An Information-Theoretic Definition of Similarity

Dekang Lin
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada R3T 2N2

Abstract

Similarity is an important and widely used concept. Previous definitions of similarity are tied to a particular application or a form of knowledge representation. We present an information-theoretic definition of similarity that is applicable as long as there is a probabilistic model. We demonstrate how our definition can be used to measure the similarity in a number of different domains.

1 Introduction

Similarity is a fundamental and widely used concept. Many similarity measures have been proposed, such as information content [Resnik, 1995b], mutual information [Hindle, 1990], Dice coefficient [Frakes and Baeza-Yates, 1992], cosine coefficient [Frakes and Baeza-Yates, 1992], distance-based measurements [Lee et al., 1989; Rada et al., 1989], and feature contrast model [Tversky, 1977]. McGill *etc.* surveyed and compared 67 similarity measures used in information retrieval [McGill et al., 1979].

A problem with previous similarity measures is that each of them is tied to a particular application or assumes a particular domain model. For example, distance-based measures of concept similarity (e.g., [Lee et al., 1989; Rada et al., 1989]) assume that the domain is represented in a network. If a collection of documents is not represented as a network, the distance-based measures do not apply. The Dice and cosine coefficients are applicable only when the objects are represented as numerical feature vectors.

Another problem with the previous similarity measures is that their underlying assumptions are often not explicitly stated. Without knowing those assumptions, it is impossible to make theoretical arguments for or against any particular measure. Almost all of the comparisons and evaluations of previous similarity measures have been based on empirical results.

This paper presents a definition of similarity that achieves two goals:

Universality: We define similarity in information-theoretic terms. It is applicable as long as the domain has a probabilistic model. Since probability theory can be integrated with many kinds of knowledge representations, such as first order logic [Bacchus, 1988] and semantic networks [Pearl, 1988], our definition of similarity can be applied to many different domains where very different similarity measures had previously been proposed. Moreover, the universality of the definition also allows the measure to be used in domains where no similarity measure has previously been proposed, such as the similarity between ordinal values.

Theoretical Justification: The similarity measure is not defined directly by a formula. Rather, it is derived from a set of assumptions about similarity. In other words, if the assumptions are deemed reasonable, the similarity measure necessarily follows.

The remainder of this paper is organized as follows. The next section presents the derivation of a similarity measure from a set of assumptions about similarity. Sections 3 through 6 demonstrate the universality of our proposal by applying it to different domains. The properties of different similarity measures are compared in Section 7.

2 Definition of Similarity

Since our goal is to provide a formal definition of the intuitive concept of similarity, we first clarify our intuitions about similarity.

Intuition 1: The similarity between A and B is related to their commonality. The more commonality they share, the more similar they are.

Intuition 2: The similarity between A and B is related to the differences between them. The more differences they have, the less similar they are.

Intuition 3: The maximum similarity between A and B is reached when A and B are identical, no matter how much commonality they share.

Our goal is to arrive at a definition of similarity that captures the above intuitions. However, there are many alternative ways to define similarity that would be consistent with the intuitions. In this section, we first make a set of additional assumptions about similarity that we believe to be reasonable. A similarity measure can then be derived from those assumptions.

In order to capture the intuition that the similarity of two objects are related to their commonality, we need a measure of commonality. Our first assumption is:

Assumption 1: The commonality between A and B is measured by

$$I(\text{common}(A, B))$$

where $\text{common}(A, B)$ is a proposition that states the commonalities between A and B; $I(s)$ is the amount of information contained in a proposition s.

For example, if A is an orange and B is an apple. The proposition that states the commonality between A and B is "fruit(A) and fruit(B)". In information theory [Cover and Thomas, 1991], the information contained in a statement is measured by the negative logarithm of the probability of the statement. Therefore,

$$I(\text{common}(A, B)) = -\log P(\text{fruit}(A) \text{ and fruit}(B))$$

We also need a measure of the differences between two objects. Since knowing both the commonalities and the differences between A and B means knowing what A and B are, we assume:

Assumption 2: The differences between A and B is measured by

$$I(\text{description}(A, B)) - I(\text{common}(A, B))$$

where $\text{description}(A, B)$ is a proposition that describes what A and B are.

Intuition 1 and 2 state that the similarity between two objects are related to their commonalities and differences. We assume that commonalities and differences are the only factors.

Assumption 3: The similarity between A and B, $\text{sim}(A, B)$, is a function of their commonalities and differences. That is,
$\text{sim}(A, B) = f(I(\text{common}(A, B)), I(\text{description}(A, B)))$
The domain of f is $\{(x, y)|x \geq 0, y > 0, y \geq x\}$.

