Lexical Semantic Relatedness with Random Graph Walks

Thad Hughes and Daniel Ramage
Computer Science Department
Stanford University
Stanford, CA 94305
{thughes, dramage}@cs.stanford.edu

Abstract

Many systems for tasks such as question answering, multi-document summarization, and information retrieval need robust numerical measures of lexical relatedness. Standard thesaurus-based measures of word pair similarity are based on only a single path between those words in the thesaurus graph. By contrast, we propose a new model of lexical semantic relatedness that incorporates information from every explicit or implicit path connecting the two words in the entire graph. Our model uses a random walk over nodes and edges derived from WordNet links and corpus statistics. We treat the graph as a Markov chain and compute a word-specific stationary distribution via a generalized PageRank algorithm. Semantic relatedness of a word pair is scored by a novel divergence measure, ZKL, that outperforms existing measures on certain classes of distributions. In our experiments, the resulting relatedness measure is the WordNet-based measure most highly correlated with human similarity judgments by rank ordering at $\rho = .90$.

1 Introduction

Several kinds of Natural Language Processing systems need measures of semantic relatedness for arbitrary word pairs. For example, document summarization and question answering systems often use similarity scores to evaluate candidate sentence alignments, and information retrieval systems use relatedness scores for query expansion. Several popular algorithms calculate scores from information contained in WordNet (Fellbaum, 1998), an electronic dictionary where word senses are explicitly connected by zero or more semantic relationships. The central challenge of these algorithms is to compute reasonable relatedness scores for arbitrary word pairs given that few pairs are directly connected.

Most pairs in WordNet share no direct semantic link, and for some the shortest connecting path can be surprising—even pairs that seem intuitively related, such "furnace" and "stove" share a lowest common ancestor in the hypernymy taxonomy (is-a links) all the way up at "artifact" (a man-made object). Several existing algorithms compute relatedness only by traversing the hypernymy taxonomy and find that "furnace" and "stove" are relatively unrelated. However, WordNet provides other types of semantic links in addition to hypernymy, such as meronymy (part/whole relationships), antonymy, and verb entailment, as well as implicit links defined by overlap in the text of definitional glosses. These links can provide valuable relatedness information. If we assume that relatedness is transitive across a wide variety of such links, then it is natural to follow paths such as *furnace–crematory–gas oven–oven–kitchen appliance–stove* and find a higher degree of relatedness between "furnace" and "stove."

This paper presents the application of random walk Markov chain theory to measuring lexical semantic relatedness. A graph of words and concepts is constructed from WordNet. The random walk model posits the existence of a particle that roams this graph by stochastically following local semantic relational links. The particle is biased toward exploring the neighborhood around a target word, and is allowed to roam until the proportion of time it visits each node in the limit converges to a stationary distribution. In this way we can compute distinct, word-specific probability distributions over how often a particle visits all other nodes in the graph when "starting" from a specific word. We compute the relatedness of two words as the similarity of their stationary distributions.

The random walk brings with it two distinct advantages. First, it enables the similarity measure to have a principled means of combination of multiple types of edges from WordNet. Second, by traversing all links, the walk aggregates local similarity statistics across the entire graph. The similarity scores produced by our method are, to our knowledge, the WordNet-based scores most highly correlated with human judgments.

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 581–589, Prague, June 2007. ©2007 Association for Computational Linguistics

2 Related work

Budanitsky and Hirst (2006) provide a survey of many WordNet-based measures of lexical similarity based on paths in the hypernym taxonomy. As an example, one of the best performing is the measure proposed by Jiang and Conrath (1997) (similar to the one proposed by (Lin, 1991)), which finds the shortest path in the taxonomic hierarchy between two candidate words before computing similarity as a function of the information content of the two words and their lowest common subsumer in the hierarchy. We note the distinction between word similarity and word relatedness. Similarity is a special case of relatedness in that related words such as "cat" and "fur" share some semantic relationships (such as meronymy), but do not express the same likeness of form as would similar words such as "cat" and "lion." The Jiang-Conrath measure and most other measures that primarily make use of of hypernymy (is-a links) in the WordNet graph are better categorized as measures of similarity than of relatedness.

Other measures have been proposed that utilize the text in WordNet's definitional glosses, such as Extended Lesk (Banerjee and Pedersen, 2003) and later the Gloss Vectors (Patwardhan and Pedersen, 2006) method. These approaches are primarily based on comparing the "bag of words" of two synsets' gloss text concatenated with the text of neighboring words' glosses in the taxonomy. As a result, these gloss-based methods measure relatedness. Our model captures some of this relatedness information by including weighted links based on gloss text.

