An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources

Yuhua Li, Zuhair A. Bandar, and David McLean

Abstract—Semantic similarity between words is becoming a generic problem for many applications of computational linguistics and artificial intelligence. This paper explores the determination of semantic similarity by a number of information sources, which consist of structural semantic information from a lexical taxonomy and information content from a corpus. To investigate how information sources could be used effectively, a variety of strategies for using various possible information sources are implemented. A new measure is then proposed which combines information sources nonlinearly. Experimental evaluation against a benchmark set of human similarity ratings demonstrates that the proposed measure significantly outperforms traditional similarity measures.

Index Terms—Semantic similarity, lexical database, information content, corpus statistics.

1 Introduction

The study of semantic similarity between words has been a part of natural language processing and information retrieval for many years. Semantic similarity is a generic issue in a variety of applications in the areas of computational linguistics and artificial intelligence, both in the academic community and industry. Examples include word sense disambiguation [18], detection and correction of word spelling errors (malapropisms) [4], text segmentation [10], image retrieval [22], multimodal documents retrieval [23], and automatic hypertext linking [8]. Similarity between two words is often represented by similarity between concepts associated with the two words. A number of semantic similarity methods have been developed in the previous decade, different similarity methods have proven to be useful in some specific applications of computational intelligence. For extensive comparisons of some representative similarity measures, please refer to [4] and [12]. Generally, these methods can be categorized into two groups: edge counting-based (or dictionary/thesaurus-based) methods and information theory-based (or corpus-based) methods.

Assuming a lexical taxonomy is structured in a tree like hierarchy with a node for a concept, Rada et al. [16] has proven that the minimum number of edges separating concepts c_1 and c_2 is a metric for measuring the conceptual distance of c_1 and c_2. Their work forms the basis of edge counting-based similarity methods. The edge counting method is useful for specific applications with highly constrained taxonomies, for example, medical semantic nets. Rada's edge counting method considers hyponymy/hypernymy (IS-A) links only, but excludes other link types such as part-whole (HASA), antonymy, etc. Moreover, lexical taxonomy may have irregular densities of links between concepts due to its broad domain. This problem of nonuniformity can be corrected by using the depth in the hierarchy where the word is found, the density of the subhierachies and the type of link [9], [19], [24].

The information theory-based method for semantic similarity was first proposed by Resnik [17]. He defines the similarity of two concepts as the maximum of the information content of the concept that subsumes them in the taxonomy hierarchy. The information content of a concept depends on the probability of encountering an instance of the concept in a corpus. That is, the probability of a concept is determined by the frequency of occurrence of the concept and its subconcept in the corpus. The information content is then defined as negative the log likelihood of the probability [17]. Since the information content is calculated from the corpus, this similarity measure can be adapted to a particular application provided that the corpus approximates that application area well. Following the original work of Resnik, some modifications have been made to enhance the pure information content method. Jiang and Conrath [9] proposed a combined method that is derived from the edge-based notion by adding the information content as a decision factor. They take into account the fact that edges in the taxonomy may have unequal link strength, so link strength of an edge between two adjacent nodes is determined by local density, node depth, information content, and link type. The similarity between two words is simply the summation of edge weights along the shortest path linking two words. Lin [11] derived a theoretically well-motivated measure that is similar to Resnik's information content. Lin's modification consisted of normalizing by the combination of information content of the compared concepts and assuming their independence.

As introduced above, different methods use different information sources and, thus, result in different levels of performance. The commonly used information sources in

- *Y. Li is with the Manchester School of Engineering, University of Manchester, Oxford Road, Manchester M13 9PL, UK. E-mail: yuhua.li@man.ac.uk.*
- *Z.A. Bandar and D. McLean are with the Intelligent System Group, Department of Computing and Mathematics, Oxford Road, Manchester Metropolitan University, Manchester M1 5GD, UK. E-mail: {z.bandar, d.mclean}@mmu.ac.uk.*

Manuscript received 20 Dec. 2001; revised 26 July 2002; accepted 22 Oct. 2002.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 115611.

1041-4347/03/$17.00 © 2003 IEEE Published by the IEEE Computer Society

previous similarity measures are shortest path length between compared words, information content, depth in the taxonomy hierarchy, and semantic density of compared words. A major problem with these similarity measures is that either information sources are directly used as a metric of similarity or a method uses a particular information source without considering the contribution of others. Since semantic similarity is influenced by a number of information sources that are interlaced with each other, we argue that semantic similarity depends not only on multiple information sources, but also that the information sources should be properly processed and combined.

In Section 2, we introduce our definition of semantic similarity between words. The similarity is constructed from a number of functional building blocks which are transferred from the information sources. Section 3 describes the extraction of information sources from a lexical knowledge base and corpus statistics. The choice and organization of a benchmark data set for evaluating similarity method is then explained. Section 4 presents a number of experiments on different strategies for calculating similarity. These are based on the building blocks introduced in Section 2, the information sources and the benchmark data set prepared in Section 3. The experimental results are discussed and compared with related work in Section 5. The paper concludes in Section 6 that, based on the benchmark data set, our measure outperforms existing measures.

2 Semantic Similarity between Words

Before proceeding to the presentation of our method, it is necessary to introduce some constraints to the development of similarity measures. Evidence from psychological experiments demonstrate that similarity is context-dependent and may be asymmetric [13], [25]. Similarity between words is influenced by the context in which the words are presented. For example, if the context is "the outside covering of living objects," then *skin* and *bark* are more similar than *skin* and *hair*; however, the opposite is true if the context is body parts. Similarity may also be asymmetric with respect to direction. People may give different ratings when asked to judge the *similarity of surgeon to butcher* and the *similarity of butcher to surgeon*. Although similarity may be asymmetric, the "asymmetries are only observed under quite circumscribed conditions" [3]. Experimental results investigating the effects of asymmetry suggest that the average difference in ratings for a word pair is less than 5 percent [13]. We believe that such a small difference will have little impact on the overall performance of computational methods, so we do not consider the effects of asymmetry. This is in line with many application areas of computational linguistics and artificial intelligence [4], [8], [10], [22], [23].

2.1 The Method for Semantic Similarity

Thanks to the success of a number of computational linguistic projects, semantic knowledge bases are readily available. The knowledge bases may be constructed in a hierarchy that is commonplace in the world. The lexical hierarchy is connected by following trails of superordinate terms in "is a" or "is a kind of" (ISA) relations. The ISA hierarchical structure of the knowledge base is important in determining the semantic distance between words. Fig. 1 shows a portion of such a hierarchical semantic knowledge base.