Intuition 3 states that the similarity measure reaches a constant maximum when the two objects are identical. We assume the constant is 1.

Assumption 4: The similarity between a pair of identical objects is 1.

When A and B are identical, knowing their commonalities means knowing what they are, i.e., $I(\text{common}(A, B)) = I(\text{description}(A, B))$. Therefore, the function f must have the property: $\forall x > 0, f(x, x) = 1$.

When there is no commonality between A and B, we assume their similarity is 0, no matter how different they are. For example, the similarity between "depth-first search" and "leather sofa" is neither higher nor lower than the similarity between "rectangle" and "interest rate".

Assumption 5: $\forall y > 0, f(0, y) = 0$.

Suppose two objects A and B can be viewed from two independent perspectives. Their similarity can be computed separately from each perspective. For example, the similarity between two documents can be calculated by comparing the sets of words in the documents or by comparing their stylistic parameter values, such as average word length, average sentence length, average number of verbs per sentence, etc. We assume that the overall similarity of the two documents is a weighted average of their similarities computed from different perspectives. The weights are the amounts of information in the descriptions. In other words, we make the following assumption:

Assumption 6:

$$\forall x_1 \leq y_1, x_2 \leq y_2 : f(x_1 + x_2, y_1 + y_2) = \frac{y_1}{y_1+y_2} f(x_1, y_1) + \frac{y_2}{y_1+y_2} f(x_2, y_2)$$

From the above assumptions, we can proved the following theorem:

Similarity Theorem: The similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are:

$$\text{sim}(A, B) = \frac{\log P(\text{common}(A, B))}{\log P(\text{description}(A, B))}$$

Proof:

$$\begin{aligned}
& f(x, y) \\
= \;& f(x + 0, x + (y - x)) \\
= \;& \tfrac{x}{y} \times f(x, x) + \tfrac{y-x}{y} \times f(0, y - x) && \text{(Assumption 6)} \\
= \;& \tfrac{x}{y} \times 1 + \tfrac{y-x}{y} \times 0 = \tfrac{x}{y} && \text{(Assumption 4 and 5)}
\end{aligned}$$

Q.E.D.

Since similarity is the ratio between the amount of information in the commonality and the amount of information in the description of the two objects, if we know the commonality of the two objects, their similarity tells us how much more information is needed to determine what these two objects are.

In the next 4 sections, we demonstrate how the above definition can be applied in different domains.

3 Similarity between Ordinal Values

Many features have ordinal values. For example, the "quality" attribute can take one of the following values "excellent", "good", "average", "bad", or "awful". None of the previous definitions of similarity provides a measure for the similarity between two ordinal values. We now show how our definition can be applied here.

If "the quality of X is excellent" and "the quality of Y is average", the maximally specific statement that can be said of both X and Y is that "the quality of X and Y are between "average" and "excellent". Therefore, the commonality between two ordinal values is the interval delimited by them.

Suppose the distribution of the "quality" attribute is known (Figure 1). The following are four examples of similarity calculations:

$$\begin{aligned}
\mathrm{sim}(excellent, good) &= \frac{2\times\log P(excellent \vee good)}{\log P(excellent)+\log P(good)} \\
&= \frac{2\times\log(0.05+0.10)}{\log 0.05+\log 0.10} = 0.72 \\
\mathrm{sim}(good, average) &= \frac{2\times\log P(good \vee average)}{\log P(average)+\log P(good)} \\
&= \frac{2\times\log(0.10+0.50)}{\log 0.10+\log 0.50} = 0.34 \\
\mathrm{sim}(excellent, average) &= \frac{2\times\log P(excellent \vee good \vee average)}{\log P(excellent)+\log P(average)} \\
&= \frac{2\times\log(0.05+0.10+0.50)}{\log 0.05+\log 0.50} = 0.23 \\
\mathrm{sim}(good, bad) &= \frac{2\times\log P(good \vee average \vee bad)}{\log P(good)+\log P(bad)} \\
&= \frac{2\times\log(0.10+0.50+.20)}{\log 0.10+\log 0.20} = 0.11
\end{aligned}$$

The results show that, given the probability distribution in Figure 1, the similarity between "excellent" and "good" is much higher than the similarity between "good" and "average"; the similarity between "excellent" and "average" is much higher than the similarity between "good" and "bad".

