R. Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

L. Finkelstein, E. Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

D. Freitag, M. Blume, J. Byrnes, E. Chow, S. Kapadia, R. Rohwer, and Z. Wang. 2005. New experiments in distributional representations of synonymy. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 25–32. Association for Computational Linguistics.

R. L. Goldstone. in press. Similarity. In R.A. Wilson Wilson and F. C. Keil, editors, *MIT Encylcopedia of Cognitive Sciences*. MIT Press, Cambridge, MA.

S. Jimenez, C. Becerra, and A. Gelbukh. 2012. Soft cardinality: A parameterized similarity function for text comparison. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 449–453. Association for Computational Linguistics.

D. Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, pages 3–10, Cambridge, MA. MIT Press.

L. Lee. 1997. *Similarity-based approaches to natural language processing*. Ph.D. thesis, Harvard University.

L. Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32. Association for Computational Linguistics.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Annual Meeting-Association for Computational Linguistics*, volume 36, pages 768–774. Association for Computational Linguistics.

C.D. Manning and H. Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge.

G.A. Miller and W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

J. Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03)*, pages 149–160.

E. Rosch and C. B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605.

H. Rubenstein and J.B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8:627–633.

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

C. J. van Rijsbergen. 1979. *Information retrieval*. Butterworth-Heinemann, Oxford.

J. Weeds and D. Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475.