Semantic Distance in WordNet: A Simplified and Improved Measure of Semantic Relatedness

by

Aaron D. Scriver

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2006

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Measures of semantic distance have received a great deal of attention recently in the field of computational lexical semantics. Although techniques for approximating the semantic distance of two concepts have existed for several decades, the introduction of the WordNet lexical database and improvements in corpus analysis have enabled significant improvements in semantic distance measures.

In this study we investigate a special kind of semantic distance, called *semantic relatedness*. Lexical semantic relatedness measures have proved to be useful for a number of applications, such as word sense disambiguation and real-word spelling error correction. Most relatedness measures rely on the observation that the shortest path between nodes in a semantic network provides a representation of the relationship between two concepts. The strength of relatedness is computed in terms of this path.

This dissertation makes several significant contributions to the study of semantic relatedness. We describe a new measure that calculates semantic relatedness as a function of the shortest path in a semantic network. The proposed measure achieves better results than other standard measures and yet is much simpler than previous models. The proposed measure is shown to achieve a correlation of $r = 0.897$ with the judgments of human test subjects using a standard benchmark data set, representing the best performance reported in the literature. We also provide a general formal description for a class of semantic distance measures — namely, those measures that compute semantic distance from the shortest path in a semantic network. Lastly, we suggest a new methodology for developing path-based semantic distance measures that would limit the possibility of unnecessary complexity in future measures.

Acknowledgments

The completion of this thesis was possible only with the assistance of several very supportive and kind individuals. First I must thank my supervisor, Dr. Chrysanne DiMarco. Dr. DiMarco's help and friendship throughout my time at the University of Waterloo has been tremendously valuable and is greatly appreciated.

My wife, Charlotte, has been my principal source of inspiration and support. I could not have completed my work without her. My parents, Pat and Ron, offered their encouragement and love throughout my studies, and I have much to thank them for.

I would also like to thank Dr. Randy Harris for some very helpful discussions relating to this work and Dr. Robin Cohen for many helpful comments on a later draft of this thesis.

Dedication

I dedicate this thesis to my father, Ron W. Scriver.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

The associative nature of memory has excited the interest of researchers since at least the time of Aristotle. In his book, *On Memory and Reminiscence* [1], Aristotle describes recollection as a partially involuntary procedure that involves following the mental connections between memories. When a person recalls a memory, "they pass swiftly in thought from one point to another, e.g. from milk to white, from white to mist, and thence to moist, from which one remembers Autumn [the 'season of mists'], if this be the season he is trying to remember." (p. 614) The basic idea of associative memory has changed surprisingly little since Aristotle's time, but the techniques of modern experimental science, and more recently of computational modelling, have deepened our understanding of this model of memory. For example, computational linguists have had considerable success with models of lexical memory that represent information about concepts in terms of their semantic relations to other concepts. These models are instances of *semantic networks* and represent an area of very active research [12].

This study explores an application of semantic networks that has remarkable similarity to Aristotle's notion of recollection. *Semantic relatedness* describes the strength of the cognitive association between two concepts. For example, *man* and *woman* are very strongly related, as are *monkey* and *banana*. The concepts *screwdriver* and *truth*, however, seem to be unrelated. Other pairs of concepts often fall somewhere in between these extremes, such as *book* and *computer* or *skyrise* and *window*. A very straightforward technique for determining the strength of relatedness between two concepts is to find the sequence of links that connects them in a semantic network. The 'closer' the concepts are to one another, i.e., the shorter the path that connects them, the more strongly they are related.

Early work in semantic networks proposed techniques quite similar to the short-

est path length approach described above. For example, Collins and Loftus [6] described a technique for determining semantic relatedness using the paths between nodes in a semantic network. However, with the availability of WordNet [12] — a large-scale semantic network for English — a great variety of techniques for measuring semantic relatedness, and for the associated problem of measuring semantic similarity, has emerged. These new measures have provided many refinements to the approach of computing the strength of relatedness from a path in a semantic network. The goal of this study is to resolve the confusion over exactly which of the many new techniques are effective, and which are not. Although our principal interest is in semantic relatedness measures, many of our observations and analyses apply equally to semantic similarity. The term *semantic distance* will be used as a general concept that encompasses both relatedness and similarity.

This dissertation makes several significant contributions to the study of semantic relatedness. We describe a new measure of semantic relatedness that matches human relatedness judgments more closely than any previous measure and yet is much simpler than other models. We provide a general formal description for a class of semantic distance measures — namely those measures that compute semantic distance from the shortest path in a semantic network. We also suggest a new methodology for developing path-based semantic distance measures that would limit the possibility of unnecessary complexity in future measures.

1.1 Motivation

Measures of semantic relatedness are interesting both in terms of the applications that they enable, and in terms of their theoretical implications for the study of lexical memory. The most common application of semantic distance measures is word sense disambiguation. Algorithms for word sense disambiguation have been proposed by Lesk [19], Sussna [44], Resnik [35], Patwardhan et al. [30], and McCarthy et al. [22], among others. We will describe the technique of Patwardhan et al. [30] in the next chapter in some detail, as a typical example of word sense disambiguation using semantic distance measures.

We will also describe an ingenious application of semantic relatedness to the problem of real-word spelling error correction. Hirst and St-Onge [15] proposed an algorithm that uses semantic relatedness measures to detect and correct misspellings of words that result in the correct spelling of another, though unintended, word.

There are many other applications of semantic distance measures. To name a

few recent applications, semantic distance measures have been used by: Kohomban and Lee [17] for learning coarse-grained concepts from a sense-tagged corpus, Corley and Mihalcea [7] for computing the similarity of texts, and Stevenson and Greenwood [41] for the acquisition of patterns for information extraction. Although we will not propose any novel applications in this study, we demonstrate significant improvements to semantic distance measures and these improvements will benefit all of the applications listed above.

Along with the practical value of improving semantic relatedness measures, the research presented in this study has some interesting psycholinguistic implications. The class of semantic distance measures that we are exploring use the WordNet [12] semantic network. WordNet was inspired by psycholinguistic theory, and is based on semantic network models of human lexical memory. In this study we compare the performance of semantic distance measures to the relatedness judgments of human subjects. This sort of comparison offers evidence for the model of human lexical memory that underlies WordNet. That is, if it is possible to closely mimic the behaviour of humans using the network model then this may be taken as evidence, albeit indirect, in favour of this model over other models of lexical memory.

Aside from simply validating the network model of lexical memory, successful semantic distance measures may provide insight into particular details of the model and how it is used by humans. For example, it will be shown in this study that variations in the 'weight' of links in the semantic network do not affect semantic distance measurements as strongly as had been believed. This has implications for the network model, as some have assumed that link weights have an important role in lexical memory [6].

1.2 Methodology

Perhaps the most significant contribution of this study is a new methodology for the development and evaluation of semantic distance measures. There are currently a large number of measures in the literature, many of which represent minor variations of one another. In particular, many measures compute semantic distance on the basis of the shortest path connecting concepts in a semantic network. Such measures often differ in relatively small ways and in some cases simply recombine the techniques of other measures. However, progress in improving path-based semantic distance measures has been slowed by the difficulty of identifying exactly which techniques are effective and which are not.

Over the last several years a *de facto* standard evaluation framework has emerged

for semantic distance measures. This framework relies on the comparison of the results of semantic distance measures to the relatedness judgments of humans that were collected in experiments by Miller and Charles [25], and by Rubenstein and Goodenough [38]. Studies that have adopted this evaluation framework include Resnik [34], Yang and Powers [47], Jiang and Conrath [16], and Budanitsky and Hirst [15].

Given a standard evaluation framework, it is possible to take a systematic approach to the development of new measures. We propose that authors of new semantic distance measures engage in regression-testing against simpler baseline measures. That is, any additions to measures that use the path-based approach should be compared with previous baseline measures and proven to improve their performance.

In the current study, we will lay the foundation for this new methodology by showing how semantic distance measures can be decomposed so that their constituent parts may be tested individually. We then test a number of the most important semantic distance measures against our simplified baseline measure and show that many of the features of these measures weaken their performance.

1.3 Overview

This dissertation consists of five chapters. Chapter 1 introduces the motivation and methodology for our study. Chapter 2 reviews relevant research, including the state-of-the-art of WordNet-based semantic distance measures as well as other essential background material. In Chapter 3, we analyze the semantic distance measures described in Chapter 2 and present the details of a new measure of semantic relatedness. In Chapter 4, we conduct an experimental evaluation of the new measure, and compare its performance to other measures that are prominent in the literature. Finally, Chapter 5 summarizes the results of this dissertation and describes areas for future study. A more detailed outline of each chapter follows below.

Chapter 2 surveys the current research relating to semantic distance measures. We begin by making an important distinction between two kinds of semantic distance: *semantic similarity* and *semantic relatedness*. While semantic similarity may be viewed as the degree to which concepts share common features, semantic relatedness can be any kind of semantic association. Although some of our findings pertain to both similarity and relatedness, relatedness is the primary object of our study. Also, as we are considering primarily WordNet-based techniques, a general

overview of the WordNet lexical database is provided in this chapter. The most important relatedness and similarity measures are described, including the relatedness measures of Sussna [44], Hirst and St-Onge [15], and Banerjee and Pedersen [2], and the similarity measures of Leacock and Chodorow [18], Resnik [34], Jiang and Conrath [16], Lin [20], and Yang and Powers [47].

Chapter 3 includes an analysis of the path-based similarity and relatedness measures that were described in Chapter 2. On the basis of these analyses, we propose a general description for these measures, and show how each measure is a special case of the generalization. A new semantic relatedness measure is presented that follows from some simplifying assumptions.

The simplified measure provides an opportunity to systematically examine two facets of semantic relatedness. First, we look at the mathematical relationship between the length of the shortest path between concepts and the strength of their relatedness. We propose five candidate functions for mapping path length to relatedness. These functions are evaluated against empirical evidence of human performance in Chapter 4. Also, we look at the effect of allowing different subsets of semantic relations in the paths between concepts when determining relatedness.

In Chapter 4 we describe a two-part experiment that compares the new semantic relatedness measure to previous relatedness and similarity measures. The measures are evaluated on the basis of correlation with human judgments of relatedness using two widely used data sets. In the first part of the experiment we examine the effect of using different sets of allowable semantic relationship types on the correlation of path length with human judgments. We find that relations other than IS-A have little effect on the results, as IS-A relations are by far the most common links between nouns in WordNet.

In the second part of the experiment we compare the five functions for mapping path length to relatedness that were outlined in Chapter 3. The functions are 'tuned' with a subset of the experimental data using statistical curve-fitting techniques. The remaining data is used to evaluate the functions and to compare their performance to that of other relatedness and similarity measures. The new measure reaches a correlation of $r = 0.897$ with the human ratings collected by Rubenstein and Goodenough [38]. Only a measure by Yang and Powers [47] has comparable performance, but their measure is much more complex than ours.

At the end of Chapter 4 some tentative results are described for another type of semantic distance that we call *semantic contrast*. Semantic contrast denotes the type of conceptual distance that separates semantic opposites, such as *love* and *hate*. This sense of semantic distance is not captured by either similarity or relatedness

and we believe that a computational model of semantic contrast would lead to some very interesting applications. Finally, Chapter 5 summarizes the results of our study and describes areas for future research.

Chapter 2

Survey of Lexical Semantic Distance Measures

2.1 Semantic Distance

The notion of *semantic distance* — sometimes called *conceptual distance* — has received a great deal of attention in the field of lexical semantics in recent years. In general, semantic distance denotes the degree of semantic association between concepts. However, many authors, including Resnik [34] and Budanitsky and Hirst [4] distinguish two kinds of semantic distance: *semantic similarity* and *semantic relatedness*. Whereas similarity expresses the degree to which two concepts resemble one another, relatedness encompasses a wide variety of semantic relationships.

Although semantic similarity and semantic relatedness have received the most study, these senses do not exhaust the range of possible types of semantic distance. For example, Budanitsky and Hirst [4] argue that distributional similarity describes a phenomenon that is distinct from both semantic similarity and semantic relatedness. Later in this study, we will introduce another sense of semantic distance that we are calling *semantic contrast* which differs from both similarity and relatedness in important ways.

Although the current study is concerned primarily with semantic relatedness, it has been argued that in many cases semantic similarity is an adequate proxy for relatedness. In fact, in a recent study by Budanitsky and Hirst [4] that evaluated the performance of a number of similarity and relatedness measures for relatedness tasks, the authors found that similarity measures achieved better results than the relatedness measures. In this chapter, we will therefore review both relatedness

and similarity measures, including all of the measures compared by Budanitsky and Hirst. Two other promising measures that were not included in Budanitsky and Hirst's study will also be described, including one by Yang and Powers [47] and another by Banerjee and Pedersen [2].

Semantic distance measures have been developed using a variety of lexical resources. However, the scope of this study will be limited to measures that employ the WordNet lexical database. There are two reasons for restricting the study to only WordNet-based measures. First, as all of the measures to be compared share a common primary resource, the validity of comparisons between the measures will not be compromised by the quality of the lexical resources that they use.

Second, most of the major approaches to measuring either similarity or relatedness are represented by WordNet-based measures. The notable exception to this are measures that employ corpus statistics to determine *distributional similarity*. Such measures rely on the observation that words that occur in similar contexts are likely to be semantically similar. Mohammad and Hirst [26] provide a theoretical comparison between corpus-based measures of distributional similarity and taxonomy-based relatedness and similarity measures, and conclude that an experimental comparison is also required.

However Mohammad and Hirst also conclude that to a certain extent the two types of measure are incommensurable. While taxonomy-based approaches measure the similarity of concepts, corpus-based approaches measure the similarity of words. Mohammad and Hirst suggest that it may be more reasonable to view distributional similarity as a phenomenon distinct from conceptual similarity. As a result of these concerns, corpus-based measures of distributional similarity are excluded from the scope of this study.

The scope of this study is also limited to measures of the semantic distance between lexicalized concepts, which is to say, concepts that are expressed by individual words in the English language. Insofar as the primary use of semantic distance measures lies in natural language processing, lexicalized concepts deserve the most attention from a practical point of view. For the rest of this study, any reference to 'concepts' may be assumed to refer specifically to 'lexicalized concepts'. To avoid redundancy, the terms 'lexical' and 'semantic' will often be dropped so that, for example, 'lexical semantic relatedness' will be simply 'relatedness'.

In order to demonstrate the utility of semantic relatedness measures, two applications will be described in some detail at the end of this chapter. The first is a technique described by Hirst and St-Onge [15] for the detection and correction of malapropisms using lexical chains. The second demonstrates the application of

semantic relatedness and similarity measures to the problem of word sense disambiguation.

2.1.1 Similarity and Relatedness

Although the difference between lexical semantic similarity and lexical semantic relatedness can sometimes be subtle, it is nevertheless significant. Similarity can be understood to denote a kind of familiar resemblance. It is sometimes described in terms of featural overlap [45]. Under this view, the similarity of two concepts is the degree to which they share features in common. Features that are common to two concepts indicate greater similarity, and features that are peculiar to one or the other indicate reduced similarity. In this study we are not committed to a feature-based representation of concepts, but features provide a useful way of talking about similarity.

In contrast to similarity, relatedness describes the degree to which concepts are associated via any kind of semantic relationship. These relationships can include the classical lexical relations such as synonymy, hypernymy (IS-A), and meronomy (HAS-A), and also what Morris and Hirst [28] have called "non-classical relations". In fact, even the relation of similarity is encompassed by relatedness. As a result, all similar concepts are also related — by virtue of their similarity — such that similarity may be viewed as a special case of relatedness.

The difference between similarity and relatedness is often illustrated with examples. Resnik [34] provides the widely used example of *car* and *gasoline*. Cars and gasoline are not very similar; they have very few features in common. Whereas a car is a solid mechanical device, gasoline is a combustible liquid. An itemization of the properties of cars and gasoline would have little overlap. In spite of their differences, however, *car* and *gasoline* are very closely related through their functional association, namely that *cars* use *gasoline*. Thus, while in terms of similarity *car* and *gasoline* are semantically distant, in terms of relatedness they are semantically close.

2.2 Computational Resources

2.2.1 WordNet

All of the computational measures of semantic distance that will be discussed in this study employ the WordNet [12] lexical database. WordNet is a lexical reference

system that was created by a team of linguists and psycholinguists at Princeton University. The purpose of WordNet is to model the English lexicon according to psycholinguistic theories of human lexical memory. WordNet may be distinguished from traditional lexicons in that lexical information is organized according to word meanings, and not according to word forms. As a result of the shift of emphasis toward word meanings, the core unit in WordNet is something called a *synset*. Synsets are sets of words that have the same meaning, that is, synonyms. A synset represents one concept, to which different word forms refer. For example, the set {*car, auto, automobile, machine, motorcar*} is a synset in WordNet and forms one basic unit of the WordNet lexicon. Although there are subtle differences in the meanings of synonyms — often differences of connotation rather than of denotation — these are ignored in WordNet.

WordNet synsets are linked together by semantic relations to form a network. These relations include hypernymy (IS-A) and meronomy (HAS-A), among others. Some relations that hold between word forms have also been included in WordNet, such as derivational relatedness. WordNet synsets are divided into nouns, adjectives, verbs, and adverbs. Although there is some interconnectivity between the different speech categories, it is quite limited. The portions of WordNet for each part of speech also have different properties, and may therefore require special treatment. For example, while the hypernymy relation is central to the organization of the noun portion of WordNet, adjectives are organized primarily in terms of the antonymy and similarity relations. Table 2.1 provides a complete list of WordNet 2.0 relations and their frequency count by category.

Many of the similarity measures discussed in this study apply to nouns exclusively and rely closely on the special properties of the noun subgraph of WordNet. The primary organizing relations in the noun part of WordNet are hypernymy and hyponymy. A concept is a hyponym if it is a specific type of a more general class. For example, a *robin* is a kind of *bird* and is therefore a hyponym of *bird*. The inverse of a hyponym is a hypernym, which denotes a more general class with respect to a more specific one. Thus *bird* is a hypernym of *robin*. Part/whole relations, including meronomy and holonomy, also play an important role in the noun portion of WordNet. A concept is a meronym if it is part of a whole, whereas a concept is a holonym with respect to its constituent parts. However, nearly 80% of semantic relations between nouns are hypernymy or hyponymy [4]. The hierarchical nature of the IS-A relation results naturally in a tree-like structure. The developers of WordNet have paid careful attention to the coherence and completeness of the IS-A hierarchy of nouns.

Although earlier versions of WordNet contained several separate IS-A hierar-

Relations	Noun	Adjective	Verb	Adverb
Antonym	2074	4118	1079	722
Hypernym (IS-A)	81857		12985	
Hyponym (SUBSUMES)	81857		12985	
Member holonym (PART-OF)	12205			
Substance holonym	787			
Part holonym	8636			
Member meronym (HAS-A)	12205			
Substance meronym	787			
Part meronym	8636			
Attribute	648			
Derivation	21491		21497	3209
Category domain	3789	1125	1215	37
Category member	6166			
Region domain	1200	76	2	2
Region member	1280			
Usage domain	654	237	18	74
Usage member	983			
Entailment			409	
Cause			218	
Verb group			1748	
Similar to		22196		
Participle of verb		124		
Pertainym		4711		
Attribute		648		
Also see		2697	597	
Totals	249927	52753	35971	4044

Table 2.1: WordNet 2.0 relations and frequency count by type, reproduced from [39]

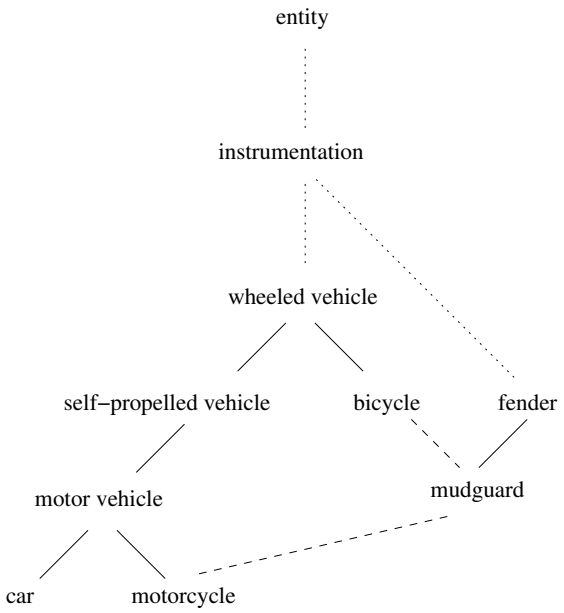


Figure 2.1: Fragment of WordNet graph for wheeled vehicles and related concepts. Solid lines represent IS-A/SUBSUMES relations, dashed lines represent HAS-A/PART-OF relations and dotted lines represent a series of omitted IS-A/SUBSUMES relations.

chies, the number of separate hierarchies was reduced in successive versions. The top node of each noun hierarchy is called a *unique beginner*. As of WordNet 2.1, the hierarchies have been merged into a single hierarchy headed by the unique beginner {*entity*}. The noun portion of WordNet may be treated as an ontology of lexicalized concepts. The similarity measures by Resnik [34], Jiang and Conrath [16], Leacock and Chodorow [18], and Lin [20] each exploit the ontology formed by the hierarchy of nouns. To illustrate the WordNet noun hierarchy, a small part of the network surrounding concepts relating to wheeled vehicles is reproduced in Figure 2.1. The small network in Figure 2.1 will be used in examples throughout this chapter.

There are a few notational conventions related to WordNet that will be adopted in this study. First, the taxonomy of nouns formed by hypernymy in WordNet will be referred to as a conceptual taxonomy, an IS-A hierarchy, or a subsumption hierarchy. However, when dealing with the entire WordNet network, including all types of semantic relations, we will instead refer to the WordNet graph, network, or semantic network. Also, whenever referring to WordNet synsets we will use italicized

text between curly braces, such as {*car, auto, automobile, machine, motorcar*}. Individual concepts will be denoted using italicized text, such as *automobile*, and word forms will be denoted with single quotations, such as 'car'. The informal names of semantic relations will be marked by small capitals, such as IS-A.