4 Feature Vectors

Feature vectors are one of the simplest and most commonly used forms of knowledge representation, especially in case-based reasoning [Aha et al., 1991; Stanfill and Waltz, 1986] and machine learning. Weights are often assigned to features to account for the fact that the dissimilarity caused by more important features is greater than the dissimilarity

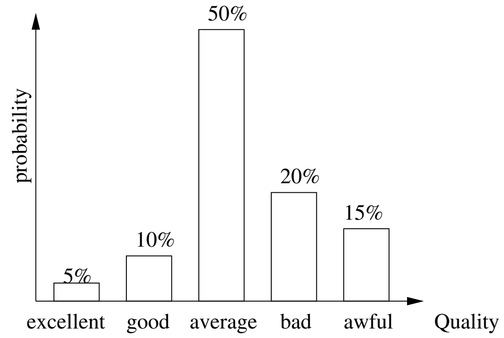

Figure 1: Example Distribution of Ordinal Values

caused by less important features. The assignment of the weight parameters is generally heuristic in nature in previous approaches. Our definition of similarity provides a more principled approach, as demonstrated in the following case study.

4.1 String Similarity—A case study

Consider the task of retrieving from a word list the words that are derived from the same root as a given word. For example, given the word "eloquently", our objective is to retrieve the other related words such as "ineloquent", "ineloquently", "eloquent", and "eloquence". To do so, assuming that a morphological analyzer is not available, one can define a similarity measure between two strings and rank the words in the word list in descending order of their similarity to the given word. The similarity measure should be such that words derived from the same root as the given word should appear early in the ranking.

We experimented with three similarity measures. The first one is defined as follows:

$$\mathrm{sim}_{\mathrm{edit}}(x, y) = \frac{1}{1 + \mathrm{editDist}(x, y)}$$

where $\mathrm{editDist}(x, y)$ is the minimum number of character insertion and deletion operations needed to transform one string to the other.

The second similarity measure is based on the number of different trigrams in the two strings:

$$\mathrm{sim}_{\mathrm{tri}}(x, y) = \frac{1}{1 + |\mathrm{tri}(x)| + |\mathrm{tri}(y)| - 2 \times |\mathrm{tri}(x) \cap \mathrm{tri}(y)|}$$

where $\mathrm{tri}(x)$ is the set of trigrams in x. For example, tri(eloquent) = $\{$elo, loq, oqu, que, ent$\}$.

Table 1: Top-10 Most Similar Words to "grandiloquent'

Rank	sim_{edit}		sim_{tri}		sim	
1	grandiloquently	1/3	grandiloquently	1/2	grandiloquently	0.92
2	grandiloquence	1/4	grandiloquence	1/4	grandiloquence	0.89
3	magniloquent	1/6	eloquent	1/8	eloquent	0.61
4	gradient	1/6	grand	1/9	magniloquent	0.59
5	grandaunt	1/7	grande	1/10	ineloquent	0.55
6	gradients	1/7	rand	1/10	eloquently	0.55
7	grandiose	1/7	magniloquent	1/10	ineloquently	0.50
8	diluent	1/7	ineloquent	1/10	magniloquence	0.50
9	ineloquent	1/8	grands	1/10	eloquence	0.50
10	grandson	1/8	eloquently	1/10	ventriloquy	0.42

Table 2: Evaluation of String Similarity Measures

Root	Meaning	$\|W_{root}\|$	11-point average precisions		
			sim_{edit}	sim_{tri}	sim
agog	leader, leading, bring	23	37%	40%	70%
cardi	heart	56	18%	21%	47%
circum	around, surrounding	58	24%	19%	68%
gress	to step, to walk, to go	84	22%	31%	52%
loqu	to speak	39	19%	20%	57%

The third similarity measure is based on our proposed definition of similarity under the assumption that the probability of a trigram occurring in a word is independent of other trigrams in the word:

$$\text{sim}(x, y) = \frac{2 \times \sum_{t \in \text{tri}(x) \cap \text{tri}(y)} \log P(t)}{\sum_{t \in \text{tri}(x)} \log P(t) + \sum_{t \in \text{tri}(y)} \log P(t)}$$

Table 1 shows the top 10 most similar words to "grandiloquent" according to the above three similarity measures.

To determine which similarity measure ranks higher the words that are derived from the same root as the given word, we adopted the evaluation metrics used in the Text Retrieval Conference [Harman, 1993]. We used a 109,582-word list from the AI Repository.[1] The probabilities of trigrams are estimated by their frequencies in the words. Let W denote the set of words in the word list and W_{root} denote the subset of W that are derived from $root$. Let $(w_1, \ldots, w_n)$ denote the ordering of $W - \{w\}$ in descending similarity to w according to a similarity measure. The precision of $(w_1, \ldots, w_n)$ at recall level N% is defined as the maximum value of $\frac{|W_{root} \cap \{w_1, \ldots, w_k\}|}{k}$ such that $k \in \{1, \ldots, n\}$ and $\frac{|W_{root} \cap \{w_1, \ldots, w_k\}|}{|W_{root}|} \geq N\%$. The quality of the sequence $(w_1, \ldots, w_n)$ can be measured by the 11-point average of its precisions at recall levels 0%, 10%, 20%, ..., and 100%. The average precision values are then averaged over all the words in W_{root}. The results on 5 roots are shown in Table 2. It can be seen that much better results were achieved with sim than with the other similarity measures. The reason for this is that sim_{edit} and sim_{tri} treat all characters or trigrams equally, whereas sim is able to automatically take into account the varied importance in different trigrams.

