Measuring Semantic Similarity Using WordNet-based Context Vectors

Shen Wan

Rafal A. Angryk

***Abstract*— Semantic relatedness between words or concepts is a fundamental problem in many applications of computational linguistics and artificial intelligence. In this paper, a new measure based on the semantic ontology database WordNet is proposed which combines gloss information of concepts with semantic relationships, and organizes concepts as high-dimensional vectors. Other relatedness measures are compared and an experimental evaluation against several benchmark sets of human similarity ratings is presented. The Context Vector measure is shown to have one of the best performances.**

***Index Terms*— semantic networks, concept hierarchies, ontology, word similarity, word relatedness, context, WordNet.**

I. Introduction

In many areas of artificial intelligence such as natural language processing and information retrieval, it is essential to interrelate semantic concepts. Consider the sentence: *I deposited a check at the bank*. An English speaker can easily recognize that the word *bank* here means *a financial institution*. Whereas in sentence *She sits on the bank of the river watching currents*, *bank* means *slope beside a body of water*. We can determine the correct sense of an ambiguous word from its context because words in a context tend to be interrelated. In the first example, the words *deposit*, *check* and *bank* are all financial concepts. And in the second example, the words *bank*, *river* and *current* are all related to body of water. The idea of word/concept relatedness has been applied in many areas, such as word sense disambiguation [1], automatic spelling error detection and correction [2], segmenting narratives into scenes [3], image retrieval [4], multimode document retrieval [5], and automatic hypertext generation [6].

Consequently, researchers have proposed many techniques [1][2][7]–[9] for solving this problem over the years. The goal is to automatically measure the relatedness of two words or concepts, and to match human judgments as closely as possible. Description and comparison of some of the techniques are discussed in section III.

In this paper, we propose a new, WordNet-based [24] semantic similarity measure. Our work is inspired by the extended gloss vector measure [19] and based on a popular hypothesis [18][19] that similar concepts tend to appear in similar linguistic contexts. It has also been shown [19] that contexts are important in defining the meaning of words, and context vectors are useful representations of word meanings.

Shen Wan is currently a Ph. D. student in Dept. of Computer Science, Montana State University at Bozeman, Bozeman, MT 59717–3880 USA phone: (406)994–7034; e-mail: swan@cs.montana.edu

Dr. Rafal A. Angryk is a faculty member in Dept. of Computer Science, Montana State University at Bozeman, Bozeman, MT 59717–3880 USA phone: (406)994–4440; e-mail: angryk@cs.montana.edu

Our *Context Vector* measure of semantic relatedness is based on 2^{nd} order co-occurrence vectors [23] and uses WordNet as a semantic ontology. It evaluates the similarity of words or concepts from the semantic information of the corpora and the semantic relations in the ontology. It has good accuracy when compared to other measures, with respect to benchmarks of human judgments. It can compare content words/concepts of any *Part Of Speech* (POS). It is adaptable to any corpus. It is also normalized and complies with the triangle inequality.

The remainder of this paper is organized as follows. The next section introduces the WordNet semantic database and its expansion—XWN. Section III describes and compares other relatedness measures. In section IV, we discuss our measurement algorithm in detail. The evaluation and comparison of our measure against other measures are in section V. Finally, the paper concludes and discusses tasks remaining as future work.

II. WordNet and eXtended WordNet

WordNet (http://wordnet.princeton.edu) is a semantic lexicon for English. It groups words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets.

In WordNet, each synset consists of synonymous words and belongs to one of the four POS (nouns, verbs, adjectives or adverbs). Nouns and verbs are organized into hierarchies based on the hypernymy/hyponymy relation. For two concepts X and Y, if every X is a (kind of) Y, then Y is X's hypernym and X is Y's hyponym. Holonymy/meronymy is another important relation for nouns. If X is a part/member of Y, then Y is X's holonym and X is Y's meronym. As an example, Fig. 1 depicts the hypernym hierarchy of four senses of the noun *cherry* in WordNet 2.0.