2.3 Relatedness Measures

There are two general approaches taken by the different relatedness measures that we will describe. The first approach relies on an examination of the shortest path in the WordNet graph that connects two synsets. This approach is represented by the measures of Sussna [44], and Hirst and St-Onge [15]. The second approach exploits the definitions provided for synsets in WordNet, called *glosses*, and is represented by the measure from Banerjee and Pedersen [2].

2.3.1 Sussna

Sussna [44] described one of the first WordNet-based relatedness measures. The measure was developed for the purpose of word sense disambiguation in an information retrieval system. Sussna's measure determines the strength of relatedness between two concepts by first finding the shortest path between their corresponding synsets in the WordNet graph. The edges (the semantic relations) in the path are assigned weights, with higher weight indicating greater semantic distance, and the sum of these weights gives the total semantic distance between the concepts.

For example, to compute the relatedness of the concepts *bicycle* and *motorcycle* using Figure 2.1, we would first find the shortest path between these nodes. In this case the path would be:

bicycle HAS-A *mudguard* PART-OF *motorcycle*

The semantic distance between *bicycle* and *motorcycle* would therefore be the sum of the distances between *bicycle* and *mudguard*, and between *mudguard* and *motorcycle*. The technique of using the sum of distances on the shortest path between concepts is repeated in many other similarity and relatedness measures, and we will refer to these types of measures as *path-based* measures.

A central problem for path-based measures is determining the distances represented by particular semantic relations in the semantic network. Sussna proposed

two schemes for estimating the semantic distances (the 'weights') of individual edges in WordNet. He observed that the more concepts a given concept is related to, the less strongly it is associated with each one. More specifically, the semantic distance of a relation is proportional to the number of other semantic relations of the same type emerging from a concept. Sussna calls this the *type-specific fanout* (TSF) factor. For example, the concept for *computer* in WordNet 2.0 has 14 meronym (HAS-A) relations, corresponding to 14 different parts of a computer, such as *keyboard*. The synset including *keyboard*, on the other hand, has only two meronym relations, one of which is *key*. Since *keyboard* has fewer parts than *computer*, *keyboard* will be more strongly associated with each of its parts. Sussna's measure would therefore assign a greater semantic distance value to the meronym link connecting *computer* and *keyboard* than to that connecting *keyboard* and *key*.

The second edge-weighting scheme in Sussna's measure is called *depth-relative scaling*, and is based on the observation that siblings deep in the taxonomy tend to be more closely related than those closer to the top. General, abstract concepts are assumed to represent broad distinctions, and therefore the differences between them cover greater semantic distance than do the finer distinctions found lower in the taxonomy.

To calculate the strength of relatedness between concepts in Sussna's measure, each relationship type is assigned a weight range between min_r and max_r, for each relationship type r. The semantic distance value for a relation of type r from the source node c_1 is:

$$wt(c_1 \longrightarrow_r) = max_r - \frac{max_r - min_r}{edges_r(c_1)} \tag{2.1}$$

where $edges_r(c_1)$ is the number of relations of type r originating from c_1. For the hypernymy, hyponymy, holonymy, and meronomy relationships the values min_r and max_r are one and two, respectively. Antonymy links always have a weight of 2.5.

For the purpose of determining the weight of an edge in the path, each edge is assumed to consist of two inverse relations. For example, if *robin* IS-A *bird*, then it is also the case that *bird* SUBSUMES *robin*. However, it is possible for the inverse relations to be assigned a different weight by Equation 2.1. For example, the weight for *keyboard* HAS-A *key* is not necessarily the same as for *key* PART-OF *keyboard* as we cannot assume that the number of meronyms of *keyboard* and the number of holonyms of *key* are the same. Sussna assumed that the semantic distance between concepts should be a symmetrical relationship and so takes the average of the two weights.

The semantic distance weight of an edge is also scaled by the depth of the relation in the taxonomy. The final semantic distance value for the edge between

two adjacent synsets c_1 and c_2 is given by:

$$dist_S(c_1, c_2) = \frac{wt(c_1 \longrightarrow_r) + wt(c_2 \longrightarrow_{r'})}{2 \times max(depth(c_1), depth(c_2))} \tag{2.2}$$

In the preceding equation, r is the type of relation that holds between c_1 and c_2, and r' is the inverse of r (the type of relation that holds between c_2 and c_1). To determine the semantic distance between any pair of synsets, Sussna takes the sum of the distances between the nodes in the shortest path between the synsets in WordNet.

2.3.2 Hirst and St-Onge

Hirst and St-Onge [15] proposed a semantic relatedness measure for WordNet that was an adaptation of an earlier measure by Morris and Hirst [27]. The measure was previously based on Roget's thesaurus [37]. Their measure was developed in the context of a system for the automatic detection and correction of malapropisms using lexical chains. Hirst and St-Onge define a malapropism as "the confounding of an intended word with another word of similar sound or spelling that has a quite different and malapropos meaning." (p. 305) For example, accidentally substituting the word 'prostate' for 'prostrate' would result in a malapropism.

For their measure, Hirst and St-Onge defined three categories of WordNet relationship types: 'upward', 'downward' and 'horizontal'. For example, hypernymy (IS-A) is classified as an upward link, as it leads toward the root of the WordNet taxonomy, whereas hyponymy (SUBSUMES) is a downward link. In general, the up and down categories are used to separate inverse relations, whereas horizontal link types correspond to relations that do not have inverses. The complete list of classifications used by Hirst and St-Onge is given in Table 2.2.

Hirst and St-Onge distinguish two strengths of semantic relations: strong and medium-strong. Two words, w_1 and w_2, are strongly related if one of three conditions holds:

1. They are synonyms (there is a synset with both w_1 and w_2).
2. They are antonyms (w_1 and w_2 belong to the synsets c_1 and c_2, and c_1 and c_2 are related by antonymy).
3. One is a compound word that includes the other one, and there exists a semantic relation (of any kind) between synsets containing the words. For

Relation	Direction
Also see	Horizontal
Antonymy	Horizontal
Attribute	Horizontal
Cause	Down
Entailment	Down
Holonymy	Down
Hypernymy	Up
Hyponymy	Down
Meronomy	Up
Pertinence	Horizontal
Similarity	Horizontal

Table 2.2: Hirst and St-Onge's classification of WordNet relations into directions

example, *school* and *private school* are strongly related, because *private school* IS-A *school*, and the compound word *private school* contains *school*.

Medium-strong relations hold between words that have corresponding synsets that are connected in the WordNet graph by an allowable path. A path is allowable if it conforms to one of eight patterns, which are defined in terms of the three directions of semantic links. The motivation for these patterns is the observation that changes in direction often result in reduced overall relatedness. For example, some semantic relations may be viewed as transitive. If A IS-A B IS-A C, then A IS-A C. Similarly, if A PART-OF B PART-OF C, then A PART-OF C. However, when a path includes a change in direction, the transitivity of the relations is compromised. The eight allowable patterns are shown in Figure 2.2. It should be noted that each vector in the patterns in Figure 2.2 represents any number of links in the given direction.

Unlike strong relations, medium-strong relations have a range of relatedness values. The strength of relatedness for a medium-strong relation between the concepts c_1 and c_2 is given by:

$$rel_{HS}(c_1, c_2) = C - len(c_1, c_2) - k \times turns(c_1, c_2) \tag{2.3}$$

where C and k are constants, $length(c_1, c_2)$ is the length, measured in nodes, of the shortest allowable path connecting the synsets c_1 and c_2, and $turns(c_1, c_2)$ is the number of changes in direction in the shortest allowable path. Budanitsky and

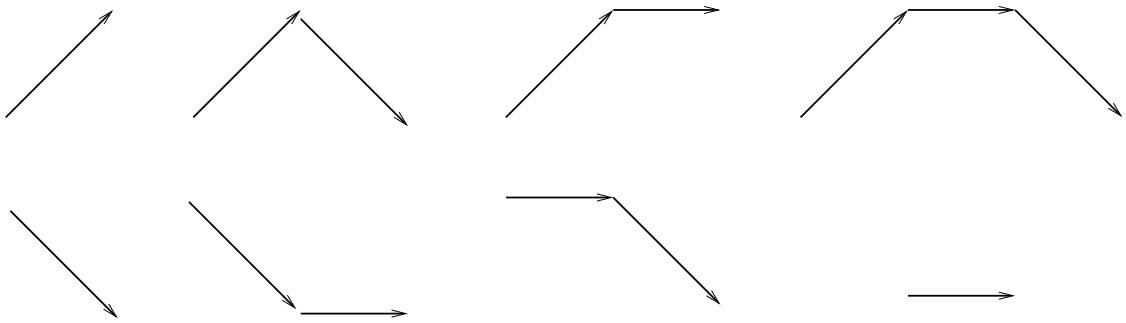


Figure 2.2: Patterns of semantic relations allowed in medium-strong relationships for Hirst and St-Onge relatedness measure. Each arrow represents any number of relations of the given direction.

Hirst [4] employed the value eight for C and one for k in their evaluation of the measure.

Finally, while the measure described above applies to word senses, in the form of synsets, Hirst and St-Onge also required relatedness values for non-disambiguated word forms. The nodes in WordNet correspond to word senses, but most words have multiple meanings. If it is not known which particular sense of a word is the correct one for the context, then the measure cannot be used as described above. To solve this problem, Hirst and St-Onge assume that the relatedness of word forms is equal to that of their most related senses. Where $S(w_i)$ denotes the set of all senses of the word w_i, the relatedness of the words w_1 and w_2 is:

$$rel(w_1, w_2) = \max_{c_1 \in S(w_1), c_2 \in S(w_2)} [rel(c_1, c_2)] \tag{2.4}$$

2.3.3 Banerjee and Pedersen

Banerjee and Pedersen [2] adopt an alternative approach to that of path-based measures, based on a technique by Lesk [19]. Rather than examining paths of semantic relations between word senses, as most other measures do, they compare the text of the definitions provided in WordNet for each synset. Relatedness is computed in terms of the overlap of words in these definitions.

The distinguishing feature of WordNet is the organization of concepts into a semantic network. However, WordNet also supplies short definitions, called *glosses*,

for each synset such as might be found in a traditional dictionary. For example, in WordNet 2.0 the gloss for the synset $\{apple\}$ is "fruit with red or yellow or green skin and sweet to tart crisp whitish flesh." Banerjee and Pedersen calculate relatedness by counting the number of words that co-occur in the glosses of different synsets.

Banerjee and Pedersen note that phrasal overlaps — sequences of words that appear in different glosses — are often indicative of a strong relationship between concepts. They therefore assign a higher value to a phrasal overlap of n words than to an overlap of n words that are not in sequence. Specifically, a phrasal overlap of n words is assigned the value n^2, whereas n shared words that do not belong to a phrasal overlap are assigned the value n. For example, the gloss for *drawing paper* is "paper that is specially prepared for use in drafting" and the gloss for *decal* is "the art of transfering designs from specially prepared paper to a wood or glass or metal surface." As the phrase "specially prepared" appears in both glosses, it contributes a score of $2^2 = 4$. The word 'paper' also appears in both glosses, and contributes a score of one, for a total score of five.

Gloss overlap is a technique that could be applied to any dictionary or lexicon with textual definitions. However, Banerjee and Pedersen exploit WordNet by comparing the glosses of not only the target synsets, but also of their nearest neighbours in the semantic network. For each relation type r, they define a function $r(s_1)$ that returns the gloss of the synset related by r to s_1. For example, the function $hypernym(s_1)$ returns the gloss of the hypernym of the synset s_1. If s_1 is connected to more than one synset by the relation type r, then $r(s_1)$ returns the concatenation of the glosses of each related synset. In addition, they also define a function named $gloss(s_1)$ that returns the gloss for the synset s_1.

Banerjee and Pedersen observe that not every relation is equally helpful for determining relatedness, and suggest that different relations may be more or less useful depending on the particular application. They therefore suggest a general formula for calculating relatedness that can use any arbitrary subset of semantic relations. Let $RELPAIRS$ be a set of pairs of gloss functions, as defined above. The pairs indicate which relations will be compared to one another when computing relatedness. In order to maintain the symmetry of the measure, for any pair $(r_1, r_2) \in RELPAIRS$, the set must also contain (r_2, r_1). This constraint ensures that $rel_{BP}(s_1, s_2) = rel_{BP}(s_2, s_1)$. Given the set of pairs $RELPAIRS$, and two synsets s_1 and s_2, relatedness is calculated using the following equation:

$$rel_{BP}(s_1, s_2) = \sum score(r_1(s_1), r_2(s_2)), \quad \forall (r_1, r_2) \in RELPAIRS \tag{2.5}$$

In the equation above, $score(t_1, t_2)$ is a function that returns the overlap score of two strings t_1 and t_2. As an illustration, given the set $RELPAIRS = \{$*(gloss,gloss), (hype,hype), (hypo,hypo), (hype,gloss), (gloss,hype)*$\}$, the relatedness function would be:

$$\begin{aligned} rel_{BP}(s_1, s_2) = {} & score(gloss(s_1), gloss(s_2)) + score(hype(s_1), hype(s_2)) \\ & + score(hypo(s_1), hypo(s_2)) + score(hype(s_1), gloss(s_2)) \\ & + score(gloss(s_1), gloss(s_2)) \end{aligned} \tag{2.6}$$

2.4 Similarity Measures

Although our principal interest is in semantic relatedness, semantic similarity measures have been shown to be very effective proxies for relatedness measures. We will therefore describe several important WordNet-based similarity measures. Jiang and Conrath [16] distinguish between three approaches in the literature to measuring similarity. They call these *edge-counting*, *node-counting*, and *combined*, or *hybrid*, approaches.

The edge-counting approach relies entirely on the IS-A hierarchy. These measures compute similarity in terms of the shortest path between the target synsets in the taxonomy. The degree of similarity is determined on the basis of this path, and generally will correspond inversely with the path length. The first application of this technique to WordNet is typically attributed to Rada et al. [33]. The edge-counting technique offers a very intuitive representation of similarity. The principal criticism of the edge-counting approach is that it is sensitive to the quality of the taxonomy that is employed. In particular, many authors have noted the inconsistent conceptual density of the WordNet graph, and the problems that this introduces for the reliability of edge-counting measures. The edge-counting method is equivalent to the path-based approach used in many relatedness measures, except that it is applied to the IS-A taxonomy exclusively, and ignores other semantic relationship types.

In order to address the criticisms of the edge-counting measures some authors have preferred to use taxonomies to determine the relationships between concepts, but to employ external resources (usually corpus statistics) to calculate the value of similarity. These sorts of measures are called node-counting, since they discard information about the edges connecting synsets and focus on a few key nodes, which typically includes the two target nodes and their most specific common subsumer

in the taxonomy. Resnik and Lin's measures will be described as examples of the node-counting approach.

Finally, while the node-counting approach eliminated certain problems that arose from inconsistencies in the taxonomy, it also ignored much useful information that is contained in the paths between synsets. Jiang and Conrath therefore proposed a measure that calculates similarity using the edges in the shortest path, but also uses corpus statistics in a secondary, corrective role.

2.4.1 Leacock and Chodorow

Leacock and Chodorow [18] proposed a semantic similarity measure that typifies the edge-counting approach. In their measure, the similarity between two concepts is determined by first finding the length of the shortest path that connects them in the WordNet taxonomy. The length of the path that is found is scaled to a value between zero and one and similarity is then calculated as the negative logarithm of this value. The measure by Leacock and Chodorow may be expressed as follows:

$$sim_{LC}(c_1, c_2) = -\log \frac{length(c_1, c_2)}{2D} \tag{2.7}$$

where $length(c_1, c_2)$ denotes the length, counted in nodes, of the shortest path between the concepts c_1 and c_2 and D denotes the maximum depth of the WordNet subsumption hierarchy.

The measure by Leacock and Chodorow can be illustrated with reference to the WordNet subgraph given in Figure 2.1. The shortest taxonomic path between *motorcycle* and *bicycle* is:

> *motorcycle* IS-A *motor vehicle* IS-A *self-propelled vehicle* IS-A *wheeled vehicle* SUBSUMES *bicycle*

It should be noted that the taxonomic path length differs from the network path length, as only hypernymy and hyponymy relations are considered. Assuming an arbitrary maximum depth of 10 in the WordNet taxonomy, the value of similarity between *motorcycle* and *bicycle* would be computed as:

$$\begin{aligned} sim_{LC}(motorcycle, bicycle) &= -\log \frac{length(motorcycle, bicycle)}{2 \times 10} \\ &= -\log \frac{5}{20} \\ &= 0.60 \end{aligned}$$

2.4.2 Resnik

Resnik [34] introduced the first similarity measure to combine corpus statistics with a conceptual taxonomy. Resnik's hybrid approach has received considerable attention, and a number of other measures have incorporated his technique. Resnik defines similarity in terms of information theory, and derives the necessary probability information from a corpus of text. The key intuition in Resnik's measure is that for any two concepts, the most specific concept that subsumes them both in the conceptual taxonomy represents the information that the concepts share in common. For example, in Figure 2.1 the most specific common subsumer of *car* and *bicycle* is *wheeled vehicle*. The concept *wheeled vehicle* is assumed to represent the information that is common to both *car* and *bicycle*. Resnik determines similarity by calculating the information content of the shared subsumers. That is, higher information content means that the concepts share more in common, and so are more similar.

First, Resnik defined $P(c)$ as the probability of encountering an instance of a concept c. In order to determine $P(c)$, Resnik relied on frequency information from a text corpus. When counting the instances of concepts in the corpus, any instances of subsumed concepts are also counted as instances of their subsuming concept. For example, any instances of the words for *apple*, *orange*, *banana*, etc. also count as instances of *fruit*. The concept *fruit* will necessarily have a higher frequency than any concepts it subsumes, including every concept subsumed by its children, and so on. Therefore, the probabilities of encountering concepts increases monotonically for concepts higher in the taxonomy.

In order to compute the probability function $P(c)$, we must first calculate the number of occurrences of the concept c and the occurrences of all concepts subsumed by c. Where $words(c)$ denotes the set of words that correspond to all of the concepts subsumed by c, the total frequency of c is given by:

$$freq(c) = \sum_{n \in words(c)} count(n) \tag{2.8}$$

The probability of encountering a concept c may be defined as the relative frequency of c, where N is the total number of words observed in the corpus:

$$P(c) = \frac{freq(c)}{N} \tag{2.9}$$

For his experiments, Resnik employed the Brown Corpus of American English

[14]. He counted only the nouns in this corpus, and only those nouns that are associated with concepts in WordNet.

According to the axioms of information theory, the information content of a concept c is the negative log of its likelihood: $-\log P(c)$. As mentioned above, Resnik argued that the similarity of two concepts is proportional to the amount of information that they share, and that the shared information is represented by their most specific common subsumer. For example, the most specific shared subsumer of *car* and *motorcycle* in Figure 2.1 is *motor vehicle*. Therefore *motor vehicle* is assumed to represent all of the information that is common to the concepts *car* and *motorcycle*. The amount of information conveyed by the concept *motor vehicle*, as determined by information theory, corresponds to the degree of similarity between *car* and *motorcycle*.

Formally, where $S(c_1, c_2)$ denotes the set of concepts that subsume both c_1 and c_2, the degree of similarity is:

$$sim_R(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log P(c)] \tag{2.10}$$

A few features of the preceding formula are worth noting. First, similarity always decreases lower in the taxonomy, as information content correlates inversely with $P(c)$. As the root node of the conceptual hierarchy subsumes every concept, it has a probability of exactly one and therefore has an information content of zero. In other words, knowing that two concepts share the root node as a subsumer provides no information, as this is true of any two concepts. If the only common subsumer of two concepts is the root node, they have the least possible similarity. Second, Resnik's equation uses the common subsumer with the maximum information content. This will always be the most specific, i.e. the 'lowest', concept in any sequence of superordinates in the taxonomy.[1]

In order to calculate the similarity of words, as opposed to that of word senses, Resnik adopts an analogous solution to that of Hirst and St-Onge. The similarity of words is assumed to be equivalent to the maximum similarity of their possible senses. Where $S(w_i)$ denotes the set of all of the senses of the word w_i, the similarity between words is:

[1]Budanitsky and Hirst [4] reformulated Resnik's measure to explicitly refer to the lowest superordinate in the taxonomy. Although Budanitsky and Hirst's formulation is more intuitive than Resnik's, it introduces ambiguity in cases of multiple inheritance. In these cases, it may not be possible to identify a 'lower' subsumer, but the information content gives an indication of the most specific concept.

$$sim(w_1, w_2) = \max_{c_1 \in S(w_1), c_2 \in S(w_2)} [sim(c_1, c_2)] \tag{2.11}$$

2.4.3 Jiang and Conrath

Jiang and Conrath [16] sought to combine the advantages of the edge-counting and node-counting approaches. In order to compensate for the unreliability of edge-distances, Jiang and Conrath weigh each edge by associating probabilities based on corpus statistics. Their approach is similar to Resnik's, in that it employs information from both a conceptual taxonomy and from a text corpus. However, whereas Resnik bases the value of similarity on the information content of one node — the most informative common subsumer — Jiang and Conrath use information theory to determine the weight of each link in a path.

Jiang and Conrath argue that the degree of similarity between a parent and its child in the noun hierarchy of WordNet is proportional to the probability of encountering the child, given an instance of the parent: $P(c \mid par(c))$. By definition, the quantity $P(c \mid par(c))$ is:

$$P(c \mid par(c)) = \frac{P(c \cap par(c))}{P(par(c))} \tag{2.12}$$

Like Resnik, Jiang and Conrath consider every instance of a child to be an instance of its parent, and thus $P(c \cap par(c)) = P(c)$. That is, it is redundant to require both a child c and its parent $par(c)$, as every instance of c is also an instance of $par(c)$. The equation for the probability of a child, given an instance of its parent, can therefore be simplified to:

$$P(c \mid par(c)) = \frac{P(c)}{P(par(c))} \tag{2.13}$$

Jiang and Conrath define the semantic distance between a child c and parent $par(c)$ as the information content of the conditional probability of c given $par(c)$, and using the basic properties of information theory obtain the following semantic distance equation:

$$\begin{aligned} dist_{JC}(c, par(c)) &= -\log P(c|par(c)) \\ &= IC(c \cap par(c)) - IC(par(c)) \\ &= IC(c) - IC(par(c)) \end{aligned} \tag{2.14}$$

The semantic distance between a parent and its child concept is therefore the difference in their information content. This seems a plausible conclusion, as the difference in information content should reflect the information required to distinguish a concept from all of its sibling concepts. For example, if a parent has only a single child, then the conditional probability $P(c \mid par(c)) = 1$. In this case, taking the negative logarithm gives $dist_{JC} = 0$. If no additional information is required to distinguish a child from its parent, then the semantic distance between them ought to be zero; they are effectively the same concept.