[1] http://www.cs.cmu.edu/afs/cs/project/ai-repository

5 Word Similarity

In this section, we show how to measure similarities between words according to their distribution in a text corpus [Pereira et al., 1993]. Similar to [Alshawi and Carter, 1994; Grishman and Sterling, 1994; Ruge, 1992], we use a parser to extract dependency triples from the text corpus. A dependency triple consists of a head, a dependency type and a modifier. For example, the dependency triples in "I have a brown dog" consist of:

(1) (have subj I), (have obj dog), (dog adj-mod brown), (dog det a)

where "subj" is the relationship between a verb and its subject; "obj" is the relationship between a verb and its object;

"adj-mod" is the relationship between a noun and its adjective modifier and "det" is the relationship between a noun and its determiner.

We can view dependency triples extracted from a corpus as features of the heads and modifiers in the triples. Suppose (avert obj duty) is a dependency triple, we say that "duty" has the feature obj-of(avert) and "avert" has the feature obj(duty). Other words that also possess the feature obj-of(avert) include "default", "crisis", "eye", "panic", "strike", "war", etc., which are also used as objects of "avert" in the corpus.

Table 3 shows a subset of the features of "duty" and "sanction". Each row corresponds to a feature. A 'x' in the "duty" or "sanction" column means that the word possesses that feature.

Table 3: Features of "duty" and "sanction"

Feature	duty	sanction	$I(f_i)$
f_1: subj-of(include)	x	x	3.15
f_2: obj-of(assume)	x		5.43
f_3: obj-of(avert)	x	x	5.88
f_4: obj-of(ease)		x	4.99
f_5: obj-of(impose)	x	x	4.97
f_6: adj-mod(fiduciary)	x		7.76
f_7: adj-mod(punitive)	x	x	7.10
f_8: adj-mod(economic)		x	3.70

Let $F(w)$ be the set of features possessed by w. $F(w)$ can be viewed as a description of the word w. The commonalities between two words w_1 and w_2 is then $F(w_1) \cap F(w_2)$.

The similarity between two words is defined as follows:

(2) $\text{sim}(w_1, w_2) = \frac{2 \times I(F(w_1) \cap F(w_2))}{I(F(w_1)) + I(F(w_2))}$

where $I(S)$ is the amount of information contained in a set of features S. Assuming that features are independent of one another, $I(S) = -\sum_{f \in S} \log P(f)$, where $P(f)$ is the probability of feature f. When two words have identical sets of features, their similarity reaches the maximum value of 1. The minimum similarity 0 is reached when two words do not have any common feature.

The probability $P(f)$ can be estimated by the percentage of words that have feature f among the set of words that have the same part of speech. For example, there are 32868 unique nouns in a corpus, 1405 of which were used as subjects of "include". The probability of subj-of(include) is $\frac{1405}{32868}$. The probability of the feature adj-mod(fiduciary) is $\frac{14}{32868}$ because only 14 (unique) nouns were modified by "fiduciary". The amount of information in the feature adj-mod(fiduciary), 7.76, is greater than the amount of information in subj-of(include), 3.15. This agrees with our intuition that saying that a word can be modified by "fiduciary" is more informative than saying that the word can be the subject of "include".

The fourth column in Table 3 shows the amount of information contained in each feature. If the features in Table 3 were all the features of "duty" and "sanction", the similarity between "duty" and "sanction" would be:

$$\frac{2 \times I(\{f_1, f_3, f_5, f_7\})}{I(\{f_1, f_2, f_3, f_5, f_6, f_7\}) + I(\{f_1, f_3, f_4, f_5, f_7, f_8\})}$$

which is equal to 0.66.

We parsed a 22-million-word corpus consisting of Wall Street Journal and San Jose Mercury with a principle-based broad-coverage parser, called PRINCIPAR [Lin, 1993; Lin, 1994]. Parsing took about 72 hours on a Pentium 200 with 80MB memory. From these parse trees we extracted about 14 million dependency triples. The frequency counts of the dependency triples are stored and indexed in a 62MB dependency database, which constitutes the set of feature descriptions of all the words in the corpus. Using this dependency database, we computed pairwise similarity between 5230 nouns that occurred at least 50 times in the corpus.