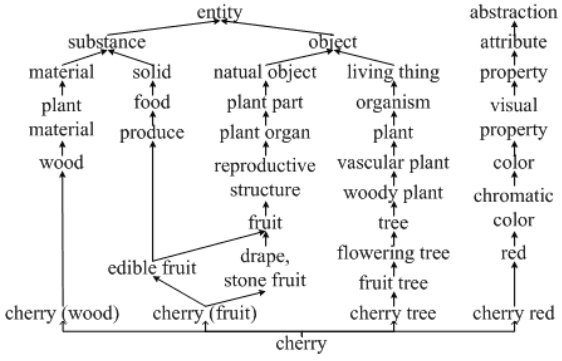

Fig. 1. Example of the Hypernymy Hierarchy of *cherry* in WordNet 2.0

In WordNet 2.0, there are 114 648 nouns in 79 689 synsets, 11 306 verbs in 13 508 synsets, 21 436 adjectives in 18 563 synsets, and 4 669 adverbs in 3 664 synsets, in a total of 144 309 words in 115 424 synsets. Detailed introduction to WordNet can be found in [24].

eXtended WordNet (XWN, http://xwn.hlt.utdallas.edu) provides enhancements to the WordNet database. In XWN, the WordNet glosses are syntactically parsed and transformed into logic forms and content words are semantically disambiguated. For each synset in WordNet 2.0, XWN provides its POS, synonym words, gloss in text, gloss in disambiguated senses, and other information that we are not interested in.

III. Related Works

To measure the relatedness of two objects, two approaches are usually used—define the similarity or the distance. Similarity is defined as a non-negative number with 0 representing most dissimilar, and a maximum value representing identical objects. Some similarity measures normalize the similarity to $[0, 1]$. Distance, sometimes referred as dissimilarity, is also defined as a non-negative number, but with 0 for identical objects and usually no upper limit. Given distance or similarity, the other can be easily calculated. e.g., $dist(A, B) \triangleq -\log sim(A, B)$. Generally, researchers define either distance or similarity at their convenience.

Rada *et al.* [7] defined the conceptual distance between two nodes in an *IS-A* semantic network as the length of the shortest path connecting the two nodes. Their work formed the basis for edge counting-based measures. The simplest edge counting method works well for domain specific applications with highly constrained taxonomies, such as medical semantic nets. However, it does not take the non-uniformity in the density of links into account, nor does it consider other semantic relationships like antonymy, holo-/meronymy, etc. Therefore, it does not perform well in a general semantic ontology like WordNet. Many measures were designed based on this edge counting principle, with consideration for other types of relationship and other attributes of the taxonomy.

In the following description of measures, we will use the notations illustrated in Fig. 2. For two synsets S_1 and S_2 in WordNet, *NCH* denotes their Nearest Common Hypernym; h is the length of the shortest path from *NCH* to the root node; l_1 and l_2 are lengths of shortest paths from S_1 and S_2 to *NCH* respectively; $l = l_1 + l_2$ is the length of the shortest path from S_1 to S_2; and $IC(S)$ is the information content of synset S calculated from some corpus: $IC(S) \triangleq -\log P(S)$, where $P(S)$ stands for the probability of S in that corpus. In WordNet 2.0 or any other multi-rooted taxonomy, a virtual root is added as hypernym of all top level concepts.

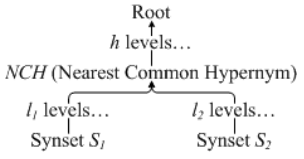

Fig. 2. Notations for a Hypernymy Hierarchy

Sussna [12] extended the edge counting measure with weighted edges. For a relation of type r, the weight of a directional edge $S_1 \rightarrow_r S_2$ is defined as $w(S_1 \rightarrow_r S_2) \triangleq \max_r - \frac{\max_r - \min_r}{n_r(S_1)}$, where $\max_r$ and $\min_r$ are the maximum and minimum weights for all edges of relation r respectively, and $n_r(S)$ is the number of edges of relation r leaving node S. The weight of an unidirectional edge $\overline{S_1 S_2}$ is defined as $w(\overline{S_1 S_2}) \triangleq \frac{w(S_1 \rightarrow_r S_2) + w(S_2 \rightarrow_{r'} S_1)}{2d}$, where r' is the inverse relation of r, and d is the depth of the deeper of the two nodes S_1 and S_2. The distance is defined as the shortest weighted path between the two nodes. To apply the measure to WordNet, synonymy and antonymy get weights of 0 and 2.5, whereas hypernymy, hyponymy, holonymy, and meronymy all have weights in $[1, 2]$.