To compute the total semantic distance between any two concepts in the taxonomy, Jiang and Conrath's measure uses the sum of the individual distances between the nodes in the shortest path. As the shared subsumer (denoted by $lso(c_1, c_2)$ for the lowest super-ordinate shared by c_1 and c_2) does not have a parent in the path, this node is excluded from the summation. The semantic distance between any two concepts c_1 and c_2 in the taxonomy is therefore:

$$dist_{JC}(c_1, c_2) = \sum_{c \in path(c_1,c_2) \setminus lso(c_1,c_2)} dist_{JC}(c, par(c)) \tag{2.15}$$

By substituting the expression in Equation 2.14 into Equation 2.15 and expanding the summation, we obtain:

$$\begin{aligned} dist_{JC}(c_1, c_2) &= IC(c_1) + IC(c_2) - 2 \times IC(lso(c_1, c_2)) \\ &= 2 \log P(lso(c_1, c_2)) - (\log P(c_1) + \log P(c_2)) \end{aligned} \tag{2.16}$$

2.4.4 Lin

Lin [20] attempted to provide a more general and theoretically sound basis for determining the similarity between concepts than previous work had provided. He argued that similarity measures should not depend on the domain of application, nor on the details of the resources that they use. Lin begins by proposing three key intuitions about similarity:

- Intuition 1: The similarity between A and B is related to their commonality. The more commonality they share, the more similar they are.
- Intuition 2: The similarity between A and B is related to the differences between them. The more differences they have, the less similar they are.

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

Lin argued that as there are different ways of capturing the intuitions above, an additional set of assumptions are required. Lin therefore proposed a set of five assumptions that capture these intuitions, and from which a measure of similarity may be derived. The five assumptions are stated in terms of information theory. In the following assumptions, $common(A, B)$ is a proposition that states the commonality of the objects A and B, and $description(A, B)$ is a proposition that states what A and B are.

- Assumption 1: The commonality between A and B is measured by: $IC(common(A, B))$
- Assumption 2: The difference between A and B is measured by: $IC(description(A, B)) - IC(common(A, B))$
- Assumption 3: The similarity between A and B is a function of the commonalities and differences of A and B. Formally: $sim(A, B) = f(IC(common(A, B), IC(description(A, B))$
- Assumption 4: The similarity between a pair of identical objects is always one. Thus: $sim(A, A) = 1$
- Assumption 5: The similarity between a pair of objects with no commonality is always zero. Thus: $\forall y > 0, f(0, y) = 0$
- Assumption 6: If the similarity between A and B can be computed using two independent sets of criteria, then the overall similarity is the weighted average of the two similarity values: $\forall x_1 \leq y_1, x_2 \leq y_2 : f(x_1 + x_2, y_1 + y_2) = \frac{y_1}{y_1+y_2} f(x_1, y_1) + \frac{y_2}{y_1+y_2} f(x_2, y_2)$

Using the six assumptions listed above, Lin proves the following similarity theorem:

$$sim_L(A, B) = \frac{\log P(common(A, B))}{\log P(description(A, B))} \tag{2.17}$$

In order to apply the similarity theorem above to a conceptual taxonomy, Lin follows similar reasoning to that of Resnik. The concept in a taxonomy that corresponds to the statement of the commonalities between the concepts c_1 and c_2 is the lowest super-ordinate, denoted $lso(c_1, c_2)$. Similarly, the statement that describes

the concepts c_1 and c_2 is the union of the two concepts. The information content of the statement "c_1 and c_2" is the sum of the information content of c_1 and c_2. According to the basic premise of information theory the information content of a message is the negative log of its probability, and therefore the sum of the information content of c_1 and c_2 is $-\log P(c_1) + -\log P(c_2)$. Substituting into Lin's similarity theorem, we obtain:

$$sim_L(c_1, c_2) = \frac{2 \times \log P(lso(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \tag{2.18}$$

Lin's measure is therefore the ratio of the information shared in common to the total amount of information possessed by two concepts. It is quite similar to Resnik's measure except that Resnik's measure considers only the information that is shared by concepts, and does not take into account the total amount of information that they represent. Due to this, Resnik's measure cannot distinguish between different pairs of concepts that have the same most informative subsumer. For example, in the small semantic network in Figure 2.1, the concept pair *car/bicycle* has exactly the same similarity as the pairs *motor vehicle/bicycle* and *self-propelled vehicle/bicycle* according to Resnik's measure.

2.4.5 Yang and Powers

Yang and Powers [47] recently published the details of a new similarity measure that they reported to achieve significantly improved results over previous efforts. Their measure combines many of the techniques found in other path-based models, such as assigning weights to edges and restricting the allowable paths based on patterns of semantic relations.

Unlike other edge-counting measures of similarity, the measure by Yang and Powers does not rely exclusively on the IS-A hierarchy of WordNet, and incorporates several other relationship types. Specifically, holonymy (PART-OF), meronymy (HAS-A), and antonymy (OPPOSITE) are considered, along with hyponymy (SUBSUMES) and hypernymy (IS-A). When determining the shortest path between concepts in WordNet, edges of each of the preceding types are explored. However, only one change in relationship type along a path is permitted. For example, a path may consist of any number of meronym links followed by any number of hypernym links, but these couldn't be followed by another relationship type.

In the partial WordNet graph shown in Figure 2.1, the path *motorcycle* HAS-A *mudguard* PART-OF *bicycle* would be permitted, since there is only one change

of relationship type, from HAS-A to PART-OF. However, the path *motor vehicle* SUBSUMES *motorcycle* HAS-A *mudguard* PART-OF *bicycle* would not be permitted.

In contrast to other path-based measures, Yang and Powers's measure calculates similarity as the product of link weights rather than as their sum. As the weights of links in Yang and Powers' measure range from zero to one, the value of similarity decreases as the path between concepts increases, approaching zero for large path lengths.

The measure by Yang and Powers uses different weight constants for each relationship type, as well as overall weighting factors that are applied to the whole path, depending on which types of links the path contains. The variety of weight constants provides a great deal of flexibility for tuning the model. The model by Yang and Powers computes similarity as follows:

$$sim_{YP}(c_1, c_2) = \begin{cases} \alpha_t \prod_{i=1}^{dist(c_1,c_2)} \beta_{t_i} & \text{if } dist(c_1, c_2) < \gamma \\ 0 & \text{if } dist(c_1, c_2) \geq \gamma \end{cases} \tag{2.19}$$

where $0 < sim(c_1, c_2) <= 1$ and

- $t = hh$ (hyper/hyponym), hm (holo/meronym), sa (syn/antonym), id (identity)
- α_t: a link type factor applied to a sequence of links of type t ($0 < \alpha_t \leq 1$).
- β_t: the depth factor, which also depends on the link type.
- γ: an arbitrary threshold on the distance introduced for efficiency, representing human cognitive limitations.
- c_1, c_2: concept node 1 and concept node 2.
- $dist(c_1, c_2)$: the distance (the shortest path) between c_1 and c_2.

The particular values for α_t and β_t are: $\alpha_{id} = 1.0$, $\alpha_{sa} = 0.9$, $\alpha_{hh} = \alpha_{hm} = 0.85$ and $\beta_{hm} = \beta_{hh} = 0.7$. Note that no value for β_{id} or β_{sa} has been provided. This is because paths with the identity, synonym or antonym relation cannot have any other relation in them, according to Yang and Powers' measure. In these cases, the path always has a length of one, and so the product in Equation 2.19 may be ignored, leaving $sim_{YP}(c_1, c_2) = \alpha_t$. Yang and Powers determined the values of α_t

and β_t by varying each of the values by small increments and then comparing the correlation of the measure against human judgments of relatedness.

The fragment of WordNet shown in Figure 2.1 may be used to illustrate the measure. The shortest path between *car* and *mudguard* is: *car* IS-A *motor vehicle* SUBSUMES *motorcycle* HAS-A *mudguard*. The path has a length of three (counted in edges) and thus $dist(car, mudguard) = 3$. As all of the relations in this path are either hypernym/hyponym or meronym/holonym relations, the value of β_{t_i} is $\beta_{hh} = \beta_{hm} = 0.7$ and the value of α_{t_i} is $\alpha_{hh} = \alpha_{hm} = 0.85$. The similarity value is calculated using Equation 2.19:

$$\begin{aligned} sim_{YP}(car, mudguard) &= 0.85 \prod_{i=1}^{3} 0.7 \\ &= 0.85 \times 0.7 \times 0.7 \times 0.7 \\ &= 0.29155 \end{aligned}$$

Like Resnik and Hirst and St-Onge, Yang and Powers are interested in measuring the semantic distance between polysemous word forms, as well as word senses. Yang and Powers suggest several possible functions for determining the semantic distance between words. Specifically, the similarity of a word pair could be the maximum, the sum, or the mean of the similarities of the possible word senses. Where $S(w)$ denotes the set of word senses associated with the word w, the relatedness between the words w_1 and w_2 are:

$$sim_{max}(w_1, w_2) = \max_{s_1 \in S(w_1), s_2 \in S(w_2)} sim(s_1, s_2) \tag{2.20}$$

$$sim_{sum}(w_1, w_2) = \sum_{s_1 \in S(w_1), s_2 \in S(w_2)} sim(s_1, s_2) \tag{2.21}$$

$$sim_{mean}(w_1, w_2) = \frac{sim_{sum}(w_1, w_2)}{|S(w_1)| \times |sense(w_2)|} \tag{2.22}$$

In their evaluation, Yang and Powers found that sim_{max} yielded the best results, confirming the intuitions of other researchers.

The measure by Yang and Powers raises some interesting questions about the difference between similarity and relatedness measures. Unlike most other path-based similarity measures, Yang and Powers do not restrict their measure to IS-A

relations, but include PART-OF relations as well as antonymy. Sussna's [44] measure is similar to theirs in many respects. Sussna used the same set of semantic relations that Yang and Powers used, but considered his distance measure a measure of relatedness. The use of semantic relations other than IS-A relations might be justified by the observation that although PART-OF and OPPOSITE relations do not entail similarity, in most cases concepts connected by these relations are in fact similar. For example, the parts of a mechanical device are likely to also be mechanical devices, and the parts of a biological system are also likely to be biological systems.[2]

2.5 Relatedness Applications

2.5.1 Word Sense Disambiguation

Several authors have applied relatedness measures to the problem of word sense disambiguation, including Sussna [44], Lesk [19], and Banerjee and Pedersen [2]. Recently, Patwardhan et al. [30] described a technique for word sense disambiguation and used it to compare the performance of several different similarity and relatedness measures, including the measures by Hirst and St-Onge [15], Jiang and Conrath [16], Leacock and Chodorow [18], Lin [20], and Resnik [34].

The algorithm described by Patwardhan et al. moves a *window of context* across a text. This window consists of a target word and some number of words to both the left and right of the target word. As several of the semantic distance measures can only be applied to nouns, the window of context includes only nouns, and only those nouns that are included in WordNet. For each word in the window, candidate senses are identified. These senses consist of all WordNet synsets associated with the surface form of the current word, and all WordNet synsets associated with the base form of the word.

Once all of the candidate senses have been identified, each candidate sense is assigned a score based on its relatedness to the other nouns in the window. The total score for each candidate sense is the sum of its relatedness to each of the senses of the other nouns in the window. The equation $score(c)$ then gives the total score

[2]Alternatively, Yang and Powers may simply be using the terms 'similarity' and 'relatedness' differently than other researchers. While they call their measure a similarity measure when comparing concepts, they call it a relatedness measure for word forms.

for a candidate sense c:

$$score(c) = \sum_{w \in W} \sum_{c_i \in S(w)} rel(c, c_i) \tag{2.23}$$

where W is the set of words in the window of context excluding the target word, $S(w)$ denotes the set of synsets associated with the word w, and $rel(c_1, c_2)$ denotes a function that gives the strength of relatedness between the synsets c_1 and c_2. The candidate sense with the highest score is selected as the correct sense of the target word:

$$sense(w) = \arg\max_{c \in S(w)} [score(c)] \tag{2.24}$$

It should be noted that the technique described above is compatible with any semantic distance measure that gives a numerical estimation of the semantic distance of two synsets. Patwardhan et al. evaluated the technique using several different measures, including the measures by Resnik [34], Jiang and Conrath [16], Lin [20], Hirst and St-Onge [15] and a gloss-based measure. They found that the Jiang-Conrath measure and the gloss-based measure achieved the best results overall.

2.5.2 Malapropism Detection

Hirst and St-Onge [15] describe an ingenious application of a lexical semantic relatedness measure for detecting and correcting malapropisms. A difficult problem for automatic spell-checking is the case of a spelling error that results in a correctly spelled, but unintended, word. These sorts of errors can occur either as the result of a typing error, or can be the result of confounding the spellings of two similar words. For example, in the phrase "an ingenuous machine for peeling oranges," the word 'ingenuous' ought to have been spelled 'ingenious'. The author of this phrase may have typed a 'u' instead of an 'i', or may have confused the meanings of the two words. In either case, traditional spelling checkers would not recognize the error. Hirst and St-Onge call these sorts of mistakes *malapropisms*, or real-word spelling errors, and propose a solution to this problem using lexical chains.

Lexical chains are representations of the cohesive relations between words in a text. That is, if a text is cohesive, there are likely to be words in successive sentences that refer to similar or related concepts. By stringing together the words with these cohesive relationships, we obtain a representation of the threads of meaning

```
description: top-level lexical chaining algorithm
input      : a text consisting of a sequence of words
output     : a set of lexical chains

1 chains ← ∅
2 foreach word ∈ Nouns(text) do
3     if InsertStrong(word,chains) then
4         continue
5     else if InsertMedStrong(word,chains) then
6         continue
7     else
          /* There were no related chains, so create a new chain
             */
8         chains ← chains ∪ {{(word,Senses(word))}}
9 return chains
```

Algorithm 2.1: Hirst and St-Onge lexical chaining algorithm

in the text. Lexical chains are relevant to the problem of malapropisms because malapropisms, in general, do not cohere with the text surrounding them.

For their lexical chains, Hirst and St-Onge considered only nouns, because of the lack of interconnectivity between the verb, adverb, and adjective portions of WordNet. The lexical chaining algorithm processes all of the nouns in a text in sequence. It maintains a set of current chains, each of which is a set of structures that contain a word and a subset of that word's senses (synsets). Each noun in the text is compared to the previously formed chains. When one of the senses of the current word is found to be related to a sense in a chain, the word is added to the chain. The new word is added with only those senses that are related to the chain. In this way, the words added to a chain are disambiguated by the words that preceded them.

The lexical chaining algorithm makes two passes of the existing chains for each new word. To help illustrate Hirst and St-Onge's algorithm, we have provided a simplified pseudocode description of their technique based on their explanation in [15]. The top-level lexical chaining algorithm is provided in Algorithm 2.1.

The relatedness measure used by Hirst and St-Onge distinguishes two degrees of semantic relatedness: strong and medium-strong. In the first pass of the lexical chaining algorithm, only strong relations are considered. The new word is inserted

```
description: algorithm for inserting a word into lexical chains using
             strong relations
input      : a set of lexical chains and a single word to be inserted
output     : true if the word is successfully inserted, else false
1  relatedsenses ← ∅
2  foreach chain ∈ chains do
       /* Compare the senses of newword to the senses in chain  */
3      foreach newsense ∈ Senses(newword) do
4          foreach (word,senses) ∈ chain do
5              foreach sense ∈ senses do
6                  if sense.StrongRelation(newsense) then
7                      relatedsenses ← relatedsenses ∪ { newsense }
       /* Add newword to the first chain that it is strongly
          related to, along with the senses of newword that are
          related to the chain                                  */
8      if relatedsenses ≠ ∅ then
9          chain.Add((newword,relatedsenses))
10         return true
11 return false
```

Algorithm 2.2: Hirst and St-Onge lexical chaining algorithm: Insert-Strong(newword, chains)

into the first chain to which it is strongly related. The word is added in a structure that contains the word form and the set of senses of that word that have a strong relationship to other senses in the chain. The algorithm for inserting words into lexical chains to which they are strongly related is provided in Algorithm 2.2.

In a second pass, the medium-strong relationships between the senses of the new word and the senses of words in the lexical chains are examined. If any medium-strong relations are found, the word is added to the chain containing the synset that it is the most strongly related to. Only the most strongly related sense of the new word is added to the chain. The algorithm for inserting words with medium-strong relationships to lexical chains is provided in Algorithm 2.3.

The description of the lexical chaining process given so far is incomplete, but should suffice for the purpose of understanding the nature of lexical chains. For ex-

ample, the full algorithm includes steps for removing word senses in chains that are not connected to other senses in the chain, in order to disambiguate words that were previously added on the basis of the words that follow them. The comprehensive description of the lexical chaining procedure is provided by St-Onge [40].

To apply the lexical chains described above to the problem of malapropisms, the target text is first processed by the lexical chainer. A set of potential malapropisms is constructed from all of the atomic chains, which represent the words that could not be added to any chain. For each of the potential errors, a search is performed for similarly spelled words. If any similarly spelled words are found that would in fact have been successfully added to a chain, then that word is flagged as a probable malapropism, and the user is informed of the potential error along with suggestions for the correct spelling.

2.6 Chapter Summary

Of the three varieties of taxonomy-based measures of semantic similarity — edge-counting, node-counting, and hybrid — the hybrid approach has met with the most success. Jiang and Conrath's measure was found to achieve the best results in studies by Budanitsky and Hirst [4] and by Patwardhan et al. [30]. Hybrid approaches exploit the rich semantic information available in semantic networks, while compensating for the inconsistencies in such resources with corpus statistics.

The node-counting and hybrid approaches turn on the intuition that similarity can be expressed in terms of quantities of information. That is, one may reasonably ask how much information concepts have in common, and how much information is peculiar to one concept with respect to another. The formulation of similarity in terms of quantity of information enables an information-theoretic treatment of the problem.

Semantic relatedness, on the other hand, does not lend itself to a formulation in terms of information quantity. The strength of relatedness between concepts seems to be a matter of the quality of their relationship, rather than of any quantity of information. For example, despite the strong relationship between *gasoline* and *car*, the fact that gasoline is used to power cars is a very small amount of information relative to all there is to know about gasoline. However, this small amount of information is in some way sufficiently important that it results in a high degree of relatedness. In this case, a plausible explanation for the importance of the relationship may be the very frequent occurrence in everyday use of the functional association between *gasoline* and *car*.

Measures of semantic relatedness have therefore been restricted to the path-based approach, with a few notable exceptions such as the gloss-based measure of Banerjee and Pedersen [2]. In the next chapter, we will examine the features of path-based measures of relatedness and similarity, and will propose a general formal description for measures of this type. Two simplifying assumptions will then be made in order to derive a new, simplified, path-based measure.

```
description: algorithm for inserting a word into lexical chains using
             medium-strong relations
input      : a set of lexical chains and a single word to be inserted
output     : true if the word is successfully inserted, else false

1  relatedsenses ← ∅
2  max ← 0
3  foreach chain ∈ chains do
       /* Compare the senses of newword to the senses in chain  */
4      foreach newsense ∈ Senses(newword) do
5          foreach (word,senses) ∈ chain do
6              foreach sense ∈ senses do
7                  if sense.MedStrongRelation(newsense) and
                   sense.RelationStrength(newsense) > max then
8                      max ← sense.RelationStrength(newsense)
9                      relatedsenses ← { newsense }
10                     relatedchain ← chain
   /* Add newword to the most strongly related chain, along with
      the most strongly related sense of newword                  */
11 if relatedsenses ≠ ∅ then
12     relatedchain.Add((newword,relatedsenses))
13     return true
14 else
15     return false
```

Algorithm 2.3: Hirst and St-Onge lexical chaining algorithm: InsertMedStrong(newword, chains)

Chapter 3

Semantic Relatedness as a Function of Network Path Length

The semantic relatedness measures described in the last chapter have met with some success. For example, Hirst and St-Onge [15] successfully demonstrated an application of their measure to the problem of real-word spelling errors. Similarly, Sussna [44], and Banerjee and Pedersen [2] reported positive results when applying their relatedness measures to the task of word sense disambiguation. Also, the measures by Hirst and St-Onge, and Banerjee and Pedersen were shown to correlate with the judgments of human test subjects, achieving correlations of $r = 0.79$ and $r = 0.67$, respectively.

However, there is also much room for improvement. One indication of the inadequacy of current measures of relatedness is their relatively poor performance compared to measures of similarity. In a comparison of semantic distance measures by Budanitsky and Hirst [4], the authors found that similarity measures performed better than relatedness measures, even though they were examining tasks believed to rely on measuring relatedness. Budanitsky and Hirst compared five similarity and relatedness measures in two different tasks. In the first, they compared the results of each measure against the relatedness judgments of human subjects, which had been collected in previous experiments. While similarity measures achieved correlations as high as $r = 0.85$, the relatedness measures did not surpass $r = 0.79$. In a second experiment, Budanitsky and Hirst compared the performance of the five measures for the task of real-word spelling error correction using lexical chains, as described by Hirst and St-Onge [15]. Once again, similarity measures outperformed relatedness measures, even though the application is believed to rely on measuring relatedness.

There are several possible interpretations of Budanitsky and Hirst's results. On the one hand, it could be that the distinction between similarity and relatedness is, in practice, unimportant. That is, given that similarity constitutes a particular type of relatedness, it may be the case that other kinds of relatedness occur very infrequently and for the purposes of most applications may be safely ignored.

On the other hand, it is possible that relatedness is significantly different from similarity, but that current relatedness measures are fundamentally flawed. In this case it would be appropriate to reject previous approaches to measuring relatedness and to seek new directions for future work. For example, as similarity measures have proved to be rather good proxies for relatedness measures, it may be reasonable to take these measures as starting points and then to modify them to accommodate the unique properties of relatedness.