The words with similarity to "duty" greater than 0.04 are listed in (3) in descending order of their similarity.

(3) responsibility, position, sanction, tariff, obligation, fee, post, job, role, tax, penalty, condition, function, assignment, power, expense, task, deadline, training, work, standard, ban, restriction, authority, commitment, award, liability, requirement, staff, membership, limit, pledge, right, chore, mission, care, title, capability, patrol, fine, faith, seat, levy, violation, load, salary, attitude, bonus, schedule, instruction, rank, purpose, personnel, worth, jurisdiction, presidency, exercise.

The following is the entry for "duty" in the Random House Thesaurus [Stein and Flexner, 1984].

(4) **duty** *n.* 1. obligation, responsibility; onus; business, province; 2. function, task, assignment, charge. 3. tax, tariff, customs, excise, levy.

The shadowed words in (4) also appear in (3). It can be seen that our program captured all three senses of "duty" in [Stein and Flexner, 1984].

Two words are a pair of respective nearest neighbors (RNNs) if each is the other's most similar word. Our program found 622 pairs of RNNs among the 5230 nouns that

Table 4: Respective Nearest Neighbors

Rank	RNN	Sim
1	earnings profit	0.50
11	revenue sale	0.39
21	acquisition merger	0.34
31	attorney lawyer	0.32
41	data information	0.30
51	amount number	0.27
61	downturn slump	0.26
71	there way	0.24
81	fear worry	0.23
91	jacket shirt	0.22
101	film movie	0.21
111	felony misdemeanor	0.21
121	importance significance	0.20
131	reaction response	0.19
141	heroin marijuana	0.19
151	championship tournament	0.18
161	consequence implication	0.18
171	rape robbery	0.17
181	dinner lunch	0.17
191	turmoil upheaval	0.17
201	biggest largest	0.17
211	blaze fire	0.16
221	captive westerner	0.16
231	imprisonment probation	0.16
241	apparel clothing	0.15
251	comment elaboration	0.15
261	disadvantage drawback	0.15
271	infringement negligence	0.15
281	angler fishermen	0.14
291	emission pollution	0.14
301	granite marble	0.14
311	gourmet vegetarian	0.14
321	publicist stockbroker	0.14
331	maternity outpatient	0.13
341	artillery warplanes	0.13
351	psychiatrist psychologist	0.13
361	blunder fiasco	0.13
371	door window	0.13
381	counseling therapy	0.12
391	austerity stimulus	0.12
401	ours yours	0.12
411	procurement zoning	0.12
421	neither none	0.12
431	briefcase wallet	0.11
441	audition rite	0.11
451	nylon silk	0.11
461	columnist commentator	0.11
471	avalanche raft	0.11
481	herb olive	0.11
491	distance length	0.10
501	interruption pause	0.10
511	ocean sea	0.10
521	flying watching	0.10
531	ladder spectrum	0.09
541	lotto poker	0.09
551	camping skiing	0.09
561	lip mouth	0.09
571	mounting reducing	0.09
581	pill tablet	0.08
591	choir troupe	0.08
601	conservatism nationalism	0.08
611	bone flesh	0.07
621	powder spray	0.06

occurred at least 50 times in the parsed corpus. Table 4 shows every 10th RNN.

Some of the pairs may look peculiar. Detailed examination actually reveals that they are quite reasonable. For example, the 221 ranked pair is "captive" and "westerner". It is very unlikely that any manually created thesaurus will list them as near-synonyms. We manually examined all 274 occurrences of "westerner" in the corpus and found that 55% of them refer to westerners in captivity. Some of the bad RNNs, such as (avalanche, raft), (audition, rite), were due to their relative low frequencies,[2] which make them susceptible to accidental commonalities, such as:

(5) The {avalanche, raft} {drifted, hit} ....
To {hold, attend} the {audition, rite}.
An uninhibited {audition, rite}.

6 Semantic Similarity in a Taxonomy

Semantic similarity [Resnik, 1995b] refers to similarity between two concepts in a taxonomy such as the WordNet [Miller, 1990] or CYC upper ontology. The semantic similarity between two classes C and C' is not about the classes themselves. When we say "rivers and ditches are similar", we are not comparing the set of rivers with the set of ditches. Instead, we are comparing a generic river and a generic ditch. Therefore, we define $\text{sim}(C, C')$ to be the similarity between x and x' if all we know about x and x' is that $x \in C$ and $x' \in C'$.