A variety of other measures of semantic relatedness have been proposed, including distributional similarity measures based on co-occurrence in a body of text—see (Weeds and Weir, 2005) for a survey. Other measures make use of alternative structured information resources than WordNet, such as Roget's thesaurus (Jarmasz and Szpakowicz, 2003). More recently, measures incorporating information from Wikipedia (Gabrilovich and Markovitch, 2007; Strube and Ponzetto, 2006) have reported stronger results on some tasks than have been achieved by existing measures based on shallower lexical resources. The results of our algorithm are competitive with some Wikipedia algorithms while using only WordNet 2.1 as the underlying lexical resource. The approach presented here is generalizable to construction from any underlying semantic resource.

PageRank is the most well-known example of a random walk Markov chain—see (Berkhin, 2005) for a survey. It uses the local hyperlink structure of the web to define a graph which it walks to aggregate popularity information for different pages. Recent work has applied random walks to NLP tasks such as PP attachment (Toutanova et al., 2004), word sense disambiguation (Mihalcea, 2005; Tarau et al., 2005), and query expansion (Collins-Thompson and Callan, 2005). However, to our knowledge, the literature in NLP has only considered using one stationary distribution per specially-constructed graph as a probability estimator. In this paper, we introduce a measure of semantic relatedness based on the divergence of the distinct stationary distributions resulting from random walks centered at different positions in the word graph. We believe we are the first to define such a measure.

3 Random walks on WordNet

Our model is based on a random walk of a particle through a simple directed graph $G = (V, E)$ whose nodes V and edges E are extracted from WordNet version 2.1. Formally, we define the probability $n_i^{(t)}$ of finding the particle at node $n_i \in V$ at time t as the sum of all ways in which the particle could have reached n_i from any other node at the previous time-step:

$$n_i^{(t)} = \sum_{n_j \in V} n_j^{(t-1)} P(n_i \mid n_j)$$

where $P(n_i \mid n_j)$ is the conditional probability of moving to n_i given that the particle is at n_j. In particular, we construct the transition distribution such that $P(n_i \mid n_j) > 0$ whenever WordNet specifies a local link relationship of the form $j \rightarrow i$. Note that this random walk is a Markov chain because the transition probabilities at time t are independent of the particle's past trajectory.

The subsections that follow present the construction of the graph for our random walk from WordNet and the mathematics of computing the stationary distribution for a given word.

3.1 Graph Construction

WordNet is itself a graph over synsets. A synset is best thought of as a concept evoked by one sense of one or more words. For instance, different senses of the word "bank" take part in different synsets (e.g. a river bank versus a financial institution), and a single synset can be represented by multiple synonymous words, such as "middle" and "center." WordNet explicitly marks semantic relationships between synsets, but we are additionally interested in representing relatedness between words. We therefore extract the following types of nodes from WordNet:

Synset Each WordNet synset has a corresponding node. For example, one node corresponds to the synset referred to by "dog#n#3," the third sense of dog as noun, whose meaning is "an informal term for a man." There are 117,597 *Synset* nodes.

TokenPOS One node is allocated to every word coupled with a part of speech, such as "dog#n" meaning dog as a noun. These nodes link to all the synsets they participate in, so that "dog#n" links the *Synset* nodes for canine, hound, hot dog, etc. Collocations—multi-word expressions such as "hot dog"—that take part in a synsets are also represented by these nodes. There are 156,588 *TokenPOS* nodes.

Token Every *TokenPOS* is connected to a *Token* node corresponding to the word when no part of speech information is present. For example, "dog" links to "dog#n" and "dog#v" (meaning "to chase"). There are 148,646 *Token* nodes.

Synset nodes are connected with edges corresponding to many of the relationship types in WordNet. We use these WordNet relationships to form edges: hypernym/hyponym, instance/instance of, all holonym/meronym links, antonym, entails/entailed by, adjective satellite, causes/caused by, participle, pertains to, derives/derived from, attribute/has attribute, and topical (but not regional or usage) domain links. By construction, each edge created from a WordNet relationship is guaranteed to have a corresponding edge in the opposite direction.

Edges that connect a *TokenPOS* to the *Synsets* using it are weighted based on a Bayesian estimate drawn from the SemCor frequency counts included in WordNet but with a non-uniform Dirichlet prior. Our edge weights are the SemCor frequency counts for each target *Synset,* with pseudo-counts of .1 for all *Synsets*, 1 for first sense of each word, and .1 for the first word in each *Synset*. Intuitively, this causes the particle to have a higher probability of moving to more common senses of a *TokenPOS*; for example, the edges from "dog#n" to "dog#n#1" (canine) and "dog#n#5" (hotdog) have un-normalized weights of 43.2 and 0.1, respectively. The edges connecting a *Token* to the *TokenPOS* nodes in which it can occur are also weighted by the sum of the weights of the outgoing *TokenPOS→Synset* links. Hence a walk starting at a common word like "cat" is far more likely to follow a link to "cat#n" than to rarities like "cat#v" (to vomit). These edges are uni-directional; no edges are created from a *Synset* to a *TokenPOS* that can represent the *Synset*.