Wu and Palmer [9] proposed the formula $sim(S_1, S_2) \triangleq \frac{2h}{l+2h}$ to calculate the conceptual similarity between two verbs and used the result for machine translation. Their measure can also be applied to nouns.

Resnik [1] first used information content to calculate semantic similarity in an *IS-A* taxonomy: $sim(S_1, S_2) \triangleq IC(NCH)$. This measure can be adapted to any domain, given a properly selected corpus. Following Resnik, several measures were proposed to take more factors into account. Some of those measures are introduced below.

Jiang and Conrath [14] pointed out that simply relying on either the taxonomy structure or the information content is inferior to combining these two factors together. They reached the formula: $dist(S_1, S_2) \triangleq IC(S_1) + IC(S_2) - 2IC(NCH)$.

Lin [11] also extended Resnik's measure by considering the information content of individual concepts together with the information content of the *NCH*. But Lin used a different formula: $sim(S_1, S_2) \triangleq \frac{2IC(NCH)}{IC(S_1)+IC(S_2)}$.

Leacock and Chodorow [13] determined the similarity as: $sim(S_1, S_2) \triangleq -\log \frac{l}{2d_{\max}}$, where $d_{\max}$ is the maximum depth of the taxonomy.

Hirst and St-Onge [15] used lexical chains to get similarity of words in WordNet. For paths with multiple edges, only some patterns are considered as valid lexical chains. They proposed the formula $sim(S_1, S_2) \triangleq 8 - l - N$, where N is the number of direction changes in the path.

Li, Bandar, and McLean [16] proposed several parameterized formulas considering multiple factors: l, h, and $IC(S)$ (local semantic density). After tuning each formula, they found the one that best correlates with human judgments: $sim(S_1, S_2) \triangleq e^{-\alpha l} \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}$, where $\alpha = 0.2$ and $\beta = 0.6$. However, this is just an empirical formula and lacks a theoretical basis.

All the measures mentioned above have a commonality: they all depend on a hierarchical structure, such as the *IS-A* hierarchy in WordNet. We believe that taxonomy structure is not enough to represent all possible relations between concepts. For instance, all physical entities will be far away from abstract entities, whereas they might have fairly strong relations in between. As an example, consider two obviously related words *nose* and *smell*. All measures above return a small relatedness between these terms because they

TABLE I

SIMILARITY OF NOSE/SMELL AND NOSE/ANYTHING

Measure	nose/smell	nose/anything
Hirst–St-Onge	0	0
Jiang–Conrath	0.054	0
Wu–Palmer	0.191	0.533
Path Length	0.056	0.125
Lin	0	0
Resnik	0	3.969
Leacock–Chodorow	0.747	1.558
Extended Gloss Overlaps	13	0
Gloss Vector	0.212	0.054

Results calculated using the WordNet::Similarity script. http://marimba.d.umn.edu/cgi-bin/similarity.cgi

are far away in the WordNet ontology. Even *anything* is evaluated more related to *nose* than *smell* because both *anything* and *nose* are hyponyms of *entity* whereas *smell* is a hyponym of *abstraction*. The next two measures, extended gloss overlaps[18] and gloss vector[17][19], avoided this problem by deciding similarity from glosses of concepts. Table I shows the similarities of *nose*/*smell* and *nose*/*anything* generated by different measures.

Banerjee and Pedersen [18] presented a measure based on the number of shared words in extended glosses of concepts. They extended the gloss of a concept to include glosses of all other related concepts and emphasized a common sequence of multiple words to the same number of scattered words.

Patwardhan and Pedersen [17][19] used extended glosses of concepts in WordNet as corpora to retrieve co-occurrence information for words and create vectors for each concept accordingly. Vectors are defined in a word hyperspace with about 20 000 dimensions. The similarity is defined as the cosine of the angle between gloss vectors of two concepts.

Patwardhan and Pederson showed [17][19] that their extended gloss vector measure was most correlated with human judgments. Our measure is actually inspired by their method of using the gloss vectors. However, we think there are some weaknesses in the original measure, that can be improved:

1) Considering example sentences as part of the gloss: Some example sentences in WordNet are related to the specific domain of the concept. However, example sentences are generally used to show the usage of the concept. Consider the example sentence for the verb *love* in WordNet 2.0: *I love French food.* It does not intend to show any relation between *love* and *French* or *love* and *food*. Treating *French* and *food* as gloss of *love* will add noise to the measure in this case.