Finally, it is also possible that previous relatedness measures have taken the correct general approach, but have failed in their details. It is this interpretation that will be adopted in this study. In particular, a number of researchers have developed measures that compute relatedness from the shortest path between concepts in a semantic network, including Sussna [44], and Hirst and St-Onge [15]. We will attempt to show that the path-based approach is viable, in spite of the mediocre results of previous measures of this type.

3.1 New Methodology for Path-Based Relatedness Measures

In order to improve the quality of path-based measures, we propose a more systematic way of constructing and evaluating path-based models. Previously, the authors of relatedness and similarity measures put forward complex and sophisticated models, and compared them to the equally complex and sophisticated models of other researchers. However, these complex models often consisted of numerous independent elements.

For example, many authors have proposed different weighting factors for the edges in the semantic network. In some cases, several different factors have been combined in a single measure. In Sussna's [44] measure, the weight of edges is scaled by both the type-specific fanout factor, and also by a depth factor. It may be the case that one of these techniques improves the accuracy of the measure, and the other one reduces its accuracy. In order to determine whether this is the case, the two techniques must be evaluated separately.

Our conjecture is that while some of the features of previous path-based relatedness measures are beneficial, others are not. In order to evaluate this conjecture, we will first enumerate the various features of path-based relatedness measures that have appeared in the literature. It will be shown that these features are both independent and compatible with one another, so that any combination of these features in a measure is possible.

Once the set of possible features has been identified, a general description of path-based semantic distance measures will be provided. This description will demonstrate how the different possible features may be combined, and will also identify the core elements shared by all path-based measures. A general description along these lines provides a modular view of relatedness measures. Given such a description, it will be possible to produce new measures with different subsets of features. These permutations can be evaluated in order to identify the best measure possible using available features. Essentially, we are proposing that current measures be reduced to their constituent parts so that these may be recombined to construct more successful measures.

After we establish a general description of relatedness measures, we will describe a baseline path-based relatedness measure derived from the general formula. This new measure is based on two simplifying assumptions. Namely, that:

1. The edges in a semantic network represent uniform semantic distance.
2. The sum of edge weights in a path maps directly to relatedness, i.e., no other properties of the path need be considered.

All of the path-based measures that have been described in this study reduce to the same simplified measure when these two assumptions are adopted. In Chapter 4, the simplified measure will be used as a baseline for examining previous relatedness and similarity measures. It will be shown that current measures do not outperform the baseline version when compared to human judgment, calling into question some of the basic assumptions of these models.

3.2 Features of Path-Based Measures

There are several categories of techniques used by semantic distance measures to improve upon the basic shortest path length approach. First, many authors have proposed means of determining the semantic distance that is represented by individual edges in the WordNet graph. A second category of techniques modifies the

weight of a whole path, rather than that of individual links. Some measures also employ mapping functions that transform the sum of edge weights of the shortest path between concepts into a value of similarity or relatedness. Finally, some measures restrict the types of semantic relations, or the combinations of types of relations, that may appear in paths between concepts when calculating semantic distance.

The purpose of weighting techniques is worth considering at this point. On the one hand, techniques for weighting either edges or paths in a semantic network may reflect fundamental properties of relatedness. It could be that a weighting technique captures something about the way in which humans determine relatedness, for example. On the other hand, it may be that weighting techniques are necessary only to compensate for deficiencies in the lexical resource that is being used.

Semantic networks can vary in quality in several ways. Not every network is equally comprehensive either in terms of its coverage of possible concepts, or in terms of the richness of its connectivity. For example, Roget's thesaurus includes a much wider range of semantic relationship types than WordNet does [5]. Also, even the same semantic network can be internally inconsistent in its quality. For example, the part of the WordNet noun taxonomy pertaining to biological organisms is quite rich [34], owing to the extensive taxonomy that has already been developed in biological science. Other parts of the WordNet taxonomy may be less well-developed, leading to variations in the length of paths between concepts of seemingly equal semantic distance. For example, the shortest path connecting the concepts for *horsefly* and *insect* in WordNet 2.0 has a length of four. This is an unusually long path given the apparently close relationship between these concepts. By way of contrast, the concepts *philosopher's stone* and *entity* are connected by a path of only three relations.

The distinction between techniques that compensate for a deficient resource and those that capture essential properties of relatedness is important because it suggests that not every technique will continue to be effective when moved between resources. A measure that is successful when using a particular semantic network may be less successful using a different one. This suggests that it may be necessary to revise and re-evaluate semantic distance measures as the lexical resources that they use change.

3.2.1 Edge-Weighting Techniques

A key assumption made by the authors of path length–based similarity and relatedness measures is that the edges in semantic networks do not represent uniform

semantic distances. Therefore many authors have proposed schemes for more accurately determining the weight — in terms of similarity or relatedness — of edges in the graph. These techniques include depth-relative scaling, type-specific fanout factor, weighting by relationship type, and information content scaling.

Depth-Relative Scaling

Several authors [44, 34, 47] have noted that concepts that are higher in a conceptual taxonomy are separated by greater conceptual distance. They argue that general concepts near the top of the hierarchy represent very broad distinctions, whereas concepts lower in a taxonomy represent much finer distinctions. For example, the concepts for *physical entity* and *abstract entity* seem to be further apart from one another in meaning than *redheaded woodpecker* and *downy woodpecker*. However, in the WordNet taxonomy both of these pairs are sibling concepts that are separated by the same number of edges.

At first glance, depth-relative scaling seems to apply more naturally to the problem of measuring similarity than it does to measuring relatedness. Relatedness does not depend on the IS-A taxonomy in the same way that similarity does. For relations that do not constitute a taxonomy there is no evident way of computing the depth of nodes in the semantic network.

Nevertheless, some authors have attempted to generalize the notion of depth-scaling to encompass other semantic relations. For example, for the purpose of adjusting the weights of links based on taxonomic depth, Sussna computes depth based on hypernym, hyponym, meronym, holonym, and antonym relations.

The evidence offered for depth-relative scaling consists primarily of motivating examples. Patwardhan et al. [30] provide a fairly typical argument, when they point to the distance between the concepts *fire iron* and *implement* and compare it to the distance between *mouse* and *rodent*. Both of these pairs are related directly in the WordNet IS-A hierarchy, but the difference between *fire iron* and *implement* seems to be much greater than that between *mouse* and *rodent*. The former pair is also higher in the taxonomy, with a depth of seven compared to a depth of eleven for the latter.

While it is easy to provide examples that support depth-relative scaling, it is also easy to provide examples that undermine it. The relation *written material* IS-A *written communication*, at a depth of five, is near the top of the hierarchy and represents a narrow semantic distance. Similarly, the relation *product* IS-A *creation* has a depth of only five, and the relation *creation* IS-A *artifact* has a depth of four.

These relations represent subtle distinctions even though they are quite high in the taxonomy.

In spite of any apparent counterexamples, it may be that relations between general concepts *tend* to be semantically distant compared to relations between more specific concepts. The usefulness of depth-relative scaling depends only on a significant correlation between depth and semantic distance, and does not require perfect correspondence. Whether a significant correlation exists is a matter for empirical study.

Type-Specific Fanout

Sussna [44] proposed a means of weighting edges based on the number of edges of the same type that originate from a node. In short, the more relations a concept has of the same type, the less each of these relations contributes to relatedness.

Sussna's type-specific fanout (TSF) factor has some interesting similarities to the information content approach of Resnik [34], Lin [20], and Jiang and Conrath [16]. In information content models, the greater the likelihood of encountering a concept given a second concept, the less distance there is between the two concepts. For example, if it were the case that most instances of *bird* are *robins*, then these concepts would be considered semantically close. For information content models the likelihood of concepts is determined by counting the number of instances in a corpus. In TSF, the likelihood of concepts is determined by counting the sibling concepts, i.e., concepts that are connected by the same relations to another concept. For example, in the case of the *keyboard* HAS-A *key* relation, the chance of encountering *key* given *keyboard* and the HAS-A relation is very high, as *keyboard* has very few other meronyms.

A disadvantage of TSF with respect to the information-content model is that it cannot make the fine-grained distinctions that are possible with the use of a large text corpus. This is a disadvantage because TSF cannot distinguish between relationships of the same type. For example, if the relation *keyboard* HAS-A *cord* existed, it would have the same weight as *keyboard* HAS-A *key*. In this case, TSF would not offer a very good approximation of the semantic distance. That is, the proximity of *keyboard* and *key* does not depend on how many other keyboard parts have been encoded in the semantic network. The number of relations for each node can be arbitrarily large in a semantic network — it is limited primarily by the time and patience of the lexicographers that encode the relations — and the inclusion of less salient relations should not diminish the importance of more salient ones.

Weight by Relationship Type

Perhaps the most straightforward means of weighing relations is to assign semantic distance weights on the basis of relation types. Given that the different edges in the WordNet graph represent a diverse set of semantic relations, this may seem a plausible technique. For example, antonyms have consistently been found to register very highly on word association tests [9]. This may indicate that antonymy represents a stronger degree of relatedness than other semantic relations.

Sussna's [44] measure incorporates a kind of relationship-type weighting scheme. In his measure, each relationship type is constrained to a different range of possible values. For example, hypernymy or hyponymy relations can have values between one and two, whereas antonymy relations have a value of 2.5. Yang and Powers [47] take a more direct route and assign a constant weight to each type of semantic relation in WordNet. Also, Jiang and Conrath [16] provided for a relation-type factor in their measure, although they did not implement or evaluate this feature.

Unfortunately, relationship type represents a very coarse filter. In the noun portion of WordNet, for example, the vast majority of relations are either hypernym/hyponym or meronym/holonym. Assigning weights on the basis of only two types of relation is not likely to offer a significant improvement in the success of a measure.

Information Content

Jiang and Conrath's major contribution was showing how information content can be used to weigh the semantic distance of edges in WordNet's conceptual taxonomy. Their technique does not generalize to semantic relations other than hypernymy and hyponymy, and so may not be a useful technique for relatedness measures. However, it is possible to weigh only the IS-A links in a path of mixed relationship types. As Jiang and Conrath's measure has been found to be very effective, adopting information content scaling for semantic relatedness measures may be worth investigating.

3.2.2 Path-Weighting Techniques

Although edge weighting is one of the most common approaches to enhancing path-based measures, some authors have also introduced methods that scale the value of relatedness using properties of whole paths instead of individual relations. The

path-weighting techniques of the measures that we are considering include depth-relative scaling, weighting by relationship types, and functions that map semantic distance to similarity or relatedness.

Depth-Relative Scaling

Leacock and Chodorow [18] suggested a kind of depth scaling for paths on the basis of the maximum depth of the semantic network. Their technique is of limited usefulness, as it is meaningful only when comparing measurements that use different semantic networks. For example, when comparing only WordNet-based measures, this technique has no effect. The purpose of scaling by the maximum depth of the taxonomy in Leacock and Chodorow seems to be primarily to obtain values between zero and one, in order to apply a logarithmic transformation.

Weight by Relationship Type

In the measure by Yang and Powers [47], a weight is applied not only to the individual edges in the shortest path connecting concepts, but also to the entire path, depending on the types of relations that it contains. For their tests, they scaled any paths consisting of only synonyms or antonyms by a factor of 0.9. Paths that consist of a mixture of hypernyms, hyponyms, meronyms and holonyms are scaled by a factor of 0.85. In Yang and Powers' model these are the only possible types of path. The close coupling of allowable paths and path-weighting make the technique difficult to generalize. In particular, if fewer constraints are placed on the types of paths that are allowed, it would be necessary to supply scaling factors for a very large number of potential path types.

However, the path-weighting technique by Yang and Powers plays an extremely minor role in their model, and it is possible to eliminate it entirely with a small change to their measure. As described in the last chapter, Yang and Powers' formula for similarity is:

$$sim_{YP}(c_1, c_2) = \begin{cases} \alpha_t \prod_{i=1}^{dist(c_1,c_2)} \beta_{t_i} & \text{if } dist(c_1, c_2) < \gamma \\ 0 & \text{if } dist(c_1, c_2) \geq \gamma \end{cases} \tag{3.1}$$

where α_t is the path weight factor for the path type t, and β_{t_i} is the edge weight factor for the ith edge in the shortest path between c_1 and c_2 with edge type t.

To eliminate α_t with minimal impact on the measure, we first add an additional term to the product, by starting with i at zero instead of one. Paths with a length of one are thus equal to the link weight factor of the first, and only, edge in the path. Removing the path-weighting factor α_t, their measure becomes:

$$sim'_{YP}(c_1, c_2) = \begin{cases} 1 \quad \prod_{i=0}^{dist(c_1,c_2)} \beta_{t_i} & \text{if } dist(c_1, c_2) < \gamma \\ 0 & \text{if } dist(c_1, c_2) \geq \gamma \end{cases} \tag{3.2}$$

Paths that contain a synonym or antonym cannot contain any other relations, and therefore always have a length of one. If we set $\beta_{sa} = \alpha_{sa}$, then antonym and synonym paths will have the same value as in the original measure. That is, using the original similarity equation (Equation 3.1) for a path between a_1 and a_2 where a_1 is an antonym of a_2, gives:

$$\begin{aligned} sim_{YP}(a_1, a_2) &= \alpha_{sa} \prod_{i=1}^{1} \beta_{sa} \\ &= \alpha_{sa} \end{aligned}$$

With the new Equation 3.2, and using $\beta_{sa} = \alpha_{sa}$, we have:

$$\begin{aligned} sim'_{YP}(a_1, a_2) &= 1 \prod_{i=0}^{1} \beta_{sa} \\ &= \beta_{sa} \\ &= \alpha_{sa} \end{aligned}$$

As for paths of hypernym, hyponym, meronym, and holonym relations, the new formula in Equation 3.2 results in a product of the same number of terms as previously, but the first would be β_{hm} instead of α_{hm}. For example, a path of three nodes and two edges would have previously been $\alpha_{hm} \times \beta_{hm} \times \beta_{hm} = 0.85 \times 0.7 \times 0.7$, but now will be $\beta_{hm} \times \beta_{hm} \times \beta_{hm} = 0.7 \times 0.7 \times 0.7$. The value of α_{hm} can be adjusted to reduce the difference even further.

Hirst and St-Onge [15] also include a sort of relationship type path-weighting scheme. They subtract the number of changes in 'direction' from the relatedness value. The premise behind this technique is that hypernym/hyponym and meronym/holonym relations are transitive when they are not mixed. That is, a

part of a part is still a part. Similarly, a subtype of a subtype is also a subtype. However, when other relationship types are introduced, the transitivity no longer holds. In this case, the strength of relatedness diminishes, as there is no longer a meaningful relationship between the first and last concepts in the path.

The limitation of having only three categories of semantic relation leads to some strange results, however. For example, paths of similarly directed, but different, relations will be treated no differently than paths composed of only one relationship type. A path that consists of a mixture of meronyms and hypernyms will not receive any penalty for changes in direction because these are both 'upward' relations. However, it would seem that transitivity is not preserved in this situation.

3.2.3 Mapping Functions

Several measures employ functions to transform the sum of edge weights in the shortest path to a final similarity or relatedness value. The purpose of this step may simply be to scale the results of measures to a reasonable and convenient range of values. On the other hand, these mapping functions vary considerably between measures, and likely do have some impact on the success of the measures.

The measures by Hirst and St-Onge [15], and Leacock and Chodorow [18] offer examples of these types of mapping functions. Hirst and St-Onge's measure derives relatedness from path length through a simple linear transformation. They subtract the length of the shortest path between concepts from a constant value, which is sufficiently large to ensure a positive result. Leacock and Chodorow elect a logarithmic function to map the taxonomic path length to similarity.

3.2.4 Pruning Techniques

The final category of techniques found in path-based relatedness measures does not directly adjust the value of relatedness, but instead manipulates the types of path that are considered.

Allowable Patterns

Hirst and St-Onge [15] permit eight patterns of semantic relations in a path, which are defined in terms of their three 'directions' of semantic relations. Yang and Powers [47] offer a similar technique, allowing only a single change of relationship type in any path.

Both of these techniques are motivated by the observation that some semantic relations are transitive. Sequences of transitive relations are assumed to correspond to stronger relatedess or similarity than sequences of different relations. For example, given that *robin* IS-A *bird* IS-A *animal* it follows that a *robin* IS-A *animal*. The semantic distance between *robin* and *animal* is small, owing to the fact that these concepts have a clear semantic relationship to one another.

Edge-Type Restrictions

A more straightforward method for restricting the types of allowable paths is to exclude certain types of semantic relations entirely. For example, most path-based similarity measures allow only IS-A links. Relatedness measures, by contrast, typically also allow meronym, holonym, and antonym relations and some relatedness measures allow additional relationship types.

3.3 Generalized Path-Based Measure of Relatedness

The features of path-based measures described above, including edge-weighting techniques, path-weighting techniques, mapping functions and pruning techniques, represent the elements of path-based measures that vary from measure to measure. These features are largely compatible with one another, and most could in principle be combined in a single relatedness measure. While it is doubtful that such a measure would be effective, a general template that describes how the features could be combined in a single measure would be helpful. This template could be used to generate measures with different subsets of the available features. A general description of semantic distance measures would also make it easier to identify a baseline measure, which would consist of only those elements that are shared by all other path-based measures.

In order to define such a generalized measure, we must first define some notation:

- p is an ordered list of concepts, representing a path between nodes in a semantic network.
- p_i is the ith element of the path p.
- $path(c_1, c_2)$ denotes the set of all paths connecting the concepts c_1 and c_2 in a semantic network.

- $wt_{edge}(c_1, c_2)$ denotes the semantic distance between the adjacent nodes c_1 and c_2.

For most of the measures under consideration, the total semantic distance of a path is calculated in terms of the sum of the semantic distances between each node in the path. These edges are first given weights using whatever edge-weighting schemes are provided for in the measure. In some measures the sum of edge weights is transformed into a similarity or relatedness value by some function. A simple formulation therefore captures the majority of features of existing models. Formally, the total semantic distance of a given path p is given by the following formula:

$$dist(p) = dist_{map}\left(\sum_{i=0}^{|p|-1} wt_{edge}(p_i, p_{i+1})\right) \tag{3.3}$$

In the equation above, $dist_{map}(x)$ is a function that maps the sum of edge weights to similarity or relatedness, and $wt_{edge}(c_1, c_2)$ is a function that determines the semantic distance represented by the edge connecting the concepts c_1 and c_2. The preceding equation calculates the semantic distance of a given path, but it is necessary to select which path in particular should be used, as concepts may be connected by many paths. One option is to choose the path that minimizes the weighted semantic distance:

$$rel(c_1, c_2) = \min_{p \in \{path(c_1, c_2)\}} [dist(p)] \tag{3.4}$$

Although this may be the most principled approach for determining the total semantic distance, it is computationally intractable. To determine the path with the least semantic distance, it is necessary to find every possible path between the nodes.

If the measure does not include any path-weighting schemes, and semantic distance is calculated as the sum of edge distances, then Dijkstra's [11] algorithm may be used to find the minimum weighted path. However, for a graph as large as WordNet, Dijkstra's algorithm is still quite inefficient. The worst-case complexity of Dijkstra's algorithm is $O(n^2)$, where n is the total number of nodes in the graph. WordNet 2.0 has more than 100000 nodes, which results in a very long running time for Dijkstra's algorithm. As the WordNet graph is fairly sparse — nodes are connected to only a limited number of other nodes — more efficient variations of Dijkstra's algorithm may be used. However, even optimized versions of the algorithm have a worst-case running time that grows with the size of the network.

As a result of the computational complexity of finding the path that minimizes the weighted graph distance, it is conventional to choose the path with the fewest edges and to compute semantic distance using this path. Identifying the path with the fewest edges is much easier than identifying the path with the minimum total weight. It can be accomplished using a simple breadth-first search. Using the shortest path, instead of the minimum weighted path, the equation for the relatedness between the concepts c_1 and c_2 is:

$$rel(c_1, c_2) = dist\left(\underset{p \in \{path(c_1,c_2)\}}{\arg\min} |p|\right) \tag{3.5}$$

where $dist(p)$ is Equation 3.3 for calculating the semantic distance of a path p, and $path(c_1, c_2)$ denotes the set of paths connecting the concepts c_1 and c_2 using a subset of the WordNet semantic relationship types, $Rel \subset Rel_{WordNet}$.

The many possible edge-weighting techniques must be combined in some way in the general measure. In all of the measures that have employed multiple edge-weighting techniques, including Sussna [44] and Jiang and Conrath [16], the edge-weighting factors are combined as a simple product. Thus, where Wt_{edge} is a set of weighting functions for two adjacent concepts c_1 and c_2, the total edge weight factor is:

$$wt_{edge}(c_1, c_2) = \prod_{wt_{edge_i} \in Wt_{edge}} wt_{edge_i}(c_1, c_2) \tag{3.6}$$

Combining Equations 3.3, 3.5 and 3.6 gives the full general semantic distance measure. The measure has three parameters, including the mapping function $dist_{map}$ in Equation 3.3, the set of allowable WordNet relations $Rel \subset Rel_{WordNet}$ in Equation 3.5, and the set of edge weight factors Wt_{edge} in Equation 3.6.

3.3.1 Fitting Measures to Generalization

The generalization described above in Equations 3.3, 3.5 and 3.6 accommodates the measures by Leacock and Chodorow [18], Jiang and Conrath [16] and Sussna [44]. For example, Sussna's relatedness measure is equivalent to the general measure with the following parameters:

- $dist_{map}(x) = x$
- $Wt_{edge} = \{wt_{tsf}, wt_{depth}\}$
- Rel ={hypernym,hyponym,meronym,holonym,antonym}

For Leacock and Chodorow's measure of similarity the parameters are:

- $dist_{map}(x) = -log(x/2D)$
- $Wt_{edge} = \{\}$
- Rel ={hypernym,hyponym}

Finally, for Jiang and Conrath's hybrid measure of similarity, we have:

- $dist_{map}(x) = x$
- $Wt_{edge} = \{wt_{ic}\}$
- Rel ={hypernym,hyponym}

The final two path-based measures that are being investigated do not fit as cleanly into our generalization. The measures by Hirst and St-Onge [15], and by Yang and Powers [47], are the only measures to use properties of whole paths, as opposed to individual links, to determine semantic distance. Hirst and St-Onge's technique of discounting the degree of relatedness for each change of 'direction' cannot be expressed in terms of edge-weighting. Similarly, the link type factor by Yang and Powers is applied only once to a whole path, and so cannot be formulated as an edge-weighting technique.