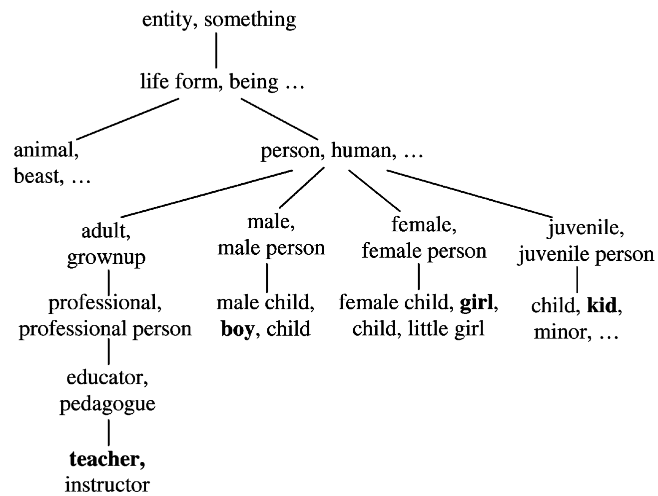


Fig. 1. Hierarchical semantic knowledge base. "..." indicates that some words in the class were omitted to save space.

Given two words w_1 and w_2, we need to find the semantic similarity of $s(w_1, w_2)$ for these two words. We can do this by analysis of the knowledge base, as follows: Words are associated with concepts in the ISA hierarchy. Therefore, we can find the first concept in the hierarchical semantic network that subsumes the concepts containing the compared words. One direct method for similarity calculation is to find the minimum length of path connecting the two concepts containing the two words [16]. For example, Fig. 1 illustrates a fragment of the semantic hierarchy of WordNet [15]. The shortest path between *boy* and *girl* is *boy-male-**person**-female-girl*, the minimum length of path is 4, the synset of ***person*** is called the subsumer for the words *boy* and *girl*; while the minimum path length between *boy* and *teacher* is 6. Thus, we could say *girl* is more similar to *boy* than *teacher* to *boy*. If a word is polysemous (i.e., a word having many meanings), multiple paths may exist between the two words. Only the shortest path is then used in calculating semantic similarity between words.

Rada et al. [16] demonstrated that this method works well on their much constrained medical semantic nets (with 15,000 medical terms). However, this method may be not so accurate if it is applied to larger and more general semantic nets such as WordNet [15]. For example, the minimum length from *boy* to *animal* is 4, less than from *boy* to *teacher*, but we should not say *boy* is more similar to *animal* than to *teacher* (unless you are cursing the boy). To address this weakness, the direct path length method must be modified by utilizing more information from the hierarchical semantic nets. It is intuitive that concepts at upper layers of the hierarchy have more general semantics and less similarity between them, while concepts at lower layers have more concrete semantics and stronger similarity. Therefore, the depth of concept in the hierarchy should be taken into account. Moreover, local density of the semantic nets is also a factor that affects the similarity between words. In summary, similarity between words is determined not only by path length but also by depth and density.

We propose that the similarity of $s(w_1, w_2)$ between w_1 and w_2 be a function of the attributes path length, depth, and local density as follows:

$$s(w_1, w_2) = f(l, h, d), \tag{1}$$

where, l is the shortest path length between w_1 and w_2.

h is the depth of subsumer in the hierarchy semantic nets.

d is the local semantic density of w_1 and w_2.

We assume that (1) can be rewritten using three independent functions as:

$$s(w_1, w_2) = f(f_1(l), f_2(h), f_3(d)) \tag{2}$$

f_1, f_2, and f_3 are transfer functions of path length, depth, and local density, respectively.

The independent assumption in (2) comes from the following considerations: Path length and depth are derived from a lexical database, while the local semantic density is computed from a corpus. These are detailed in the following sections. The independence assumption between path length and depth is a useful simplification because this can make pure edge-based method as in [16] a specific case of (2) by assigning proper function forms to individual attributes. The independence assumption in (2) enables us to investigate the contribution of individual attributes to overall similarity through combining them.

2.2 Properties of Transfer Functions

Values of an attribute in (2) may cover a large range up to infinity, while the interval of similarity should be finite with extremes of exactly the same or nothing similar at all. If we assign exactly the same with a value of 1 and no similarity as 0, then the interval of similarity is [0, 1]. The direct use of information sources as a metric of similarity is inappropriate due to its infinite property. Therefore, it is intuitive that the transfer function from information sources to semantic similarity is a nonlinear function. Taking path length as an example, when the path length decreases to zero, the similarity would monotonically increase toward the limit 1. While path length increases infinitely, the similarity should monotonically decrease to 0. Therefore, to meet these constraints, the transfer function must be a nonlinear function. The nonlinearity of the transfer function is taken into account in the derivation of the formula for semantic similarity between two words as in the following subsections.

To obtain $s(w_1, w_2)$, from (2), the task is split into three subtasks: the contributions of path length, depth, and local density.

2.3 Contribution of Path Length

For a semantic net organized hierarchically as in Fig. 1, the path length between two words, w_1 and w_2, can be determined from one of three cases:

1. w_1 and w_2 are in the same concept.
2. w_1 and w_2 are not in the same concept, but the concept for w_1 and the concept for w_2 contain one or more of the same words. For example, in Fig. 1, the concept for *boy* and the concept for *girl* contain one matching word *child*.
3. w_1 and w_2 are not in the same concept nor do their concepts contain the same words.

Case 1 implies that w_1 and w_2 have the same meaning; we assign the semantic path length between w_1 and w_2 to 0. Case 2 indicates that w_1 and w_2 partially share the same features, we assign the semantic path length between w_1 and w_2 to 1. This assignment is derived from the observation that, in the lexical hierarchy, when the concepts for w_1 and w_2 contain one or more of the same words, they are usually very similar (e.g., some concepts in Fig. 1). There are some exceptions to this assignment. For example,[1] the word *deposit* (in WordNet's sense of "put into a bank account") has very little in common with *swear* (in WordNet's sense of "have confidence or faith in"), but both words are in concepts that contain the verb *bank*. As a result, the words will have a path length of 1. Although this sort of inaccuracy appears to occur quite rarely, a comprehensive study has not yet been carried out; this may require further investigation. For Case 3, we count the actual path length between w_1 and w_2.

Taking the above considerations into account, we set $f_1(l)$ in (2) to be a monotonically decreasing function of l:

$$f_1(l) = \mathrm{e}^{-\alpha l}, \tag{3}$$

where α is a constant. The selection of the function in exponential form ensures that f_1 satisfies the constraints discussed in Section 2.2 and the value of f_1 is within the range from 0 to 1.

2.4 Scaling Depth Effect

The depth of the subsumer is derived by counting the levels from the subsumer to the top of the lexical hierarchy. For polysemous words, the subsumer on the shortest path is considered in deriving the depth of the subsumer. Words at upper layers of hierarchical semantic nets have more general concepts and less semantic similarity between words than words at lower layers. This behavior must be taken into account in calculating $s(w_1, w_2)$. We therefore need to scale down $s(w_1, w_2)$ for subsuming words at upper layers and to scale up $s(w_1, w_2)$ for subsuming words at lower layers. Moreover, the similarity interval is finite, say [0, 1] as stated earlier. As a result, $f_2(h)$ should be a monotonically increasing function with respect to depth h. To satisfy these constraints, we set f_2 as:

$$f_2(h) = \frac{\mathrm{e}^{\beta h} - \mathrm{e}^{-\beta h}}{\mathrm{e}^{\beta h} + \mathrm{e}^{-\beta h}}, \tag{4}$$

where $\beta > 0$ is a smoothing factor. As $\beta \to \infty$, the depth of a word in the semantic nets is not considered. This function form can be considered as an extension of Shepard's law [21], which claims that exponential-decay functions are a universal law of stimulus generalization for psychological science. The extension is that (4) employs an exponential-growth function of similarity with depth rather than an exponential-decay function because similarity increases with depth.