2) Word based dimensions and content vectors: All the vectors defined in the measure are in a word space. However, due to existence of polysemy (a word with multiple meanings) in languages, it is imprecise to represent semantic relations between words. A sense space seems to be better.

3) Ignoring semantic relationships: As a result, the similarity of *green* to *blue* is 0.548, whereas the similarity of *green* to *emerald* is 0.179, because *green*'s gloss resembles *blue*'s gloss. However, since *green* is the direct hypernym of *emerald* because emerald color is a kind of green color, we usually think *green* is more like *emerald* than *blue*.

IV. CONTEXT VECTORS MEASURE

A. Definition of Context of a Sense

In our measure, the sense of a concept is considered highly related to the concepts in its context. Concepts frequently occurring in similar contexts are often conceptually similar. We use WordNet to retrieve contexts for senses because it is a general ontology with rich information about senses and their relations. Other corpus or lexicon can be used as well.

The most obvious context information in WordNet is the gloss of senses. For each sense, WordNet gives a definition like a dictionary does. XWN further analyzes the gloss and parses it into a list of senses. So it is convenient to use XWN to get context for any concept. However, glosses in WordNet (or any other dictionary) are usually rather short, and even shorter after all stop words [17][19] and function words (words other than nouns, verbs, adjectives and adverbs) are removed. The brevity of gloss makes it possible for two closely related concepts to have dissimilar definitions, and therefore degrades the performance of our measure. This problem was already reported in [18][19]. To solve it, the authors [18][19] expanded a concept's gloss with glosses of concepts directly linked by a WordNet relation, and used the augmented gloss as the context of the original concept. We propose another approach: use related concepts themselves instead of their glosses to augment the gloss of a concept.

In our measure, a context consists of the gloss synsets in XWN and all related synsets in WordNet. The related synsets are defined as synsets having direct semantic relations to the concerned synset, together with all direct and inherited hypernyms, and the concerned synset itself. By considering all hypernyms as part of the context, our measure returns high similarity for senses from the same part of the hypernymy hierarchy. Counting the sense itself as its context is also important. For instance, sense S_1 and S_2 are semantically related directly and therefore in each other's context, but their contexts do not intersect yet. With the senses themselves being counted into contexts, they will now have two senses in overlapped context and therefore have the close relation reflected in their context vectors.

For example, consider $smell_{n_2}$ (the second noun sense of *smell*) in WordNet. Its gloss is *"any property detected by the olfactory system"*; its hypernyms are: *smell* $\rightarrow$ *property* $\rightarrow$ *attribute* $\rightarrow$ *abstraction*; it has derivatively related forms $smell_{v_1}$ and $smell_{v_2}$; it also has two attributes *odorous* and *odorless*. Including itself, its context contains $smell_{n_2}$, $property_{n_3}$, $detect_{v_1}$, $olfactory_{a_1}$, $system_{n_1}$, $property_{n_3}$, $attribute_{n_2}$, $abstraction_{n_6}$, $smell_{v_1}$, $smell_{v_2}$, $odorous_{a_1}$, and $odorless_{a_1}$. Notice that $property_{n_3}$ occurs twice in the context. This is normal because it is the direct hypernym and therefore is closely related to $smell_{n_2}$.

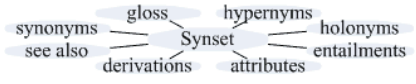

Fig. 3. Context of a Synset in WordNet 2.0

B. Creating Concept Space

In order to define context vector for each sense, we need to define a concept space first. Intuitively, every distinct sense can be one dimension, and projection of a vector on a dimension is the co-occurrence information in between. However, different senses have different impacts on the quality of the final similarity measure result. Common senses of specific topics, such as *food*, *color*, make good dimensions. Uncommon and highly domain-specific senses, such as *horsepower*, *sweat off*, can hardly reflect relatedness to other senses and will increase dimensions of the space and the computational complexity. General and widespread senses, such as *object*, *have*, and *etc.*, add noises to the vectors. In conclusion, we should include topic-related senses with decent frequency, and exclude general senses and over-specific senses.

Similar to the idea of stop words [17][19], we introduce the idea of stop senses. Stop senses are senses that appear frequently in the corpus and do not have any specific meaning to a domain. For example, the verb *be* and all of its forms belong to one sense, and is considered a stop sense. All stop senses are decided by humans.