In the case of Yang and Powers, we showed that the link type factor plays a minor role in their measure, and can be eliminated with only a small change. Aside from the link type factor, the measure by Yang and Powers also differs from others in that similarity is calculated as the product of edge weights, rather than as their sum. However, with the link type factor removed this issue is also easily resolved. The Yang and Powers measure is equivalent to the general description with the following parameters:

- $dist_{map}(x) = 0.7^x$
- $Wt_{edge} = \{wt_{yp}\}$
- Rel ={hypernym,hyponym,meronym,holonym,antonym}

The edge weight function $wt_{yp}(c_1, c_2)$ for the concepts c_1 and c_2 is:

$$wt_{yp}(c_1, c_2) = \begin{cases} 1 & \text{if } c_1 \text{ and } c_2 \text{ are related by antonymy or synonymy} \\ 0.295 & \text{otherwise} \end{cases} \quad (3.7)$$

The formulation above is possible because paths in Yang and Powers' measure cannot contain mixed edge weights. There are only two types of allowable path in Yang and Powers' measure: paths consisting of a single synonym or antonym relation, and paths consisting of a combination of hypernym, hyponym, meronym, and holonym relations. In both cases, the edge weights are constant for all of the edges in the path. The edge weights for hypernym, hyponym, meronym, and holonym relations are the same, and the weights of synonyms and antonyms are also the same. In Yang and Powers' notation, the weight of a hypernym or hyponym is denoted β_{hh} and the weight of a meronym or holonym is β_{hm}. The value for both β_{hh} and β_{hm} is 0.7. The weight for a synonym or antonym relation, denoted by β_{sa}, is 0.9.

Because the weights of edges in paths are always constant, it is possible to reformulate the product of edge weights as a power of edge weight. In order to do this and maintain identical output to the original measure, the values of the link weighting factors will be changed. The new values of the link weight factors β_t, for each relationship type t, will be denoted by β'_t. As mentioned above, every path with a length greater than one consists of hypernyms, hyponyms, meronyms, and holonyms. All of these relations have the same weight of $\beta_{hm} = \beta_{hh}$, and therefore the product of the weights is $(\beta_{hm})^n$, for a path of n edges. This quantity can be expressed in terms of the sum of edge weights if we set $\beta'_{hm} = \beta'_{hh} = 1$. With this change, the sum of edge weights is equal to the length of the path. The total path weight can now be expressed as $(\beta_{hm})^n$, where n is the sum of edge weights of the path. Substituting the value of 0.7 for β_{hm}, the semantic similarity of the Yang and Powers measure is 0.7^n.

It is also possible to express the total weight of synonym and antonym paths in terms of the sum of edge weights. Synonym and antonym paths always have a length of one, and so the sum of the edge weights is equal to the weight of a single synonym or antonym edge, which is β'_{sa}. The new formula for similarity is $(\beta_{hm})^n$, where n is the sum of edge weights. The sum of edge weights for a synonym or antonym path is β'_{sa} and therefore using the new formula, the total similarity for such a path is $(\beta_{hm})^{\beta'_{sa}}$. The value of $(\beta_{hm})^{\beta'_{sa}}$ must equal β_{sa} to be consistent with the original version of the measure. We must therefore determine β'_{sa} by solving

the equation:

$$\begin{aligned}(\beta_{hm})^{\beta'_{sa}} &= \beta_{sa} \\ \beta'_{sa} &= \log_{\beta_{hm}} \beta_{sa} \\ &= \frac{\log \beta_{sa}}{\log \beta_{hm}} \\ &= \frac{\log 0.9}{\log 0.7} \\ &= 0.295\end{aligned}$$

There are only two types of path in Yang and Powers' measure, and we have shown that a formulation of their measure in terms of our general description of path-based semantic distance measures is equivalent to the original formulation for both types of path.

The Hirst and St-Onge discount factor for changes in 'direction' cannot be accommodated by the generalization in its current form. It is possible to introduce another term to Equation 3.3 that would provide for Hirst and St-Onge's technique. However, as the principal purpose of the generalization is to show the common elements of the measures, we prefer to exclude the unique feature of Hirst and St-Onge's measure. Excluding the direction change discount factor, Hirst and St-Onge's measure corresponds to the following parameters of the general measure:

- $dist_{map}(x) = 8 - x$
- $Wt_{edge} = \{\}$
- $Rel = Rel_{WordNet}$

3.4 Simplified Path-Based Measure of Relatedness

We wish to establish a baseline relatedness measure, which represents a simplification of the path-based measures that are being investigated in this study. This baseline measure will be used to test some of the core assumptions shared by many relatedness measures. We will demonstrate that existing path-based measures are overly complex, and that they include elements that impede their performance.

Ultimately, all of the features found in semantic distance measures should be evaluated individually. However, in this study, only two types of feature will be

examined: the mapping function from sum of edge weights to relatedness, and the allowable edge types for paths. These two factors have been chosen in part because it is not possible to eliminate them through any simplifying assumptions. One of the goals of this study is to determine a baseline measure of relatedness by simplifying previous path-based measures. For most features that we have considered, what constitutes a simplification is more or less evident. For example, a measure that does not use any edge weighting schemes is simpler than a measure that does. However, it is not clear that a measure that uses any particular set of edge types is simpler than a measure that uses any other set of edge types.

The baseline measure will be derived from two simplifying assumptions. The assumptions that we are adopting are that:

1. The edges in a semantic network represent uniform semantic distance.
2. The sum of edge weights in a path maps directly to relatedness, i.e., no other properties of the path need be considered.

The first assumption means that no edge-weighting techniques will be used in the simplified measure. The weights of edges will be assumed to be constant, and will be assigned a value of one. Thus the sum of edge weights will be equivalent to the length of the shortest path between concepts for our simplified measure. The second assumption is targeted at the path-weighting schemes of Hirst and St-Onge, and of Yang and Powers. These schemes assume that certain combinations, or patterns, of relationship types affect the semantic distance represented by a path. We will adopt the alternative assumption that semantic distance is not affected by any combinations of relation types in the shortest path.

Of course, it is likely that semantic distance does in fact depend on the types of relations, and their combinations, in a path. The purpose of assuming that these factors are not significant is to demonstrate that previous attempts to estimate these factors have not been successful. If our simplified measure achieves results that are as good as the more complex measures, then the additional features of these measures should be rejected until new evidence is offered in their support.

In terms of the generalized relatedness measure, the simplified measure has $Wt_{edge} = \{\}$. However, the other parameters are not constrained by our two simplifying assumptions. That is, we must decide what set of relationship types should be allowed in paths connecting concepts, and which function ought to be used to map the path length to relatedness. To allow a valid comparison to other measures, these parameters will be varied to match each of the other measures that we are examining.

The possible values for the set of allowable relationship types will include Rel ={hypernym, hyponym}, Rel ={hypernym, hyponym, meronym, holonym, antonym}, and $Rel = Rel_{WordNet}$. The performance of the measure for each of these possibilities will be evaluated in the next chapter. When comparing our baseline measure to other measures, we will use the same set of relationships for both measures. The optimal set of allowable edge types will be determined separately from any comparisons between the proposed measure and previous measures.

The other parameter of the generalization that will vary in the simplified measure is the function that is used to relate the sum of path weights to a relatedness value. Again, we will evaluate all of the functions used by other measures, including the linear model of Hirst and St-Onge [15] and the logarithmic model of Leacock and Chodorow [18]. In addition to these, several other mapping functions will be considered. Each of these functions will be evaluated in the next chapter.

To summarize, we are proposing a simplified relatedness measure that computes relatedness as a function of the shortest path connecting two concepts in a semantic network. This measure is a simplification of previous work, and will serve as a baseline for evaluating other measures. The simplified measure will also serve as a starting point for the systematic development of new measures. In the next chapter, different subsets of semantic relations to be used when searching for paths will be compared. Several different functions for mapping path length to relatedness will also be evaluated.

The equation for the simplified relatedness measure for two concepts c_1 and c_2 can be given as follows:

$$rel(c_1, c_2) = rel_{map}\left(\min_{p \in path(c_1, c_2)} |p|\right) \tag{3.8}$$

where rel_{map} denotes a function that maps path lengths to relatedness values, and $path(c_1, c_2)$ denotes the set of all paths connecting the concepts c_1 and c_2 using the set of relationship types $Rel \subset Rel_{WordNet}$.

3.5 Relatedness Functions

In the simplified relatedness measure described above, the function $rel_{map}(x)$ transforms the length, denoted by x, of the shortest path between two concepts in a semantic network into a relatedness value. There are infinitely many functions that could accomplish this, but we will try to select a few plausible candidates. Five

functions will be described, three of which have been used in other relatedness measures. The functions will be described in general terms at this stage, with unspecified constant values. In the next chapter, further details of these functions will be provided for the purposes of evaluation.

3.5.1 Linearly Decreasing

The first class of functions for mapping path length to relatedness value that we will consider are linear functions. Adopting a linear function implies that the strength of relatedness decreases from some initial value (the maximum possible relatedness value) at a regular interval for each link in the path. The size of the interval will be represented by the constant k, and the initial value as m:

$$rel_{Linear}(x) = \begin{cases} m - kx & \text{if } m - kx >= 0 \\ 0 & \text{if } m - kx < 0 \end{cases} \tag{3.9}$$

Hirst and St-Onge use a linear function to map the length of the shortest allowable path to relatedness. Specifically, excluding the discount factor for the number of changes in direction, they use the above function with $m = 8$ and $k = 1$.

3.5.2 Exponentially Decreasing

A second possible relationship between path length and relatedness is that relatedness decreases as a power of path length. If the power is positive and greater than one, then in this model the relatedness decreases at an exponentially increasing interval for each link in the path. Let a represent the maximum relatedness value, k is a constant value greater than 1, and a and b are constants. The relatedness function is then given in Equation 3.10, and the plot of a sample exponentially decreasing function may be found in Figure 3.1(c).

$$rel_{Exp}(x) = \begin{cases} a - bx^k & \text{if } a - bx^k >= 0 \\ 0 & \text{if } a - bx^k < 0 \end{cases} \tag{3.10}$$

A model in which relatedness decreases at a growing rate may be plausible from a psycholinguistic point of view. When searching for the shortest path length between senses, each increase in path length corresponds to an exponential increase

in the number of concepts that must be considered. If humans must undertake a graph search similar to that used in our path search algorithm, then the relatedness of a pair of concepts may correlate with the effort required in finding their relationship. Although the assumption of a graph-like cognitive arrangement of lexical information in humans is contentious, this possibility lends some additional incentive for investigating the exponentially decreasing function.

3.5.3 Exponential Decay

A third function for relatedness is that of exponential decay. In this model, the relatedness value decreases at increasingly small intervals as path length increases, approaching but never reaching zero. The function for exponential decay is given in Equation 3.11, and a sample plot is given in Figure 3.1(b).

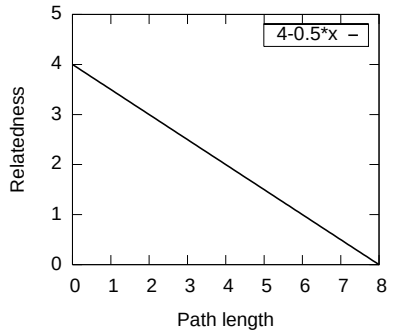
$$rel_{Decay}(x) = ae^{-kx} \tag{3.11}$$

Exponential decay is somewhat difficult to reconcile with psycholinguistic theory, since it is unlikely that a human could detect arbitrarily weak relationships between word senses. On the other hand, exponential decay possesses some desirable mathematical properties. Exponential decay handles cases of very long path lengths more elegantly than the other measures, and can distinguish between cases of very low relatedness and cases of no relatedness whatsoever. Even for very large path length values, the function continues to decrease without ever reaching zero.

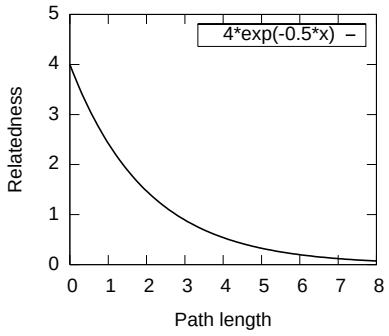
The Yang and Powers measure employs a mapping function that is mathematically equivalent to exponential decay. The similarity function of the simplified version of their measure is $rel_{YP}(x) = b^x$. The constant b can be any arbitrary value, so we let $b = e^{-k}$. In this case $b^x = (e^{-k})^x = e^{-kx}$. Then we let $a = 1$ so that $e^{-kx} = ae^{-kx}$. Thus the simplified Yang and Powers measure uses an exponential decay model that is a special case of Equation 3.11.

3.5.4 Logarithmic

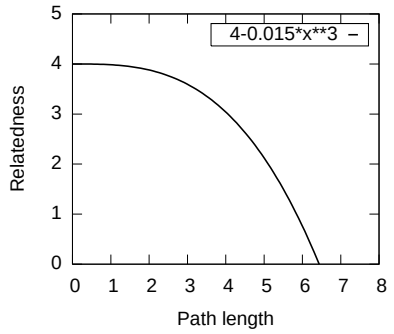
Leacock and Chodorow [18] employed a logarithmic function to map taxonomic path length to similarity. Resnik [34] pointed out that this approach makes their measure "information like," insofar as similarity is calculated using the negative log of a quantity, and thus resembles the formula for information content. However,



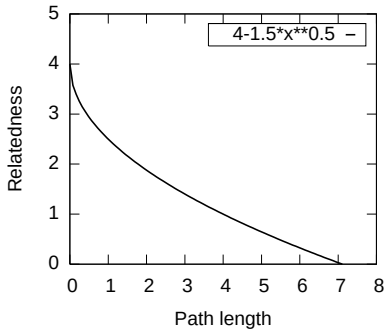
(a) Linear



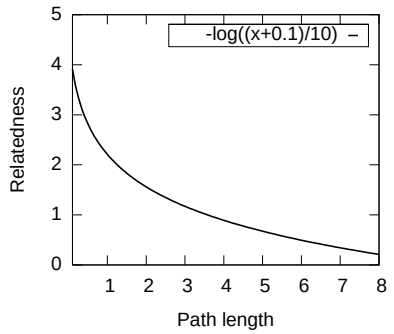
(b) Exponential decay



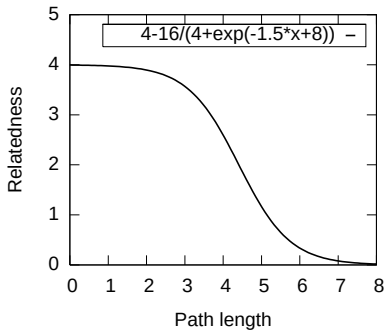
(c) Exponentially decreasing ($k > 1$)



(d) Exponentially decreasing ($k < 1$)



(e) Logarithmic



(f) Sigmoid

Figure 3.1: Examples of functions for mapping shortest path length to relatedness

as they are not applying the formula to a probability, the resemblance is likely not significant. Nevertheless, a log-based function can provide a reasonable curve for mapping path-length to relatedness. A sample logarithmic function is shown in Figure 3.1(e). The equation for the logarithmic mapping function is:

$$rel_{Log}(x) = -\log\left(\frac{x+a}{b}\right) \tag{3.12}$$

3.5.5 Sigmoid

The sigmoid function has been adopted because it combines some of the potentially desirable properties of different mapping functions. This function has an asymptote at the x-axis, ensuring that negative values are impossible, and also allowing for distinctions between very long paths. For short path lengths, however, the sigmoid function behaves like the exponentially decreasing function, with an accelerating rate of decrease as the path length grows longer. A sample sigmoid function is displayed in Figure 3.1(e), and the equation for the sigmoid function is:

$$rel_{Sigmoid}(x) = a + \frac{b}{c + e^{-dx+f}} \tag{3.13}$$

3.6 Chapter Summary

In this chapter, a general description of path-based similarity and relatedness measures was proposed. The general description offers a view of the similarities and differences between path-based measures that have been proposed in the literature. The generalization views each of these measures as consisting of a common model that is augmented with some subset of additional features. Given this modular view of relatedness measures, we have argued that the various features that have been proposed in the literature should be evaluated independently of one another to determine their effectiveness. We conjecture that since some of these features have never been isolated and tested, it is possible that some will be found to be ineffective. Furthermore, as it has not been the custom to evaluate new relatedness measures against any sort of baseline it may be the case that some measures are unnecessarily complex, and do not improve on the success of a simpler measure. In the next chapter it will be shown that, surprisingly, none of the measures evaluated

achieve better results than corresponding baseline measures when compared with human judgments. The next chapter will also examine the effect of different sets of allowable semantic relations in paths between concepts, and the effect of different functions for mapping path length to relatedness.

Chapter 4

Evaluation of Lexical Semantic Relatedness Measures

4.1 Methodology

4.1.1 Evaluation Approaches

As a result of the intense research activity surrounding semantic distance, a fairly consistent evaluation methodology has emerged. Beginning with Resnik [34], the primary basis for the comparison of different measures has been correlation with human judgment. Two experiments, one by Rubenstein and Goodenough [38] and a second by Miller and Charles [25], have provided human semantic distance ratings that have been used in many evaluations. For example, Resnik [34], Yang and Powers [47], Jiang and Conrath [16] and Budanitsky and Hirst [4] all employ comparisons to human ratings for evaluating semantic distance measures. Budanitsky and Hirst's study is particularly useful, as it compares most of the important WordNet-based measures using a common evaluation framework.

Although comparison to human judgments seems to be the most popular approach to evaluating semantic distance measures, Budanitsky and Hirst [4] have noted two other approaches that appear in the literature. Some authors, such as Lin [20] and Rada et al. [33], have attempted to determine the desirable mathematical properties of distance measures in order to provide a purely theoretical evaluation. For example, measures may be analyzed to determine whether they satisfy the properties of metrics, whether they may be projected as smooth functions, and so on. Another evaluation technique identified by Budanitsky and Hirst

is to evaluate performance with respect to a particular application. For example, a study by Patwardhan et al. [30] examined the success of a word sense disambiguation algorithm that relies on relatedness measurements using different relatedness and similarity measures. The measures were rated based on the success of the word sense disambiguation task. The application-based method has the benefit of demonstrating the usefulness of a measure at the same time as offering evidence for its accuracy.

Of the three evaluation approaches, the first is the most meaningful, given that human judgments of semantic distance are presumed correct by definition [4]. In our view, examining the mathematical properties of measures is primarily useful in guiding the development of new measures. For example, theoretical analysis might reveal serious flaws in a measure that would lead to nonsensical output, such as negative semantic distance values. However, when it comes to assessment, mathematical properties are trumped by human judgment. The presence or absence of any theoretical properties are insignificant if they do not impact the performance of the measure compared with the behaviour of human subjects.

Application-based evaluations are valuable in assessing semantic distance measures in cases where direct comparison to human subjects is impossible. However, the indirect nature of application-based evaluation makes it a less reliable approach than direct comparison to human judgments. For example, it is difficult to demonstrate that an application relies exclusively on semantic distance. Given the success or failure of an application in which a semantic distance measure is incorporated, an additional assumption is necessary — namely, that the degree of success of the application correlates with the accuracy of the distance measure. The requirement for this additional assumption weakens the application-based approach compared to that of direct comparison to human judgment.

Due to the weaknesses of the application-based and theoretical approaches, and because an evaluation framework already exists for direct comparison to human judgments, we will use direct comparison to human judgments in this study.

4.1.2 Experimental Data

Rubenstein-Goodenough Data Set

Two data sets in the literature are commonly used for the validation of relatedness measures. The first is a list of 65 word pairs, each rated by 51 human test subjects in an experiment conducted by Rubenstein and Goodenough [38]. The subjects were asked to rate each word pair according to their similarity of meaning, on a

scale from 0.0 ("semantically unrelated") to 4.0 ("highly synonymous"). Only the mean values of human ratings are available, and we will follow Resnik [34] and others in using the mean values for all our comparisons.

Although the instructions given by Rubenstein and Goodenough do not make it clear whether the subjects were meant to rate the semantic relatedness or the semantic similarity of the word pairs, the results indicate that the test subjects rated the word pairs for their relatedness. For example, the word pair *bird/woodland* was given a higher average rating than word pairs that are clearly more similar, such as *lad/wizard* and *monk/slave*. While birds are often associated with woodlands, birds and woodlands are very different things, with little overlap in their properties.

Miller-Charles Data Set

A second data set was collected by Miller and Charles [25], and contains 30 word pairs chosen from the Rubenstein-Goodenough set. Each word pair was rated by 38 subjects, who were given instructions identical to those used in the Rubenstein-Goodenough experiment. The smaller Miller-Charles set will be used to tune the new measure prior to evaluation, and the larger Rubenstein-Goodenough set will be reserved for validation. The complete set of word pairs with mean human ratings for the Miller-Charles and Rubenstein-Goodenough data sets may be found in section $A.1$ of the Appendix.

Resnik Data Set

Resnik [34] replicated the Miller-Charles experiment with 10 human subjects, in order to determine a theoretical limit to the performance of semantic distance measures. He found that the average correlation of the human subjects with the results previously obtained by Miller and Charles was 0.8848, with a standard deviation of 0.08. Resnik therefore determined that a correlation coefficient of 0.9 represents an upper bound on the performance of any computational measure, as this would match the performance of a typical human judge.

However, Resnik also found that the correlation of the mean ratings of his test subjects with the Miller-Charles ratings was 0.96. As others have pointed out [47], it is at least theoretically possible for a measure to approach not only the performance of an individual human, but the combined effort of several humans. Given that the results that are presented below approach the 0.9 upper bound very closely, and in a few cases even exceed it, it may be more reasonable to view the value of 0.96 as the theoretical limit to performance.

4.1.3 Experiment Description

There are several distinct goals for this evaluation. First, we wish to evaluate which subset of relations is the most effective for determining relatedness using network path length. Second, we wish to determine the mathematical relationship between network path length and relatedness. Finally, we wish to compare previous measures of relatedness against corresponding simplified measures, in order to determine whether the techniques used in these measures are effective. These three goals can be achieved through two experiments.