2.5 Considering Local Semantic Density

Local semantic density is a rather difficult issue to tackle, we cannot obtain it solely from the semantic nets. However, we can calculate it from semantic nets with the help of a large corpus. One measure of local semantic density is the information content of a concept [18]. Information content was originally developed to measure semantic similarity between words. Here, it is borrowed to represent the local semantic density of a concept in a corpus to assist in developing our similarity method.

1. This example is adopted from the comments of one of the anonymous reviewers.

Let the probability of encountering an instance of concept c in a corpus be $p(c)$ (see Section 3.2 for the calculation of $p(c)$), then the information content of concept c is quantified by following the notation of information theory as

$$IC(c) = -\log p(c). \tag{5}$$

For two concepts c_1 and c_2, the similarity measure is determined from the subsumer in the semantic net with the maximum information content.

$$sim(c_1, c_2) = \max_{c \in sub(c_1,c_2)} [-\log p(c)], \tag{6}$$

where $sub(c_1, c_2)$ is the set of concepts that subsume both c_1 and c_2.

The information shared by words w_1 and w_2 is defined as

$$wsim(w_1, w_2) = \max_{c_1,c_2}[sim(c_1, c_2)], \tag{7}$$

where c_1 and c_2 range over possible senses of w_1 and w_2, respectively. In order to take context into account, Resnik [18] further suggested a modified form of (7) when applied to a syntactic disambiguation task. He explained that a similarity measure could be improved by a weighted sum over concepts by using more information from the context. However, as in previously published results of Resnik's and others, we will use maximum form as (7) for similarity computation. This is because Resnik's original evaluation of his method did not take context into account and the employed word pairs were judged without context.

Finally, as in setting the function form for transferring depth in (4), local semantic density is defined as a monotonically increasing function of $wsim(w_1, w_2)$:

$$f_3(wsim) = \frac{e^{\lambda \cdot wsim(w_1,w_2)} - e^{-\lambda \cdot wsim(w_1,w_2)}}{e^{\lambda \cdot wsim(w_1,w_2)} + e^{-\lambda \cdot wsim(w_1,w_2)}}, \tag{8}$$

where $\lambda > 0$. If $\lambda \to \infty$, then the information content of words in the semantic nets is not considered.

3 Implementation

Two databases are used in the implementation of the proposed method of semantic similarity; they are WordNet [15] and the Brown Corpus [7]. Both databases are publicly available and widely used in previously published works. The use of WordNet and the Brown Corpus in this paper makes our results comparable with other published work. This section first provides a brief description of these two databases, then presents the search in the lexical taxonomy and the statistics from the corpus.

3.1 The Databases

WordNet is an on-line semantic dictionary—a lexical database, developed at Princeton by a group led by Miller [15]. The version used in this study is WordNet 1.6 which has 121,962 words organized in 99,642 synonym sets. WordNet partitions the lexicon into nouns, verbs, adjectives, and adverbs. Nouns, verbs, adjectives, and adverbs are organized into synonym sets, called synsets. A synset represents a concept in which all words have similar meaning. Thus, words in a synset are interchangeable in some syntax. Knowledge in a synset includes the definition of these words as well as pointers to other related synsets.

The Brown Corpus [7] of Standard American English was the first of the modern, computer readable, general corpora. It was compiled by Francis and Kucera of Brown University. The corpus consists of one million words of American English texts printed in 1961. The texts for the corpus were sampled from 15 different text categories to make the corpus a good standard reference. The number of texts in each category varies. There are a total of 500 texts, each consisting of just over 2,000 words. Much research within the field of corpus linguistics has been made using these data.

In this study, WordNet is the main semantic knowledge base for the calculation of semantic similarity, while the Brown Corpus assists this calculation.

3.2 Obtaining Information Sources

The implementation of semantic similarity measures consists of two subtasks concerning preparation of the information sources. First, a search of the semantic net for the shortest path length between the synsets containing the compared words and the depth of the first synset subsuming the synsets corresponding to the compared words. Second, the calculation of the statistical information from the Brown Corpus.

Synsets in WordNet are designed in a tree-like hierarchical structure going from many specific terms at the lower levels to a few generic terms at the top. WordNet provided rich relations linking synsets. Most previous similarity measures only use the shortest path length from this ISA search. It is commonly accepted that other semantic relations also contribute to the determination of semantic similarity. One important such relation is the part-whole (or HASA) relation. Thus, we also search for HASA relations in WordNet in obtaining the shortest path length.

The statistics from the Brown Corpus are used to obtain the information content of a concept. There are some slightly different methods of calculating the concept probabilities in a corpus [1]. In this work, we use Resnik's [18] method, particularly for noun probability. Each noun that occurred in the corpus was counted as an occurrence of each taxonomic class containing it. For example, in Fig. 1, an occurrence of the noun *professional* would be counted toward the frequency of *professional-professional person*, *adult-grownup*, *person-human-...*, *life form-being ...*, and *entity-something*. Formally,

$$freq(c) = \sum_{w \in words(c)} count(w),$$

where $words(c)$ is the set of words subsumed by concept c. Concept probability is computed as relative frequency:

$$\hat{p}(c) = \frac{freq(c)}{N}, \tag{9}$$

where N is the total number of nouns observed in the corpus (excluding those not subsumed by any WordNet class). The information content of a concept is then calculated from (5).

3.3 The Benchmark Data Set

The quality of a computational method for calculating word similarity can only be established by investigating its performance against human common sense. In evaluating our method against others, we compute word similarity on a benchmark word set with human ratings. A commonly used word set is from an experiment by Miller and Charles [14].