The two statements "$x \in C$" and "$x' \in C'$" are independent (instead of being assumed to be independent) because the selection of a generic C is not related to the selection of a generic C'. The amount of information contained in "$x \in C$ and $x' \in C'$" is

$$-\log P(C) - \log P(C')$$

where $P(C)$ and $P(C')$ are probabilities that a randomly selected object belongs to C and C', respectively.

Assuming that the taxonomy is a tree, if $x_1 \in C$ and $x_2 \in C_2$, the commonality between x_1 and x_2 is $x_1 \in C_0 \wedge w_2 \in C_0$, where C_0 is the most specific class that subsumes both C_1 and C_2. Therefore,

$$\text{sim}(x_1, x_2) = \frac{2 \times \log P(C_0)}{\log P(C_1) + \log P(C_2)}$$

For example, Figure 2 is a fragment of the WordNet. The number attached to each node C is $P(C)$. The similarity

[2] They all occurred 50–60 times in the parsed corpus.

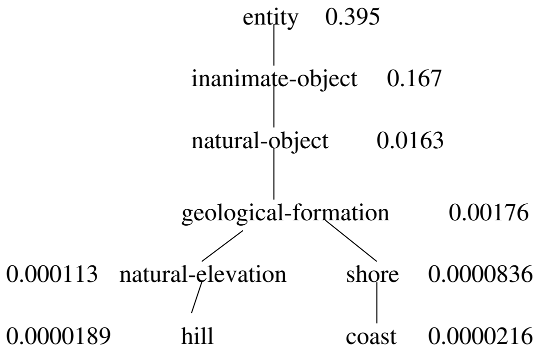

Figure 2: A Fragment of WordNet

between the concepts of Hill and Coast is:

$$\text{sim}(\text{Hill}, \text{Coast}) = \frac{2 \times \log P(\text{Geological-Formation})}{\log P(\text{Hill}) + \log P(\text{Coast})}$$

which is equal to 0.59.

There have been many proposals to use the distance between two concepts in a taxonomy as the basis for their similarity [Lee et al., 1989; Rada et al., 1989]. Resnik [Resnik, 1995b] showed that the distance-based similarity measures do not correlate to human judgments as well as his measure. Resnik's similarity measure is quite close to the one proposed here: $\text{sim}_{\text{Resnik}}(A, B) = \frac{1}{2} I(\text{common}(A, B))$. For example, in Figure 2, $\text{sim}_{\text{Resnik}}(\text{Hill}, \text{Coast}) = -\log P(\text{Geological-Formation})$.

Wu and Palmer [Wu and Palmer, 1994] proposed a measure for semantic similarity that could be regarded as a special case of $\text{sim}(A, B)$:

$$\text{sim}_{\text{Wu\&Palmer}}(A, B) = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3}$$

where N_1 and N_2 are the number of IS-A links from A and B to their most specific common superclass C; N_3 is the number of IS-A links from C to the root of the taxonomy. For example, the most specific common superclass of Hill and Coast is Geological-Formation. Thus, $N_1 = 2$, $N_2 = 2$, $N_3 = 3$ and $\text{sim}_{\text{Wu\&Palmer}}(\text{Hill}, \text{Coast}) = 0.6$.

Interestingly, if $P(C|C')$ is the same for all pairs of concepts such that there is an IS-A link from C to C' in the taxonomy, $\text{sim}_{\text{Wu\&Palmer}}(A, B)$ coincides with $\text{sim}(A, B)$.

Resnik [Resnik, 1995a] evaluated three different similarity measures by correlating their similarity scores on 28 pairs of concepts in the WordNet with assessments made by human subjects [Miller and Charles, 1991]. We adopted the same data set and evaluation methodology to compare $\text{sim}_{\text{Resnik}}$, $\text{sim}_{\text{Wu\&Palmer}}$ and sim. Table 5 shows the similarities between 28 pairs of concepts, using three different similarity measures. Column Miller&Charles lists the average similarity scores (on a scale of 0 to 4) assigned by human subjects in Miller&Charles's experiments [Miller and Charles, 1991]. Our definition of similarity yielded slightly higher correlation with human judgments than the other two measures.