In order for our graph construction to incorporate textual gloss-based information, we also create uni-directional edges from *Synset* nodes to the *TokenPOS* nodes for the words and collocations used in that synset's gloss definition. This requires part-of-speech tagging the glosses, for which we use the Stanford maximum entropy tagger (Toutanova et al., 2003). It is important to correctly weight these edges, because high-frequency stop-words such as "by" and "he" do not convey much information and might serve only to smear the probability mass across the whole graph. Gloss-based links to these nodes should therefore be down-weighted or removed. On the other hand, up-weighting extremely rare words such as by `tf-idf` scoring might also be inappropriate because such rare words would get extremely high scores, which is an undesirable trait in similarity search. (Haveliwala et al., 2002) and others have shown that a "non-monotonic document frequency" (NMDF) weighting can be more effective in such a setting. Because the frequency of words in the glosses is distributed by a power-law, we weight each word by its distance from the mean word count in log space. Formally, the weight w_i for a word appearing r_i times is

$$w_i = exp\left(-\frac{(log(r_i) - \mu)^2}{2\sigma^2}\right)$$

where μ and σ are the mean and standard deviation of the logs of all word counts. This is a smooth approximation to the high and low frequency stop lists used effectively by other measures such as (Patwardhan and Pedersen, 2006). We believe that because non-monotonic frequency scaling has no parameters and is data-driven, it could stand to be more widely adopted among gloss-based lexical similarity measures.

We also add bi-directional edges between *Synsets* whose word senses overlap with a common *TokenPOS*. These edges have raw weights given by the number of *TokenPOS* nodes shared by the *Synsets*. The intuition behind adding these edges is that WordNet often divides the meanings of words into fine-grained senses with similar meanings, so there is likely to be some semantic relationship between *Synsets* sharing a common *TokenPOS*.

The final graph has 422,831 nodes and 5,133,281 edges. This graph is very sparse; fewer than 1 in 10,000 node pairs are directly connected. When only the unweighted WordNet relationship edges are considered, the largest degree of any node is "city#n#1" with 667 edges (mostly connecting to particular cities), followed by "law#n#2" with 602 edges (mostly connecting to a large number of domain terms such as "dissenting opinion" and "freedom of speech"), and each node is on average connected to 1.7 other nodes. When the gloss-based edges are considered separately, the highest degree nodes are those with the longest definitions; the maximum out-degree is 56 and the average out-degree is 6.2. For the edges linking *TokenPOS* nodes to the *Synsets* in which they participate, *TokenPOS* nodes with many senses are the most connected; "break#v" with 59 outgoing edges and "make#v" with 49 outgoing edges have the highest out-degrees, with the average out-degree being 1.3.

3.2 Computing the stationary distribution

Each of the K edge types presented above can be represented as separate transition matrix $E_k \in \mathbb{R}^{N \times N}$ where

N is the total number of nodes. For each matrix, column j contains contains a normalized outgoing probability distribution,[1] so the weight in cell (i, j) contains $P_K(n_i \mid n_j)$, the conditional probability of moving from node n_j to node n_i in edge type K. For many of the edge types, this is either 0 or 1, but for the weighted edges, these are real valued. The full transition matrix M is then the column normalized sum of all of the edge types:

$$\hat{M} = \sum_k E_k$$

$$M = \left(\left\|\hat{M}\right\|_\infty\right)^{-1} \cdot \hat{M}$$

$M$is a distillation of relevant relatedness information about all nodes extracted from WordNet and is not tailored for computing a stationary distribution for any specific word. In order to compute the stationary distribution $v_{dog\#n}$ for a walk centered around the *TokenPOS* "dog#n," we first define an initial distribution $v^{(0)}_{dog\#n}$ that places all the probability mass in the single vector entry corresponding to "dog#n." Then at every step of the walk, we will return to $v^{(0)}$ with probability β. Intuitively, this return probability captures the notion that nodes close to "dog#n" should be given higher weight, and also guarantees that the stationary distribution exists and is unique (Bremaud, 1999). The stationary distribution v is computed via an iterative update algorithm:

$$v^{(t)} = \beta v^{(0)} + (1 - \beta) M v^{(t-1)}$$

Because the walk may return to the initial distribution $v^{(0)}$ at any step with probability β, we found that $v^{(t)}$ converges to its unique stationary distribution $v^{(\infty)}$ in a number of steps roughly proportional to β^{-1}. We experimented with a range of return probabilities and found that our results were relatively insensitive to this parameter. Our convergence criteria was $\left\|v^{(t-1)} - v^{(t)}\right\|_1 < 10^{-10}$, which, for our graph with a return probability of $\beta = .1$, was met after about two dozen iterations. This computation takes under two seconds on a modern desktop machine.

Note that because M is sparse, each iteration of the above computation is linear in the total number of nonzero entries in P, i.e. linear in the total number of edges. Introducing an edge type that is dense would dramatically increase running time.

