An Empirical Evaluation of Models of Text Document Similarity

Michael D. Lee (michael.lee@adelaide.edu.au)
Department of Psychology, University of Adelaide
South Australia, 5005, AUSTRALIA

Brandon Pincombe (brandon.pincombe@dsto.defence.gov.au)
Intelligence Surveillance and Reconnaissance Division, Defence Science and Technology Organisation
PO Box 1500, Edinburgh SA 5111 AUSTRALIA

Matthew Welsh (matthew.welsh@adelaide.edu.au)
Australian School of Petroleum Engineering, University of Adelaide
South Australia, 5005, AUSTRALIA

Abstract

Modeling the semantic similarity between text documents presents a significant theoretical challenge for cognitive science, with ready-made applications in information handling and decision support systems dealing with text. While a number of candidate models exist, they have generally not been assessed in terms of their ability to emulate human judgments of similarity. To address this problem, we conducted an experiment that collected repeated similarity measures for each pair of documents in a small corpus of short news documents. An analysis of human performance showed inter-rater correlations of about 0.6. We then considered the ability of existing models—using word-based, n-gram and Latent Semantic Analysis (LSA) approaches—to model these human judgments. The best performed LSA model produced correlations of about 0.6, consistent with human performance, while the best performed word-based and n-gram models achieved correlations closer to 0.5. Many of the remaining models showed almost no correlation with human performance. Based on our results, we provide some discussion of the key strengths and weaknesses of the models we examined.

Introduction

Modeling the semantic similarity between text documents is an interesting problem for cognitive science, for both theoretical and practical reasons. Theoretically, it involves the study of a basic cognitive process with richly structured natural stimuli. Practically, search engines, text corpus visualizations, and a variety of other applications for filtering, sorting, retrieving, and generally handling text rely fundamentally on similarity measures. For this reason, the ability to assess semantic similarity in an accurate, automated, and scalable way is a key determinant of the effectiveness of most information handling and decision support software that deals with text.

A variety of different approaches have been developed for modeling text document similarity. These include simple word-based, keyword-based and n-gram measures (e.g., Salton, 1989; Damashek, 1995), and more complicated approaches such as Latent Semantic Analysis (LSA: Deerwester et al., 1990; Landauer and Dumais, 1997). While all of these approaches have achieved some level of practical success, they have generally not been assessed in terms of their ability to model human judgments of text document similarity. The most likely reason for this failure is that no suitable empirical data exist, and considerable effort is involved in collecting pairwise ratings of text document similarity for even a moderate number of documents. This paper reports the collection of data that give ten independent ratings of the similarity of every pair of 50 short text documents, and so represents an attempt to establish a 'psychological ground truth' for evaluating models. Using the new data, we report a first evaluation of the ability of word-based, n-gram and LSA approaches to model human judgments.

Experiment

Materials

The text corpus evaluated by human judges contained 50 documents selected from the Australian Broadcasting Corporation's news mail service, which provides text e-mails of headline stories. The documents varied in length from 51 to 126 words, and covered a number of broad topics. A further 314 documents from the same were collected to act as a larger 'backgrounding' corpus for LSA.

Both document sets were assessed against a standard corpus of five English texts using four models of language. These were the log-normal, generalized inverse Gauss-Poisson (with $\gamma = -0.5$), Yule-Simon and Zipfian models (Baayen, 2001). Both document sets were within the normal range of English text for word frequency spectrum and vocabulary growth and were therefore regarded as representative of normal English texts.

Subjects

The subjects were 83 University of Adelaide students (29 males and 54 females), with a mean age of 19.7 years. They were each paid with a ten (Australian) dollar gift voucher for every 100 document pair ratings made.

Procedure

Subjects were asked to read and judge the similarity of documents presented in pairs displayed side by side. The full text of each document was always displayed. For each pair, a subject indicated how similar they felt the documents were on a five-point scale (with one indicating "highly unrelated" and five indicating "highly related"). Once a judgement had been made, another pair of documents was presented and the process repeated. Each possible pair of documents (excluding self-comparisons) was presented between eight and twelve times[1]. The pairings were presented in a random order, and which documents were shown on the left and right was also randomly determined.