TABLE 1
Semantic Similarity of Human Ratings and Basic Measures for Test Set D_0

Word Pair	RG Rating	MC Replica	Resnik Replica	Information Content	Length	Depth
cord-smile	0.02	0.13	0.1	1.1762	12	0
rooster-voyage	0.04	0.08	0	0	30	0
noon-string	0.04	0.08	0	0	30	0
glass-magician	0.44	0.11	0.1	1.0105	8	0
monk-slave	0.57	0.55	0.7	2.9683	4	2
coast-forest	0.85	0.42	0.6	0	6	1
monk-oracle	0.91	1.1	0.8	2.9683	7	2
lad-wizard	0.99	0.42	0.7	2.9683	4	2
forest-graveyard	1.00	0.84	0.6	0	7	1
food-rooster	1.09	0.89	1.1	1.0105	12	0
coast-hill	1.26	0.87	0.7	6.2344	4	3
car-journey	1.55	1.16	0.7	0	30	0
crane-implement	2.37	1.68	0.3	2.9683	4	3
brother-lad	2.41	1.66	1.2	2.9355	4	2
bird-crane	2.63	2.97	2.1	9.3139	3	5
bird-cock	2.63	3.05	2.2	9.3139	1	5
food-fruit	2.69	3.08	2.1	5.0076	4	3
brother-monk	2.74	2.82	2.4	2.9683	1	5
asylum-madhouse	3.04	3.61	3.6	15.666	1	7
furnace-stove	3.11	3.11	2.6	1.7135	2	2
magician-wizard	3.21	3.5	3.5	13.666	0	4
journey-voyage	3.58	3.84	3.5	6.7537	1	5
coast-shore	3.60	3.7	3.5	10.808	1	4
implement-tool-	3.66	2.95	3.4	6.0787	1	4
boy-lad	3.82	3.76	3.5	8.424	1	4
automobile-car	3.92	3.92	3.9	8.0411	0	7
midday-noon	3.94	3.42	3.6	12.393	0	7
gem-jewel	3.94	3.84	3.5	14.929	0	6

Miller-Charles and Resnik used chord-smile rather than cord-smile in their replications.

Miller and Charles gave a group of 38 human subjects 30 word pairs and asked the subjects to rate them for similarity in meaning on a scale from 0 (no similarity) to 4 (perfect synonymy). These 30 word pairs were a subset of Rubenstein-Goodenough's original data set of 65 word pairs. Miller-Charles' set consisted of 10 high level, 10 intermediate level, and 10 low level word pairings of semantic similarity. Rubenstein and Goodenough obtained similarity from 51 human subjects. The similarities between words were also rated on the scale of 0.0 to 4.0 for "semantically unrelated" to "highly synonymous." The Miller-Charles experiment was carried out 25 years later than Rubenstein-Goodenough's, but correlation coefficient between the two sets of ratings is 0.97 [14]. This indicates that human knowledge about semantic similarity between words is remarkably stable over a large time span and human ratings can be reliably used as a reference for evaluating computational methods.

Most researchers (e.g., [9], [11], [12], [18]) have used only 28 word pairs of the Miller-Charles set and ignore the remaining pairs in the Rubenstein-Goodenough set. We use Rubenstein-Goodenough's original word set, but consider the popularity of Miller-Charles' in our experiment. Therefore, Rubenstein-Goodenough's 65 word pairs were divided into two sets: One contains the commonly used 28 word pairs (denoted as D_0), another contains the remainder, which has 37 word pairs (denoted as D_1). We use D_1 to design our method and D_0 to test our method. The inclusion of 28 word pairs in D_0 is to make our results comparable with the work of other researchers who used 28 of the 30 pairs of words in the Miller-Charles set, due to the omission of two word pairs from earlier versions of WordNet.

Word pairs in D_0 and D_1 are listed in Table 1 and Table 2, respectively. Both tables list Rubenstein-Goodenough's ratings (in column *RG Rating*). Table 1 also lists the Miller-Charles' replication (in column *MC Replica*) and Resnik's replication (in column *Resnik Replica*). The shortest path length and the depth of the subsumer for these word pairs using ISA and HASA are obtained from WordNet and are listed in columns *length* and *depth*, respectively. Both tables also list information content (in column *Information Content*) for the word pairs. These two tables provide the benchmark data set for our experiments. Information content, shortest path length and subsumer depth are considered to be the information sources for the experiments in deriving our similarity measure in the next section.

4 Experiments on Similarity Methods

As discussed above, useful information in measuring word similarity includes the path length of the two words, depth of the subsumer, and information content of a concept. We argue that, to achieve a good similarity measure, all of the information should be taken into account in a reliable strategy. This section investigates the effectiveness of our method by exploring a variety of strategies for using possible information, which are presented as follows:

Word pairs in Table 1 and Table 2 are used in investigating the suitability of individual strategies. For each of the proposed strategies, we carried out the experiments with two steps. First, we tune the strategy parameters on the training data set D_1. Given the value of a parameter, semantic similarity values of the word pairs are calculated. Then, the correlation coefficient[2] between the computed semantic similarity values and the human ratings of Rubenstein-Goodenough's is calculated. Thus, a set of correlation coefficients is obtained by changing the value of the strategy parameters. The parameters resulting in the greatest correlation coefficient are considered as the optimal parameters for that particular strategy. Second, the identified optimal parameters are used to calculate semantic similarity for word pairs in test data set D_0. Again, the correlation coefficient between computed similarity values and human ratings of Rubenstein-Goodenough's is calculated for words pairs in D_0. This correlation coefficient is used to judge the suitability of the particular strategy comparing to other strategies and previously published results.

Strategy 1. Similarity measure is linear and exclusively based on the shortest path length between the two words [18], using both IS-A and HAS-A links as described in Section 3.2.

$$S_1(w_1, w_2) = f_0(l) = 2 \cdot M - l, \tag{10}$$

where M is the maximum depth of the semantic hierarchy. WordNet was used in our study and, thus, M was set to 15. This value could be any other number provided that $2M \geq \max(l)$. This ensures that similarity is not negative, as a positive similarity measure is more intuitive.

This strategy directly uses the shortest path length as the measure of similarity between two words. Since a linearly transformed variable does not change the magnitude of its correlation coefficient with another variable, this strategy is an equivalent form of the semantic distance method originally proposed by researchers of [16]. The conversion of shortest path length in (10) has no effect, but keeps it in accordance with the meaning of similarity, i.e., a larger value of S_1 for more similar words. This changes the sign of its correlation coefficient with human ratings from negative as in [16] to positive here. This strategy does not have any parameters to tune, we simply calculate the similarities for word pairs in test set D_0. The correlation coefficient between S_1 and human similarity judgments of Rubenstein-Goodenough's was 0.664 on D_0.

Strategy 2. Similarity measure is a linear combination of shortest path length and depth.

TABLE 2
Semantic Similarity of Human Ratings and Basic Measures for Design Set D_1

Word Pair	RG Rating	Information Content	Length	Depth
fruit- furnace	0.05	1.8563	6	2
autograph- shore	0.06	0	30	0
automobile- wizard	0.11	0.9764	11	0
mound- stove	0.14	2.9062	6	2
grin- implement	0.18	0	30	0
asylum- fruit	0.19	1.8563	6	2
asylum- monk	0.39	0.9764	10	0
graveyard- madhouse	0.42	0	12	1
boy- rooster	0.44	2.3852	11	1
cushion- jewel	0.45	1.8563	6	2
asylum- cemetery	0.79	0	9	1
grin- lad	0.88	0	30	0
shore- woodland	0.90	1.5095	5	1
boy- sage	0.96	2.5349	5	2
automobile- cushion	0.97	2.9062	7	3
mound- shore	0.97	6.1974	4	3
cemetery- woodland	1.18	0	7	1
shore- voyage	1.22	0	30	0
bird- woodland	1.24	1.5095	7	1
furnace- implement	1.37	1.8563	5	2
crane- rooster	1.41	8.8872	7	5
hill- woodland	1.48	1.5095	5	1
cemetery- mound	1.69	0	8	1
glass- jewel	1.78	1.8563	7	2
magician- oracle	1.82	13.5898	2	4
sage- wizard	2.46	2.5349	5	2
oracle- sage	2.61	2.5349	7	2
hill- mound	3.29	12.0807	0	7
cord- string	3.41	9.2513	1	4
glass- tumbler	3.45	11.3477	1	5
grin- smile	3.46	10.4198	0	7
serf- slave	3.46	5.2844	3	3
autograph- signature	3.59	14.2902	1	5
forest- woodland	3.65	11.2349	0	3
cock- rooster	3.68	14.2902	0	9
cushion- pillow	3.84	13.5898	1	4
cemetery- graveyard	3.88	13.7666	0	6