3.3 Model variants

For this paper, we consider three model variants that differ based on which subset of the edge types are included in the transition matrix M.

MarkovLink This variant includes the explicit WordNet relations such as hypernymy and the edges representing overlap between the *TokenPOS* nodes contained in *Synsets*. A particle walking through this graph reaches only *Synset* nodes and can step from one *Synset* to another whenever WordNet specifies a relationship between the *Synsets* or when the *Synsets* share a common word. There is a single connected component in this model variant. This model is loosely analogous to a smoothed version of the path-based WordNet measures surveyed in (Budanitsky and Hirst, 2006) but differs in that it integrates multiple link types and aggregates relatedness information across all paths in the graph.

MarkovGloss This variant includes only the weighted uni-directional edges linking *Synsets* to the *TokenPOS* nodes contained in their gloss definitions, and the edges from a *TokenPOS* node to the *Synsets* containing it. The intuition behind this model variant is that the particle can move as if it were recursively looking up words in a dictionary, stepping from *Synsets* to the *Synsets* used to define them. Because WordNet's gloss definitions are not sense-tagged, the particle must make an intermediate step to a *TokenPOS* contained in the gloss definition and then to a *Synset* representing a particular sense of that *TokenPOS*. The availability of sense-tagged glosses would eliminate the noise introduced by this intermediate step. The particle can reach both *Synsets* and *TokenPOS* nodes in this variant, but some parts of the graph are not reachable from other parts. This model incorporates much of the same information as the gloss-based WordNet measures (Banerjee and Pedersen, 2003; Patwardhan and Pedersen, 2006) but differs in that it considers many more glosses than just those in the immediate neighborhoods of the candidate words.

MarkovJoined This variant is the natural combination of the above two; we construct the graph containing WordNet relation edges, *Synset* overlap edges, and gloss-based *Synset* to *TokenPOS* edges.

Many of the characteristics of the model variants can be understood in terms of how much probability mass they assign to each node for a particular word-specific stationary distribution. Table 1 shows the highest scoring nodes in the word-specific stationary distributions centered around the *Token* node for "wizard," as computed by the *MarkovLink* and *MarkovGloss* variants. In both variants, the "wizard" *Token*'s only neighbors are the "wizard#n" and "wizard#a" *TokenPOS* nodes, and "wizard#n"

[1] The frequency-count derived edges are normalized by the largest column sum. This effectively preserves relative term frequency information across the graph and causes some columns to sum to less than one. We interpret this lost mass as a link to "nowhere."

MarkovLink		*MarkovGloss*	
Node	Probability	Node	Probability
wizard	1.0E-1	wizard	1.3E-01
wizard#n	2.5E-3	wizard#n	2.9E-02
wizard#a	7.8E-5	wizard#a	9.1E-04
ace#n#3	4.2E-5	ace#n#3	1.1E-06
sorcerer#n#1	2.2E-6	sorcerer#n#1	5.8E-07
charming#a#2	2.2E-6	dazzlingly#r	2.4E-08
expert#n#1	1.1E-6	charming#a#2	1.6E-09
track star#n#1	1.1E-6	sorcery#n	2.6E-10
occultist#n#1	5.7E-7	magic#n	6.8E-12
Cagliostro#n#1	5.7E-7	magic#a	6.8E-12
star#v#2	5.5E-7	dazzlingly#r#1	4.3E-14
breeze_through#v#1	5.4E-7	dazzle#n	9.4E-16
magic#n#1	2.1E-8	beholder#n	9.4E-16
sorcery#n#1	2.1E-7	dazzle#v	9.4E-16
magician#n#1	1.9E-7	magic#n#1	5.1E-16

Table 1: Highest scoring nodes in the stationary distributions for "wizard#n" as generated by the *MarkovLink* model and the *MarkovGloss* model with return probability 0.1.

has a higher probability mass because of its higher SemCor usage counts. Likewise, the only possible steps permitted in either variant from "wizard#n" and "wizard#a" are to the *Synsets* that can be expressed with those nodes: "ace#n#3," "sorcerer#n#1," and "charming#a#1." Again, the amount of mass given to these nodes depends on the strength of these edge weights, which is determined by the SemCor usage counts.

The highest probability nodes in the table are common because both model variants share the same initial links. However, the orders of the remaining nodes in the stationary distributions are different. In the *MarkovLink* variant, the random walk can only proceed to other *Synsets* using WordNet relationship edges; "track star#n#1" and "expert#n#1" are first reached by following hyponym and hypernym edges from "ace#n#1," and "occultist#n#1" and "Cagliostro#n#1" are first reached with hypernym and instance edges from "sorcerer#n#1." The node "breeze through#v#1" is reached through a path following derivational links with "ace#n" and "ace#v."