Basic Results

The distribution of ratings over all trials revealed a heavy skew towards low similarity values, with frequencies of about 0.64, 0.18, 0.10, 0.06 and 0.02 for the similarity responses 'one', 'two', 'three', 'four' and 'five' respectively.

To test for individual differences in similarity ratings, the difference between every rating made by a subject and the overall mean for that document pair was calculated. The distribution of these difference scores is shown in Figure 1. The mean absolute difference is about 0.46 on the five-point scale and about 90% of the differences are less than one. We also produced a measure of 'inter-rater' correlation, by choosing one rating for each document pair at random, and measuring its correlation with the average of the remaining human judgments. The average of 1,000 such correlations was 0.605.

To test whether the left-right positioning of documents affected similarity judgment, the difference between the average similarity for both positionings was calculated. The average difference was 0.37 on the five-point scale, and more than 95% of all the pairs were within one point on that scale.

These results suggest that similarity judgments do not vary significantly across subjects or because of left-right positioning, and so similarity values for all presentations of each document pair were averaged. The resultant five-point similarity scores were then normalized to lie on a 0-1 scale for ease of comparison with the various models of similarity.

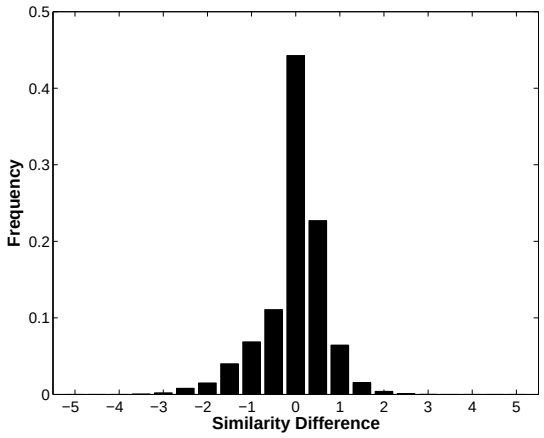

Figure 1: Distribution of differences between individual subjects ratings, and the overall mean for each document pair.

Evaluation of Automated Measures

Document Representation

After removing all punctuation and capitalization from the text, words were defined as unique strings separated by spaces. The corpus representations using the complete set of words, and using that subset of the words not included in a standard set of common words (a stoplist) were both generated. In addition, n-gram representations (Damashek, 1995), based on sequences of n successive characters occuring in the text, were generated for $n = 3, 4, \ldots, 10$. Formally, each approach represents the corpus as a $u \times v$ matrix of counts $\mathbf{X} = [x_{ik}]$ where x_{ik} counts the number of times the i-th word or n-gram occurs in the k-th document. The value u is the number of words, number of words not in the stop list, or number of n-grams in the corpus, and v is the number of documents.

Binary Similarity Models

Measures In cognitive science, considerable attention has been given to the problem of modeling human similarity judgments for featural stimuli, where only the presence or absence of features is used to represent objects (e.g., Navarro and Lee, ress; Tversky, 1977; Tenenbaum and Griffiths, 2001). In the current context, this corresponds to a representation that does not count the number of times words occur in documents, but simply denotes whether they occur at all. Defining $t_{ik} = 1$ if $x_{ik} > 0$ and $t_{ik} = 0$ if $x_{ik} = 0$ allows different similarity models for binary representations to be defined in terms of four counts. For the i-th and j-th documents, the count $a_{ij} = \sum_k t_{ik} t_{jk}$ is the number of words or n-grams in the corpus representation that are common to both documents, the counts $b_{ij} = \sum_k t_{ik} (1 - t_{jk})$ and $c_{ij} = \sum_k (1 - t_{ik}) t_{jk}$ are the distinctive words or n-grams that one document has but the other does not, and the count $d_{ij} = \sum_k (1 - t_{ik})(1 - t_{jk})$ is the number contained within neither document.

[1]The intention was to present each pair exactly ten times, but an error in running the program resulted in about 10% of the pairings being presented eight, nine, eleven or twelve times

Previous similarity modeling suggests three theoretically important alternatives based on these counts. The most widely used (e.g., Lee and Navarro, 2002; Shepard and Arabie, 1979) is the Common Features Model, which is a special case of Tversky's (1977) Contrast Model, and assumes simply that similarity is measured by the proportion of common features, so that:

$$s_{ij}^{com} = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij} + d_{ij}}.$$

An alternative is Tversky's (1977) Ratio Model:

$$s_{ij}^{rat} = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}},$$

which measures similarity as the ratio of common to common and distinctive features. Finally, the Distinctive Features special case of the Contrast Model, which is equivalent to the similarity model used in discrete multidimensional scaling (e.g., Rohde, 2002), assumes that two stimuli become more dissimilar to the extent that one stimulus has a feature that the other does not, so that:

$$s_{ij}^{dis} = \frac{a_{ij} + d_{ij}}{a_{ij} + b_{ij} + c_{ij} + d_{ij}}.$$

Beyond these psychologically motivated measures, Cox and Cox (1994, p. 11) list another nine similarity measures based on the a, b, c, and d counts. We also evaluated these measures, but found that none outperformed the best of the three psychologically motivated measures.

Results Figure 2 shows the correlations between the human similarity measures, and those predicted by the Ratio, Common Features and Distinctive Features similarity models. These correlations are shown for complete binary and stopped binary word-based representations, and binary 3-gram through to 10-gram representations.

Figure 2 shows at least four clear results. First, the Ratio Model outperforms the Common Features Model for most representations, and both are significantly better than the Distinctive Features Model. Secondly, for the Ratio and Common Features Models, the stopped representation leads to better performance than using the complete word representation. Thirdly, the Ratio and Common Features Models achieve their best correlation using 7-, 8-, or 9-grams, with worse performance for smaller and larger lengths. In the best case, the models have a correlation of about 0.5 with human judgments.

Count Similarity Models

Measures For the corpus representations using counts, we tested the four symmetric similarity models

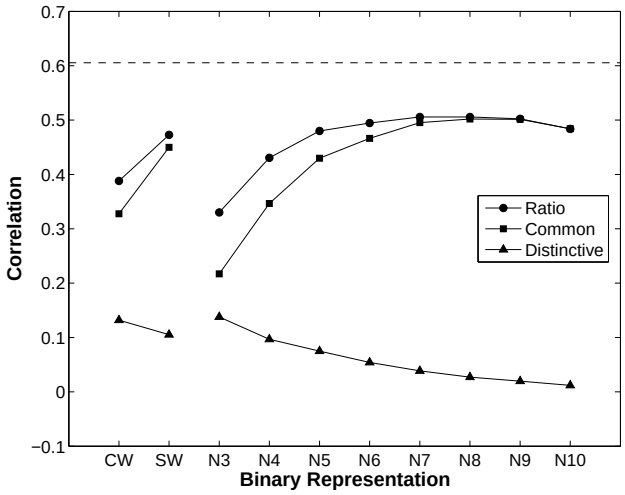

Figure 2: Correlations between the human similarity measures and all binary representations using three similarity models. CW=complete word, SW=stopped word, N3=3-gram, and so on. The dashed line shows the inter-rater correlation.

considered by Rorvig (1999). Using his terminology, these are the Correlation model:

$$s_{ij}^{cor} = \frac{\sum_k x_{ik} x_{jk}}{\sum_y c_{ik} + \sum_k x_{jk}},$$

the Jaccard model:

$$s_{ij}^{jac} = \frac{\sum_k x_{ik} x_{jk}}{\sum_k x_{ik} + \sum_k x_{jk} - \sum_k x_{ik} x_{jk}},$$

the Cosine model:

$$s_{ij}^{cos} = \frac{\sum_k x_{ik} x_{jk}}{\left(\sum_k x_{ik}^2 \sum_k x_{jk}^2\right)^{\frac{1}{2}}},$$

and the Overlap model:

$$s_{ij}^{ove} = \frac{\sum_k x_{ik} x_{jk}}{\min\left(\sum_k x_{ik}^2, \sum_k x_{jk}^2\right)}.$$

Results Figure 3 shows the correlations between the human similarity measures, and those predicted by the Jaccard, Cosine, Correlation and Overlap similarity models. Once again, these correlations are shown for complete and stopped word-based representations, and 3-gram through to 10-gram representations. Figure 3 shows that the differences between the four similarity models are very small, but that there are important differences in the performance supported by the underlying document representations. Stopped representation leads to better performance than the complete

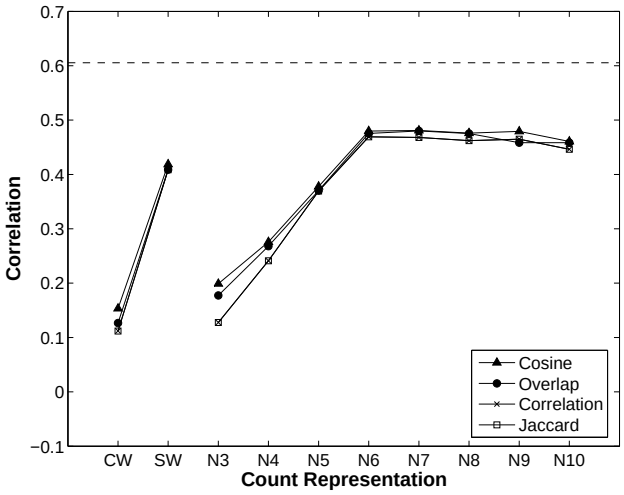

Figure 3: Correlations between the human similarity measures and all count representations using four similarity models. CW=complete word, SW=stopped word, N3=3-gram, and so on. The dashed line shows the inter-rater correlation.

word representation, as with the binary representation analysis. n-grams with length six or above are better performed than smaller lengths. None of the correlations reach the 0.5 level and, perhaps surprisingly, the count representations generally led to worse correlations with human performance than the binary representations.

LSA Similarity Models

Measures LSA begins with a $n \times v$ matrix $\mathbf{C} = [c_{ik}]$ where n is the number of words, v is the number of documents in the corpus, and c_{ik} is the frequency of the i-th word in the k-th document. There are three conceptually different components to the way LSA uses this document representations to measure similarity. These are the local weighting function, which measures the importance of a word within a document, the global weighting function, which measures the importance of a word across the entire corpus of documents, and the number of dimensions retained during the singular value decomposition, which makes assumptions about the complexity of the underlying semantic regularities expressed by the corpus.

The three local weighting functions we considered are based on the frequency of the i-th word in the k-th document. The first local weighting function was just this term frequency, the second was logarithmic in the frequency, and the third was binary, taking the value one if the frequency was non-zero. Intuitively, the first two local weights give increasing importance to more frequent words, but the logarithmic gives progressively smaller additional emphasis to larger frequencies, while the third measure is sensitive only to whether the word is in the document. The first global weighting function we considered normalized each word using the local weighting function, the second was an inverse document frequency measure, and the third global was an entropy measure. More details are provided by Pincombe (2004).

Local and global weighting functions are used to generate a weighted corpus representation $\mathbf{W} = [w_{ik}]$. In LSA, this weighted representation is subjected to singular value decomposition. This involves choosing a dimensionality $d \leq m$ for the subspace representation, and finding the $n \times d$ orthonormal matrix $\mathbf{U}$, the $d \times d$ diagonal matrix $\mathbf{D}$ and the $m \times d$ orthonormal matrix $\mathbf{V}$ that minimize the squared difference $\|\mathbf{W} - \mathbf{UDV^T}\|$.