In order to remove highly domain-specific senses, we simply use a frequency cutoff, although other methods are possible. For instance, the multiplication of *term frequency (tf)* and *inverse document frequency (idf)* can be used as a cutoff value [19], where $idf = \log \frac{number\ of\ documents}{document\ frequency}$. The *document frequency* is the number of documents in which the term occurs. Given a large corpus with documents of different domains, this $tf*idf$ value can be a good indication of a term's specificity. However, it does not fit well into our approach. If we treated a context as a document, for most concepts, the *term frequency* would be equal to the *document frequency* because of the brevity of contexts in our approach.

Considering the discussion above, we define the concept hyperspace by creating a dimension for every non-stop sense that occurs above a minimum frequency threshold in our corpus. For WordNet 2.0, we have approximately 21 000 dimensions using a cutoff frequency of 6.

C. Creating Concept Vectors

Context vectors are widely used in information retrieval and natural language processing, where 1^{st} order context vectors represent 1^{st} order co-occurrences in a corpus.

With concept space defined, we calculate the concept vectors as 1^{st} order context vectors for every sense $\vec{s}$ by:

1) Initialize the concept vector of $\vec{s}$ to a zero vector $\vec{w}$.
2) In the context of $\vec{s}$, for each sense $\vec{x}$ that is a dimension, increment the dimension of $\vec{w}$ that corresponds to $\vec{x}$.
3) Normalize $\vec{w}$.

Despite of 21 000 dimensions, a context of a concept usually contains only tens of concepts. Consequently, most coordinates of a concept vector are all zeros, and concept vectors are distributed in the concept space sparsely.

The concept vectors need to be normalized, otherwise senses with larger contexts will have greater magnitudes. This difference in magnitudes will add noise to later calculations of context vectors and similarities.

Since we only treat some synsets in WordNet as dimensions, there are synsets which cannot be projected to any dimension according to their contexts. For example, $thing_{n_2}$ has the gloss *"an action"*, whereas both *an* and *action* are stop senses. For these synsets, we define their concept vector as a unit vector with the same coordinates for all dimensions. This is reasonable because these synsets usually represent general concepts that are related to all remaining concepts.

D. Creating Context Vectors

We define the context vector of a concept as the 2^{nd} order context vector in the concept space, i.e., the summation of every sense's concept vector in the concept's context. They do not need to be normalized because only angles between them are used in later steps.

It is very hard to imagine vectors and their relations in a hyperspace with tens of dimensions (the average number of non-zero coordinates of a context vector is 52 in our approach). Fig. 4 attempts to depict concept vectors and context vectors in a 2-dimensional space. The noun *map*'s gloss in WordNet 2.0 is *"a diagrammatic representation of the earth's surface (or part of it)"*. The words *a*, *of*, *the*, *'s*, *or*, and *it* are function words and therefore not considered in the context of *map*. Other concepts are drawn as vectors in a space with two dimensions *planet* and *creation*. We can see that the concepts *earth* and *surface* are closer to the dimension *planet*, the concepts *diagrammatic* and *representation* are closer to the dimension *creation*, whereas the concept *part* is more general and located in between. The context vector of *map* is the summation of all the concept vectors.

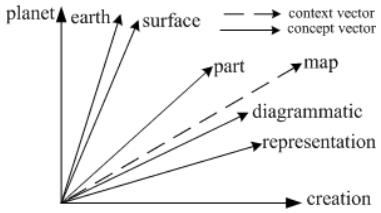

Fig. 4. Concept Vectors and Context Vector in a Concept Space

E. Measuring Similarity of Concepts

Having context vectors calculated for every concept, we are now ready to calculate the similarity between any two concepts. In our concepts space, we expect similar concepts to be adjacent, because their contexts often have a large intersection. Therefore, the angle between two context vectors can represent the conceptual distance between the corresponding two concepts. We can use the cosine of the angle to get a normalized measure of similarity of two senses S_1, S_2, $sim(S_1, S_2) \triangleq \cos\angle(\vec{v_1}, \vec{v_2}) = \frac{\vec{v_1} \cdot \vec{v_2}}{|\vec{v_1}||\vec{v_2}|}$.