The *MarkovGloss* variant in table 1 shows how information can be extracted solely from the textual glosses. Once the random walk reaches the first *Synset* nodes, it can step to the *TokenPOS* nodes in their glosses; for example, "ace#n#1" has the gloss "someone who is dazzlingly skilled in any field." Links to *TokenPOS* nodes that are very common in glosses are down-weighted with NMDF weighting, so "someone#n" receives little mass while "dazzlingly#r" receives more. From there, the random walk can step to another *Synset* such as "dazzlingly#r#1," and then on to other *TokenPOS* nodes used in its definition: "in a manner or to a degree that dazzles the beholder."

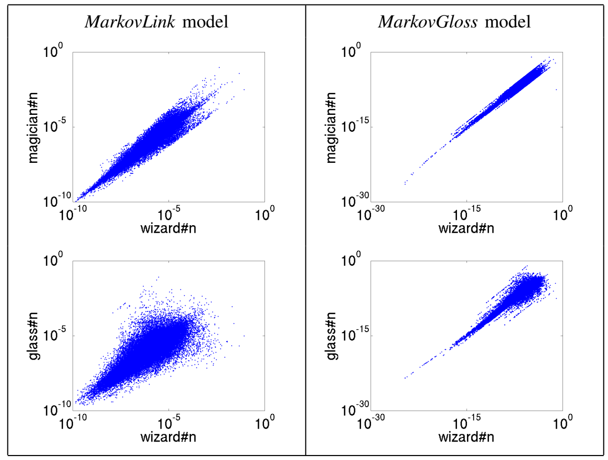

Figure 1: Example stationary distributions plotted against each other for similar (top) and dissimilar (bottom) word pairs, using the *MarkovLink* (left) and *MarkovGloss* (right) model variants.

Figure 1 demonstrates how two word-specific stationary distributions are more highly correlated if the words are related. In both model variants, random walks for related words are more likely to visit the same parts of the graph, and so assign higher probability to the same nodes. Figure 1 also shows that the *MarkovGloss* variant produces distributions with a much wider range of probabilities than the *MarkovLink*, which might be a source of difficulty in integrating the two model variants.

Figure 2 shows the correlation between the stationary distributions produced by the two model variants for the same word. The log-log scale makes it possible to see the entire range of probabilities on the same axes, and shows that distributions produced by these two model variants share many of the same highest-probability words.

A noteworthy property of the constructed graphs is that word relatedness can be computed directly by comparing walks that start at *Token* nodes. By contrast, existing WordNet-based measures require independent similarity judgments for all word senses relevant to a target word pair (of which the maximum relatedness value is usually taken). Our algorithm lends itself to comparisons between walks centered at a *Synset* node, or a *TokenPOS* node, or a *Token* node, or any mixed distribution thereof. And because the *Synset* nodes are strongly connected, the model also admits direct comparison across parts of speech.

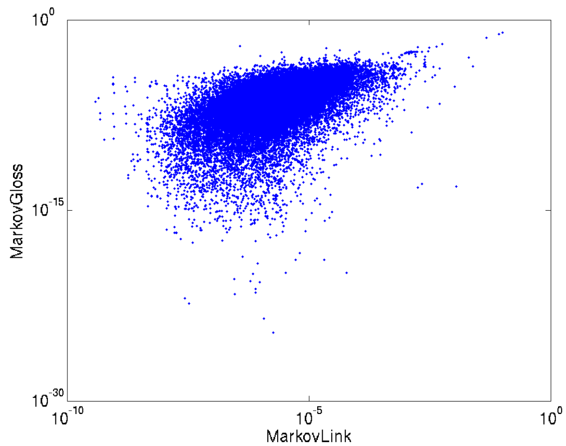

Figure 2: Correlation of the stationary distributions for "wizard#n," produced by the *MarkovLink* variant (x-axis) and the *MarkovGloss* variant (y-axis).

4 Similarity judgments

We have shown how to compute the word-specific stationary distribution from any starting distribution in the graph. Now consider the task of deciding similarity between two words. Intuitively, if the random walk starting at the first word's node and the random walk starting at the second word's node tend to visit the same nodes, we would like to consider them semantically related. Formally, we measure the divergence of their respective stationary distributions, p and q.

A wide literature exists on similarity measures between probability distributions. One standard choice is to consider p and q to be vectors and measure the cosine of the angle between them, which is rank equivalent to Euclidean distance.

$$sim_{cos}(p, q) = \frac{\sum_i p_i q_i}{\|p\| \; \|q\|}$$

Because p and q are probability distributions, we would also expect a strong contender from the information-theoretic measures based on Kullback-Leibler divergence, defined as:

$$D_{KL}(p \parallel q) = \sum_i p_i \, log \frac{p_i}{q_i}$$

Unfortunately, KL divergence is undefined if any q_i is zero because those terms in the sum will have infinite weight. Several modifications to avoid this issue have been proposed in the literature. One is Jensen-Shannon divergence (Lin, 1991), a symmetric measure based on KL-divergence defined as the average of the KL divergences of each distribution to their average distribution. Jensen-Shannon is well defined for all distributions because the average of p_i and q_i is non-zero whenever either number is.