2. Let x and y be variables with covariance σ_{xy} and standard deviations σ_x and σ_y, respectively. The correlation coefficient x and y is: $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$. For a linearly derived variable: $x' = ax + b$, a and b are constants and $a \neq 0$, we then have $\rho_{x'y} = \begin{cases} \rho_{xy} & \text{for} \quad a > 0 \\ -\rho_{xy} & \text{for} \quad a < 0 \end{cases}$, see [6].

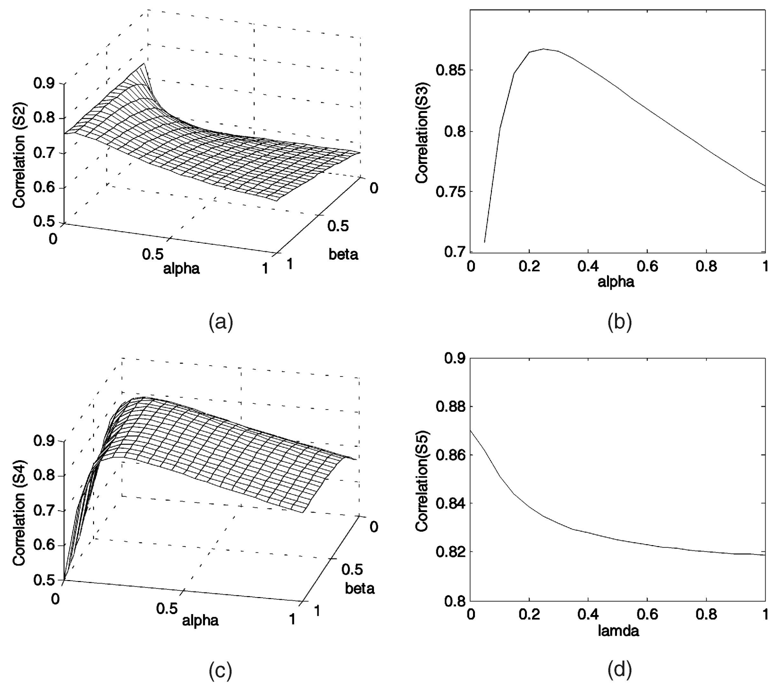


Fig. 2. Correlation coefficient versus strategy parameters on design word set D_1. (a) Correlation between S_2 and human judgments versus α and β. (b) Correlation between S_3 and human judgments versus α. (c) Correlation between S_4 and human judgments versus α and β. (d) Correlation between S_5 and human judgments versus λ.

$$S_2(w_1, w_2) = \alpha S_1(w_1, w_2) + \beta d, \tag{11}$$

where $\alpha \in [0, 1], \beta \in [0, 1]$ and α, β cannot be zero at the same time. This strategy is plausible because the depth of the subsumer carries useful information about where the two words possess the same features. The higher the subsumer is in the semantic hierarchy, the more abstract meaning the two words share and vice versa. As a result, it is straightforward to combine this information with the shortest path length in calculating the semantic similarity of words. Using (11), we search for the optimal values of the parameters α and β using word pairs in D_1. Our experiments demonstrated that the correlation coefficient between S_2 and human similarity judgments reach the maximum value at $\alpha = 0.05$ and $\beta = 1$. Fig. 2a shows the correlation value against α and β on D_1. Using the optimal parameters $\alpha = 0.05$ and $\beta = 1$, the similarities for word pairs in the test set were calculated. The correlation coefficient between S_2 and human similarity judgments is 0.8315 on D_0.

Fig. 2a gives an interesting observation: Depth contributes more than shortest path length if these two sources are linearly combined. This may be because depth carries more information than shortest path length in WordNet in terms of similarity. To verify this observation, we calculate the correlation between depth and human judgement on D_0, it gives a value of 0.8218, which is greater than 0.664 of S_1. Thus, in WordNet, depth is a better indicator of human similarity than path length.

Strategy 3. The similarity measure is a nonlinear function of the shortest path length. The exponential function as defined in (3) is considered in this study, so the similarity is defined as

$$\begin{aligned} S_3(w_1, w_2) &= f_1(l) \\ &= \mathrm{e}^{-\alpha l}. \end{aligned} \tag{12}$$

Compared to Strategy 1, this strategy also uses the shortest path length alone, but make use of a nonlinear function. For different values of α, the correlation to human similarity judgments for word pairs in D_1 is shown in Fig. 2b.

It is observed that the strongest correlation is reached at $\alpha = 0.25$. Using this optimal α, the similarities for word pairs in the test set were calculated. The correlation coefficient between S_2 and human similarity judgments is 0.8911 on D_0. As presented in Strategy 1, the correlation is 0.664 if the shortest path length is directly used as the measure of semantic similarity. This experiment illustrates that a simple transformation of the shortest path length using a nonlinear function can significantly increase the accuracy of the similarity measure.

Strategy 4. Semantic similarity is considered to be governed by the shortest path length as well as the depth of the subsumer, that is

$$\begin{aligned} S_4(w_1, w_2) &= f_1(l) \cdot f_2(h) \\ &= \mathrm{e}^{-\alpha l} \cdot \frac{\mathrm{e}^{\beta h} - \mathrm{e}^{-\beta h}}{\mathrm{e}^{\beta h} + \mathrm{e}^{-\beta h}}, \end{aligned} \tag{13}$$

where $\alpha \geq 0$ and $\beta > 0$ are parameters scaling the contribution of shortest path length and depth, respectively.

Differing from Strategy 2, the shortest path length and depth are transferred by a nonlinear function, respectively, and then combined by multiplication. The correlation of S_4 with different parameters against human similarity judgments on design word set D_1 is shown in Fig. 2c.