Since the coordinates of concept vectors and context vectors are always non-negative, the angle between any two vectors is in $[0, \frac{\pi}{2}]$, and the cosine of the angle is in $[0, 1]$.

F. Measuring Similarity of Words

Our context vector measure is sense-based. However, it is more often required to measure relatedness of words. Even the relatedness measure benchmark is based on how close the measured word relatedness is to the relatedness judged by humans. So we need to provide a similarity measure for two words based on our measure for senses.

It is possible to disambiguate a word to find its POS and sense in a context. However, we have to assume that a standalone word could mean either of its senses in different POS. Here we define the similarity of two words W_a, W_b as: $sim(W_a, W_b) \triangleq \max_{1 \le i \le m, 1 \le j \le n} \big(sim(S_{a_i}, S_{b_j})\big)$, where $S_{a_1}, \ldots, S_{a_m}$ are the senses of W_a, and $S_{b_1}, \ldots, S_{b_n}$ are the senses of W_b. For example, the word *magician* and *wizard* are considered to have a similarity of 1 because they can both mean *"one who practices magic or sorcery"* and therefore belong to the synonym group (*sorcerer, magician, wizard, necromancer*), although they are seldom considered identical by humans because the primary sense of *magician* is *"someone who performs magic tricks to amuse an audience"* whereas the primary sense of *wizard* is *"someone who is dazzlingly skilled in any field"*. With statistic data about the probabilities of different senses of each word, we may calculate the similarity of two words as the weighted average of similarities between all of their senses.

V. Evaluations

The most obvious and commonly used metric for evaluating the performance of a semantic relatedness measure is to compare the result of measure to the human judgments. After all, semantic relatedness is subjective and all depends on human understanding of words and concepts in the world.

A. Comparison with Human Judgments

In the area of research, semantic relatedness measuring results are usually compared with several experiments on human judgments. Words and relatedness decided by humans in these experiments have been considered benchmarks for relatedness measures. A detailed evaluation and comparison of measures [1][11][13]–[15] can be found in [22].

Rubenstein and Goodenough [20] did the first experiment in 1965. They paid several groups of college students to decide the similarity of word pairs selected from 65 pairs of ordinary English nouns. A score of 4.0 was assigned to words considered synonyms and a score of 0.0 was assigned to words considered totally unrelated. The scores of all human judges were averaged and analyzed.

From then on, researchers used the same 65 noun pairs or their subsets to test their measures consistently, and redid the experiment to check the consistency between human judgments. Miller and Charles [21] did the experiment again in 1991 using a subset of 30 pairs of nouns, and obtained a correlation of 0.97 to the original experiment. Resnik [1] replicated the experiment of Miller and Charles in 1999, and obtained a correlation of 0.96. The average correlation over the 10 subjects who did the judgments was 0.88. This showed the variance of human judgments and is considered the

