

# Random Walk on WordNet to Measure Lexical Semantic Relatedness

Yanbo Xu

June 30, 2011

with Professor Kang L James, Professor Ted Pedersen  
Department of Mathematics and Statistics  
University of Minnesota, Duluth  
Duluth, MN 55812

## Acknowledgments

I would like to take this opportunity to give my sincere thanks to my advisors, Dr. Kang L James from Department of Mathematics and Statistics, and Dr. Ted Pedersen from Department of Computer Science, for their encouragement, support and guidance from the initial to the final level, which enabled me to develop an understanding of my project.

Thanks also to Dr. Barry James for being on my degree committee.

Lastly, I offer my regards and blessings to my friends, my family and Tuo Zhao, who have supported me in many respects during the completion of the project.

# Random Walk on WordNet to Measure Lexical Semantic Relatedness

Yanbo Xu

with Professor Kang L James, Professor Ted Pedersen

Department of Mathematics and Statistics

University of Minnesota, Duluth

Duluth, MN 55812

## Abstract

The need to determine semantic relatedness or its inverse, semantic distance, between two lexically expressed concepts is a problem that pervades much of natural language processing such as document summarization, information extraction and retrieval, word sense disambiguation and the automatic correction of word errors in text. Standard ways of measuring similarity between two words on a thesaurus, such as WordNet, are usually based on the path between those words in the thesaurus graph. By contrast, Hughes and Ramage (2007) proposed a Markov chain model of lexical semantic relatedness based on WordNet, which uses a random walk over nodes and edges derived from WordNet link and corpus statistics to compute a word-specific stationary distribution. Based on this model, we propose to apply  $\alpha$ -divergence to score the semantic relatedness of a word pair. In our experiments, the resulting relatedness measure with  $\alpha = .9$  outperforms the existing measures on certain classes of distributions, that is correlated with human similarity judgements by rank ordering at  $\rho = .937$ .

*keywords:* Random Walk, WordNet, Semantic Relatedness Measure, Similarity Measure,  $\alpha$ -divergence, Markov Chain, PageRank

# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>                             | <b>5</b>  |
| <b>2</b> | <b>WordNet</b>                                  | <b>7</b>  |
| <b>3</b> | <b>Related Work</b>                             | <b>9</b>  |
| <b>4</b> | <b>Random Walk on WordNet</b>                   | <b>10</b> |
| 4.1      | Graph Construction . . . . .                    | 11        |
| 4.2      | Computing the stationary distribution . . . . . | 13        |
| <b>5</b> | <b>Relatedness Measures</b>                     | <b>15</b> |
| 5.1      | Zero-KL Divergence . . . . .                    | 15        |
| 5.2      | Hellinger Similarity Measure . . . . .          | 16        |
| 5.3      | $\alpha$ -Divergence . . . . .                  | 17        |
| <b>6</b> | <b>Evaluation</b>                               | <b>19</b> |
| <b>7</b> | <b>Conclusion</b>                               | <b>22</b> |

# 1 Introduction

Many problems in Natural Language Processing require a numerical measure of semantic relatedness between two arbitrary words. Information retrieval needs these scores to expand the query words; word sense disambiguation uses relatedness measures to choose appropriate meaning of a word according to its context; question answering systems often evaluate candidate sentence alignments by using the similarity scores. It is important to distinguish the concepts of semantic relatedness and similarity, where the former one should be a more general concept than the later. Two semantically similar words are intuitively related because of the similarity between their meanings, such as *happy* and *graceful* (synonymy). But two dissimilar concepts can also be semantically related, such as *window* and *house*, since *window* is a part of *house* (meronymy). Moreover, *hot* and *cold* are also related just because of their semantic dissimilarity (antonymy). So a pair of words can be semantically related by any kind of functional relations or their frequent co-occurrence, like *pen* and *paper*. Many applications typically require semantic relatedness rather than similarity. Several popular algorithms compute the lexical semantic relatedness measure based on the knowledge from WordNet [7], an electronic dictionary indicating semantic relationships among word senses.

A main challenge for such algorithms is to find the relatedness for two arbitrary words even if they share few direct relationships. Actually most words in WordNet share no direct semantic links, or some of the pairs can be very far away on WordNet even though intuitively they are highly related. For example, *furnace* and *stove* are connected by going up to the same ancestor *artifact* along a series of *IS-A* (hypernymy/hyponymy) relationship. Several existing algorithms use the shortest path or only the hypernym/hyponymy relationship to define relatedness measure and find *furnace* and *stove* are not related. However, WordNet provides lots of other relationships besides hypernymy/hyponymy, such as meronymy, antonymy, verb entailment etc. In addition, some implicit links can also be obtained from the text of the definitional gloss information in WordNet. If all of the above relationships can be considered, the *furnace* will be easily connected to *stove* by following the path *furnace - crematory - gas oven - oven - kitchen appliance - stove*. Another problem is that many pairs are connected in WordNet through several paths, such as *moon* is connected to *satellite* by two paths: one is from the sense of *moon* as *the natural satellite of the Earth*, the other is from the sense of *moon* as *any natural satellite of a planet*. *moon* is also connected to *religious leader* by only one path which is from the sense of *moon* as *Sun Myung Moon: a United States religious leader*. Both *religious leader* and *satellite* are at the same shortest

path distance from *moon*, however, the connectivity structure of the graph would suggest *satellite* to be "more" similar than *religious leader* as there are multiple senses, and hence multiple paths, connecting *satellite* and *moon*. These problems motivated people to find an alternative way to measure semantic relatedness on WordNet rather than the path-based methods.