The strongest correlation between S_4 and human similarity judgments on D_1 is at $\alpha = 0.2$ and $\beta = 0.6$. Using the optimal parameters $\alpha = 0.2$ and $\beta = 0.6$, the similarities

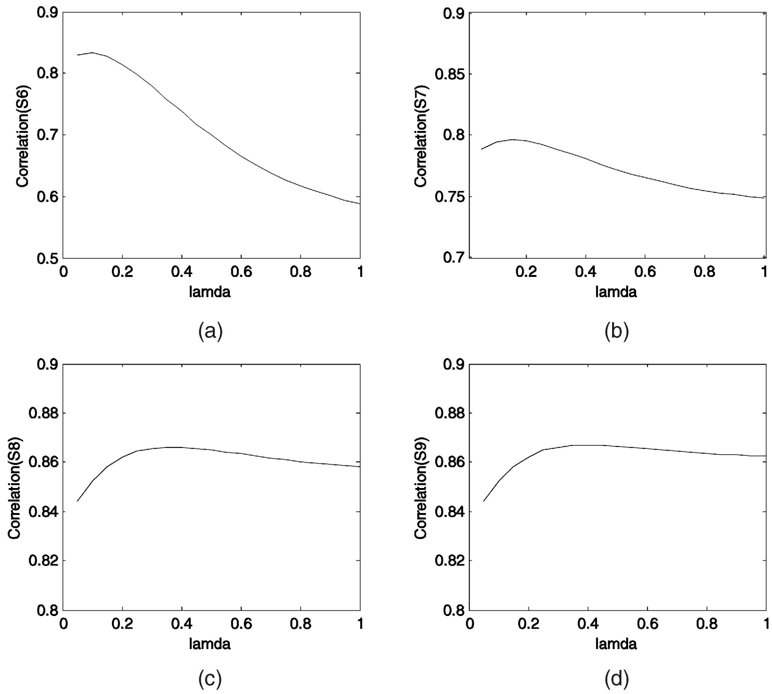


Fig. 3. Correlation coefficient versus strategy parameters on design word set D_1. (a) Correlation between S_6 and human judgments versus λ. (b) Correlation between S_7 and human judgments versus λ. (c) Correlation between S_8 and human judgments versus λ. (d) Correlation between S_9 and human judgments versus λ.

for word pairs in the test set D_0 were calculated. The correlation coefficient between S_4 and human similarity judgments is 0.8914 on D_0. This measure performs nearly at a level of human replication (where correlation between individuals is 0.9015 [17]).

Strategy 5. Information content is linearly combined with Strategy 4, that is,

$$S_5(w_1, w_2) = S_4(w_1, w_2) + \lambda \cdot wsim(w_1, w_2), \quad (14)$$

where $\lambda \in [0, 1]$ is the weight for information content. Here, we consider S_4 with fixed optimal parameters of $\alpha = 0.2$ and $\beta = 0.6$. The correlation against human judgments on design word set D_1 is shown in Fig. 2d.

Fig. 2d illustrates that the performance of similarity measure decreases with the amount of contribution of information content, so this combination of information content with length and depth cannot produce a better similarity measure.

Strategy 6. Information content is nonlinearly combined with shortest path length, that is,

$$S_6(w_1, w_2) = S_1(w_1, w_2) \cdot f_3(wsim). \quad (15)$$

For different values of λ, the correlation to human similarity judgments for word pairs in D_1 is shown in Fig. 3a. From the figure, the strongest correlation is obtained at $\lambda = 0.1$. Applying (15) on test set D_0, we obtain a correlation coefficient of 0.8012 between S_6 and human ratings. This is better than the direct use of shortest path length ($\rho = 0.664$) and information content ($\rho = 0.745$).

Strategy 7. Information content is nonlinearly combined with shortest path length and depth, that is,

$$S_7(w_1, w_2) = S_2(w_1, w_2) \cdot f_3(wsim). \quad (6)$$

Fig. 3b shows the correlation versus parameter λ on design set D_1. The strongest correlation is obtained at $\lambda = 0.15$. Using this value, the similarities for word pairs in a test set were calculated. The correlation coefficient between S_7 and human similarity judgments is 0.8023 on D_0.

Strategy 8. Information content is nonlinearly combined with nonlinearly transferred shortest path length, that is,

$$S_8(w_1, w_2) = S_3(w_1, w_2) \cdot f_3(wsim). \quad (17)$$

Fig. 3c shows correlation versus parameter λ on the design set D_1. The strongest correlation is found at $\lambda = 0.35$, which does not bring a better correlation than using S_3 alone. Taking $\lambda = 0.35$ gives a correlation coefficient of 0.8741 for S_8 against human similarity judgments on D_0.

Strategy 9. Information content is nonlinearly combined with nonlinear transferred shortest path length and depth (S_4), that is,

$$S_9(w_1, w_2) = S_4(w_1, w_2) \cdot f_3(wsim). \quad (18)$$

Fig. 3d shows correlation versus parameter λ. The strongest correlation against human similarity judgments is obtained at $\lambda = 0.4$. Like Strategy 8, this combination strategy does not bring a better correlation than using S_4 alone. The correlation coefficient between S_9 and human similarity judgments is 0.8772 on D_0.

Strategy 10. Similarity measure is the transferred depth of the subsumer through a nonlinear function as defined in Section 2.4, that is,

$$S_{10}(w_1, w_2) = \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}, \quad (19)$$

where $\beta \in (0, 1]$.

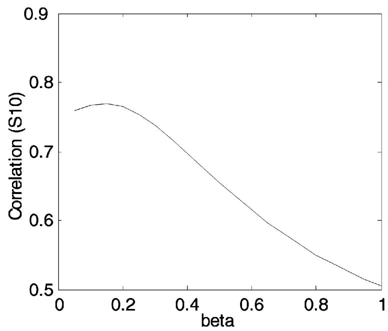


Fig. 4. Correlation between S_{10} and human judgments versus β.

Fig. 4 shows correlation versus parameter β on design set D_1. The strongest correlation against human similarity judgments is at $\beta = 0.15$. The correlation coefficient between S_{10} and human similarity judgments is 0.8356 on D_0.

5 Discussion

Before discussing our results, we provide a look at the performance of the similarity measures reported by previous researchers. Performance of the individual measures is presented in terms of the correlation coefficient against the benchmark set of word pairs in D_0, as listed in Table 1. In order to get a baseline for comparison, Resnik replicated the ratings for Miller-Charles word pairs with a different group of subjects [17]. Table 3 summarizes the experimental results in the literature [5], all computational methods are measured against Rubenstein-Goodenough's ratings as listed in Table 1. The correlation between the mean of human ratings of Miller-Charles' and Resnik's is 0.9583 and the average correlation between individual subjects of human replication is 0.9015. Considering these two correlation coefficients of 0.9583 and 0.9015, all researchers (e.g., [17], [12]) have taken 0.9015 to be the upper-bound to what one should expect from a computational method performing the same task, we argue that 0.9583 is a more reasonable upper-bound because semantic similarity should be considered a collective property of groups of people rather than individuals.