Experiment 1: Evaluating Subsets of Semantic Relations for Path Determination

In the first experiment, we will determine which subset of allowable semantic relations in the paths between synsets results in the best correlation to human judgments. As discussed in the previous chapter, three subsets will be considered: {hypernymy, hyponymy}, {hypernymy, hyponymy, meronymy, holonomy, antonymy} and the set of all WordNet relations. These sets were selected because they are those that have been used in previous semantic distance measures.

In order to evaluate the three sets of relations, we will examine the correlation of the length of the shortest path between concepts with human judgments of relatedness. That is, for the Rubenstein-Goodenough (RG) and Miller-Charles (MC) data sets, the shortest path between each pair of words will be computed using each of the three sets of relations that we are evaluating. The correlation of the calculated path lengths to human judgments will be determined, and the three sets of relations will be compared on the basis of the strength of correlation.

Experiment 2: Evaluating Relatedness Functions

In the previous chapter we suggested five models that can be used to determine relatedness, given path length. These included linear, exponentially decreasing, exponential decay, logarithmic, and sigmoid functions. In our second experiment, we will evaluate these functions with respect to human relatedness judgments. However, these functions include variables representing unspecified constant values. In order to evaluate these models, these constants must be determined in a systematic way that does not privilege one model over any other.

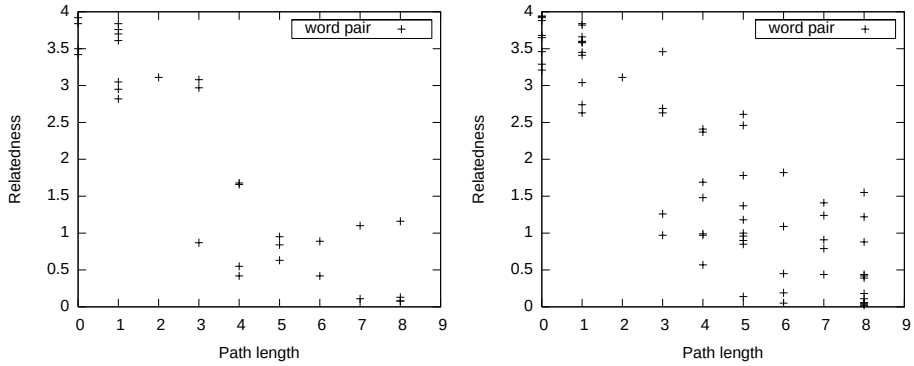
For many previous measures, it is impossible to know to what degree the measures have been 'tuned' by their authors. For example, Yang and Powers derived

the values of constants in their measure very systematically, and they explain the procedure that they used in detail. Hirst and St-Onge's measure also includes constant values, but the authors do not appear to have determined these values in a systematic way. Without a standard tuning framework, it is likely that the performance of measures will be partly attributable to the quality of tuning, and not wholly to the merits of the models themselves.

In our experiment, all of the models will be tuned using the techniques of statistical data analysis. Specifically, we will use regression to fit the models to a subset of our data. Even if a model is capable of a close fit to available data, it is essential to revalidate the models on data that was not used for tuning. The problem of over-fitting can arise because a model becomes tuned to satisfy incidental properties of a particular data set that do not generalize to other data. This problem is particularly dangerous with highly complex models (such as high-order polynomial functions) applied to small data sets, since the model is sufficiently flexible to fit itself to minor features of the data.

In order to avoid the problem of over-fitting, we will reserve part of the data for evaluation. For the regression calculations, the Miller-Charles data set will be used, and the Rubenstein-Goodenough data set will be reserved for evaluation. However, the Rubenstein-Goodenough word pairs are a superset of the Miller-Charles word pairs. Therefore, we will also examine the subset of Rubenstein-Goodenough word pairs that are not in Miller-Charles (i.e. the set-theoretic difference between the Rubenstein-Goodenough and Miller-Charles sets). This set contains word pairs that were not involved in tuning, and the results for this set will be considered to be the most significant.

The most widely used curve-fitting techniques are linear and nonlinear regression. These are both iterative techniques that seek to minimize the sum of squared errors for a function (the curve) over a set of data points. In the case of the proposed relatedness measures, the curves are functions that map path length to relatedness value, specifically rel_{Linear}, rel_{Decay}, rel_{Exp}, rel_{Log} and $rel_{Sigmoid}$. By determining the shortest path length between each word pair in the data set, regression may be used to fit the relatedness functions to the experimentally determined values. To illustrate, Figure 4.1 plots the calculated shortest path lengths against the average relatedness values collected in the Rubenstein-Goodenough and the Miller-Charles experiments. Word pairs for which no path could be found have been assigned the maximum search depth plus one, which in this case is eight. The goal of regression is to tune the relatedness functions so as to achieve the best possible fit to the Miller-Charles data points shown in Figure 4.1(a). Once the measures have been tuned to the Miller-Charles data, they will be evaluated using the Rubenstein-Goodenough



(a) Miller-Charles data (b) Rubenstein-Goodenough data

Figure 4.1: Shortest calculated path lengths against average human ratings for the Rubenstein-Goodenough and the Miller-Charles experiments

data shown in Figure 4.1(b).

Basis for Comparison

In order to determine the strength of correlation between computational measures and human judgments, most authors have followed Resnik [34] in employing Pearson's coefficient of correlation. As the coefficient of correlation measures the strength of the linear association between two variables, it is a convenient way to summarize the accuracy of measures.

However, some authors have argued that the problem of relatedness is essentially a ranking problem. That is, the important question that must be answered by relatedness measures is whether a given pair of concepts is more or less semantically related than another pair. If this is the case, then the value of relatedness is only useful insofar as it distinguishes a more related concept pair from a less related pair. A better metric for the performance of a measure would then be Spearman's rank correlation coefficient. The rank correlation is, essentially, the correlation of the ranks of the data points. That is, the most related word pair is assigned the value one, the second is assigned the value two, and so on. These rank values are then compared to the rank values of the validation data using Pearson's coefficient of correlation. We will provide both the correlation and the rank correlation for

our experiments.

Each measure will be evaluated against three data sets: the Rubenstein and Goodenough (RG) set, the Miller and Charles (MC) set, and the set theoretic difference of the RG and MC data sets ($RG \setminus MC$). The third of these has been selected to provide a validation set that does not use any word pairs that were involved in tuning the models. Although the RG data set was not used for tuning, the MC word pairs are a subset of the RG word pairs. This means that some of the word pairs in RG were involved in tuning, albeit with independent human ratings. The $RG \setminus MC$ data set will use the RG human ratings, and contains only word pairs that were not used for tuning.

The results for the $RG \setminus MC$ data will be considered the most meaningful basis for comparison. However, as the RG data set has the most data points, and as the RG ratings have not been used for tuning, the results for the RG set will also be taken into consideration.

4.2 Implementation

The relatedness measures were implemented in Perl, using the WordNet::QueryData package for access to the WordNet 2.0 lexical database. Perl was chosen principally because of the existence of Pedersen's popular WordNet::Similarity [31] package, which contains up-to-date Perl implementations of the most important WordNet-based similarity and relatedness measures. The proposed measures were implemented as additional modules for the WordNet::Similarity package, allowing convenient, uniform access to all of the measures. An additional advantage of implementing the new measures as modules of Pedersen's software is that they could be easily distributed in this form to interested researchers.

4.2.1 Search Algorithms

The proposed relatedness measures rely on a search for the shortest path connecting two WordNet synsets. The search has been implemented both as a unidirectional breadth-first search, and as a bidirectional asymmetric breadth-first search. The computational complexity of the two algorithms is equivalent — the worst-case complexity in each case is $O(B^n)$, where B is the maximum number of relations of any node in the graph, and n is the depth of the search.

The unidirectional breadth-first search finds a path from a source node to a target node by examining the neighbours of the source node, and then examining

the neighbours of the neighbours, and so on until the target node is reached. At most, each node in the search 'fringe' — the nodes that we are currently examining — will have B neighbours, as this is the maximum number of neighbours of any node in the graph. Starting from the source node, the number of nodes visited will increase by a factor of at most B for each expansion. To reach a target node that is separated from the source node by n edges, the algorithm requires n iterations, resulting in $B^n + B^{n-1} + \cdots + 1$ visited nodes. That is, in the worst case, we first visit the B neighbours of the source node, then for every one of the B neighbours we visit B more nodes ($B \times B = B^2$), and then for each of the B^2 nodes we visit B more, and so on n times.

A bidirectional search employs the same procedure, but concurrently expands the neighbours of both the source and target nodes. If the two fringes are expanded at the same rate, then each side will require $n/2$ expansions to find a path with a length of n. The total size of the fringe in this search is thus $B^{n/2} + B^{n/2} = 2B^{n/2}$, which is considerably smaller than B^n. However, in the bidirectional search it is much harder to recognize when the search has completed successfully. Instead of testing newly expanded nodes against the target node, we need to test against every node in the opposite fringe. At a given depth n, each node in one fringe must be compared against each node in the other fringe, resulting in $B^{n/2} \times B^{n/2} = B^n$ comparisons. Given the requirement for these comparisons, the order of the worst-case running time is $O(B^n)$, which is the same as that of the unidirectional search. Although the space requirements are less for the bidirectional search, as fewer nodes are visited, there are more operations in total.

However, as a result of the wide range in the degree of connectivity of WordNet synsets, the bidirectional search can be made much more efficient for the average case. Devitt and Vogel [10] have determined that the branching factor in the WordNet noun hierarchy varies between one and 573. The average branching factor, excluding leaf nodes, is 5.793. These metrics include only IS-A relations, and it is therefore possible to encounter even higher branching factors in our search when other relationship types are admitted.

Our bidirectional search takes advantage of the properties of WordNet's topology by selectively expanding the smallest of the two fringes. The algorithm searches from both endpoints, but always elects to expand the side that has visited the fewest nodes so far. The variable number of relations of WordNet nodes results in one fringe growing more rapidly than the other, making the bidirectional algorithm much more efficient than the unidirectional one on average.

Processing the Rubenstein-Goodenough set of word pairs using the unidirectional search took an average of 39.8 minutes of CPU time for each word pair. The

bidirectional version of the search took an average of 3.7 seconds of CPU time for each word pair. The unidirectional search visited an average of 94481 nodes per word pair, and the bidirectional search visited an average of 5535 nodes per word pair. These tests were executed on an AMD Athlon XP 2500+ processor with 768 MB ram. The search was limited to a maximum depth of seven.

The extreme variance in average performance is the result, in part, of a few very long searches dominating the average runtime for the unidirectional search algorithm. The maximum number of nodes visited in the unidirectional search was 1390147, for the word pair *crane/implement*. This word pair alone therefore accounted for nearly a fifth of the total nodes visited in all 65 unidirectional searches. The maximum number of nodes visited for one word pair by the bidirectional search algorithm was only 43894, by comparison.

Neither of the two algorithms that we have described is strictly complete for the task of finding the shortest path between nodes in a directed graph. While most relations in the WordNet graph are reciprocated — meronyms by holonyms, hypernyms by hyponyms, etc. — not all of them are. For example, the CAUSE-TO relation has no corresponding CAUSED-BY relation in WordNet. In practice, non-reciprocated relations do not seem to pose a major problem. This is not entirely surprising, as the vast majority of relations between nouns in WordNet are either undirected, such as antonymy, or reciprocal, such as hypernymy/hyponymy and meronymy/holonymy. The WordNet graph for nouns is therefore effectively an undirected graph for our purposes. Of the 65 word pairs in the Rubenstein-Goodenough data set, the unidirectional and bidirectional searches differed in their results in only three cases. In each case the length of the shortest path found differed by only one. The results of the bidirectional search have been used for the analyses that follow in this chapter.

The success of the bidirectional asymmetric search is may be the result of WordNet satisfying the properties of a *small-world* graph. Small-world graphs were originally identified by Milgram [24] in experiments on social networks. Milgram found that although most of the people (nodes) in social networks are only connected to a few others, it is possible to find surprisingly short paths between any two random people. This phenomenon is now commonly referred to as the "six degrees of separation." More recently, Watts and Strogatz [46] and Strogatz [43] identified many other naturally occurring networks that are instances of small-world graphs. A few examples include the World Wide Web, the power grid of the western United States, and the collaboration graph of film actors. Steyvers [42] demonstrated that many semantic networks are also small-world graphs, including word association networks, Roget's thesaurus, and WordNet.

Small-world graphs have unusually short paths between nodes in spite of being very sparse. This is because small-world graphs are generally *scale-free* graphs, meaning that they have a few very highly connected 'hubs' that serve to connect many different parts of an otherwise sparse graph. The implication for a breadth-first search of such a graph is that once the search reaches a hub, the size of the fringe grows dramatically. An asymmetric bidirectional search is able to circumvent the problem of hubs, to a certain degree. In the case of a path that passes through a highly connected node, the bidirectional search will avoid growing the fringe that has reached the hub. Instead, the second fringe will grow until it intersects the first one, meeting it quite close to the hub. Thus the very expensive expansions that occur after a hub has been reached are minimized.

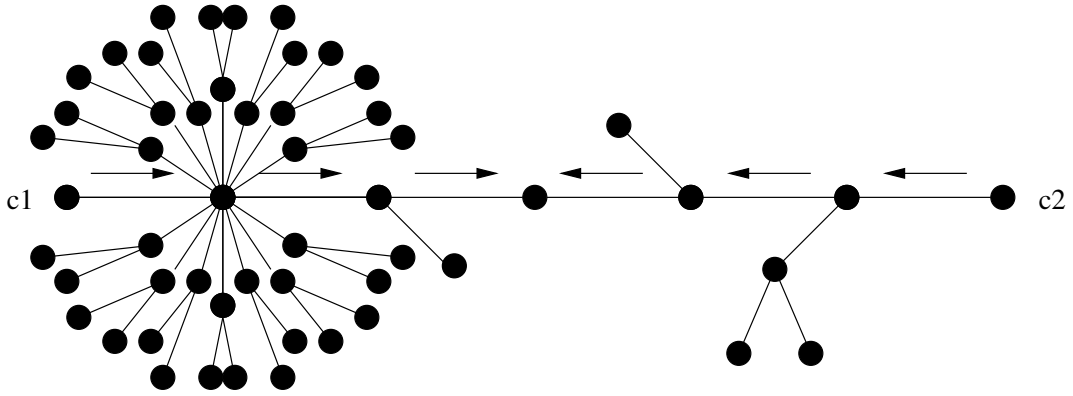
Figure 4.2 illustrates the advantage of an asymmetric expansion in a scale-free network. In Figure 4.2(a), the nodes expanded in a bidirectional symmetric search for a path between two nodes (c_1 and c_2) is shown for a hypothetical semantic network. Figure 4.2(b) shows the network for the same search using an asymmetric expansion of nodes. Fewer nodes in total are expanded in the asymmetric version.

To make any strong claims about the suitability of our algorithm to small-world graphs in general will require an average-case complexity analysis. Although such an analysis should be possible using the formal properties of scale-free networks, it is left to future work.

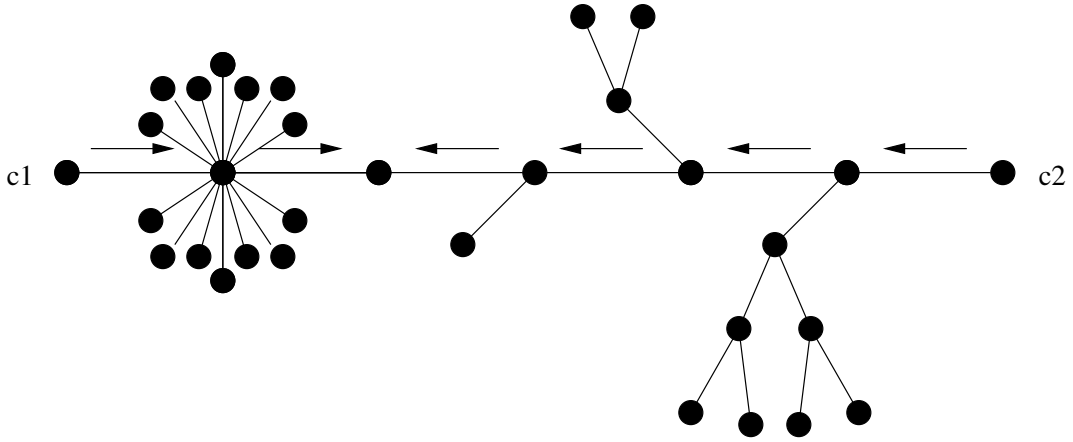
4.2.2 Previous Measures

Implementations of the measures by Hirst and St-Onge [15], Resnik [34], Lin [20], Leacock and Chodorow [18], and Jiang and Conrath [16] are available in Pedersen's [31] WordNet::Similarity Perl module. This package has been widely used and tested. There is no advantage to reimplementing these measures for the purposes of this evaluation. Also, the study by Budanitsky and Hirst [4] provided detailed results from their own tests. As it has been found that the results from Budanitsky and Hirst's implementations are slightly better that those from WordNet::Similarity, we will use Budanitksy and Hirst's results. These differences are likely the result of the particular version of WordNet that was used, as Budanitksy and Hirst used the older WordNet 1.5.

Sussna's [44] measure has not been implemented recently, however. This measure is widely cited, and offers some innovative features not found in other measures. However, Budanitsky [3] encountered two problems when implementing Sussna's technique, and subsequent researchers have not attempted an implemen-



(a) Node expansions in symmetric bidirectional search



(b) Node expansions in asymmetric bidirectional search

Figure 4.2: Example of asymmetric bidirectional search procedure

tation. Both of the problems pertained to Sussna's method of calculating the depth of nodes in WordNet, which uses IS-A and PART-OF relations, as well as antonymy.

The first problem encountered by Budanitksy was the possibility of synsets with multiple antonyms. Although Sussna's measure explicitly assumes no more than one antonym, cases of multiple antonyms exist in WordNet. For example, Budanitsky points out that the concept *introversion* is an antonym of both *extroversion* and *ambiversion* in WordNet. However, Budanitksy suggested a plausible revision to Sussna's method that allows for concepts with multiple antonyms. In Sussna's original formula for the depth of nodes that do not have antonyms, the depth of a node is the average depth of the node's parents, plus one. Formally, where $par_i(c)$ is the ith parent (holonym or hypernym) of the concept c, the depth of c is:

$$d(c) = \frac{1}{m} \sum_{i=1}^{m} d(par_i(c)) + 1 \tag{4.1}$$

For nodes that have antonyms, Sussna calculates the depth of both the node and the node's antonym, as in Equation 4.1. The depth of the antonym is incremented to account for the edge connecting it to the target node, and the total depth is the average of the two values:

$$d(c) = \frac{1}{2} \left[\left(\frac{1}{m} \sum_{i=1}^{m} d(par_i(c)) + 1 \right) + \left(\frac{1}{n} \sum_{j=1}^{n} d(par_j(ant(c))) + 2 \right) \right] \tag{4.2}$$

In order to account for nodes with multiple antonyms, Budanitksy replaced the depth of the antonym in Equation 4.2 with the average depth of all antonyms. Where $ant_j(c)$ is the jth antonym of the concept c, Budanitksy's new equation for depth is:

$$d(c) = \frac{1}{2} \left[\left(\frac{1}{m} \sum_{i=1}^{m} d(par_i(c)) + 1 \right) + \left(\frac{1}{s} \sum_{k=1}^{s} \frac{1}{n_k} \sum_{j=1}^{n_k} d(par_j(ant_k(c))) + 2 \right) \right] \tag{4.3}$$

Although Budanitksy was satisfied with the solution above for the problem of multiple antonyms, he encountered a second problem in Sussna's use of both IS-A and PART-OF relations when searching for the top node of the taxonomy. In recent versions of WordNet, it is possible to encounter cycles in paths to the root node when searching with these relations. Budanitsky abandoned Sussna's measure because of this problem. However, it is not difficult to exclude cycles. We have implemented the depth formula in Equation 4.3, with the change that nodes can be

Model	Function	Function with derived constants
rel_{Linear}	$4 - ax$	$4 - 0.5868x$
rel_{Decay}	$4e^{kx}$	$4e^{-0.2537x}$
rel_{Exp}	$4 - ax^k$	$4 - 0.7573x^{0.8415}$
rel_{Log}	$-\log(a + x/b)$	$-\log(0.0188 + x/21.4130)$
$rel_{Sigmoid}$	$a - \frac{b}{c+exp(dx+e)}$	$3.5474 - \frac{3.6806}{1.2604+exp(-1.7198x+5.5905)}$

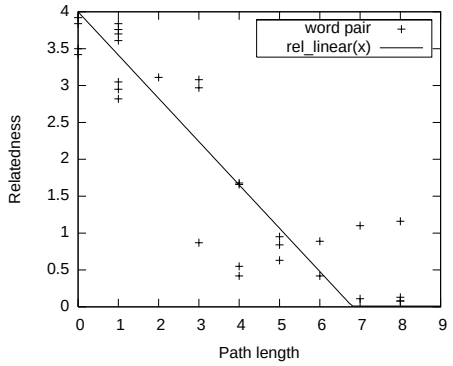
Table 4.1: Relatedness functions with constant values derived using regression

included only once in paths to the top node. This is ensured by maintaining a list of previously visited nodes that is passed to successive calls to the recursive function given in Equation 4.3. No node may be visited that has already been visited in the current search for the top node.

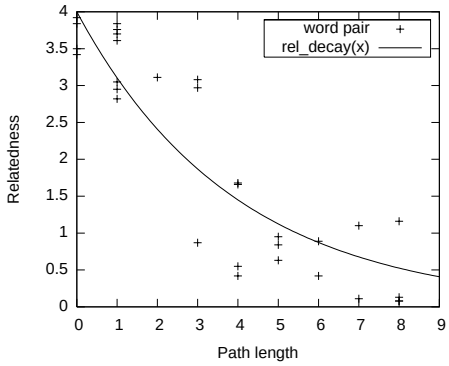
4.3 Tuning the Models Using Regression

The constant values in the five relatedness functions described in the last chapter were derived using Marquardt-Levenburg [21] nonlinear regression with the Miller-Charles data set. Word pairs for which no path could be found have been excluded from the regression calculations. Although we know that these word pairs correspond to path lengths of eight and above, it is impossible to accurately quantify their length. Assigning an arbitrary path length introduces unnecessary error. The final relatedness functions with derived constants are given in Table 4.1. Also, the relatedness functions and the Miller-Charles average human ratings are plotted against path length in Figure 4.3. This figure provides a graphical representation of how relatedness varies as path length increases. The purpose of regression is to minimize the difference between the curves (the relatedness functions) and the data points that are shown in Figure 4.3.