TABLE II
Comparison with Human Judgments Using Benchmarks

Word Pairs		Context Vector	R&G 1965	M&C 1991	Resnik 1999
chord*	smile	0.120		0.13	0.1
cord*	smile	0.011	0.02		
rooster	voyage	0.009	0.04	0.08	0
noon	string	0.038	0.04	0.08	0
fruit	furnace	0.070	0.05		
autograph	shore	0.089	0.06		
automobile	wizard	0	0.11		
mound	stove	0.382	0.14		
grin	implement	0.083	0.18		
asylum	fruit	0.055	0.19		
asylum	monk	0.064	0.39		
graveyard	madhouse	0	0.42		
glass	magician	0.050	0.44	0.11	0.1
boy	rooster	0.178	0.44		
cushion	jewel	0.082	0.45		
monk	slave	0.339	0.57	0.55	0.7
asylum	cemetery	0	0.79		
coast	forest	0.196	0.85	0.42	0.6
grin	lad	0.115	0.88		
shore	woodland	0.169	0.90	0.63	
monk	oracle	0.270	0.91	1.10	0.8
boy	sage	0.325	0.96		
automobile	cushion	0.342	0.97		
mound	shore	0.402	0.97		
lad	wizard	0.348	0.99	0.42	0.7
forest	graveyard	0.132	1.00	0.84	0.6
food	rooster	0.034	1.90	0.89	1.1
cemetery	woodland	0.044	1.18	0.95	
shore	voyage	0.093	1.22		
bird	woodland	0.018	1.24		
coast	hill	0.562	1.26	0.87	0.7
furnace	implement	0.073	1.37		
crane	rooster	0.497	1.41		
hill	woodland	0.027	1.48		
car	journey	0.060	1.55	1.16	0.7
cemetery	mound	0.070	1.69		
glass	jewel	0.091	1.78		
magician	oracle	0.286	1.82		
crane	implement	0.276	2.37	1.68	0.3
brother	lad	0.530	2.41	1.66	1.2
sage	wizard	0.299	2.46		
oracle	sage	0.395	2.61		
bird	crane	0.396	2.63	2.97	2.1
bird	cock	0.489	2.63	3.05	2.2
food	fruit	0.069	2.69	3.08	2.1
brother	monk	0.625	2.74	2.82	2.4
asylum	madhouse	0.799	3.04	3.61	3.6
furnace	stove	0.240	3.11	3.11	2.6
magician	wizard	1	3.21	3.50	3.5
hill	mound	1	3.29		
cord	string	0.805	3.41		
glass	tumbler	0.510	3.45		
grin	smile	1	3.46		
serf	slave	0.795	3.46		
journey	voyage	0.834	3.58	3.84	3.5
autograph	signature	0.838	3.59		
coast	shore	0.666	3.60	3.70	3.5
forest	woodland	1	3.65		
implement	tool	0.611	3.66	2.95	3.4
cock	rooster	1	3.68		
boy	lad	0.803	3.82	3.76	3.5
cushion	pillow	0.352	3.84		
cemetery	graveyard	1	3.88		
automobile	car	1	3.92	3.92	3.9
midday	noon	1	3.94	3.42	3.6
gem	jewel	1	3.94	3.84	3.5

*Miller–Charles and Resnik used the word *chord* instead of *cord* in their replications.

TABLE III
CORRELATION TO HUMAN PERCEPTION

Relatedness Measure	M&C	R&G	Resnik
Gloss Vector [17][19]	0.91	0.90	
Other Human Perception			0.88
Context Vector (ours)	0.80	0.83	0.85
Extended Gloss Overlaps [18]	0.81	0.83	
Hirst–St-Onge [15]	0.78	0.81	
Leacock–Chodorow [13]	0.74	0.77	
Jiang–Conrath [14]	0.73	0.75	
Resnik [1]	0.72	0.72	
Lin [11]	0.70	0.72	

upper bound on what one should expect from a measurement algorithm.

We used the same 65 pairs of nouns to test our measure. The results are listed in Table II.

B. Comparison with Other Measures

We compared our Context Vector measure with other measures using the benchmark tests of Miller–Charles [21] (denoted as M&C in Table II and III), Rubenstein–Goodenough [20] (denoted as R&G in Table II and III), and Resnik [1]. The correlation results are shown in Table III.

Table III shows that our measure has a decent (second best) performance compared against other measures. It is only inferior to the Gloss Vector measure, which seems currently the best measure against human judgment. It is close to the Extended Gloss Overlaps measure and superior to all other measures. It is only slightly inferior to the correlation between human perceptions, which means our measure can decide word similarity almost as well as an individual person does.

Analyzing the difference of relatedness returned by our measure with that judged by the humans in the benchmark, we thought there was a weakness in the benchmark data. Among 48 nouns used in the test, only 8 of them are abstract entities. And among 65 pairs of nouns, only 7 pairs are between physical entities and abstract entities. As we pointed out in section IV, measures dependent on the hierarchical structure of WordNet perform well on concept pairs that are close together in the hierarchy but can hardly find relations between concepts far away in the hierarchy. The lack of physical–abstract entity pairs in the benchmark might disguise the low performance of those measures.

VI. CONCLUSION AND FUTURE WORK

We introduced a novel measure of semantic relatedness based on combining the semantic ontology with the corpus to create context vectors in concept hyperspace. Our measure correlates well with human judgments and can be easily applied to different domains.

In the future, we may tune the measure with different cutoff methods to improve the concept space construction. We may also assign different weights to concepts in gloss and concepts of different relations. We may also try functions besides cosine to compute the similarity between two vectors.

REFERENCES

[1] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", *J. of Artificial Intell. Research*, No. 11, 1999, pp. 95–130.

[2] A. Budanitsky, G. Hirst, "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures", *Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Assoc. for Comput. Linguistics*, 2001.