The straight edge counting approach gives poorest performance against human ratings. Resnik's information content method provides a better similarity measure with a correlation of 0.745. Jiang and Conrath [9] also use information content to measure semantic similarity, but they combined it with edge counting using a formula that also took into consideration local density, node depth, and link type. Their method obtained a correlation of 0.8484. Although Jiang and Conrath's method uses multiple information sources, our method is different from theirs in two major aspects: 1) Information sources are transferred nonlinearly in our method. 2) They weight edges using information sources before obtaining similarity between words, we focus directly on a similarity measure. Lin [11] proposed an alternative information-theoretic similarity measure from a set of assumptions. Lin's measure is essentially similar to Resnik's information content measure, but normalized by the combined information content of the two concepts being compared. Lin's method has achieved a correlation of 0.8213. It is clear that each of the successive measures is approaching human ratings gradually. However, if we take the correlation of 0.9583 to be an upper-bound for a computational similarity method, there is still large room for improvement.

TABLE 3
Reference Results on Test Set D_0

Similarity method	Correlation
Resnik Replication to Miller-Charles	0.9583
Human replication	0.9015
Edge-counting	0.664
Information content	0.745
Jiang & Conrath	0.8484
Lin	0.8213

One observation is that all reported methods make direct use of the first-hand information linearly as a measure for similarity. As discussed before, the first hand information sources may be processed in the human brain to produce a human judgement. We assumed this processing is nonlinear, hence the information source is transferred by a suitable nonlinear function. To search for an optimal measure for semantic similarity, we investigated a variety of strategies to combine the information sources. Table 4 lists the similarity values for word pairs in the benchmark set from different strategies with optimal parameters as explored in the last section. The correlation coefficients of our experiments against human ratings of Rubenstein-Goodenough's are listed in Table 5. It is noted that, in Table 4 and Table 5, S_1 is a duplicate of the traditional edge-counting measure. To investigate how much information is carried by the depth of the subsumer, we also calculated the correlation coefficient of depth against Rubenstein-Goodenough's human similarity ratings, the result is indicated as S_d in Table 5.

Referring to results in Table 4 and Table 5, we reach the following observations:

1. The similarity measure can be improved by a suitable combination of information sources.
2. The similarity measure can be improved by nonlinearly transferring information sources.
3. The depth of the subsumer is more similar to human ratings than the shortest path length. We can obtain a much better measure than the traditional edge counting measure, even though we only use the depth of subsumer as in S_{10} and S_d, with correlations of 0.8356 and 0.8218, respectively. This is better than most reported methods. To some extent, Resnik's information content has the same root as S_{10} because both of them utilize information derived from the subsumer.
4. Against our expectation, we did not obtain an improved measure by combining information content as in S_5 to S_9.
5. The best performance was obtained from Strategy 4, with correlation coefficient of 0.8914. This similarity

TABLE 4
Similarity Results from Different Measures on Test Set D_0

Word Pair	S_1	S_2 α=0.05 β=1	S_3 α=0.25	S_4 α=0.2 β=0.6	S_5 λ=0	S_6 λ=0.1	S_7 λ=0.15	S_8 λ=0.35	S_9 λ=0.4	S_{10} β=0.15
cord-smile	18	0.90	0.050	0	0	2.107	0.157	0.019	0	0
rooster-voyage	0	0	0.001	0	0	0	0	0	0	0
noon-string	0	0	0.001	0	0	0	0	0	0	0
glass-magician	22	1.10	0.135	0	0	2.141	0.16	0.045	0	0
monk-slave	26	3.20	0.368	0.355	0.355	6.453	1.161	0.261	0.272	0.291
coast-forest	24	2.15	0.223	0.170	0.170	3.596	0.479	0.108	0.092	0.149
monk-oracle	23	3.05	0.174	0.168	0.168	5.709	1.107	0.123	0.129	0.291
lad-wizard	26	3.20	0.368	0.355	0.355	6.453	1.161	0.261	0.272	0.291
forest-graveyard	23	2.10	0.174	0.132	0.132	0	0	0	0	0.149
food-rooster	18	0.90	0.050	0	0	1.752	0.131	0.016	0	0
coast-hill	26	4.15	0.368	0.366	0.366	14.325	3.031	0.358	0.361	0.422
car-journey	0	0	0.001	0	0	0	0	0	0	0
crane-implement	26	4.15	0.368	0.366	0.366	7.350	1.703	0.283	0.301	0.422
brother-lad	26	3.20	0.368	0.355	0.355	6.453	1.161	0.261	0.272	0.291
bird-crane	27	6.10	0.472	0.472	0.472	19.191	5.307	0.471	0.472	0.635
bird-cock	29	6.20	0.779	0.779	0.779	20.612	5.394	0.776	0.778	0.635
food-fruit	24	2.15	0.223	0.170	0.170	3.596	0.479	0.108	0.092	0.149
brother-monk	29	6.20	0.779	0.779	0.779	7.198	2.250	0.553	0.598	0.635
asylum-madhouse	29	8.10	0.779	0.779	0.779	26.596	7.956	0.779	0.779	0.782
furnace-stove	28	3.30	0.607	0.585	0.585	5.139	0.896	0.347	0.369	0.291
magician-wizard	30	5.30	1	0.999	0.999	26.285	5.123	1	0.999	0.537
journey-voyage	29	6.20	0.779	0.779	0.779	18.794	5.086	0.772	0.776	0.635
coast-shore	29	5.25	0.779	0.779	0.778	23.339	4.889	0.778	0.778	0.537
implement-tool-	29	5.25	0.779	0.778	0.778	15.99	3.837	0.759	0.768	0.537
boy-lad	29	5.25	0.779	0.778	0.778	19.732	4.446	0.774	0.776	0.537
automobile-car	30	8.15	1	1	1	20.923	7.009	0.995	0.998	0.782
midday-noon	30	8.15	1	1	1	27.636	8.016	1	1	0.782
gem-jewel	30	7.20	1	1	1	26.801	7.010	1	1	0.716

measure shows a dramatic improvement over published methods.

In summary, we propose a formula for similarity measure as

$$S(w_1, w_2) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}, \tag{20}$$

where $\alpha \geq 0$ and $\beta > 0$ are parameters scaling the contribution of shortest path length and depth, respectively. Based on the benchmark data set, the optimal parameters for the proposed measure are: $\alpha = 0.2, \beta = 0.6$. The experimental results demonstrated that our measure significantly outperforms published measures and is close to individual human judgement (with a correlation of 0.9015). This verified our hypothesis that human similarity judgement is a type of nonlinear process over information sources. However, since our similarity measure was intuitively and empirically derived, we believe that our measure may be improved further if the true nonlinear function is found. One way to obtain the optimal nonlinear function may be to learn it from a large set of word pairs with human ratings. Such a set is not currently available and it will need to be constructed by a carefully defined and controlled experimental process. This deserves further investigation in the future.