These measures and others are surveyed in (Lee, 2001), who finds that Jensen-Shannon is outperformed by the Skew divergence measure introduced by Lee in (1999). The skew divergence[2] accounts for zeros in q by mixing in a small amount of p.

$$\begin{aligned} s_\alpha(p, q) &= D(p \parallel \alpha q + (1-\alpha) p) \\ &= \textstyle\sum_i p_i \, log \frac{p_i}{\alpha q_i + (1-\alpha) p_i} \end{aligned}$$

Lee found that as $\alpha \to 1$, the performance of skew divergence on natural language tasks improves. In particular, it outperforms most other models and even beats pure KL divergence modified to avoid zeros with sophisticated smoothing models. In exploring the performance of divergence measures on our model's stationary distributions, we observed the same phenomenon. Note that in the limit as $\alpha \to 1$, alpha skew is identically KL-divergence.

4.1 Zero-KL Divergence

In this section we introduce a novel measure of distributional divergence based on a reinterpretation of the skew divergence. Skew divergence avoids zeros in q by mixing in some of p, but its performance on many natural language tasks improves as it better approximates KL divergence. We propose an alternative approximation to KL divergence called Zero-KL divergence, or ZKL. When q_i is non-zero, we use exactly the term from KL divergence. When $q_i = 0$, we have a problem—in the limit as $\alpha \to 1$, the corresponding term approaches infinity. We let ZKL use the Skew divergence value for these terms: $p_i \, log \frac{p_i}{\alpha q_i + (1-\alpha) p_i}$. Because $q_i = 0$ this simplifies to $p_i \, log \frac{p_i}{(1-\alpha) p_i} = p_i \, log \frac{1}{1-\alpha}$.

Lee showed skew divergence's best performance was for α near to 1, so we formalize this intuition by choosing α exponentially near to 1, i.e. we can choose our α as $1 - 2^{-\gamma}$ for some $\gamma \in \mathbb{R}^+$. Zero terms in the sum can now be written as $p_i \, log \frac{1}{2^{-\gamma}} = p_i \, log \, 2^\gamma = p_i \, \gamma$. Note here an analogy to the case with $q_j > 0$ and where p_j is exactly one order of magnitude greater than q_j, i.e. $p_j = 2 \cdot q_j$. For such a term in the standard KL divergence, we would get $p_j \, log \frac{p_j}{q_j} = p_j \, log(2) = p_j$. Therefore, the α term in skew divergence implicitly defines a parameter stating how many orders of magnitude smaller than p_j to count q_j if $q_j = 0$.

We define the Zero-KL divergence with respect to

[2] In Lee's (1999) original presentation, skew divergence is defined not as $s_\alpha(p, q)$ but rather as $s_\alpha(q, p)$. We reverse the argument order for consistency with the other measures discussed here.

gamma:

$$ZKL_\gamma(p,q) = \sum_i p_i \begin{cases} log\frac{p_i}{q_i} & q_i \neq 0 \\ \gamma & q_i = 0 \end{cases}$$

Note that this is exactly KL-divergence when KL-divergence is defined and, like skew divergence, approximates KL divergence in the limit as $\gamma \rightarrow \infty$.

A similar analysis of the skew divergence terms for when $0 < q_i \ll p_i$ (and in particular with q_i less than p_i by more than a factor of $2^{-\gamma}$) shows that such a term in the skew divergence sum is again approximated by $\gamma\, p_i$. ZKL does not have this property. Because ZKL is a better approximation to KL divergence and because they have the same behavior in the limit, we expect ZKL's performance to dominate that of skew divergence in many distributions. However, if there is a wide range in the exponent of noisy terms, the maximum possible penalty to such terms ascribed by skew divergence may be beneficial.

Figure 3 shows the relative performance of ZKL versus Jensen-Shannon, skew divergence, cosine similarity, and the Jaccard score (a measure from information retrieval) for correlations with human judgment on the *MarkovLink* model. ZKL consistently outperforms the other measures on distributions resulting from this model, but ZKL is not optimal on distributions generated by our other models. The next section explores this topic in more detail.

5 Evaluation

Traditionally, there have been two primary types of evaluation for measures of semantic relatedness: one is correlation to human judgment, the other is the relative performance gains of a task-driven system when it uses the measure. The evaluation here focuses on correlation with human judgments of relatedness. For consistency with previous literature, we use rank correlation (Spearman's ρ coefficient) rather than linear correlation when comparing sets of relatedness judgments because the rank correlation captures information about the relative ordering of the scores. However, it is worth noting that many applications that make use of lexical relatedness scores (e.g. as features to a machine learning algorithm) would better be served by scores on a linear scale with human judgments.