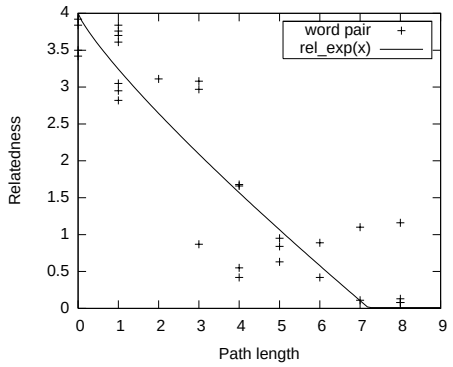
A standard measure of the "goodness of fit" when fitting models using regression is the coefficient of determination, denoted by r^2. This value can be interpreted as the percentage of the data that is accounted for by the model. The values of r^2 for each measure are provided in Table 4.2.



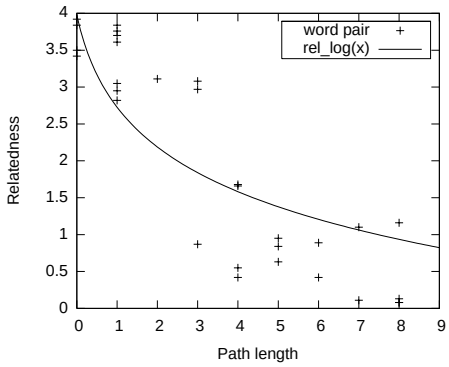
(a) Linear



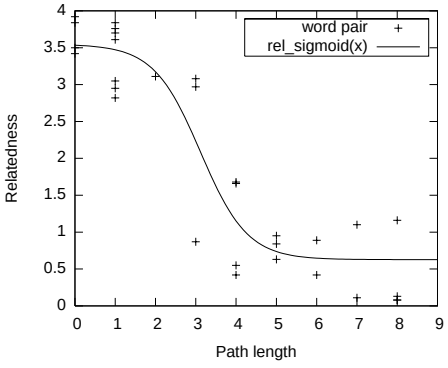
(b) Exponential decay



(c) Exponentially decreasing



(d) Logarithmic



(e) Sigmoid

Figure 4.3: Relatedness functions and average human ratings for Miller-Charles data set

Measure	r^2
rel_{Linear}	0.7939
rel_{Decay}	0.7959
rel_{Exp}	0.8050
rel_{Log}	0.6898
$rel_{Sigmoid}$	0.8573

Table 4.2: Coefficients of determination (r^2) for relatedness functions

The results for r^2 given in Table 4.2 indicate that $rel_{Sigmoid}$ was capable of the best fit to the MC data set. The rel_{Log} function had the worst fit. The other three functions, namely rel_{Linear}, rel_{Decay}, and rel_{Exp}, had nearly identical r^2 values, indicating that there was no significant difference in the quality of their fit to the data. However, these results only give an indication of how well these models explain the values in the MC data set. The evaluation of each of the models against the larger RG data set will be a more important test of their success.

4.4 Results

In this section we describe the results of our evaluation of the proposed simplified measure against previous similarity and relatedness measures. First we will give an overview of the results for each of the two experiments that we conducted, drawing attention to any interesting or unexpected features of the data. Following the overview of the results we will discuss the significance of our findings, including the implications for each of the previous similarity and relatedness measures that are being considered.

4.4.1 Experiment 1: Relationship Types

For the first experiment, the shortest path between each word pair in the RG data set was calculated using each of three subsets of lexical relations. The shortest path length connecting two concepts calculated using Rel ={hypernymy, hyponymy} will be denoted $dist_{Tax}$, path lengths calculated using Rel ={hypernymy, hyponymy, meronymy, holonomy, antonymy} will be denoted $dist_{Part}$ and path lengths calculated using all WordNet relations will be denoted $dist_{Length}$. For the RG and

	Correlation		Rank correlation	
	RG	*MC*	*RG*	*MC*
$dist_{Part}$	**-0.8888**	-0.8978	**-0.8604**	-0.8594
$dist_{Length}$	-0.8877	-0.8978	-0.8566	-0.8594
$dist_{Tax}$	-0.8665	-0.8381	-0.8116	-0.7998
rel_{YP}	**0.897**	**0.921**	—	—
$dist_{S}$	-0.8185	-0.8356	**-0.8629**	**-0.8526**
sim_{LC}	0.8382	0.8157	0.7911	0.7622
sim_{L}	0.8193	0.8292	0.7936	0.7904
sim_{R}	0.7787	0.7736	0.7573	0.7357
rel_{HS}	0.7861	0.7444	0.7917	0.7614
$dist_{JC}$	-0.7813	-0.8500	-0.7038	-0.8128

Table 4.3: Correlation coefficients for shortest path lengths using different subsets of semantic relations ($dist_{Part}$,$dist_{Length}$,$dist_{Tax}$) and previous relatedness and similarity measures

MC data sets, the correlation between path length and human ratings are given in Table 4.3. For the purposes of comparison, the correlations of previous similarity and relatedness measusures are provided in the bottom part of the table. It should be noted that negative values indicate an inverse correlation. Strong inverse correlations are just as desirable as strong correlations, as a positive correlation can be achieved by inverting the values of the data. Therefore the absolute values of the correlation coefficients should be considered when comparing the success of measures.

The results in Table 4.3 show that the difference in the correlations of path length with human ratings does not vary a great deal depending on the set of allowable semantic relations. In fact, the difference between $dist_{Part}$ and $dist_{Length}$ is negligible. The path lengths calculated using these two sets of relations differed for only one of the 65 word pairs that were tested.

Although no significant difference was found between the results of $dist_{Part}$ and $dist_{Length}$, $dist_{Tax}$ had a somewhat lower correlation with human ratings. These results can be partly explained by the distribution of semantic relationship types in the paths that are found when all types are allowed. For the 65 word pair *RG* data set, the frequencies of the relationship types that were found in the shortest paths are given in Table 4.4. Over 90% of the relations in the shortest paths were IS-A relations, and 6% were PART-OF relations. Only 1.5% of the relations in the

Relation	Count	%
hypernymy	729	79
hyponymy	125	13.5
holonymy	28	3
meronymy	25	3
antonymy	10	1
cause	2	0.2
see also	2	0.2
entailment	1	0.1

Table 4.4: Frequency of relationship types in search for shortest paths using all semantic relations

shortest paths were of other types. Thus while the addition of PART-OF relations contributed somewhat to the success of the measure, the other relations were too infrequent to affect performance.

4.4.2 Experiment 2: Relatedness Functions

For the second experiment that we conducted, the results of each of the five relatedness mapping functions were compared to human ratings of relatedness. The functions used the path lengths obtained when all WordNet semantic relations are allowed. Evaluating the relatedness functions using different sets of allowable relationship types was unnecessary, as the previous experiment showed that the set of allowed semantic relation types does not significantly affect the results.

A summary of the coefficients of correlation for all of the proposed measures against each test set is provided in Table 4.5. For the purpose of comparison, the coefficients of correlation of other semantic distance measures are also included in Table 4.5, including all of the measures evaluated in Budanitsky and Hirst's [4] study, as well as the results reported by Yang and Powers [47] for their measure (rel_{YP}), and the results of our implementation of Sussna's [44] measure ($dist_S$). The measures that are included in Budanitsky and Hirst's study include Hirst-St-Onge (rel_{HS}), Jiang-Conrath ($dist_{JC}$), Leacock-Chodorow (sim_{LC}), Lin (sim_L) and Resnik (sim_R).

The raw results that were used to calculate the coefficients of correlation reported in Table 4.5 may be found in the appendices. Yang and Powers did not

	Correlation			Rank correlation		
	RG	MC	$RG \setminus MC$	RG	MC	$RG \setminus MC$
rel_{Linear}	**0.8967**	0.9129	**0.8780**	0.8694	**0.8658**	**0.8118**
rel_{Decay}	0.8896	0.9098	0.8688	0.8694	0.8652	0.8072
rel_{Exp}	0.8956	0.9109	0.8757	0.8694	0.8652	0.8072
rel_{Log}	0.8568	0.8676	0.8383	0.8694	0.8652	0.8072
$rel_{Sigmoid}$	0.8880	**0.9329**	0.8682	0.8694	0.8652	0.8072
$dist_{Length}$	-0.8877	-0.8978	-0.8679	-0.8566	-0.8594	-0.7937
rel_{YP}	**0.897**	**0.921**	**0.877**	—	—	—
$dist_S$	-0.8185	-0.8356	-0.7898	**-0.8629**	**-0.8526**	**-0.7973**
sim_{LC}	0.8382	0.8157	0.8371	0.7911	0.7622	0.7324
sim_L	0.8193	0.8292	0.8164	0.7936	0.7904	0.7452
sim_R	0.7787	0.7736	0.8106	0.7573	0.7357	0.7414
rel_{HS}	0.7861	0.7444	0.7887	0.7917	0.7614	0.7590
$dist_{JC}$	-0.7813	-0.8500	-0.7216	-0.7038	-0.8128	-0.5398

Table 4.5: Correlation coefficients for proposed semantic relatedness measures and previous similarity and relatedness measures

provide their raw results, so only the reported correlation coefficients for their measure are given.

The values in Table 4.5 should be interpreted as the degree of linear correlation between the results of each measure and the human ratings collected by Miller and Charles (for the MC data set), or by Rubenstein and Goodenough (for the RG and the $RG \setminus MC$ data set). The highest possible correlation is one, which would indicate that the two sets of values being compared are identical.

The simplified measure with a linear relatedness function (rel_{Linear}) had the highest overall correlation for the two most significant data sets: RG and $RG \setminus MC$. The measure by Yang and Powers had the second highest correlation, trailing rel_{Linear} by only a small margin. Other previous measures had significantly lower correlations with this data. For the MC data set, $rel_{Sigmoid}$ had the highest correlation, and sim_{YP} had the second highest. Once again, other previous measures had significantly lower correlation coefficients.

The rank correlations for each measure are also provided in Table 4.5. For the most part, the rank correlations corresponded closely to the correlations, although the rank correlations were in general slightly lower. The only significant exception

is Sussna's measure, which had a higher rank correlation than correlation for all three data sets.

The rank correlations of the new measure with different relatedness functions are nearly identical. This is to be expected, as all of the functions that we are testing are monotonically decreasing as path length increases. The ranks of the results for these measures should not vary and so the rank correlation should not vary either. The small difference in the rank correlation of rel_{Linear} results from the fact that this function does not decrease for path lengths of seven and eight, which are both assigned the value zero.

4.5 Discussion of Results

4.5.1 Comparison with Previous Measures

The simplified relatedness measure proposed in this study outperforms all of the semantic distance measures surveyed by Budanitsky and Hirst by a wide margin. Only Yang and Powers' measure matches its performance. However, we are particularly interested in how previous measures compare to the version of the simplified measure that most resembles them. By comparing previous measures to similar but simpler measures it will be possible to make specific conclusions about where the previous measures fail.

Jiang and Conrath Similarity Measure

The Jiang and Conrath measure [16] computes semantic distance as the sum of weighted edges on the shortest path connecting concepts in the noun taxonomy. The most interesting comparison in our experiment is with the $dist_{Tax}$ measure, which calculates semantic distance as taxonomic path length. When the edge-weighting scheme is removed from Jiang and Conrath, and all edge weights are given a value of one, then their measure is equivalent to $dist_{Tax}$.

The $dist_{Tax}$ measure performed significantly better than sim_{JC} for the larger RG data set, with correlations of $r = -0.8665$ for $dist_{Tax}$ and $r = -0.7813$ for sim_{JC}, as shown in Table 4.3. For the smaller MC data set, sim_{JC} correlated more highly than $dist_{Tax}$, though by a smaller margin. For the MC data set, the $dist_{Tax}$ measure had a correlation of $r = -0.8381$ and sim_{JC} had a correlation of $r = -0.8500$. As $dist_{Tax}$ achieved better results on the larger data set, and

as the difference in performance was much greater for this set than for MC, we might be justified in calling $dist_{Tax}$ the more successful measure. At the least, we can conclude that sim_{JC} does not consistently improve on the results of $dist_{Tax}$. Overall, these results suggest that Jiang and Conrath's edge-weighting technique is not in fact effective.

An interesting result for sim_{JC} is the large difference between the correlation and the rank correlation. Although most of the measures receive a lower rank correlation than correlation, Jiang and Conrath's measure showed the greatest variation. For example, for the $RG \setminus MC$ data set the measure drops from a correlation of -0.7216 to the surprisingly low rank correlation of -0.5398. The reason for this appears to be that the $dist_{JC}$ measure does a good job at clustering word pairs that belong near one another, but does a poor job in identifying the relative positions of the word pairs within a cluster. When calculating the rank correlation, the word pairs are effectively de-clustered, and the exact ordering of the word pairs contributes more heavily to the correlation value.

Leacock and Chodorow Similarity Measure

The measure by Leacock and Chodorow [18] is a simple similarity measure that takes the taxonomic path length between concepts and maps the path length to a similarity value using a logarithmic function. Leacock and Chodorow's measure is interesting to us because it is the nearest thing to a baseline measure that has been included in previous evaluations, such as the one by Budanitksy and Hirst [4].

The simplified measure that provides the most appropriate comparison for sim_{LC} is the $dist_{Tax}$ measure. As the measure by Leacock and Chodorow transforms taxonomic path length to similarity, their measure should achieve better results than taxonomic path length on its own. However, for both the RG and MC data sets, $dist_{Tax}$ correlates more strongly with human ratings than their measure. This suggests that Leacock and Chodorow chose either the wrong function for transforming path length, or chose inappropriate constants for their function.

The results in Table 4.5 confirm that the logarithmic function, represented by rel_{Log}, is not a good choice for transforming path length to similarity or relatedness. It has the lowest correlation of any of the functions that were tested, and has a lower correlation than simple path length.

Hirst and St-Onge Relatedness Measure

The relatedness measure by Hirst and St-Onge [15] had the poorest performance of any measure. It had the lowest correlation of all the measures tested, and had a lower correlation than simple path length. As Hirst and St-Onge used a linear relationship between path length and relatedness, it is somewhat surprising that their results are so poor. The simplified linear measure (rel_{Linear}) had the best results overall of any measure.

The measure by Hirst and St-Onge differs from the simplified measures in several ways. First, their measure restricts the allowable paths, based on their categorization of semantic relations into three possible 'directions' as described in earlier chapters. Second, they discount the value of relatedness for each change in direction, once again relying on the directions of semantic relations that they defined. Third, they use a maximum search depth of five, whereas we used a maximum depth of seven.

An informal test showed that even at a maximum depth of five, simple path length correlates much better with human judgments than Hirst and St-Onge's measure. It is therefore likely that either the path restrictions or the discount factor for changes in direction, or both of these, reduce the effectiveness of their measure.

Sussna Relatedness Measure

Sussna's [44] measure has not been implemented previously for a major comparative evaluation, such as that by Budanitsky and Hirst [4]. This is unfortunate because, as the results in Table 4.5 show, Sussna's measure is the most successful of the measures evaluated in Budanitsky and Hirst's study. Among the previous measures that we are comparing, Sussna's is the second most successful measure, after the measure by Yang and Powers.

The most appropriate measure with which to compare Sussna's measure is $dist_{Part}$. Sussna's measure uses the same semantic relation types as $dist_{Part}$, but differs from $dist_{Part}$ in that it uses two edge-weighting techniques. The simplified measure had higher correlations than Sussna's measure, as shown in Table 4.3. However, Sussna's measure achieved very good rank correlations and had a slightly higher rank correlation than $dist_{Part}$ for the RG data set.

Given that Sussna's measure attained results that were fairly good, it is quite possible that one of the two edge-weighting techniques that it includes is a useful

technique for measuring relatedness. Future research that investigates the contributions of individual edge-weighting techniques should look closely at these two factors.

Yang and Powers Similarity Measure

Yang and Powers [47] described the best results in the literature, using an evaluation methodology that is nearly identical to our own. However, their measure is also the most complex one that we have tested, and has been systematically tuned to achieve the best correlation possible. In the last chapter, we argued that Yang and Powers's measure assumes an exponential decay model of the relationship between path length and similarity. We will therefore first compare their model to rel_{Decay}.

The measure by Yang and Powers attained a higher correlation than rel_{Decay} for all three data sets, although the difference in correlation is not large — approximately 0.01 in all cases. The Yang and Powers measure is therefore the only measure tested that consistently outperforms a corresponding simplified measure. This means that the techniques introduced by Yang and Powers were, taken together, beneficial and should be considered in the development of future measures.

However, the simplified linear measure, rel_{Linear} had identical correlation coefficients to the Yang and Powers measure for the RG data, and slightly higher correlation for the $RG \setminus MC$ data. This suggests that the Yang and Powers measure might have achieved even better results had it adopted a linear relationship between path length and similarity.

There are several advantages to the simpler model over the model proposed by Yang and Powers. For one, it is generally preferable to adopt a simpler model over a more complex one, when the models are equal in other respects. Unnecessary complexity should be avoided whenever possible, and the rel_{Linear} measure is simpler than the measure by Yang and Powers in many respects. For example, their model has many more parameters than the simplified model, including constants for every relationship type.

Also, as the simplified measure is not closely bound to semantic relation types it is more general than the Yang and Powers measure. For example, our measure can be applied without modification to the WordNet subgraphs for other parts of speech, such as verbs and adjectives. Although there is no empirical data against which to test the performance of the measure for other parts of speech, a casual analysis of a few examples is promising. Consider the following examples:

friendly/affable
```
[affable#a#1<sim>friendly#a#1]
```

friendly/kind
```
[friendly#a#1<also>congenial#a#1<also>sympathetic#a#2<sim>kind#a#4]
```

friendly/loving
```
[loving#a#1<also>lovable#a#1<also>amicable#a#1<sim>friendly#a#2]
```

friendly/casual
```
[friendly#a#1<also>amicable#a#1<also>peaceful#a#1<also>quiet#a#1
    <also>untroubled#a#1<also>unconcerned#a#1<sim>casual#a#1]
```

friendly/hungry
```
no path found
```

In the preceding examples, path length appears to correspond, more or less, with the strength of relatedness. The word pair *friendly/hungry* does not have any clear semantic relationship, and no path shorter than eight could be found in WordNet. For the other word pairs there is an intuitive association, and paths were found that connect these concepts.

4.5.2 Methodological Suggestions

We have shown that many path-based measures are unnecessarily complex. For example, when edge-weighting techniques are eliminated, most measures actually improve in performance. We believe that many of the flaws of semantic distance measures are the result of flaws in their development and evaluation methodologies.

To prevent unnecessary complexity in semantic distance measures, particularly path-based measures, they should be compared against baseline measures at every opportunity. To prove the merits of modifications to a simpler measure, the new measure should be evaluated with and without the modifications. What we are proposing is therefore a methodology akin to regression testing, as used in software engineering. Although it may not always be possible to decompose measures so as to test their components in this way, we showed in the previous chapter that many of the path-based measures that exist in the literature are in fact modular.

It is not clear why researchers have not used baseline measures to verify the success of their techniques. It could be that researchers have indeed adopted baseline measures, but that they chose poor measures to serve in this role. For example, the

Leacock and Chodorow measure is often included in semantic distance measure evaluations, and their measure takes a very straightforward approach. Unfortunately, their measure is also not very effective owing to the logarithmic transformation that they use.

Alternatively, it could be the case that researchers used a sound methodology when first developing their measure, but failed to retest against the baseline measures when external factors changed. For example, Hirst and St-Onge's measure is an adaptation of Morris and Hirst's [27] relatedness measure for Roget's thesaurus [37]. It could be that the elements of Hirst and St-Onge's measure that are not effective were helpful when using Roget's thesaurus. The same phenomenon could occur between different versions of WordNet.

As the evaluation framework for semantic distance measures is now fairly well established, at least by convention, the regression-testing approach that we are proposing should not be a serious burden to researchers. In the case of path-based relatedness measures, at the very least no measure should be outperformed by the simple measures that we have proposed in this study, such as rel_{Linear}.

4.5.3 Limitations of Evaluation

There are several limitations to the evaluation that we have described that should be noted. The Rubenstein-Goodenough set contains just 65 word pairs and the Miller-Charles set contains a subset of 30 of these. Although these sets have served as the primary means of evaluating relatedness and similarity measures for a number of years, a larger study is needed.

Also, the data appears to be fairly uniform in several respects. First, the level of abstractness of the words does not appear to vary significantly. There are no abstract concepts such as *justice* or general categories such as *thing*. Second, there is limited variety in the semantic domains of the concepts. For example, the data includes a large number of concepts that are types of people, such as *monk*, *lad*, *boy* and *wizard*. Finally, certain sorts of semantic relationships are not represented in the data. For example, while there are many pairs that are synonyms or near-synonyms, the data contains no antonyms or near-antonyms.

The uniformity of the data likely favours particular measures over others. As there were no abstract terms in the data, techniques that employ depth-scaling to account for the greater semantic distance of concepts high in the WordNet taxonomy would be less effective. Similarly, as cases of antonymy and near-antonymy were not present in the data, the accuracy of the models for these sorts of relations was

not tested. For example, the Yang and Powers measure discards paths that include antonymy links. While they do account for direct antonyms, near-antonyms would not be properly rated by their measure.

The terms *light* and *shade* are not direct antonyms, but they are nevertheless closely related via antonymy. The shortest path in WordNet between the word pair *light*/*shade* is:

```
[shade#n#1<hype>semidarkness#n#1<hype>dark#n#1<ants>light#n#10]
```

While the proposed simplified measure would find this path, the Yang and Powers measure would not and would therefore not capture the intuitively close relationship between these concepts. However, since the data does not contain cases of semantic opposition, the evaluation does not expose this shortcoming.

Another limitation of our experiment is that we have used only one means of evaluation. Some previous evaluations, including Budanitsky and Hirst's study, supplemented comparison with human judgments with an application-based evaluation. In the case of Budanitsky and Hirst, similarity and relatedness measures were integrated into an application for correcting real-word spelling errors, and the measures were rated in terms of the success of this task.