Table 5: Results of Comparison between Semantic Similarity Measures

Word Pair	Miller& Charles	Resnik	Wu & Palmer	sim
car, automobile	3.92	11.630	1.00	1.00
gem, jewel	3.84	15.634	1.00	1.00
journey, voyage	3.84	11.806	.91	.89
boy, lad	3.76	7.003	.90	.85
coast, shore	3.70	9.375	.90	.93
asylum, madhouse	3.61	13.517	.93	.97
magician, wizard	3.50	8.744	1.00	1.00
midday, noon	3.42	11.773	1.00	1.00
furnace, stove	3.11	2.246	.41	.18
food, fruit	3.08	1.703	.33	.24
bird, cock	3.05	8.202	.91	.83
bird, crane	2.97	8.202	.78	.67
tool, implement	2.95	6.136	.90	.80
brother, monk	2.82	1.722	.50	.16
crane, implement	1.68	3.263	.63	.39
lad, brother	1.66	1.722	.55	.20
journey, car	1.16	0	0	0
monk, oracle	1.10	1.722	.41	.14
food, rooster	0.89	.538	.7	.04
coast, hill	0.87	6.329	.63	.58
forest, graveyard	0.84	0	0	0
monk, slave	0.55	1.722	.55	.18
coast, forest	0.42	1.703	.33	.16
lad, wizard	0.42	1.722	.55	.20
chord, smile	0.13	2.947	.41	.20
glass, magician	0.11	.538	.11	.06
noon, string	0.08	0	0	0
rooster, voyage	0.08	0	0	0
Correlation with Miller & Charles	1.00	0.795	0.803	0.834

7 Comparison between Different Similarity Measures

One of the most commonly used similarity measure is call Dice coefficient. Suppose two objects can be described with two numerical vectors $(a_1, a_2, \ldots, a_n)$ and

Table 6: Comparison between Similarity Measures

	Similarity Measures: WP: sim$_{\text{Wu\&Palmer}}$ R: sim$_{\text{Resnik}}$ Dice: sim$_{\text{dice}}$				
Property	sim	WP	R	Dice	sim$_{\text{dist}}$
increase with commonality	yes	yes	yes	yes	no
decrease with difference	yes	yes	no	yes	yes
triangle inequality	no	no	no	no	yes
Assumption 6	yes	yes	no	yes	no
max value=1	yes	yes	no	yes	yes
semantic similarity	yes	yes	yes	no	yes
word similarity	yes	no	no	yes	yes
ordinal values	yes	no	no	no	no

$(b_1, b_2, \ldots, b_n)$, their Dice coefficient is defined as

$$\text{sim}_{\text{dice}}(A, B) = \frac{2 \times \sum_{i=1,n} a_i b_i}{\sum_{i=1,n} a_i^2 + \sum_{i=1,n} b_i^2}.$$

Another class of similarity measures is based a distance metric. Suppose $dist(A, B)$ is a distance metric between two objects, sim$_{\text{dist}}$ can be defined as follows:

$$\text{sim}_{\text{dist}}(A, B) = \frac{1}{1 + dist(A, B)}$$

Table 6 summarizes the comparison among 5 similarity measures.

Commonality and Difference: While most similarity measures increase with commonality and decrease with difference, sim$_{\text{dist}}$ only decreases with difference and sim$_{\text{Resnik}}$ only takes commonality into account.

Triangle Inequality: A distance metrics must satisfy the triangle inequality:

$$dist(A, C) \leq dist(A, B) + dist(B, C).$$

Consequently, sim$_{\text{dist}}$ has the property that sim$_{\text{dist}}(A, C)$ cannot be arbitrarily close to 0 if none of sim$_{\text{dist}}(A, B)$ and sim$_{\text{dist}}(B, C)$ is 0. This can be counter-intuitive in some situations. For example, in Figure 3, A and B are similar in their shades, B and C are similar in their shape, but A and C are not similar.

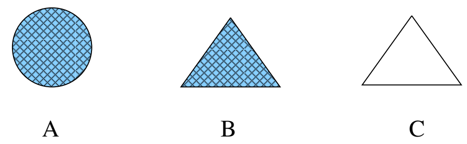

Figure 3: Counter-example of Triangle Inequality

Assumption 6: The strongest assumption that we made in Section 2 is Assumption 6. However, this assumption is not unique to our proposal. Both sim$_{\text{Wu\&Palmer}}$ and sim$_{\text{dice}}$ also satisfy Assumption 6. Suppose two objects A and B are represented by two feature vectors $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$, respectively. Without loss of generality, suppose the first k features and the rest $n - k$ features represent two independent perspectives of the objects.

$$\text{sim}_{\text{dice}}(A, B) = \frac{2 \times \sum_{i=1,n} a_i b_i}{\sum_{i=1,n} a_i^2 + \sum_{i=1,n} b_i^2} = \frac{\sum_{i=1,k} a_i^2 + \sum_{i=1,k} b_i^2}{\sum_{i=1,n} a_i^2 + \sum_{i=1,n} b_i^2} \frac{2 \times \sum_{i=1,k} a_i b_i}{\sum_{i=1,k} a_i^2 + \sum_{i=1,k} b_i^2} + \frac{\sum_{i=k+1,n} a_i^2 + \sum_{i=k+1,n} b_i^2}{\sum_{i=1,n} a_i^2 + \sum_{i=1,n} b_i^2} \frac{2 \times \sum_{i=k+1,n} a_i b_i}{\sum_{i=k+1,n} a_i^2 + \sum_{i=k+1,n} b_i^2}$$

which is a weighted average of the similarity between A and B in each of the two perspectives.