Rubenstein and Goodenough (1965) solicited human judgments of semantic similarity for 65 pairs of common nouns on a scale of zero to four. Miller and Charles (1991) repeated their experiment on a subset of 29 noun pairs (out of 30 total) and found that although individuals varied among their judgments, in aggregate the scores were highly correlated with those found by Rubenstein and Goodenough (at $\rho = .944$ by our calculation). Resnik (1999) replicated the Miller and Charles experiment and reported that the average per-subject linear correlation on the dataset was around $r = 0.90$, providing a rough upper bound on any system's linear correlation performance with respect to the Miller and Charles data. Figure 3 shows that the ZKL measure on the *MarkovLink* model has linear correlation coefficient $r = .903$—at the limit of human inter-annotator agreement.

Recently, a larger set of word relatedness judgments was obtained by (Finkelstein et al., 2002) in the WordSimilarity-353 (WS-353) collection. Despite the collection's name, the study instructed participants to score word pairs for relatedness (on a scale of 0 to 10), which is in contrast to the similarity judgments requested of the Miller and Charles (MC) and Rubenstein and Goodenough (RG) participants. For this reason, the WordSimilarity-353 data contains many pairs that are not semantically similar but still receive high scores, such as "computer-software" at 8.81. WS-353 contains pairs that include non-nouns, such as "eat-drink," one proper noun not appearing in WordNet ("Maradona-football"), and some pairs potentially subject to political bias. Again, the aggregate human judgments correlate well with earlier data sets where they overlap—the 30 judgments that WordSimilarity-353 shares with the Miller and Charles data have $\rho = .939$ and the 29 shared with Rubenstein and Goodenough have $\rho = .904$ (by our calculations).

We generated similarity scores for word pairs in all three data sets using the three variants of our walk model (*MarkovLink, MarkovGloss, MarkovJoined*) and with multiple distributional distance measures. We used the WordNet::Similarity package (Pedersen et al., 2004) to compute baseline scores for several existing measures, noting that one word pair was not processed in WS-353 because one of the words was missing from WordNet. The results are summarized in Table 2. These numbers differ slightly from previously reported scores due to variations in the exact experimental setup, WordNet version, and the method of breaking ties when computing ρ (here we break ties using the product-moment formulation of Spearman's rank correlation coefficient). It is worth noting that in their experiments, (Patwardhan and Pedersen, 2006) report that the Vector method has rank correlation coefficients of .91 and .90 for MC and RG, respectively, which are also top performing values.

In our experiments, the *MarkovLink* model with ZKL distance measure was the best performing model overall. *MarkovGloss* and *MarkovJoined* were also strong contenders but with the cosine measure instead of ZKL. One reason for this distinction is that the stationary distributions resulting from the *MarkovLink* model are non-zero for all but the initial word nodes (i.e. non-zero for all *Synset* nodes). Consequently, ZKL's re-estimate for the zero terms adds little information. By contrast, the *MarkovGloss* and *MarkovJoined* models include links that traverse from *Synset* nodes to *TokenPOS* nodes, re-

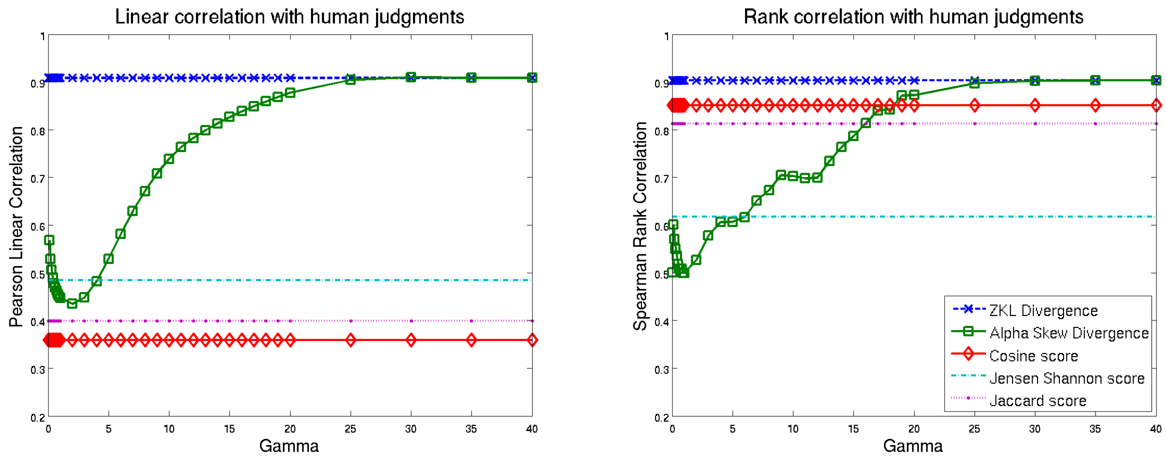

Figure 3: Correlation with the Miller & Charles data sets by linear correlation (left) and rank correlation (right) for the *MarkovLink* model. All data points were based on one set of stationary distributions over the graph; only the divergence measure between those distributions is varied. Note that ZKL_γ dominates both graphs but skew divergence does well for increasing α (computed as $1 - 2^\gamma$). Gamma is swept over the range 0 to 1, then 1 through 20, then 20 through 40 at equal resolutions.

Model	MC Rank	RG Rank	WS-353 Rank
MarkovLink (ZKL)	**.904**	.817	**.552**
MarkovGloss (cosine)	.841	.762	.467
MarkovJoined (cosine)	.841	**.838**	.547
Gloss Vectors	.888	.789	.445
Extended Lesk	.869	.829	.511
Jiang-Conrath	.653	.584	.195
Lin	.625	.599	.216

Table 2: Spearman's ρ rank correlation coefficients with human judgments using $\gamma = 2.0$ for ZKL. Note that figure 3 demonstrates ZKL's insensitivity with regard to the parameter setting for the *MarkovLink* model.

sulting in a final stationary distribution with more (and more meaningful) zero/non-zero pairs. Hence the proper setting of gamma (or alpha for skew divergence) is of greater importance. ZKL's performance improves with tuning of gamma, but cosine similarity remained the more robust measure for these distributions.

6 Conclusion

In this paper, we have introduced a new measure of lexical relatedness based on the divergence of the stationary distributions computed from random walks over graphs extracted WordNet. We have explored the structural properties of extracted semantic graphs and characterized the distinctly different types of stationary distributions that result. We explored several distance measures on these distributions, including ZKL, a novel variant of KL-divergence. Our best relatedness measure is at the limit of human inter-annotator agreement and is one of the strongest measures of semantic relatedness that uses only WordNet as its underlying lexical resource.

In future work, we hope to integrate other lexical resources such as Wikipedia into the walk. Incorporating more types of links from more resources will underline the importance of determining appropriate relative weights for all of the types of edges in the walk's matrix. Even for WordNet, we believe that certain link types, such as antonyms, may be more or less appropriate for certain tasks and should weighted accordingly. And while our measure of lexical relatedness correlates well with human judgments, we hope to show performance gains in a real-word task from the use of our measure.

Acknowledgments

Thanks to Christopher D. Manning and Dan Jurafsky for their helpful comments and suggestions. We are also grateful to Siddharth Patwardhan and Ted Pedersen for assistance in comparing against their system. Thanks to Sushant Prakash, Rion Snow, and Varun Ganapathi for their advice on pursuing some of the ideas in this paper, and to our anonymous reviewers for their helpful critiques. Daniel Ramage was funded in part by an NDSEG fellowship. This work was also supported in part by the DTO AQUAINT Program, the DARPA GALE Program, and the ONR (MURI award N000140510388).

References

S. Banerjee and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco*, pages 805–810.

P. Berkhin. 2005. A survey on pagerank computing. *Internet Mathematics*, 2(1):73–120.

P. Bremaud. 1999. *Markov chains: Gibbs fields, monte carlo simulation, and queues.* Springer-Verlag.

A. Budanitsky and G. Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

K. Collins-Thompson and J. Callan. 2005. Query expansion using random walk models. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 704–711, New York, NY, USA. ACM Press.

C. Fellbaum. 1998. *WordNet: An electronic lexical database.* MIT Press.

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

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*.

T. Haveliwala, A. Gionis, D. Klein, and P. Indyk. 2002. Evaluating strategies for similarity search on the web. In *WWW2002*.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of RANLP-03*, pages 212–219.

J. J. Jiang and D. W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING X)*, pages 19–33.

Lillian Lee. 1999. Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics 2001*, pages 65–72.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. In *IEEE Transactions on Information Theory*, volume 37(1), pages 145–151.

Rada Mihalcea. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 411–418, Morristown, NJ, USA. Association for Computational Linguistics.

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

S. Patwardhan and T. Pedersen. 2006. Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Pyscholinguistics Together*, pages 1–8.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, (11):95–130.

H. Rubenstein and J.B. Goodenough. 1965. Contextual correlates of synonymy. *Computational Linguistics*, 8:627–633.

Michael Strube and Simone Paolo Ponzetto. 2006. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1419–1424.

Paul Tarau, Rada Mihalcea, and Elizabeth Figa. 2005. Semantic document engineering with wordnet and pagerank. In *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, pages 782–786, New York, NY, USA. ACM Press.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pages 252–259.

Kristina Toutanova, Christopher D. Manning, and Andrew Y. Ng. 2004. Learning random walk models for inducing word dependency distributions. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, New York, NY, USA. ACM Press.

Julie Weeds and David Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Comput. Linguist.*, 31(4):439–475.