Hughes and Ramage [10] presented the application of random walk Markov chain theory to measure lexical semantic relatedness. A weighted and directed graph of words/-concepts is constructed from WordNet. A particle starting at an arbitrary word/concept on the graph can randomly walk towards a target word and after a long term running it will visit each node with the proportion of each node's visiting times converging to a stationary distribution. An interpretation of such word-specific probability distributions is how often the particle visits all the other nodes when starting from a specific word. Then the relatedness score between two words can be defined by the similarity of their respective stationary distributions. This random walk method enables a consistent way to combine multiple types of relationships on WordNet and extend a local similarity statistics into the entire graph.

In this project, we compute a stationary distribution for each word in WordNet referring to the random walk model proposed by Hughes and Ramage. Beyond applying their Zero-KL(ZKL) divergence as relatedness measure, however, we propose to use  $\alpha$ -divergence which derives Kullback-Leibler (KL) divergence [13] when  $\alpha \rightarrow 0$  or 1. We can show that the  $\alpha$ -divergence measure outperforms the ZKL and the existing measures on certain classes of distributions when  $\alpha = .9$ , with a correlated coefficient  $\rho = .937$  to human similarity judgements.

## 2 WordNet

WordNet is a lexical database of English with nouns, verbs, adjectives and adverbs included. All of these words are grouped into sets of synonyms, called *synsets*. *Synsets* are interlinked by types of semantic and lexical relationships. Each *synset* can present a discriminative concept and be further explained by a short gloss (usually definition and/or example sentences). For example, a *synset* in WordNet with glosses is shown as followed:

dog#n#1, domestic dog#n#1, Canis familiaris#n#1 (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) "*the dog barked all night*"

where "dog#n#1" represents the 1st sense of word *dog* as its part of speech noun, "dog#n#1, domestic dog#n#1, Canis familiaris#n#1" represents a *synset* and the sentences followed represent the gloss of this *synset*.

The latest version of WordNet, WordNet 3.0, contains 155,287 words (including 26,896 polysemous words), 206,941 word-sense pairs for a total of 117,659 *synsets*. The semantic and lexical relationships shared between *synsets* are based on the part-of-speech and few relations can cross the part-of-speech boundaries. These relations include:

- Nouns
  - *hypernyms/hyponyms* (IS-A relation): *canine* is a hypernym of *dog*, *dog* is a hyponym of *canine* (*dog* IS A kind of *canine*)
  - *coordinate terms*: *wolf* is a coordinate term of *dog*, *dog* is a coordinate term of *wolf* (*wolf* and *dog* share a hypernym *canine*)
  - *holonym/meronym* (IS-A relation): *building* is a holonym of *window*, *building* is a holonym of *window* (*window* IS A part of *building*)
- Verbs
  - *hypernyms/hyponyms* (IS-A relation): to *perceive* is an hypernym of to *listen*, to *listen* is an hyponym of to *perceive* (the activity of *perceiving* IS A kind of *listening*)
  - *troponyms*: *lisp* is a troponym of to *talk* (the activity *talking* is doing *lisp*ing in some manner)

```
dog, domestic dog, Canis familiaris
=> canine, canid
=> carnivore
=> placental, placental mammal, eutherian, eutherian mammal
=> mammal
=> vertebrate, craniate
=> chordate
=> animal, animate being, beast, brute, creature, fauna
=> ...
```

Figure 1: An example of WordNet3.0

- *entailment*: *sleep* is entailed by *to snore* (by doing *snoring* you must be doing *sleeping*)
- *coordinate terms*: *to lisp* and *to yell*

- Adjectives

- *related nouns*
- *similar to*
- *participle of verb*

- Adverbs

- *root adjectives*

The organizations of nouns and verbs more look like hierarchical trees respectively, because of their main relation *hypernym* or *IS-A* relationship. For instance, the first sense of the word *dog* as noun would have the following hypernym hierarchy, where the words at the same level are synonyms of each other:

### 3 Related Work

A simple way to measure similarity in WordNet is to treat it as a graph and then define the shorter path two nodes share, the more similar they are [23]. This shortest path length in a lexical network can actually measure conceptual distance if the path is *IS-A* relationship. Jiang and Conrath, 1997 [12] proposed to find the shortest path in the taxonomic hierarchy between two concepts (similarly to [17]), which had one of the best performance [5]. Both of the above methods primarily make use of *IS-A* relationship, i.e. hypernymy, in WordNet and therefore are more like to measure similarity than to measure relatedness.

Some alternative approaches have been proposed, such as Extended Lesk [2], Gloss Vectors [21] etc. The basic idea behind these methods is to compare the "bag of words" in WordNet *Synsets*' definitional gloss texts. Thus, these gloss-based approaches can measure relatedness. Besides, Weeds and Weir, 2005 [29] defined distributional similarity measures based on co-occurrence in body of text.

Apart from WordNet, people also use some other structured information resources. Jarmasz and Szpakowicz, 2003 [11] showed "the Roget's thesaurus and WordNet are birds of a feather". Incorporating information from Wikipedia has become more popular recently [9] [25]. These Wikipedia-based measures can perform better on some tasks than some existing measures based on shallower lexical resources. But the results of Huges and Ramage's algorithm are competitive even if using only WordNet.

Huges and Ramage proposed to apply a random walk algorithm, PageRank – see [3] for a survey, to obtain a probabilistic distribution for each concept in WordNet. Random walks have been applied to natural language processing in several works, such as PP attachment [26], word sense disambiguation [18], query expansion [6] etc. Hughes and Ramage calculated distinct stationary distributions from random walks centered at different nodes in the WordNet graph. They then introduced a novel variant of KL-divergence to measure semantic relatedness and showed this is one of the strongest measures. Rao, Yarowsky and Callison-Burch (2008) [31] used the pseudo inverse of the Laplacian to derive estimates for commute times in graph. They further improved this measure by discarding the least significant eigenvectors, corresponding to the noise in the graph construction process, using Singular Value Decomposition (SVD). But they only considered uniform weights on all edges derived from WordNet and didn't show an explicit advantage over Hughes and Ramage's.

## 4 Random Walk on WordNet

Let  $X = \{X_1, X_2, \dots, X_T\}$  be a set of finite states, then a *random walk* on  $X$  corresponds to a sequence of states, where each state represents each step of the walk. At each step, the walk will either leave its current state to a new state or remain at the current state. *Random walks* are usually *Markovian*, which means the transition at each step is independent of the previous steps and only depends on the current state.

The model proposed by Hughes and Ramage is based on a random walk through a directed graph  $G = (V, E, \mathbf{W})$  derived from WordNet, with nodes  $V$ , edges  $E$ , and weight matrix  $\mathbf{W} = [w_{ij}]$ , where  $w_{ji}$  as the weight of the edge connecting node  $i$  and  $j$ . The weight,  $w_{ij} = 0$  for nodes  $i$  and  $j$  being not neighbors;  $w_{ij} > 0$  for an indication of relationship from  $j \rightarrow i$  in WordNet. We now can obtain a matrix  $\mathbf{P}$  from  $\mathbf{W}$  as follows:

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$$

where  $\mathbf{D}$  is a diagonal matrix with  $d_{jj} = \sum_{i=1}^n w_{ij}$ ,  $n = |V|$  the order of the graph. Observe that:

- $p_{ii} = \mathbf{P}[i, i] = 0$ ,
- $p_{ij} = \mathbf{P}[i, j] \geq 0$ , and
- $\sum_{i=1}^n p_{ij} = 1$

Hence  $p_{ij}$  can be interpreted as the conditional probability of moving to node  $i$  given the current state is at node  $j$ . If  $P\{X_t = i\}$  denote the probability of being at node  $i$  at time  $t$ , then, by total probability theorem, it can be defined as the sum of probabilities over all ways to node  $i$  from any other nodes at the previous time step. Formally,

$$P\{X_t = i\} = \sum_{j \in V} P\{X_{t-1} = j\}P\{X_t = i | X_{t-1} = j\} = \sum_{j \in V} P\{X_{t-1} = j\}\mathbf{P}[i, j].$$

Note that this is a Markov chain because the transition probability at time  $t$  is dependent only on the last states at time  $t - 1$  and the matrix  $\mathbf{P}$  is the probability transition matrix

with columns normalized.

The following subsections present how to construct the weight graph  $G$  from WordNet and then compute the stationary distribution for a given word by random walking.

## 4.1 Graph Construction

As introduced in the previous section, WordNet is itself a graph over *synsets* and WordNet explicitly marks semantic relationships between *synsets*. But we are also interested in representing relatedness between words. Therefore the nodes  $V$  includes three types of nodes extracted from WordNet:

**Synset** The same *synset* node from WordNet. For example, one node corresponds to the *synset* of "dog#n#1, domestic dog#n#1, Canis familiaris#n#1". There are 117,597 **Synset** nodes.

**TokenPOS** Assign each word contained in WordNet followed by its part of speech, such as "dog#n" meaning *dog* as a noun. These nodes link to all the **Synset** nodes for "dog#n#1, domestic dog#n#1, Canis familiaris#n#1", "dog#n#2, frump#n#1", "dog#n#3", etc. There are 156,588 **TokenPOS** nodes.

**Token** Each single word in WordNet. By eliminating the part of speech information, every **TokenPOS** is connected to its corresponding **Token**. For example, "dog" is connected to "dog#n" and "dog#v" (meaning "go after with the intent to catch"). There are 148,646 **Token** nodes.

For each edge in  $E$  connecting node  $i$  and  $j$ , it has a weight  $w_{ij}$ . From the semantic and lexical relationships provided in WordNet, we can derive the edges  $E$  and weight matrix  $W$  like this:

- Weights of the uni-directional edges connecting **TokenPOS** to **Synset** are based on the *SemCor frequency counts* included in WordNet. To distinguish the zero counts from the zero weight, we add pseudo-count 0.1 for all *SemCor frequency counts* of each

target **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, domestic dog#n#1, Canis familiaris#n#1" and "dog#n#3" have weights of 42.1 and 0.1, respectively.

- Uni-directional edges that connect **Token** to **TokenPOS** are weighted as the sum of outgoing weights from this **TokenPOS** to **Synset** nodes. This also makes the walker, starting at word "dog", be more likely to follow a link to "dog#n" than to "dog#v". For example, the edges from "dog" to "dog#n" and "dog#v" have weights of 42.7 and 2.1, respectively.
- Add uni-directional edges from **Synset** to **TokenPOS** in order to exploit the textual *gloss* information from WordNet. A **Synset** node is connected to **TokenPOS** nodes who are used in that *synset's gloss* definition. This requires the glosses to be part-of-speech tagged, for which we use the Brill Tagger (Eric Brill 1993). Then we weight each **TokenPOS** by its distance from the mean word frequency counts in the gloss in log space, which is called as "Non-Monotonic Document Frequency" (NMDF). Formally, the weight  $w_{ij}$  from a **Synset** node  $j$  to a **TokenPOS**  $i$  is:

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

where  $r_i$  is the appearing times of this **TokenPOS** in the gloss,  $\mu$  and  $\sigma$  are the mean and standard deviation of the logs of all word counts. As we can see, this function can down weights the extremely rare words, which is opposite from the **tf-idf** weighting. On the other hand, the high-frequency stop words such as "by" and "the" are also assigned by low weights. This is exactly what we need to roughly weight the relatedness between two nodes. Thus NMDF weighting can be very effective here.

- We also add bi-directional edges between **Synsets** who share common **TokenPOS** nodes. These edges are weighted by how many common **TokenPOS** nodes the two **Synsets** share. Intuitively, different senses of the same word should be considered related.

- At last, **Synset** nodes are connected with bi-directional edges according to the semantic relationship types in WordNet. We use the following WordNet semantic relationships to form edges: hypernyms/hyponyms, instance of/has instance of, meronyms/holonyms, attributes, similar to, entailment, cause, domain/member of domain - all, as well as some WordNet lexical relationships: antonyms, derived forms, participle of verb, pertains to noun and verb group. The weights are uniformed by 10, which is approximately the maximum of the "overlap" weights we obtained in our experiment. The intuition to use this maximum is that edges from the direct relationship provided in WordNet should have a higher weight than those from an indirect relationship such as overlapping with a common **TokenPOS**.

In our experiment, after loading all the nouns from WordNet, we can construct a graph with 357,756 nodes and 1,287,050 weighted edges, which is very sparse; fewer than 1 in 100,000 node pairs are directly connected. This makes our experiment can quickly converge and get the stationary distribution, which will be introduced in the next subsection, in about 30 iterations.

## 4.2 Computing the stationary distribution

Now we can get a columns normalized matrix  $\mathbf{P}$  from the weight matrix  $\mathbf{W}$  presented above. A generalized PageRank algorithm is then introduced to compute a stationary distribution for any specific word based on  $\mathbf{P}$ , which contains the relatedness information about all nodes extracted from WordNet. The notion of PageRank, first introduced by Brin and Page (1998) [20], forms the basis for their Web search algorithms. Although the original version of PageRank was used for the Web graph (with all the webpages as nodes and hyperlinks as edges), PageRank is well defined for any given graph and is quite effective for capturing various relations among nodes of graphs.

In order to compute the stationary distribution  $v_{dog\#n}$  for a walk centered at the **TokenPOS** "dog#n", a generalized PageRank algorithm first defines an initial distribution  $v_{dog\#n}^{(0)}$  as an element vector with probability 1 assigned to itself and 0 to the other nodes. Then at every step of the walk, it will "jump" to  $v_{dog\#n}^{(0)}$  with probability  $\beta$ . Intuitively, this return probability captures the notion that nodes close to "dog#n" should be given higher

probabilities, and also guarantees a unique stationary distribution (Bremaud, 1999) [4]. The stationary distribution  $v_i$  for node  $i$  is computed with an iterative update algorithm:

- Let  $v_i^{(0)} = e_i$ ,  $e_i$  is a vector of all zeros except  $e_i(i) = 1$
- Repeat until  $\|v_i^{(t)} - v_i^{(t-1)}\|_1 < \epsilon$ 
  - $v_i^{(t)} = \beta v_i^{(0)} + (1 - \beta)\mathbf{P}v_i^{(t-1)}$
  - $t = t + 1$
- $v_i = v_i^{(t)}$

We decide to apply the same convergence criteria as Hughes and Ramage use:  $\epsilon = 10^{-10}$  and  $\beta = 0.1$ . Due to the sparsity of the matrix  $\mathbf{P}$ , it only takes three dozen iterations to meet the criteria in our experiment.

## 5 Relatedness Measures

Based on Hughes and Ramage’s random walk model, we can compute the word-specific stationary distribution for any word in WordNet. Now we need to consider how to define a relatedness measure between two words. For a weighted graph  $G(V, E, \mathbf{W})$ , we define a relation **sim**:  $V \times V \rightarrow \mathfrak{R}^+$  such that **sim**( $p, q$ ) is the relatedness measure between two words with their respective stationary distributions,  $p$  and  $q$ . In our case, we measure the divergence of the word-specific distributions since, intuitively, if the random walker starting at the first word’s node and another random walker starting at the second word’s node tend to visit the same nodes, they can be considered semantically related.

Many statistical similarity measures between probability distributions can be applied to this task. One standard way is to consider  $p$  and  $q$  geometrically be two vectors, and then measure the cosine value of the angle between them. This is usually called *cosine similarity* and represented as a dot product:

$$\mathbf{sim}_{\cos}(p, q) = \frac{\sum_i p_i q_i}{\|p\| \|q\|}.$$

In the following subsections, we would like to introduce some other measures such as the ZKL divergence (a modified KL-divergence proposed by Hughes and Ramage), Hellinger distance and  $\alpha$ -divergence. All of these measures give a dissimilarity score to the word pair, which actually are equivalent to measure similarities. We will later apply them on the random walk model and compare them with some existing measures, including the above cosine similarity measure.

### 5.1 Zero-KL Divergence

KL-divergence (or relative-entropy) is an asymmetric dissimilarity measure between two probability distributions. Intuitively speaking, it measures the added number of bits required for encoding events sampled from the first distribution using a code based on the second distribution. The need for such a measure arises quite naturally for probability distributions associated with language models such as measuring the distance between a query and members of a document collection [14]. Because  $p$  and  $q$  are probability distributions, we could expect to apply KL-divergence to be our relatedness measure, defined as:

$$\mathbf{sim}_{KL}(p, q) = D_{KL}(p||q) = \sum_i p_i \log \frac{p_i}{q_i}.$$

As shown in the formula, however, KL-divergence is undefined if any  $q_i$  is zero. Several modifications to smooth KL-divergence have been proposed and Lee, 2001 [16] surveyed that the Skew divergence measure [15] outperformed the others.

$$s_\alpha(p, q) = D_{KL}(p||\alpha q + (1 - \alpha)p) = \sum_i p_i \log \frac{p_i}{\alpha q_i + (1 - \alpha)p_i}.$$

To simplify the Skew divergence, Hughes and Ramage proposed a reinterpretation of the Skew divergence. While  $q_i = 0$  simplifies the Skew divergence to  $p_i \log \frac{p_i}{(1-\alpha)p_i} = p_i \log \frac{1}{1-\alpha}$ , and Skew divergence’s best performance was shown by Lee for  $\alpha$  near 1, thus Hughes and Ramage choose  $\alpha$  as  $1 - 2^{-\gamma}$  for some  $\gamma \in \mathbb{R}^+$ . Zero terms in the sum can now be written as  $p_i \log \frac{1}{2^{-\gamma}} = p_i \log 2^\gamma \approx p_i \gamma$ . The final ZKL divergence measure is defined with respect to  $\gamma$ :

$$\mathbf{sim}_{ZKL\gamma}(p, q) = \sum_i p_i \begin{cases} \log \frac{p_i}{q_i} & q_i \neq 0 \\ \gamma & q_i = 0. \end{cases}$$

Hughes and Ramage showed this ZKL-divergence measure is the WordNet-based measure most highly correlated with human similarity judgements with  $\gamma = 2$ . In our experiment, we found when  $\gamma = 5$ , ZKL-divergence gave its best performance but it didn’t outperform our resulting measure.

## 5.2 Hellinger Similarity Measure

Hellinger similarity measure is inspired from the Hellinger distance between two probability distributions. The Hellinger distance between  $p$  and  $q$  is defined as the quantity

$$D_H(p, q) = \left(\frac{1}{2} \sum_{i=1}^n (\sqrt{p_i} - \sqrt{q_i})^2\right)^{\frac{1}{2}} = \left(1 - \sum_{i=1}^n \sqrt{p_i q_i}\right)^{\frac{1}{2}}.$$

Since a distance measures dissimilarity, we define the Hellinger similarity measure as:

$$\mathbf{sim}_H(p, q) = \sum_{i=1}^n \sqrt{p_i q_i}.$$

Both KL-divergence and Hellinger distance are in the family of  $\alpha$ -divergence [1] [28] [27], with  $\alpha \rightarrow 0$  and  $\alpha = \frac{1}{2}$  respectively. So we introduce  $\alpha$ -divergence in the following subsection.

### 5.3 $\alpha$ -Divergence

$\alpha$ -divergence is a family of divergences, indexed by  $\alpha \in (-\infty, \infty)$ . The formula of  $\alpha$ -divergence between two continuous probability distributions  $p(x)$  and  $q(x)$  is:

$$D_\alpha(p||q) = \frac{\int_x \alpha p(x) + (1 - \alpha)q(x) - p(x)^\alpha q(x)^{1-\alpha} dx}{\alpha(1 - \alpha)}$$

for all  $\alpha \in (-\infty, \infty)$ .

$\alpha$ -divergence has several important properties as followed:

- $D_\alpha(p||q)$  is convex with respect to  $p(x)$  and  $q(x)$

We first simplify the definition of  $\alpha$ -divergence to  $D_\alpha(p||q) = \frac{1 - \int p(x)^\alpha q(x)^{1-\alpha} dx}{\alpha(1-\alpha)}$ . Now, by showing  $\frac{\partial^2 D_\alpha(p||q)}{\partial p^2} = \frac{\int \alpha(1-\alpha)p(x)^{\alpha-2} q(x)^{1-\alpha} dx}{\alpha(1-\alpha)} = \int p(x)^{\alpha-2} q(x)^{1-\alpha} \geq 0$  and repeating the same calculation for  $q$ , we prove  $D_\alpha(p||q)$  is convex with respect to  $p$  and  $q$ .

- $D_\alpha(p||q) \geq 0$  and  $D_\alpha = 0$  if and only if  $p(x) = q(x)$  for all  $x$

Let  $f(p, q) = \frac{\alpha p(x) + (1-\alpha)q(x) - p(x)^\alpha q(x)^{1-\alpha}}{\alpha(1-\alpha)}$ , we can easily prove that  $f(p, p) = 0$ ,  $\frac{\partial f}{\partial p}(p, q) = 0$  and  $\frac{\partial^2 f}{\partial p^2}(p, q) > 0$  for all  $p > 0$ . Thus,  $p$  has only one minimum with respect to  $p$  and it occurs at  $p = q$ . The same result works for  $q$ . Therefore,  $D_\alpha(p, q) \geq 0$  for all  $p$  and  $q$  and the equality is obtained if and only if  $p(x) = q(x)$  for all  $x$ .

- KL-divergence is a special case of an  $\alpha$ -divergence with  $\alpha \rightarrow 0$

By using L'Hopital rule, we can see  $\lim_{\alpha \rightarrow 0} D_\alpha(p||q) = \lim_{\alpha \rightarrow 0} \frac{-\int \ln \frac{p(x)}{q(x)} p(x)^\alpha q(x)^{1-\alpha} dx}{1-2\alpha} = -\int q(x) \ln \frac{p(x)}{q(x)} dx$ , which is the definition of KL-divergence between continuous distributions  $q(x)$  and  $p(x)$ ,  $D_{KL}(q||p)$ .

Note that it is symmetric between  $(p, \alpha)$  and  $(q, 1 - \alpha)$ , thus  $\lim_{\alpha \rightarrow 1} D_\alpha(p||q) = D_{KL}(p||q)$ .

Therefore,  $\alpha$ -divergence includes KL-divergence as a special case. Some other special cases are:

$$D_{-1}(p||q) = \frac{1}{2} \int \frac{(q(x) - p(x))^2}{q(x)} dx$$

$$D_2(p||q) = \frac{1}{2} \int \frac{(q(x) - p(x))^2}{p(x)} dx$$

$$D_{\frac{1}{2}} = 2 \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$$

where  $\sqrt{D_{\frac{1}{2}}}$  is the Hellinger distance we showed in the last subsection. But here it is defined between continuous distributions  $p(x)$  and  $q(x)$ .

Due to the properties of  $\alpha$ -divergence and it being a general form of KL-divergence, it is possible to assume that we can find a good  $\alpha$ , which can measure the distance well between two stationary distributions  $p$  and  $q$  in our model. Therefore, we define:

$$\mathbf{sim}_\alpha(p, q) = \frac{\sum_{i=1}^n p_i^\alpha q_i^{1-\alpha}}{\alpha(1-\alpha)}.$$

## 6 Evaluation

Now we have several methods to compute similarity scores for any pair of words in WordNet via their stationary distributions. We are going to evaluate these measures by calculating their correlations to human judgement of relatedness.

Due to a different scale of each set of scores, we will use Spearman’s rank correlation coefficient to quantify the correlation between sets of judgements. This rank coefficient can also capture some information about the relative ordering of scores. For two ranked sets of  $n$  scores  $x_i$  and  $y_i$  with the difference  $d_i = x_i - y_i$ , Spearman’s  $\rho$  coefficient is defined as follow:ed

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}.$$

If the tied ranks are considered, a corrected Spearman’s  $\rho$  coefficient is then defined:

$$\rho = \frac{C_x + C_y - \sum_{i=1}^n d_i^2}{2\sqrt{C_x C_y}},$$

where

$$C_x = \frac{n^3 - n}{12} - \sum_{t_x} \frac{t_x^3 - t_x}{12}$$

$$C_y = \frac{n^3 - n}{12} - \sum_{t_y} \frac{t_y^3 - t_y}{12},$$

and  $t_x$  and  $t_y$  stand for the number of observations with the same rank order in  $x_i$ 's and  $y_i$ 's.

There are several popular datasets by human judgements of relatedness applied in evaluating the similarity measures. The most widely used one is from Miller and Charles ,1991 [19], who repeated their experiments on 30 pairs of words. Even though individuals can vary their judgements, the results summed up in Miller and Charles’ experiments are still very close to the scores conducted by Rubenstein and Goodenough ,1965 [24], which provide human judgements of semantic similarity for 65 pairs of words. Both datasets have been scored on a scale of zero to four. Finkelstein et al., 2002 [8] offered a larger set of word relatedness judgements, WordSimilarity-353 (WS-353, on a scale of zero to ten). However, the WS-353 data contains a lot of word pairs that are not semantically related but still have higher scores, such as "computer - software". So in our experiment, we will compute the correlation to Miller and Charles’s (MC), Rubenstein and Goodenough’s (RG) and only 30 pairs judgements from WS-353 that are also included in MC.

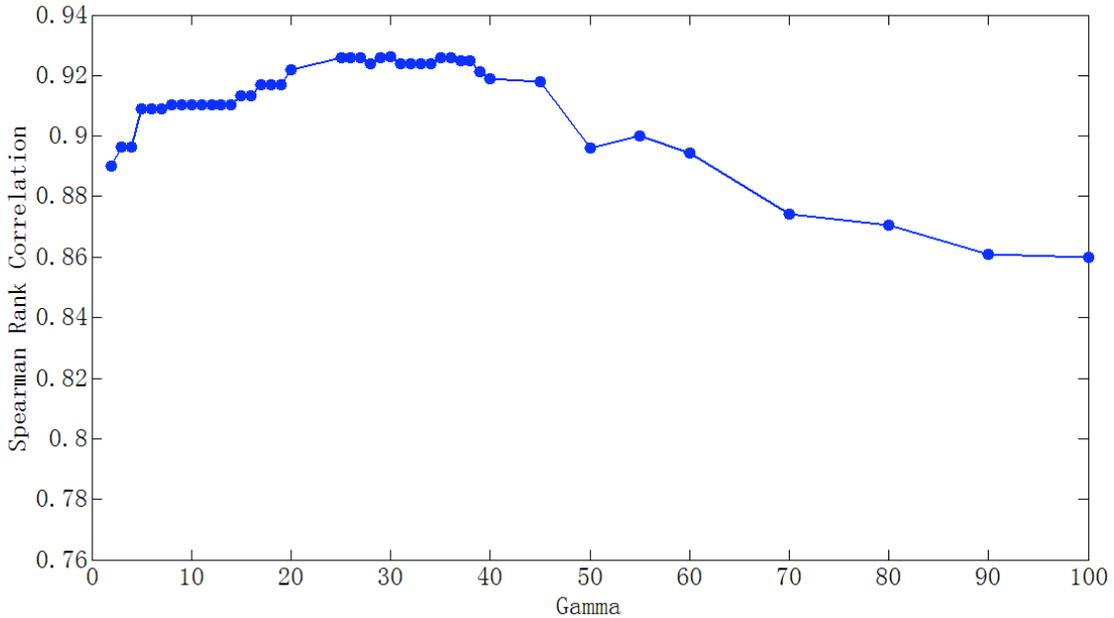


Figure 2: Spearman rank correlation between ZKL-divergence and human judgements

Although for ZKL-divergence, Hughes and Ramage recommended to set  $\gamma = 2$ , we found when  $\gamma = 30$  the ZKL-divergence will give the best Spearman correlation,  $\rho = .926$  for Miller and Charles data (shown in Figure 3). The reason causing the above difference is because, on one hand, we were using a different part of speech tagger from Hughes and Ramage’s, which was the Stanford maximum entropy tagger. On the other hand, we ourselves specified some steps that are not detailed in Hughes and Ramage’s graph construction, such as how to weight the semantic relationships directly included in WordNet and to use a Bayesian estimate based on the SemCor frequency counts to weight the edges between TokenPOS and Synset nodes.

Figure 3 shows that the  $\alpha$ -divergence provides  $\rho = .937$  for Miller and Charles when  $\alpha = .9$ , which outperforms the ZKL-divergence. We also generate some other similarity measures based on the three datasets and  $\alpha$ -divergence also shows the best performance in Table 1. We use the WordNet::Similarity package [22] to compute the scores for four existing measure: *wup* [30], *vector* [21], *lesk* [2] and *path*. *path* and *wup* are based on path lengths, whereas *vector* and *lesk* are based on the glosses and they both can cross part of speech boundaries.

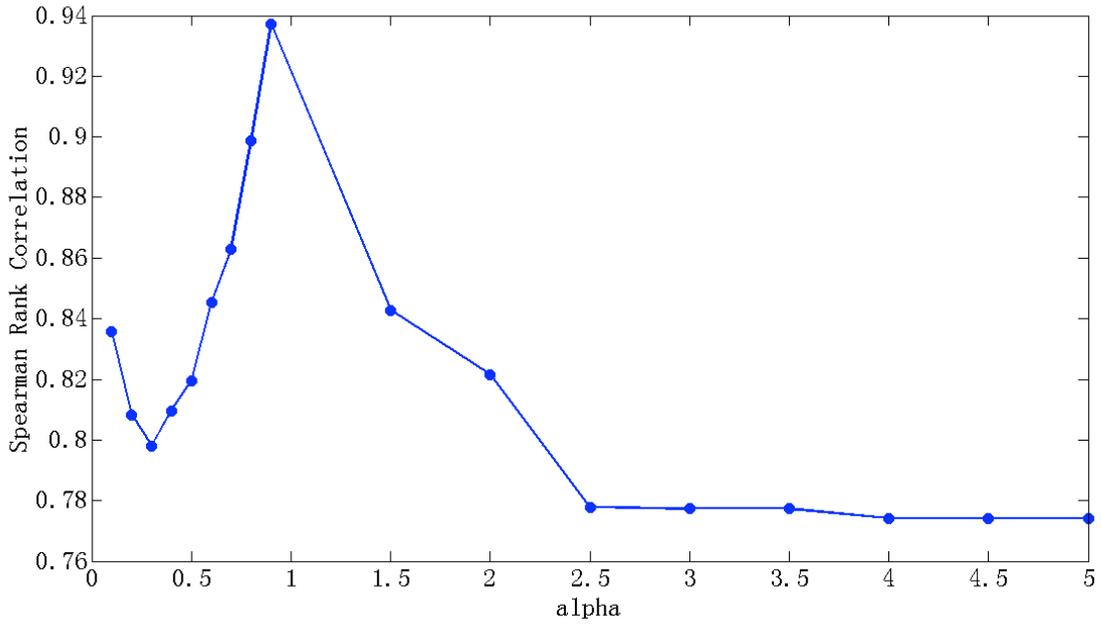


Figure 3: Spearman rank correlation between  $\alpha$ -divergence and human judgements

Table 1.

Different relatedness measures' Spearman rank correlation  $\rho$ 's with human judgements.

| Relatedness measures                   | $\rho$ with MC | $\rho$ with RG | $\rho$ with WS-353 |
|--|----------------|----------------|--------------------|
| $\alpha$ -divergence ( $\alpha = .9$ ) | <b>.937</b>    | <b>.816</b>    | <b>.852</b>        |
| ZKL-divergence ( $\gamma = 30$ )       | .926           | .775           | .842               |
| Hellinger                              | .820           | .718           | .718               |
| Cosine                                 | .883           | .736           | .842               |
| <i>wup</i>                             | .769           | .697           | .715               |
| <i>vector</i>                          | .891           | .721           | .813               |
| <i>path</i>                            | .732           | .702           | .680               |
| <i>lesk</i>                            | .826           | .664           | .814               |

## 7 Conclusion

In this project, we have introduced and re-implemented Hughes and Ramage’s Markov chain model on WordNet. According to their idea, we can compute a stationary distribution for each word by using a random walk algorithm. To measure the lexical semantic relatedness between a word pair, we propose to use the  $\alpha$ -divergence, which is computing the divergence between two distributions of the two words. We have tuned the parameter  $\alpha$  based on comparing the correlation with human judgements of relatedness. Our measure outperforms the existing measures, including the ZKL-divergence that was shown having the best score by Hughes and Ramage in 2007.

## References

- [1] S. Amari. *Differential-geometrical methods in statistics*. Springer Verlag, 1985.
- [2] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco, 2003.
- [3] P. Berkhin. A survey on pagerank computing. *Internet Mathematics*, 2(1):73–120, 2005.
- [4] P. Bremaud. *Markov chains: Gibbs field, montecarlo simulation, and queues*. Springer Verlag, 1999.
- [5] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, pages 32(1):13–47, 2006.
- [6] K. Collins-thompson and J. Callan. Query expansion using random walk models. In *CIKM*, pages 704–711, 2005.
- [7] C. Fellbaum. *WordNet: An electronic lexical database*. MIT Press, 1998.
- [8] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlina, Z. Solan, G. Wolfman, and E. Rupp. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002.
- [9] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, pages 1606–1611, 2007.
- [10] T. Hughes and D. Ramage. Lexical semantic relatedness with random graph walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 581–589, Prague, June 2007. ©2007 Association for Computational Linguistics.
- [11] Mario Jarmasz and Stan Szpakowicz. S.: Roget’s thesaurus and semantic similarity. In *In: Proceedings of the RANLP-2003*, pages 212–219, 2003.
- [12] J. Jiang and D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics (ROCLING X)*, pages 19–33, Taiwan, 1997.

- [13] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [14] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 111–119, 2001.
- [15] L. Lee. Measures of distributional similarity. In *Proceedings of ACL-1999*, pages 23–32, 1999.
- [16] L. Lee. On the effectiveness of the skew divergence for statistical language analysis. *Artificial Intelligence and Statistics*,, pages 65–72, 2001.
- [17] D. Lin. An information-theoretic definition of similarity. In *Proceedings of 15th International Conference on Machine Learning*, pages 296–304, August 1998.
- [18] R. Mihalcea. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 411–418. Association for Computational Linguistics, 2005.
- [19] G.A. Miller and W.G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.
- [20] L. Page S. Brin R. Motwani and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Technical report*, 1998.
- [21] S. Patwardhan and T. Pedersen. Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, 2006.
- [22] T. Pedersen S. Patwardhan and J. Michelizzi. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, 2004.
- [23] P. Resnik. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, Canada, August 1995.

- [24] H. Rubenstein and J.B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.
- [25] M. Strube and S.P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1419–1424, 2006.
- [26] K. Toutanova, C. D. Manning, and A. Y. Ng. Learning random walk models for inducing word dependency distributions. In *ICML*. ACM Press, 2004.
- [27] M. Trottni and F. Spezzaferri. Information geometric measurements of generalization. *Technical report NCRG/4350*, 1995.
- [28] M. Trottni and F. Spezzaferri. A generalized predictive criterion for model selection. *Technical report 702*, 1999.
- [29] J. Weeds and D. Weir. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31:439–475, 2005.
- [30] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, 1994.
- [31] D. Rao D. Yarowsky and C. Callison-Burch. Affinity measures based on the graph laplacian. In *Proceedings of 3rd Textgraphs workshop on Graph-Based Algorithms in Natural Language Processing*, pages 41–48, Manchester, 2008.