[3] H. Kozima, "Computing Lexical Cohesion as a Tool for Text Analysis", *doctoral thesis*, Computer Science and Information Mathematics, Graduate School of Electro-Communications, University of Electro-Communications, 1993.

[4] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Content-Based Image Retrieval at the End of the Early Years", *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 22, issue 12, Dec. 2000, pp. 1349–1380.

[5] R. Srihari, Z. Zhang, A. Rao, "Intelligent Indexing and Semantic Retrieval of Multimodal Documents", *Inform. Retrieval*, vol. 2, No. 2–3, May 2000, pp. 245–275.

[6] S. Green, "Building hypertext links by computing semantic similarity", *IEEE Trans. on Knowl. and Data Eng.*, vol. 11, No. 5, Sep./Oct. 1999, pp. 713–730.

[7] R. Rada, H. Mili, E. Bicknell, M. Blettner, "Development and Application of a Metric on Semantic Nets", *IEEE Trans. on Syst., Man, and Cybern.*, vol. 19, issue 1, Jan./Feb. 1989, pp. 17–30.

[8] T. Andreasen, H. Bulskov, R. Knappe, "On ontology-based querying", *Proc. 4^{th} Intl. Conf. on Flexible Query Answering Syst.*, 2000, pp. 15–26.

[9] Z. Wu, M. Palmer, "Verbs semantics and lexical selection", *Proc. 32^{nd} Annu. Meeting on Assoc. for Comput. Linguistics*, 1994, pp. 133–138.

[10] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", *Proc. 14^{th} Intl. Joint Conf. on Artificial Intell.*, 1995, pp. 448–453.

[11] D. Lin, "An Information-Theoretic Definition of Similarity", *Proc. 15^{th} Intl. Conf. on Mach. Learning*, 1998, pp. 296–304.

[12] M. Sussna, "Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network", *Proc. 2^{nd} Intl. Conf. on Inform. and Knowl. Manage.*, 1993, pp. 67–74.

[13] C. Leacock, M. Chodorow, "Combining local context and WordNet sense similarity for word sense disambiguation", *WordNet, an Electronic Lexical Database*, The MIT Press, 1998.

[14] J. Jiang, D. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", *Proc. Intl. Conf. Research on Comput. Linguistics*, 1997.

[15] G. Hirst, D. St-Onge, "Lexical chains as representations of context for the detection and correction of malapropisms", *WordNet, an Electronic Lexical Database*, The MIT Press, 1998, pp. 305–332.

[16] Y. Li, Z. Bandar, D. McLean, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", *IEEE Trans. on Knowl. and Data Eng.*, vol. 15, issue 4, Jul./Aug. 2003, pp. 871–882.

[17] S. Patwardhan, S. Banerjee, T. Pedersen, "Using Measures of Semantic Relatedness for Word Sense Disambiguation", *Proc. 4^{th} Intl. Conf. on Intell. Text Process. and Comput. Linguistics*, 2003, pp. 241–257.

[18] S. Banerjee, T. Pedersen, "Extended Gloss Overlaps as a Measure of Semantic Relatedness", *Proc. 18^{th} Intl. Joint Conf. on Artificial Intell.*, 2003, pp. 805–810.

[19] S. Patwardhan, T. Pedersen, "Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts", *Proc. EACL 2006 Workshop Making Sense of Sense—Bringing Computational Linguistics and Psycholinguistics Together*, 2006, pp. 1–8.

[20] H. Rubenstein, J. Goodenough, "Contextual Correlates of Synonymy", *Commun. of the ACM*, vol. 8, issue 10, Oct. 1965, pp. 627–633.

[21] G. Miller, W. Charles, "Contextual correlates of semantic similarity", *Language and Cognitive Processes*, vol. 6, issue 1, Feb. 1991, pp. 1–28.

[22] A. Budanitsky, G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness", *Comput. Linguistics*, vol. 32, issue 1, Mar. 2006, pp. 13–47.

[23] H. Schütze, "Automatic Word Sense Discrimination", *Comput. Linguistics*, vol. 24, issue 1, Mar. 1998, pp. 97–123.

[24] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, "Introduction to WordNet: An On-line Lexical Database", *Intl. J. of Lexicography*, vol. 3, No. 4, 1993, pp. 235–244.