Maximum Similarity Values: With most similarity measures, the maximum similarity is 1, except sim$_{\text{Resnik}}$, which have no upper bound for similarity values.

Application Domains: The similarity measure proposed in this paper can be applied in all the domains listed in Table 6, including the similarity of ordinal values, where none of the other similarity measures is applicable.

8 Conclusion

Similarity is an important and fundamental concept in AI and many other fields. Previous proposals for similarity measures are heuristic in nature and tied to a particular domain or form of knowledge representation. In this paper, we present a universal definition of similarity in terms of information theory. The similarity measure is not directly stated as in earlier definitions, rather, it is derived from a set of assumptions. In other words, if one accepts the assumptions, the similarity measure necessarily follows. The universality of the definition is demonstrated by its applications in different domains where different similarity measures have been employed before.

Acknowledgment

The author wishes to thank the anonymous reviewers for their valuable comments. This research has been partially supported by NSERC Research Grant OGP121338.

References

[Aha et al., 1991] Aha, D., Kibler, D., and Albert, M. (1991). Instance-Based Learning Algorithms. *Machine Learning*, 6(1):37–66.

[Alshawi and Carter, 1994] Alshawi, H. and Carter, D. (1994). Training and scaling preference functions for disambiguation. *Computational Linguistics*, 20(4):635–648.

[Bacchus, 1988] Bacchus, F. (1988). *Representing and Reasoning with Probabilistic Knowledge*. PhD thesis, University of Alberta, Edmonton, Alberta, Canada.

[Cover and Thomas, 1991] Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory*. Wiley series in telecommunications. Wiley, New York.

[Frakes and Baeza-Yates, 1992] Frakes, W. B. and Baeza-Yates, R., editors (1992). *Information Retrieval, Data Structure and Algorithms*. Prentice Hall.

[Grishman and Sterling, 1994] Grishman, R. and Sterling, J. (1994). Generalizing automatically generated selectional patterns. In *Proceedings of COLING-94*, pages 742–747, Kyoto, Japan.

[Harman, 1993] Harman, D. (1993). Overview of the first text retrieval conference. In *Proceedings of SIGIR'93*, pages 191–202.

[Hindle, 1990] Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of ACL-90*, pages 268–275, Pittsburg, Pennsylvania.

[Lee et al., 1989] Lee, J. H., Kim, M. H., and Lee, Y. J. (1989). Information retrieval based on conceptual distance in is-a hierarchies. *Journal of Documentation*, 49(2):188–207.

[Lin, 1993] Lin, D. (1993). Principle-based parsing without overgeneration. In *Proceedings of ACL–93*, pages 112–120, Columbus, Ohio.

[Lin, 1994] Lin, D. (1994). Principar—an efficient, broad-coverage, principle-based parser. In *Proceedings of COLING–94*, pages 482–488. Kyoto, Japan.

[McGill et al., 1979] McGill et al., M. (1979). An evaluation of factors affecting document ranking by information retrieval systems. Project report, Syracuse University School of Information Studies.

[Miller, 1990] Miller, G. A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

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

[Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Mateo, California.

[Pereira et al., 1993] Pereira, F., Tishby, N., and Lee, L. (1993). Distributional Clustering of English Words. In *Proceedings of ACL-93*, pages 183–190, Ohio State University, Columbus, Ohio.

[Rada et al., 1989] Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application ofa metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics*, 19(1):17–30.

[Resnik, 1995a] Resnik, P. (1995a). Disambiguating noun groupings with respect to wordnet senses. In *Third Workshop on Very Large Corpora*. Association for Computational Linguistics.

[Resnik, 1995b] Resnik, P. (1995b). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI-95*, pages 448–453, Montreal, Canada.

[Ruge, 1992] Ruge, G. (1992). Experiments on linguistically based term associations. *Information Processing & Management*, 28(3):317–332.

[Stanfill and Waltz, 1986] Stanfill, C. and Waltz, D. (1986). Toward Memory-based Reasoning. *Communications of ACM*, 29:1213–1228.

[Stein and Flexner, 1984] Stein, J. and Flexner, S. B., editors (1984). *Random House College Thesaurus*. Random House, New York.

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

[Wu and Palmer, 1994] Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico.