Refining the Notions of Depth and Density in WordNet-based Semantic Similarity Measures

Tong Wang
Department of Computer Science
University of Toronto
tong@cs.toronto.edu

Graeme Hirst
Department of Computer Science
University of Toronto
gh@cs.toronto.edu

Abstract

We re-investigate the rationale for and the effectiveness of adopting the notions of depth and density in WordNet-based semantic similarity measures. We show that the intuition for including these notions in WordNet-based similarity measures does not always stand up to empirical examination. In particular, the traditional definitions of depth and density as ordinal integer values in the hierarchical structure of WordNet does not always correlate with human judgment of lexical semantic similarity, which imposes strong limitations on their contribution to an accurate similarity measure. We thus propose several novel definitions of depth and density, which yield significant improvement in degree of correlation with similarity. When used in WordNet-based semantic similarity measures, the new definitions consistently improve performance on a task of correlating with human judgment.

1 Introduction

Semantic similarity measures are widely used in natural language processing for measuring distance between meanings of words. There are currently two mainstream approaches to deriving such measures, i.e., distributional and lexical resource-based approaches. The former usually explores the co-occurrence patterns of words in large collections of texts such as text corpora (Lin, 1998) or the Web (Turney, 2001). The latter takes advantage of mostly handcrafted information, such as dictionaries (Chodorow et al., 1985; Kozima and Ito, 1997) or thesauri (Jarmasz and Szpakowicz, 2003).

Another important resource in the latter stream is semantic taxonomies such as WordNet (Fellbaum, 1998). Despite their high cost of compilation and limited availability across languages, semantic taxonomies have been widely used in similarity measures, and one of the main reasons behind this is that the often complex notion of lexical semantic similarity can be approximated with ease by the distance between words (represented as nodes) in their hierarchical structures, and this approximation appeals much to our intuition. Even methods as simple as "hop counts" between nodes (e.g., that of Rada et al. 1989 on the English WordNet) can take us a long way. Meanwhile, taxonomy-based methods have been constantly refined by incorporating various structural features such as depth (Sussna, 1993; Wu and Palmer, 1994), density (Sussna, 1993), type of connection (Hirst and St-Onge, 1998; Sussna, 1993), word class (sense) frequency estimates (Resnik, 1999), or a combination these features (Jiang and Conrath, 1997). Most of these algorithms are fairly self-contained and easy to implement, with off-the-shelf toolkits such as that of Pedersen et al. (2004).

With the existing literature focusing on carefully weighting these features to construct a better semantic similarity measure, however, the rationale for adopting these features in calculating semantic similarity remains largely intuitive. To the best of our knowledge, there is no empirical study directly investigating the effectiveness of adopting structural features such as depth and density. This serves as the major motivation for this study.

The paper is organized as follows. In Section 2 we review the basic rationale for adopting depth

and density in WordNet-based similarity measures as well as existing literature on constructing such measures. In Section 3, we show the limitations of the current definitions of depth and density as well as possible explanations for these limitations.[1] We then propose new definitions to avoid such limitations in Section 4. The effectiveness of the new definitions is evaluated by applying them in semantic similarity measures in Section 5 and conclusions made in Section 6.

2 Related Work

The following are the current definitions of depth and density which we aim at improving. Given a node/concept c in WordNet, depth refers to the number of nodes between c and the root of WordNet, (i.e., the root has depth zero, its hyponyms depth one, and so on). There are more variations in the definition of density, but it is usually defined as the number of edges leaving c (i.e., its number of child nodes) or leaving its parent node(s) (i.e., its number of sibling nodes). We choose to use the latter since it is used by most of the existing literature.

2.1 The Rationale for Depth and Density

The rationale for using the notions of depth and density in WordNet-based semantic similarity measures is based on the following assumption:

Assumption 1 *Everything else being equal, two nodes are semantically closer if (a) they reside deeper in the* WordNet *hierarchy, or (b) they are more densely connected locally.*

This is the working assumption for virtually all WordNet-based semantic similarity studies using depth and/or density. For depth, the intuition is that adjacent nodes deep down the hierarchy are likely to be conceptually close, since the differentiation is based on finer details (Jiang and Conrath, 1997). Sussna (1993) termed the use of depth as *depth-relative scaling*, claiming that "only-siblings deep in a tree are more closely related than only-siblings higher in the tree". Richardson and Smeaton (1995) gave an hypothetical example illustrating this "only-siblings" situation, where *plant–animal* are the only two nodes under *living things*, and *wolfhound–foxhound* under *hound*. They claimed the reason that the former pair can be regarded as conceptually farther apart compared to the latter is related to the difference in depth.

As for the relation between density and similarity, the intuition is that if the overall semantic mass for a given node is constant (Jiang and Conrath, 1997), then the more neighboring nodes there are in a locally connected subnetwork, the closer its members are to each other. For example, *animal*, *person*, and *plant* are more strongly connected with *life form* than *aerobe* and *plankton* because the first three words all have high density in their local network structures (Richardson and Smeaton, 1995). Note that the notion of density here is not to be confused with the *conceptual density* used by Agirre and Rigau (1996), which is essentially a semantic similarity measure by itself.

In general, both observations on depth and density conform to intuition and are supported qualitatively by several existing studies. The main objective of this study is to empirically examine the validity of this assumption.

2.2 Semantic Similarity Measures Using Depth and/or Density

One of the first examples of using depth and density in WordNet-based similarity measures is that of Sussna (1993). The weight on an edge between two nodes c_1 and c_2 with relation r in WordNet is given as:

$$w(c_1, c_2) = \frac{w(c_1 \rightarrow_r c_2) + w(c_2 \rightarrow_r c_1)}{2d}$$

where d is the depth of the deeper of the two nodes. As depth increases, weight decreases and similarity in turn increases, conforming to Assumption 1. The edge weight was further defined as

$$w(c_1 \rightarrow_r c_2) = max_r - \frac{max_r - min_r}{n_r(c_1)}$$

where $n_r(X)$ is "the number of relations of type r leaving node X", which is essentially an implicit form of density, and max_r and min_r are the maximum and minimum of n_r, respectively. Note that this formulation of density contradicts Assumption

[1] Since the works we review in this section have different definitions of depth and density, we defer our formal definitions to Section 3.

1 since it is proportional to edge weight (left-hand-side) and thus negatively correlated to similarity.

Wu and Palmer (1994) proposed a concept similarity measure between two concepts c_1 and c_2 as:

$$sim(c_1,c_2) = \frac{2 \cdot dep(c)}{len(c_1,c) + len(c_2,c) + 2 \cdot dep(c)} \quad (1)$$

where c is the lowest common subsumer (LCS) of c_1 and c_2, and $len(\cdot,\cdot)$ is the number of edges between two nodes. The rationale is to adjust "hop count" (the first two terms in the denominator) with the depth of LCS: similarity between nodes with same-level LCS is in negative proportion to hop counts, while given the same hop count, a "deeper" LCS pulls the similarity score closer to 1.

Jiang and Conrath (1997) proposed a hybrid method incorporating depth and density information into an information-content-based model (Resnik, 1999):

$$\begin{aligned} w(c,p) = & \left[\frac{dep(p)+1}{dep(p)}\right]^{\alpha} \\ & \times \left[\beta + (1-\beta)\frac{\bar{E}}{den(p)}\right] \\ & \times \left[IC(c) - IC(p)\right] T(c,p) \end{aligned} \quad (2)$$

Here, p and c are parent and child nodes in WordNet, $dep(\cdot)$ and $den(\cdot)$ denote the depth and density of a node, respectively, $\bar{E}$ is the average density over the entire network of WordNet, and α and β are two parameters controlling the contribution of depth and density values to the similarity score. $IC(\cdot)$ is the information content of a node based on probability estimates of word classes from a small sense-tagged corpus (Resnik, 1999), and $T(c,p)$ is a link-type factor differentiating different types of relations between c and p.

3 Limitations on the Current Definitions of Depth and Density

To what extent do the notions of depth and density help towards an accurate semantic similarity measure? Our empirical investigation below suggests that more often than not, they fail our intuition.

A direct assessment of the effectiveness of using depth and density is to examine their correlation with similarity. Empirical results in this section

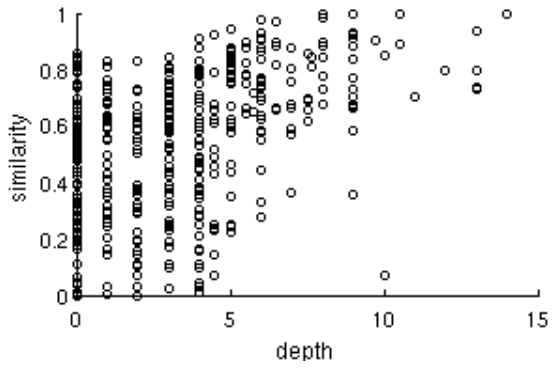

Figure 1: Correlation between depth and similarity.

are achieved by the following experimental setting. Depth is defined as the number of edges between the root of the hierarchy and the lowest common subsumer (LCS) of two nodes under comparison, and density as the number of siblings of the LCS.[2] Similarity is measured by human judgment on similarity between word pairs. Commonly used data sets for such judgments include that of Rubenstein and Goodenough (1965), Miller and Charles (1991), and Finkelstein et al. (2001) (denoted *RG*, *MC*, and *FG*, respectively). *RG* is a collection of similarity ratings of 65 word pairs averaged over judgments from 51 human subjects on a scale of 0 to 4 (from least to most similar). *MC* is a subset of 30 pairs out of the *RG* data set. These pairs were chosen to have evenly distributed similarity ratings in the original data set, and similarity judgment was elicited from 38 human judges with the same instruction as used for *RG*. *FG* is a much larger set consisting of 353 word pairs, and the rating scale is from 0 to 10. We combine the *RG* and *FG* data sets in order to maximize data size. Human ratings r on individual sets are normalized to r_n on 0 to 1 scale by the following formula:

$$r_n = \frac{r - r_{\min}}{r_{\max} - r_{\min}}$$

where $r_{\max}$ and $r_{\min}$ are the maximum and minimum of the original ratings, respectively. Correlation is evaluated using *Spearman's* ρ.

[2] We also tried several other variants of these definitions, e.g., using the maximum or minimum depth of the two nodes instead of the LCS. With respect to statistical significance tests, these variants all gave the same results as our primary definition.

3.1 Depth

The distribution of similarity of the combined data set over depth is plotted in Figure 1. For depth values under 5, similarity scores are fairly evenly distributed over depth, showing no statistical significance in correlation. For depth 5 and above, the shape of distribution resembles an upper-triangle, suggesting that (1) correlation with similarity becomes stronger in this range of depth value, and (2) data points with higher depth values tend to have higher similarity scores, but the reverse of the claim does not hold, i.e., word pairs with "shallower" LCS can also be judged quite similar by humans.

There are many more data points with lower depth values than with higher depth values in the combined data set. In order to have a fair comparison of statistical significance tests on the two value ranges for depth, we randomly sample an equal number (100) of data points from each value range, and the correlation coefficient between depth and similarity is averaged over 100 of such samplings. Correlation coefficients for depth value under 5 versus 5 and above are $\rho = 0.0881, p \approx 0.1$ and $\rho = 0.3779, p < 0.0001$, respectively, showing an apparent difference in degree of correlation.

Two interesting observations can be made from these results. Firstly, the notion of depth is relative to the distribution of number of nodes over depth value. For example, depth 20 by itself is virtually meaningless since it might be quite high if the majority of nodes in WordNet are of depth 10 or less, or quite low if the majority depth value are 50 or more. According to the histogram of depth values in WordNet (Figure 2), the distribution of number of nodes over depth value approximately conforms to a normal distribution $\mathcal{N}(8,2)$. It is visually quite noticeable that the actual quantity denoting how deep a node resides in WordNet is conflated at depth values below 5 or above 14. In other words, the distribution makes it rather inaccurate to say, for instance, that a node of depth 4 is twice as deep as a node of depth 2. This might explain the low degree of correlation between similarity and depth under 5 in Figure 1 (manifested by the long, vertical stripes across the entire range of similarity scores (0 to 1) for depth 4 and under), and also how the correlation increases with depth value. Unfortunately, we do not have enough

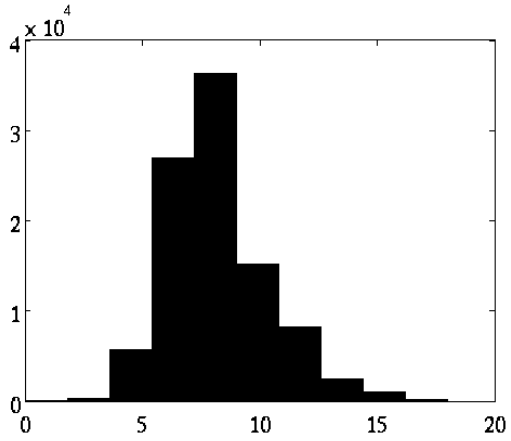

Figure 2: Histogram of depth of WordNet noun synsets.

data for depth above 14 to draw any conclusion on this higher end of the depth spectrum.

Secondly, even on the range of depth values with higher correlation with similarity, there is no definitive sufficient and necessary relation between depth and similarity (hence the upper triangle instead of a sloped line or band). Particularly, semantically more similar words are not necessarily deeper in the WordNet hierarchy. Data analysis reveals that the LCS of highly similar words can be quite close to the hierarchical root. Examples include *coast–shore*, which is judged to be very similar by humans (9 on a scale of 0–10 in both data sets). The latter is a hypernym of the former and thus the LCS of the pair, yet it is only four levels below the root node *entity* (via *geological formation*, *object*, and *physical entity*). Another situation is when the human judges confused relatedness with similarity, and WordNet fails to capture the relatedness with its hierarchical structure of lexical semantics: the pair *software–computer* can only be related by the root node *entity* as their LCS, although the pair is judged quite "similar" by humans (8.5 on 0 to 10 scale).

The only conclusive claim that can be made here is that word pairs with deeper LCS's tend to be more similar. However, since only word forms (rather than senses) are available in these psycho-linguistic experiments, the one similarity rating given by human judges sometimes fails to cover multiple senses for polysemous words. In the pair *stock–jaguar* of the *FG* set, for example, one sense of *stock* (*livestock, stock, farm animal: any animals kept for use*

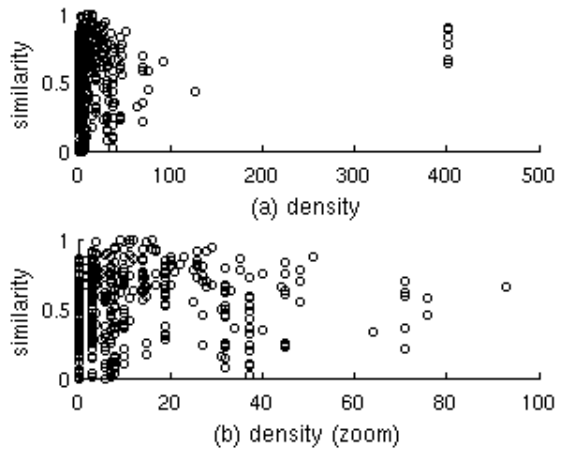

Figure 3: Correlation between density and similarity.

	MC	*RG*	*FG*
dep	0.7056***	0.6909***	0.3701***
den	0.2268	0.2660*	0.1023

Table 1: Correlation between depth/density and similarity on individual data sets. Number of asterisks indicates different confidence intervals ("*" for $p < 0.05$, "***" for $p < 0.0001$).

or profit) is closely connected to *jaguar* through a depth-10 LCS (*placental, placental mammal, eutherian, eutherian mammal*). However, the pair received a low similarity rating (0.92 on 0–10), probably because judges associated the word form *stock* with its financial sense, especially when there was an abundant presence of pairs indicating this particular sense of the word (e.g., *stock–market*, *company–stock*).

3.2 Density

Comparing to depth, density exhibits much lower correlation with similarity (Figure 3-a and 3-b). We conducted correlation experiments between density and similarity with the same setting as for depth and similarity above. Data points with extremely high density values (up to over 400) are mostly idiosyncratic to the densely connected regions in WordNet and are numerically quite harmful. We thus excluded outliers with density values above 100 in the experiment.

Evaluation on the combined data set shows no correlation between density and similarity. To confirm the result, we break the experiments down to the three individual data sets, and the results are listed in Table 1. The correlation coefficient between density and similarity ranges from 0.10 to 0.27 There is no statistical significance of correlation on two of the three data sets (*MC* and *FG*), and the significance on *RG* is close to marginal with $p = 0.0366$.

Data analysis suggests that density values are often biased by particular fine-grainedness of local structures in WordNet. Qualitatively, Richardson and Smeaton (1995) previously observed that "the irregular densities of links between concepts results in unexpected conceptual distance measures". Empirically, on the one hand, more than 90% of WordNet nodes have density values less than or equal to 3. This means that for 90% of the LCS's, there are only three integer values for density to distinguish the varying degrees of similarity. In other words, such a range might be too narrow to have any real distinguishing power over similarity. On the other hand, there are outliers with extreme density values particular to the perhaps overly fine-grained subcategorization of some WordNet concepts, and these nodes can be LCS's of word pairs of drastically different similarity. The node *person*, *individual*, for example, can be the LCS of similar pairs such as *man–woman*, as well as quite dissimilar ones such as *boy–sage*, where the large density value does not necessarily indicate high degree of similarity.

Another crucial limitation of the definition of density is the information loss on specificity. In the existing literature, density is often adopted as a proxy for the degree of specificity of a concept, i.e., nodes in densely connected regions in WordNet are taken to be more specific and thus closer to each other. This information of a given node should be inherited by its hierarchical descendants, since specificity should monotonically increase as one descends the hierarchy. For example, the node *piano* has a density value of 15 under the node *percussion instrument*. However, the density value of its hyponyms *Grand piano*, *upright piano*, and *mechanical piano*, is only 3. Due to the particular structure of this subnetwork in WordNet, the *grand–upright* pair might be incorrectly regarded as less specific (and thus less similar) than, say, between *piano–gong*, both as percussion instruments.

	MC	*RG*	*FG*
dep_u	0.7201***	0.6798***	0.3751***
den_u	0.2268	0.2660*	0.1019
den_i	0.7338***	0.6751***	0.3445***

Table 2: Correlation between new definitions of depth/density and similarity.

4 New Definitions of Depth and Density

In this section, we formalize new definitions of depth and density to correct for their current limitations discussed in Section 3.

4.1 Depth

The major problem with the current definition of depth is its failure to take into account the uneven distribution of number of nodes over the depth value. As seen in previous examples, the distribution is rather "flat" on both ends of depth value, which does not preserve the linearity of using the ordinal values of depth and thus introduces much inaccuracy.

To avoid this problem, we "re-curve" depth value to the cumulative distribution. Specifically, if we take the histogram distribution of depth value in Figure 2 as a probability density function, our approach is to project cardinal depth values onto its cumulative distribution function. The new depth is denoted dep_u and is defined as:

$$dep_u(c) = \frac{\sum_{c' \in WN} |\{c' : dep(c') \leq dep(c)\}|}{|WN|}$$

Here, $dep(\cdot)$ is the original depth value, and WN is the set of all nodes in WordNet. The resulting depth values not only reflect the flat ends, but also preserve linearity for the depth value range in the middle. In comparison with Table 1), correlation between dep_u and similarity increases over the original depth values on two of the three data sets (first row in Table 2 and decreases on the *RG* set. Later, in Section 5, we show how these marginal improvements translate into better similarity measures with statistical significance.

4.2 Density

In theory, a procedure analogous to the above cumulative definition can also be applied to density, i.e., by projecting the original values onto the cumulative distribution function. However, due to the Zipfian nature of density's histogram distribution (Figure 4, in contrast to Gaussian for depth in Figure 2), this is essentially to collapse most density values into a very small number of discrete values (which correspond to the original density of 1 to 3). Experiments show that it does not help in improving correlation with similarity scores (second row in Table 2 for den_u): correlation remains the same on *MC* and *RG*, and decreases slightly on *FG*.

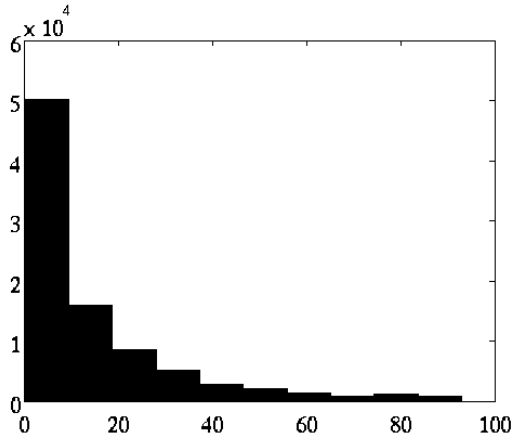

Figure 4: Histogram of density in WordNet.

We therefore resort to addressing the issue of information loss on specificity by inheritance. Intuitively, the idea is to ensure that a node be assigned no less density mass than its parent node(s). In the "piano" example (Section 3.2), the concept *piano* is highly specific due to its large number of siblings under the parent node *percussion instruments*. Consequently, the density of its child nodes *upright piano* and *grand piano* should inherit its specificity on top of their own.

Formally, we redefine density recursively as follows:

$$den_i(r) = 0$$

$$den_i(c) = \frac{\sum_{h \in hyper(c)} den_i(h)}{|hyper(c)|} + den(c)$$

where r is the root of WordNet hierarchy (with no hypernym), and $hyper(\cdot)$ is the set of hypernyms of a given concept. The first term is the inheritance part, normalized over all hypernyms of c in case of multiple inheritance, and the second term is the original value of density.

The resulting density values correlate significantly better with similarity. As shown in row 3 in Table 2, the correlation coefficients are about tripled on all three data sets with the new density definition den_i, and the significance of correlation is greatly improved as well (from non-correlating or marginally correlating to strongly significantly correlating on all three data sets).

5 Using the New Definitions in Semantic Similarity Measures

In this section, we test the effectiveness of the new definitions of depth and density by using them in WordNet-based semantic similarity measures. The two similarity measures we experiment with are that of Wu and Palmer (1994) and Jiang and Conrath (1997). The first one used depth only, and the second one used both depth and density.

The task is to correlate the similarity measures with human judgment on similarity between word pairs. We use the same three data sets as in Section 3. despite the fact that *MC* is a subset of *RG* data set, we include both in order to compare with existing studies.

Correlation coefficient is calculated using *Spearman's* ρ, although results reported by some earlier studies used parametric tests such as the *Pearson Correlation Coefficient*. The reason for our choice is that the similarity scores of the word pairs in these data sets do not necessarily conform to normal distributions. Rather, we are interested in testing whether the artificial algorithms would give higher scores to pairs that are regarded closer in meaning by human judges. A non-parametric test suits better for this scenario. And this partly explains why our re-implementations of the models have lower correlation coefficients than in the original studies.

Note that there are other WordNet-based similarity measures using depth and/or density that we opt to omit for various reasons. Some of them were not designed for the particular task at hand (e.g., that of Sussna, 1993, which gives very poor correlation in this task). Others use depth of the entire WordNet hierarchy instead of individual nodes as a scaling factor (e.g., that of Leacock and Chodorow, 1998), which is unsuitable for illustrating the improvement brought about by the new depth and density definitions.

	Best			Average		
	MC	*RG*	*FG*	*MC*	*RG*	*FG*
dep	0.7671	0.7824	0.3773	0.7612	0.7686	0.3660
dep_u	0.7824	0.7912	0.3946	0.7798	0.7810	0.3787

Table 3: Correlation between human judgment and similarity score by Wu and Palmer (1994) using two versions of depth.

	Best			Average		
	MC	*RG*	*FG*	*MC*	*RG*	*FG*
dep, *den*	0.7875	0.8111	0.3720	0.7689	0.7990	0.3583
dep_u, den	0.8009	0.8181	0.3804	0.7885	0.8032	0.3669
dep, den_i	0.7882	0.8199	0.3803	0.7863	0.8102	0.3689
dep_u, den_i	0.8065	0.8202	0.3818	0.8189	0.8194	0.3715

Table 4: Correlation between human judgment and similarity score by Jiang and Conrath (1997) using different definitions of depth and density.

Parameterization of the weighting of depth and density is a common practice to control their individual contribution to the final similarity score (e.g., α and β in Equation (2)). Jiang and Conrath already had separate weights in their original study. In order to parameterize depth used by Wu and Palmer in their similarity measure, we also modify Equation (1) as follows:

$$sim(c_1,c_2) = \frac{2 \cdot dep^{\alpha}(c)}{len(c_1,c) + len(c_2,c) + 2 \cdot dep^{\alpha}(c)}$$

where depth is raised to the power of α to vary its contribution to the similarity score.

For a number of combinations of the weighting parameters, we report both the best performance and the averaged performance over all the parameter combinations. The latter number is meaningful in that it is a good indication of numerical stability of the parameterization. In addition, parameterization is able to generate multiple correlation coefficients, on which statistical tests can be run in order to show the significance of improvement. We use the range from 0 to 5 with step 1 for α and from 0 to 1 with step 0.1 for β.

Table 3 and 4 list the experiment results. In both models, the cumulative definition of depth dep_u consistently improve the performance of the similarity measures. In the Jiang and Conrath (1997) model, where density is applicable, the inheritance-based

definition of density den_i also results in better correlation with human judgments. The optimal result is achieved when combining the new definitions of depth and density (row 4 in Table 4). For average performance, the improvement of all the new definitions over the original definitions is statistically significant on all three data sets according to paired *t-test*.

6 Conclusions

This study explored effective uses of depth and/or density in WordNet-based similarity measures. We started by examining how well these two structural features correlate with human judgment on word pair similarities. This direct comparison showed that depth correlates with similarity only on certain value ranges, while density does not correlate with human judgment at all.

Further investigation revealed that the problem for depth lies in the simplistic representation as its ordinal integer values. The linearity in this representation fails to take into account the conflated quantity of depth in the two extreme ends of the depth spectrum. For density, a prominent issue is the information loss on specificity of WordNet concepts, which gives an inaccurate density value that is biased by the idiosyncratic constructions in densely connected regions in the hierarchy.

We then proposed new definitions of depth and density to address these issues. For depth, linearity in different value ranges is realistically reflected by projecting the depth value to its cumulative distribution function. The loss of specificity information in density, on the other hand, is corrected by allowing concepts to inherit specificity information from their parent nodes. The new definitions show significant improvement in correlation of semantic similarity given by human judges. In addition, when used in existing WordNet-based similarity measures, they consistently improve performance and numerical stability of the parameterization of the two features.

The notions of depth and density pertain to any hierarchical structure like WordNet, which suggests various extensions of this work. A natural next step of the current work is to apply the same idea to semantic taxonomies in languages other than English with available similarity judgments are also available. Extrinsic tasks using WordNet-based semantic similarity can potentially benefit from these refined notions of depth and density as well.

Acknowledgments

This study was inspired by lectures given by Professor Gerald Penn of the University of Toronto, and was financially supported by the Natural Sciences and Engineering Research Council of Canada.

References

Eneko Agirre and German Rigau. Word sense disambiguation using conceptual density. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 16–22. Association for Computational Linguistics, 1996.

Martin Chodorow, Roy Byrd, and George Heidorn. Extracting semantic hierarchies from a large online dictionary. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pages 299–304, Chicago, Illinois, USA, 1985.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, pages 406–414. ACM, 2001.

Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332. 1998.

Mario Jarmasz and Stan Szpakowicz. Roget's thesaurus and semantic similarity. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, pages 212–219, Borovets, Bulgaria, 2003.

Jay Jiang and David Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of International Conference on Research in Computational Linguistics*, 33, 1997.

Hideki Kozima and Akira Ito. Context-sensitive measurement of word distance by adaptive scaling of a semantic space. *Recent Advances in Natural Language Processing: Selected Papers from RANLP*, 95:111–124, 1997.

Claudia Leacock and Martin Chodorow. Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998.

Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 768–774, Montreal, Canada, 1998.

Goerge Miller and Walter Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet::Similarity: measuring the relatedness of concepts. In *Demonstration Papers at Human Language Technologies - North American Chapter of the Association for Computational Linguistics*, pages 38–41. Association for Computational Linguistics, 2004.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30, 1989.

Philip Resnik. 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(11):95–130, 1999.

R. Richardson and A.F. Smeaton. Using WordNet in a knowledge-based approach to information retrieval. In *Proceedings of the BCS-IRSG Colloquium, Crewe*. Citeseer, 1995.

Herbert Rubenstein and John Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

Michael Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the second international conference on Information and knowledge management*, pages 67–74. ACM, 1993.

Peter Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the Twelfth European Conference on Machine Learning*, pages 491–502, 2001.

Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.