6 Conclusion

This paper presented word similarity measures from a new perspective. We argue that all first-hand information sources need to be properly processed in defining a similarity measure. First-hand information sources are infinite to some extent, for example, the information content would tend to infinity if the probability of concept approaches zero in corpus. On the other hand, humans compare word similarity with a finite interval between completely similar and nothing similar. Thus, the transformation from the infinite interval to a finite interval is intuitively nonlinear. Based on this consideration and empirical results, we proposed a new

TABLE 5
Correlation of Different Strategies against Human Similarity Judgements on Test Set D_0

Strategy	S_1	S_2 α=0.05 β=1	S_3 α=0.25	S_4 α=0.2 β=0.6	S_5 λ=0	S_6 λ=0.1	S_7 λ=0.15	S_8 λ=0.35	S_9 λ=0.4	S_{10} β=0.15	S_d
Correlation	0.664	0.8315	0.8911	0.8914	0.8914	0.8012	0.7951	0.8741	0.8772	0.8356	0.8218

similarity measure which combines the shortest path length and the depth of subsumer nonlinearly. In a similar manner to other researchers, we carried out experiments on a benchmark set of word pairs with human similarity ratings. The best correlation against Rubenstein-Goodenough's human similarity ratings in literature has been 0.8484 [5, 11], while ours is 0.8914. The experimental results demonstrated that our measure significantly outperforms previous published measures.

References

[1] S. Abney and M. Light, "Hiding a Semantic Class Hierarchy in a Markov Model," *Proc. ACL Workshop Unsupervised Learning in Natural Language Processing,* pp. 1-8, 1999.

[2] E. Agirre and G. Rigau, "Word Sense Disambiguation Using Conceptual Density," *Proc. 16th Int'l Conf. Computational Linguistics,* 1996.

[3] C.M. Aguilar and D.L. Medin, "Asymmetries of Comparison," *Psychonomic Bulletin & Rev.,* vol. 6, no. 2, pp. 328-337, 1999.

[4] A. Budanitsky and G. Hirst, "Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures," *Proc. Workshop WordNet and Other Lexical Resources, Second Meeting North Am. Chapter Assoc. for Computational Linguistics,* June 2001.

[5] A. Budanitsky, "Lexical Semantic Relatedness and Its Application in Natural Language Processing," Technical Report CSRG-390, Dept. of Computer Science, Univ. of Toronto, Aug. 1999.

[6] A.L. Edwards, *An Introduction to Linear Regression and Correlation.* San Francisco: W.H. Freeman, 1976.

[7] W.N. Francis and H. Kucera, "Brown Corpus Manual—Revised and Amplified," Dept. of Linguistics, Brown Univ., Providence, R.I., 1979.

[8] S.J. Green, "Building Hypertext Links by Computing Semantic Similarity," *IEEE Trans. Knowledge and Data Eng.,* vol. 11, no. 5, pp. 713-730, Sept./Oct. 1999.

[9] J.J. Jiang and D.W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," *Proc. ROCLING X,* 1997.

[10] H. Kozima, "Computing Lexical Cohesion as a Tool for Text Analysis," doctoral thesis, Computer Science and Information Math., Graduate School of Electro-Comm., Univ. of Electro-Comm., 1994.

[11] D. Lin, "An Information-Theoretic Definition of Similarity," *Proc. Int'l Conf. Machine Learning,* July 1998.

[12] M. McHale, "A Comparison of WordNet and Roget's Taxonomy for Measuring Semantic Similarity," *Proc. COLING/ACL Workshop Usage of WordNet in Natural Language Processing Systems,* 1998.

[13] D.L. Medin, R.L. Goldstone, and D. Gentner, "Respects for Similarity," *Psychological Rev.,* vol. 100, no. 2, pp. 254-278, 1993.

[14] G.A. Miller and W.G. Charles, "Contextual Correlates of Semantic Similarity," *Language and Cognitive Processes,* vol. 6, no. 1, pp. 1-28, 1991.

[15] G.A. Miller, "WordNet: A Lexical Database for English," *Comm. ACM,* vol. 38, no. 11, pp. 39-41, 1995.

[16] R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and Application of a Metric on Semantic Nets," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 9, no. 1, pp. 17-30, Jan. 1989.

[17] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," *Proc. 14th Int'l Joint Conf. Artificial Intelligence,* 1995.

[18] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language," *J. Artificial Intelligence Research,* vol. 11, pp. 95-130, 1999.

[19] R. Richardson, A.F. Smeaton, and J. Murphy, "Using WordNet as a Knowledge Base for Measuring Semantic Similarity," Working paper CA-1294, School of Computer Applications, Dublin City Univ., Dublin, 1994.

[20] H. Rubenstein and J.B. Goodenough, "Contextual Correlates of Synonymy," *Comm. ACM,* vol. 8, pp. 627-633, 1965.

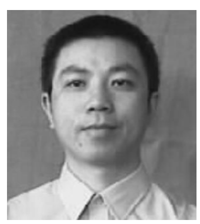
[21] R.N. Shepard, "Towards a Universal Law of Generalisation for Psychological Science," *Science,* vol. 237, pp.1317-1323, 1987.

[22] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

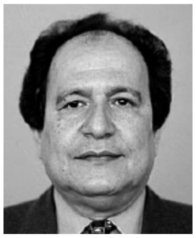
[23] R.K. Srihari, Z.F. Zhang, and A.B. Rao, "Intelligent Indexing and Semantic Retrieval of Multimodal Documents," *Information Retrieval,* vol. 2, pp. 245-275, 2000.

[24] M. Sussna, "Word Sense Disambiguation for Free-Text Indexing Using a Massive Semantic Network," *Proc. Second Int'l Conf. Information and Knowledge Management,* 1993.

[25] A. Tversky, "Features of Similarity," *Psychological Rev.,* vol. 84, no. 4, pp. 327-352, 1977.



Yuhua Li received the BEng and MSc degrees from China in 1986 and 1989, respectively. He began his PhD study in the Department of Engineering at the University of Leicester in 1997. Since 2000, he has been a senior research fellow in the Department of Computing and Mathematics at Manchester Metropolitan University. His research interests include pattern recognition, neural networks, knowledge-based systems, signal processing, and condition monitoring.



Zuhair A. Bandar received the BSc (Eng.) degree in electrical engineering from Mosul University in 1972, the MSc degree in electronics from the University of Kent at Canterbury in 1974, and the PhD degree (AI and neural networks) from Brunel University in 1981. Between periods of study and after graduation, he spent time in industry forming companies and acting as a technical director. He joined Manchester Metropolitan University as a senior lecturer in 1988. He is a founding member and the managing director of ConvAgent Ltd., a company that undertakes fundamental research and development of conversational agents. He also performs research in adaptive psychological profiling from nonverbal behavior using artificial intelligence techniques. Aspects of this work have been patented.

David McLean received the BSc (hons) degree in computer science from the University of Leeds in 1982 and the PhD degree (neural networks), "Generalization in Continuous Data Domains," from Manchester Metropolitan University in 1996. From 1996 to 1997, he worked for DERA (Malvern) and Thomson Marconi Sonar and became a lecturer at Manchester Metropolitan University in 1997. He is a member of the Intelligent Systems Group and is a founding member of Convagent Ltd., which conducts research and develops applications in the field of conversational agents. His other main interest is in automatic psychological profiling from nonverbal behavior using artificial intelligence techniques. He is part of a group that currently holds a patent for such a system.