The resulting $n \times d$ matrix $\mathbf{N} = [n_{ik}]$ is a least squares best fit to $\mathbf{W}$ produced by zeroing all but the largest d coefficients of $\mathbf{D}$. The document similarities are arrived at using a similarity measure similar to the earlier Cosine model, the exact form of which is

$$s_{kj}^{cos} = \frac{\sum_i n_{ik} n_{ij}}{(\sum_i n_{ik}^2 \sum_i n_{ik}^2)^{1/2}}.$$

For the original corpus, in both its complete and stopped forms, and using all nine possible pairings of local and global weighting functions, we considered dimensionalities of 10, 20, 30, 40, and 50. For the extended corpus, again in both complete and stopped forms, and using all weighting combinations, we considered dimensionalities of 10, 20, 30, 40, 50, 100, 150, 200, 250, and 300.

Results The results of these analyses are shown in Figure 4. It is clear that altering the local weighting function makes relatively little difference but that changing the global weighting function does make a difference. Entropy global weighting is generally superior to normalized weighting, and both are better than the inverse document frequency function. For the 50 document corpus, performance is best when there is no dimensionality reduction in the representation (i.e., when all 50 factors are used thus reducing LSA to a weighted vector space model). Peak performance for the extended 364 document corpus is better and is achieved when between 100 and 200 factors are used. Applying the stop word list leads to a significant improvement when using the (poorly performing) inverse document frequency global weighting function, and there is also a small improvement in most other cases. The best performed LSA models, correlate about 0.6 with human judgments, which is better than the keyword and n-gram vector space methods, and at the base level of inter-rater correlation.

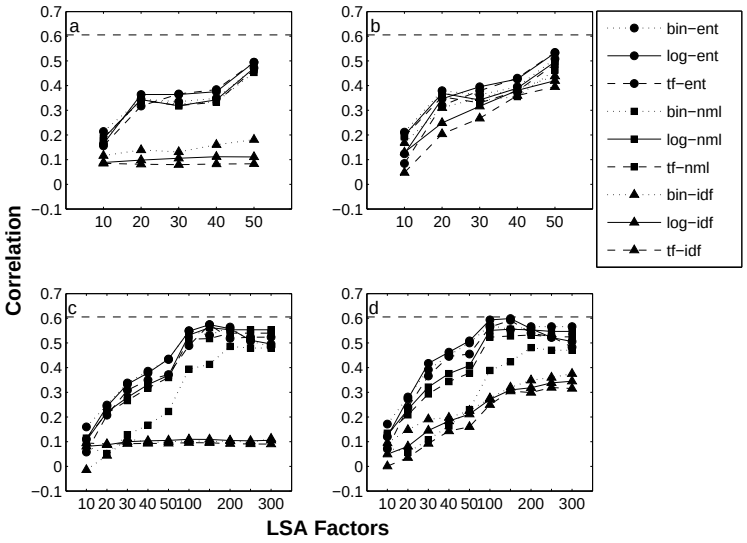

Figure 4: Correlations between the human similarity measures and nine LSA similarity models, for each of four situations corresponding to (a) the 50 document corpus; (b) the 50 document without stopwords; (c) the 364 document corpus; (b) the 364 document without stopwords. The nine similarity models consider every pairing of the binary ('bin'), logarithmic ('log') and term frequency ('tf") local weighting functions with the entropy ('ent'), normalized ('nml') and inverse document frequency ('idf') global weighting functions. The dashed lines shows the inter-rater correlation.

Conclusion

We have argued that the automated measurement of the similarity between text documents is fundamentally a psychological modeling problem. This means that various existing methods, widely used in information science applications, ought to be assessed (at least in part) in terms of their ability to model human performance. This paper presents an assessment of keyword, n-gram and LSA approaches against human data for a small corpus of short news documents. We considered a variety of existing cognitive models, using different representational assumptions—including whether or not a stopword list was applied, whether word frequency was considered, what n-gram length was used, and how many LSA factors were used—and a variety of different similarity modeling assumptions.

Using an extended corpus to retain about 100 factors, LSA under the entropy global weighting function produced correlations of about 0.6, at the base level of inter-rater correlation.. The best performed keyword and n-gram models achieved correlations closer to about 0.5. Many of the methods we considered showed almost no correlation with human performance.

An examination of the relationship between modeled and human similarity values shows two clear regularities that highlight weaknesses in the models we examined. These regularities are well characterized by the scatterplots shown in Figure 5, which show the relationship between modeled and human values for all documents pairs, using the common and distinctive similarity models, with 8-gram binary representations. The performance of the distinctive model is typical of those with near-zero correlations, showing no systematic relationship between modeled and human values. The deficiencies evident for the common model, however, are more interesting. It is clear that when the model judges two documents to be highly similar, it is correct. Its weakness is that it fails to detect other high-similarity pairings, giving them relatively low values. In information science terms, if the task is to identify highly similar documents, the model has very good precision, but poor recall. It seems that the best performed models we examined are able to detect only a subset of the highly semantically similar document pairs.

These findings suggest alternative models of text document similarity. Alternatives could arise from new representations of text documents, specifying new similarity models, or both. The performance of the common features model in Figure 5 suggests that it works well when the underlying representations of two documents share features. More sophisticated representations might be able to identify the common features between the highly similar document pairs currently being missed. Obvious candidates for improved representation include those used by the topics model (e.g.,

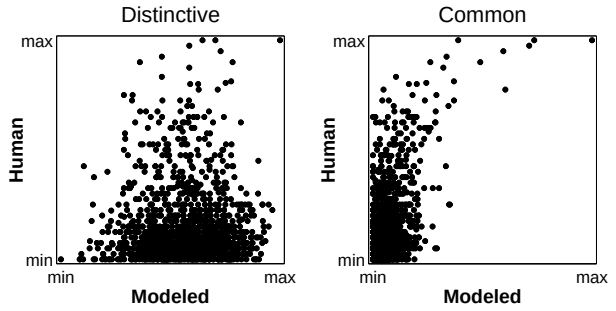

Figure 5: The relationship between model and human judgments of similarity for all document pairs, using the distinctive features (left panel) and the common features models (right panel), using the binary features representation based on 8-grams.

Griffiths and Steyvers, 2004) and the SP model (e.g., Dennis, 2004).

Alternatively, more complicated representations might not be required if a more sophisticated similarity model was used. We have in mind a similarity model that judges documents in terms of their similarity relationships to other documents in the corpus. Intuitively, there may be short paths between highly-similar documents as they are currently represented and measured, even if that similarity is not evident from examining the two documents being assessed in isolation. Recent psychologically-motivated data-analysis methods that focus on establishing global relationships using local properties, such as ISOMAP (e.g., Tenenbaum et al., 2000) provide a possibile starting point for examining this idea.

Of course, to what extent these sorts of theoretical improvements manifest themselves in applied benefits, improving the performance of search engines, decision support systems, or other text based systems, remains an open question. We believe, however, they are avenues worth exploring.

Acknowledgments

This research was supported by Australian Research Council Grant DP0211406, and the Defence Science and Technology Organisation. We thank Simon Dennis, Dan Navarro, Ben Grindlay and Lama Chandrasena.

References

Baayen, R. H. (2001). *Word Frequency Distributions*. Kluwer, London.

Cox, T. F. and Cox, M. A. A. (1994). *Multidimensional Scaling*. Chapman and Hall, London.

Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267:843–848.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Dennis, S. J. (2004). An unsupervised method for the extraction of propositional information from text. *Proceedings of the National Academy of Sciences*, 101:5206–5213.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Landauer, T. K. and 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.

Lee, M. D. and Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, 9(1):43–58.

Navarro, D. J. and Lee, M. D. (in press). Common and distinctive features in stimulus similarity: A modified version of the contrast model. *Psychonomic Bulletin & Review*.

Pincombe, B. M. (2004). Comparison of human and latent semantic analysis (LSA) judgments of pairwise document similarities for a news corpus. Defence Science and Technology Organisation Research Report DSTO–RR–0278.

Rohde, D. L. T. (2002). Methods for binary multidimensional scaling. *Neural Computation*, 14(5):1195–1232.

Rorvig, M. E. (1999). Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. *Journal of the American Society for Information Science*, 50(8):639–651.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Boston, MA.

Shepard, R. N. and Arabie, P. (1979). Additive clustering representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86(2):87–123.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323.

Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327–352.