Budanitksy and Hirst themselves argue that comparison to human judgments is the most appropriate means of evaluating semantic distance measures, and we have offered similar arguments to theirs above. Also, in Budanitsky and Hirst's [4] study, it was found that the application-based evaluation only confirmed the results of the comparison to human judgments. The results for the task of real-word spelling correction found "Jiang and Conrath leading, followed by Lin and Leacock-Chodorow together, Resnik, and then Hirst-St-Onge." (p. 27) This ranking is identical to that for the results of the MC data set, as shown in Table 4.5. It is only slightly different for the RG data set, where the only significant change is a poorer performance for the Jiang and Conrath measure.

Another possible objection to our evaluation methodology is in the tuning that we have performed. Other measures, with the exception of Yang and Powers, do not describe any systematic tuning of their models. It might be argued that this gives the simplified models an unfair advantage over the others. However, path length on its own ($dist_{Length}$) correlated more highly with human judgments than any of the previous measures except for sim_{YP}. Therefore an untuned version of the simplified model achieved better results than these measures. As for Yang and Powers, as they employed systematic tuning for their model, comparing their measure to our tuned measures is justified.

4.6 Semantic Contrast

The work in the current study arose out of our interest in another form of semantic distance, that we call *semantic contrast*. Before concluding this chapter, we will offer some motivation for the study of semantic contrast in future work, and will present some preliminary findings. Semantic contrast has not received any serious academic attention, but it is a phenomenon that is often referred to in work on antonymy and semantic opposition [29]. It is also a phenomenon that we believe could offer some very interesting applications, if it could be modeled computationally.

In general, semantic distance can be viewed as the degree to which two concepts differ along some scale. The scales that we have discussed so far include similarity and relatedness. Two concepts can be more or less semantically distant depending on their similarity or dissimilarity, or they can be more or less semantically distant depending on their relatedness or lack thereof.

However, semantic opposition illustrates another way in which concepts can differ from one another. Budanisky and Hirst [4] point out that "antonymous concepts are dissimilar, and hence distant in one sense, and yet are strongly related semantically and hence close in the other." (p. 2) Budanitksy and Hirst are not correct in asserting the dissimilarity of antonyms, however. Muehleisen [29] observes that "opposites seem as different as they can possibly be, yet they still have something in common." (p. 3) In fact, semantic opposites often have very much in common. For example, the antonyms *man* and *woman* have a great deal of conceptual overlap. Men and women are both human beings, living creatures, have two legs, two arms, etc. In a word, they are very similar. They are also closely related. However, *man* and *woman* are in a sense very semantically distant, maybe even maximally so. As Cruse [8] writes:

> The meanings of a pair of opposites are felt to be maximally separated ... the closeness of opposites, on the other hand, manifests itself, for instance, in the fact that members of a pair have almost identical distributions, that is to say, very similar possibilities of normal and abnormal occurrence ... (p. 197)

The measures of either semantic similarity or semantic relatedness that we have examined in this study would find *man* and *woman* to be very close. In order to capture the sense of semantic distance that separates antonyms, we propose a new type of semantic distance called *semantic contrast*. Antonyms have very high

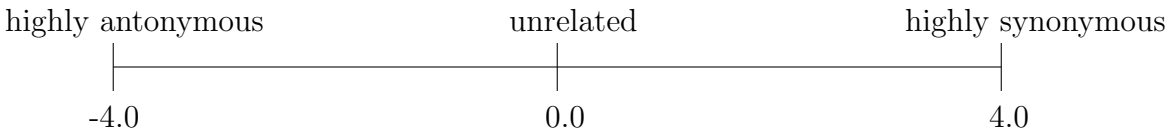


Figure 4.4: A scale for semantic contrast

semantic contrast, but there are cases of weaker contrast. The concepts *love* and *cruelty* are not opposites, but they are distant in the same sense that opposites are distant, though to a lesser degree.

4.6.1 Contrast Scale

As we are proposing that contrast is a matter of degree, it will be helpful to examine the features of the scale of contrast. For example, semantic opposites lie at one end of the scale, as they represent the most possible contrast. It is not clear, however, what sorts of relations lie at the other end of the scale. For example, the opposite of contrast could be unrelatedness, similarity, or something else.

We will not offer a detailed argument here, but will tentatively propose a scale that is an extension of the scale for semantic similarity. The original scale used by Rubenstein and Goodenough [38] ranges from "highly synonymous" to "unrelated." By extending this scale from "unrelated" to "highly antonymous," we create a scale that is symmetrical with the similarity scale and captures the full range of degrees of contrast. A visual representation of the scale may be helpful, and is provided in Figure 4.4. This scale assumes that related, but non-contrasting, word pairs have less contrast than unrelated ones. For example, the pair *car/gasoline* has less contrast than *bicycle/gasoline*, according to the scale.

By extending the scale in this way, we solve the problem that would have been encountered by Rubenstein and Goodenough's test subjects had antonyms been included in the test set for their experiment. Antonyms are neither "highly synonymous," nor "unrelated." Neither is it fair to call antonyms "somewhat related" and "somewhat synonymous" — they are in fact both highly related and minimally synonymous. It is impossible to express the semantic distance between opposites on the scale used in Rubenstein and Goodenough's experiment.

4.6.2 Preliminary Contrast Measure

We have experimented with a preliminary computational measure of contrast, that combines the techniques of path-based semantic distance measures with some observations by Fong [13]. In an interesting solution to a special form of semantic opposition described by Pustejovsky [32], Fong demonstrated that semantic opposition is to some extent transitive along paths of relations in a semantic network. He found that if the shortest path between concepts includes an antonym relation, then the endpoint concepts are likely to also be opposed. The task that was addressed by Fong required only a binary determination of opposition. However, his technique also lends itself to a gradable model of contrast.

The following chains of lexical relationships, taken from WordNet 2.1, illustrate the transitivity of contrast:

1. hate <*antonym*> love
2. hate <*antonym*> love <*sister_term*> joy
3. hate <*antonym*> love <*sister_term*> joy <*hypernym*> elation

The antonyms *hate* and *love* are highly contrasting, but as the distance from the antonym pair grows, contrast diminishes. The pairs of concepts *hate/joy* and *hate/elation* are not opposites. However, as a result of their proximity to the antonym relation *hate/love*, they are in contrast. In the computational measures of contrast that will be presented below, the quality of contrast will be determined by the presence of indirect antonymy. The strength, or quantity, of contrast will be determined using a relatedness measure.

Equation 4.4 expresses contrast as a function of the shortest path, p:

$$con(p) = \begin{cases} -rel(p) & \text{if } p \text{ contains one antonym link} \\ rel(p) & \text{otherwise} \end{cases} \tag{4.4}$$

In the preceding formula, p is the shortest path between concepts in a semantic network, and $rel(p)$ is the relatedness of the endpoints of p. Unfortunately, the formula above met with mixed results when applied to a test set of contrasting word pairs collected by Mettinger [23]. This set includes pairs of words selected from 20 English novels using syntactic frames that indicate contrast, such as "X rather than Y." Although the technique had no false positives when applied to non-contrasting words, it was only able to properly identify about 50% of the contrasting

word pairs in Mettinger's list. There was no data available to evaluate how well this measure was able to determine the strength of contrast. From the results that were obtained, it appears that this technique works well for certain types of contrasting word pairs, but not for others.

Future work in this area will likely require a typology of contrast and special consideration for the different kinds of contrast that exist. For example, the preliminary measure above works well for associative contrast such as that between *cat* and *bark*. In such cases the contrast appears to derive from the proximity of the concepts to an antonym pair. Specifically, the concept *bark* is strongly related to *dog*, and *dog* is the opposite of *cat*. However, other relationships that are intuitively contrasting are not captured by this measure. Words that imply positive and negative value judgments also seem to be contrasting, for example. The words *honesty* and *hate* seem to be in contrast, but they are not closely related to a pair of antonyms by classical semantic relations.

4.7 Chapter Summary

In this chapter we have described a two-part experiment that compared the simplified semantic relatedness measure proposed in the last chapter to previous relatedness and similarity measures. The measures were evaluated on the basis of correlation with human ratings of relatedness, using two widely used data sets. In the first part of the experiment, we examined the effect of using different sets of allowable semantic relationship types on the correlation of path length with human judgments. We found that relations other than IS-A have little effect on the results, as IS-A relations are by far the most common links between nouns in WordNet.

The second part of the experiment compared five functions for mapping path length to relatedness. We found that a linear relatedness function was the most effective, and that it obtained better results overall than any previous measure.

Also, for each previous measure that we were examining, we identified a corresponding simplified measure to serve as a baseline test. The surprising conclusion was that every measure other than the one by Yang and Powers had worse results than the corresponding simplified measure. On the basis of these results, we have concluded that many of the features of previous measures are detrimental to their performance.

In order to prevent the introduction of unnecessary features, we proposed a new methodology for the development and evaluation of future semantic distance measures. Any new techniques designed to improve upon the path-based approach to

measuring semantic relatedness or similarity should be compared to a baseline measure. Measures that cannot be demonstrated to improve upon a baseline measure should be rejected.

Chapter 5

Conclusion

5.1 Review

In this study, we have made several significant contributions to the study of semantic relatedness. We have demonstrated that previous measures of relatedness and similarity are overly complex. Many of them include elements that are unnecessary and reduce their correlation to human performance. For example, we have shown that the techniques used by previous measures to estimate the semantic distances of edges in the WordNet graph are ineffective.

We have described and evaluated a new measure of semantic relatedness which may be viewed as a simplification of current path-based measures. Despite its simplicity, the new measure achieved higher correlation with human ratings of relatedness than any of the previous measures that we compared. Compared with the five measures evaluated by Budanitksy and Hirst [4] in their recent study, our measure showed very significant improvement. Although Yang and Powers [47] demonstrated results comparable to those of our new measure, they did so with a much more complex model.

In the course of developing our new model, we were able to systematically examine two aspects of path-based semantic relatedness measures. First, we determined that due to the high percentage of IS-A relations in the noun portion of WordNet, the effect of including other types of semantic relations is negligible. This result may partly explain why similarity measures, which typically examine only IS-A relations, have served as successful proxies for semantic relatedness measures. Secondly, we examined the mathematical relationship between path length and the

strength of relatedness. We used statistical regression to test five different functions and found that a simple linear function had the best overall results.

We have also provided a general formal description of path-based similarity and relatedness measures. This general description showed that it is possible to decompose current measures into their constituent elements, and suggests how these could be recombined in future measures. We have proposed a methodology for the future development of path-based semantic distance measures that could help to avoid the introduction of the types of errors that we found in previous measures. Specifically, we recommend that any new measures be evaluated against simpler baseline measures whenever possible.

5.2 Future Work

5.2.1 Relatedness Measures

In this study we have proposed a new methodology for improving path-based relatedness measures. Although we have taken the first steps towards improving path-based measures, there remains much work to be done. For example, our methodology suggests that the elements of current relatedness and similarity measures should be examined individually in order to determine their merit. Although we demonstrated in Chapter 4 that most measures of similarity and relatedness perform more poorly than baseline measures, it may be that useful features in these measures were obscured by other detrimental features. For example, our experiments left some question as to the value of the edge-weighting techniques of Sussna [44] and Yang and Powers [47], and these should be examined in the future.

5.2.2 Experimental Data

Another important area for future work is in improving the evaluation framework. Although the field has settled on a fairly well-defined framework, it can be improved in several ways. First, the human ratings of semantic relatedness collected by Rubenstein and Goodenough [38] and by Miller and Charles [25] represent a rather small body of data. Additional experiments would be worthwhile to increase the amount of available data, and to thus increase the significance of experiments such as our own.

Also, the methodology that was used for collecting human semantic relatedness judgments should be improved for future experiments. The instructions used in the

previous experiments were ambiguous, given the distinction between relatedness and similarity. While the subjects were asked to rate word pairs according to their "similarity of meaning," they appear to have provided ratings of relatedness, and not similarity. This inconsistency may be the result of the scale that was used, with "unrelated" being the least possible value. Alternatively, it may be that judging relatedness is a more natural task for humans than judging similarity, and that the test subjects inadvertently defied the instructions that they were given. In either case, it is hard to guess what effect the discrepancy between the instructions and the behaviour of the subjects might have had on the results.

Finally, not only the number but also the variety of the word pairs in the current data could be improved. In Chapter 4 we pointed out some aspects of the data that appear to be quite uniform across the 65 word pairs of the Rubenstein-Goodenough [38] and Miller-Charles [25] data sets. For example, there were few general concepts represented in the data.

5.2.3 Contrast Applications

In Chapter 4, we described a new type of semantic distance called *semantic contrast*. We believe that semantic contrast has a number of valuable applications. In natural language generation contrast could be used, for example, in properly framing contrasting adjectives. That is, it is customary to provide special syntactic frames when asserting contrasting properties of an object. Consider the following sentences:

? Frances is friendly and abrupt.
Frances is friendly, but abrupt.

The first sentence above is unusual, but the second is much less jarring. Text generation systems could use a measure of semantic contrast to select appropriate syntax for such cases of contrasting adjectives.

Another potential application of a computational measure of contrast is for computational humour. Ritchie [36] provides a formal model of a certain class of jokes, that he calls *forced reinterpretation* jokes. An important part of this model is what Ritchie calls *contrast*. Although he is deliberately vague about the nature of contrast in his model, our notion of semantic contrast fits well with the limited description that Ritchie gives. We believe that a model of semantic contrast could be very helpful for simple jokes, such as puns, that turn on the semantic relationships between words.

5.3 Final Words

Semantic relatedness, and more generally semantic distance, is likely to remain an important area of interest for computational linguists. The many new applications that were described in papers at just one recent conference [17, 7, 41] indicate that the potential uses of semantic distance measures are far from being exhausted. We believe that this study helps to clear the way for both future improvements of path-based semantic distance measures and provides a simple and effective measure to be used in semantic distance applications.

Appendix A

Appendix

A.1 Experimental Results

The following tables of results are the raw ratings of semantic relatedness and semantic similarity for new and previous measures for the experiments described in Chapter 4.

Table A.1: Results of previous similarity and relatedness measures for the *RG* data set

Word #1	Word #2	Humans	relHS	distJC	simLC	simL	simR	distS
cord	smile	0.02	0	19.6711	1.387	0.09	1.1762	3
noon	string	0.04	0	22.6451	1.5025	0	0	3
rooster	voyage	0.04	0	26.908	0.9175	0	0	3
fruit	furnace	0.05	0	18.5264	2.2801	0.1482	1.8563	1.5691
autograph	shore	0.06	0	22.724	1.387	0	0	3
automobile	wizard	0.11	0	17.8624	1.5025	0.0986	0.9764	3
mound	stove	0.14	0	17.2144	2.2801	0.2204	2.9062	0.9963
grin	implement	0.18	0	16.6232	1.2801	0	0	3
asylum	fruit	0.19	0	19.5264	2.2801	0.1425	1.8563	1.571
asylum	monk	0.39	0	25.6762	1.628	0.0707	0.9764	3
graveyard	madhouse	0.42	0	29.7349	1.1806	0	0	3
boy	rooster	0.44	0	17.8185	1.5025	0.2112	2.3852	3
glass	magician	0.44	0	22.829	1.9175	0.0788	0.9764	2.143
cushion	jewel	0.45	0	22.9386	2.2801	0.1393	1.8563	1.5307
monk	slave	0.57	94	18.9192	2.7655	0.2113	2.535	1.0864
asylum	cemetery	0.79	0	28.1499	1.5025	0	0	3
coast	forest	0.85	0	20.2206	2.2801	0.1299	1.5095	1.6047
grin	lad	0.88	0	20.8152	1.2801	0	0	3
shore	woodland	0.9	93	19.3361	2.5025	0.1351	1.5095	1.5559
monk	oracle	0.91	0	22.7657	2.0875	0.1821	2.535	1.7251
boy	sage	0.96	93	19.934	2.5025	0.2028	2.535	1.2026
automobile	cushion	0.97	98	15.0786	2.0875	0.2782	2.9062	0.4315
mound	shore	0.97	91	12.492	2.7655	0.498	6.1974	0.8089
lad	wizard	0.99	94	16.5177	2.7655	0.2349	2.535	1.0557
forest	graveyard	1	0	24.573	1.7655	0	0	1.3458
food	rooster	1.09	0	17.4637	1.387	0.1006	0.9764	1.0435
cemetery	woodland	1.18	0	25.0016	1.7655	0	0	1.3458
shore	voyage	1.22	0	23.738	1.387	0	0	3
bird	woodland	1.24	0	18.1692	2.0875	0.1382	1.5095	2.0093
coast	hill	1.26	94	10.8777	2.7655	0.5326	6.1974	0.5442
furnace	implement	1.37	93	15.8742	2.5025	0.1895	1.8563	1.3538
crane	rooster	1.41	0	12.806	2.0875	0.5812	8.8872	1.0914
hill	woodland	1.48	93	18.2676	2.5025	0.1418	1.5095	1.0123
car	journey	1.55	0	16.3425	1.2801	0	0	3
cemetery	mound	1.69	0	23.8184	1.9175	0	0	0.7791
glass	jewel	1.78	0	22.0185	2.0875	0.1443	1.8563	1.1184
magician	oracle	1.82	98	1	3.5025	0.9645	13.5898	1.4945
crane	implement	2.37	94	15.6813	2.7655	0.2704	2.9062	0.9074
brother	lad	2.41	94	16.3583	2.7655	0.2366	2.535	1.0718
sage	wizard	2.46	93	22.8275	2.5025	0.1817	2.535	1.2552
oracle	sage	2.61	0	26.2251	2.0875	0.162	2.535	1.0381
bird	cock	2.63	150	5.403	4.0875	0.7669	8.8872	0.179

Continued on Next Page...

Table A.1 – Continued

Word #1	Word #2	Humans	relHS	distJC	simLC	simL	simR	distS
bird	crane	2.63	97	7.403	3.0875	0.706	8.8872	0.484
food	fruit	2.69	0	10.2695	2.2801	0.2272	1.5095	0.698
brother	monk	2.74	93	19.2087	2.5025	0.2088	2.535	0.1809
asylum	madhouse	3.04	150	0.263	4.0875	0.9917	15.7052	0.1026
furnace	stove	3.11	0	20.5459	2.0875	0.1342	1.8563	0.2814
magician	wizard	3.21	200	0	5.0875	1	13.5898	0
hill	mound	3.29	200	0	5.0875	1	12.0807	0
cord	string	3.41	150	2.2707	4.0875	0.8907	9.2513	0.1911
glass	tumbler	3.45	150	5.9425	4.0875	0.7925	11.3477	0.1889
grin	smile	3.46	200	0	5.0875	1	10.4198	0
serf	slave	3.46	0	19.8021	2.2801	0.348	5.2844	0.5692
journey	voyage	3.58	150	5.2133	4.0875	0.7476	7.7194	0.2092
autograph	signature	3.59	150	2.415	4.0875	0.9221	14.2902	0.1514
coast	shore	3.6	150	0.8845	4.0875	0.9618	11.1203	0.1975
forest	woodland	3.65	200	0	5.0875	1	11.2349	0
implement	tool	3.66	150	1.1777	4.0875	0.9133	6.2034	0.2197
cock	rooster	3.68	200	0	5.0875	1	14.2902	0
boy	lad	3.82	150	5.3942	4.0875	0.7285	8.2987	0.1851
cushion	pillow	3.84	150	0.7004	4.0875	0.9749	13.5898	0.213
cemetery	graveyard	3.88	200	0	5.0875	1	13.7666	0
automobile	car	3.92	200	0	5.0875	1	8.6231	0
gem	jewel	3.94	200	0	5.0875	1	14.3833	0
midday	noon	3.94	200	0	5.0875	1	15.9683	0
correlation:		1	0.7861	-0.7813	0.8382	0.8193	0.7787	-0.8185
rank correlation:		1	0.7917	-0.7038	0.7911	0.7936	0.7573	-0.8526

Table A.2: Results of proposed relatedness measures for the *RG* data set

Word #1	Word #2	Humans	distLen	relLinear	relDecay	relExp	relLog	relSig
cord	smile	0.02	8	0	0.5255	0	0.9355	0.6279
noon	string	0.04	8	0	0.5255	0	0.9355	0.6279
rooster	voyage	0.04	8	0	0.5255	0	0.9355	0.6279
fruit	furnace	0.05	6	0.48	0.8728	0.5794	1.2074	0.6476
autograph	shore	0.06	8	0	0.5255	0	0.9355	0.6279
automobile	wizard	0.11	8	0	0.5255	0	0.9355	0.6279
mound	stove	0.14	5	1.07	1.1249	1.0659	1.3772	0.7374
grin	implement	0.18	7	0	0.6772	0.1056	1.0623	0.6309
asylum	fruit	0.19	6	0.48	0.8728	0.5794	1.2074	0.6476
asylum	monk	0.39	8	0	0.5255	0	0.9355	0.6279
graveyard	madhouse	0.42	8	0	0.5255	0	0.9355	0.6279
boy	rooster	0.44	7	0	0.6772	0.1056	1.0623	0.6309
glass	magician	0.44	8	0	0.5255	0	0.9355	0.6279
cushion	jewel	0.45	6	0.48	0.8728	0.5794	1.2074	0.6476
monk	slave	0.57	4	1.65	1.4498	1.5683	1.582	1.1513
asylum	cemetery	0.79	8	0	0.5255	0	0.9355	0.6279
coast	forest	0.85	6	0.48	0.8728	0.5794	1.2074	0.6476
grin	lad	0.88	8	0	0.5255	0	0.9355	0.6279
shore	woodland	0.9	5	1.07	1.1249	1.0659	1.3772	0.7374
monk	oracle	0.91	7	0	0.6772	0.1056	1.0623	0.6309
boy	sage	0.96	5	1.07	1.1249	1.0659	1.3772	0.7374
automobile	cushion	0.97	3	2.24	1.8685	2.0911	1.8396	2.2327
mound	shore	0.97	4	1.65	1.4498	1.5683	1.582	1.1513
lad	wizard	0.99	4	1.65	1.4498	1.5683	1.582	1.1513
forest	graveyard	1	5	1.07	1.1249	1.0659	1.3772	0.7374
food	rooster	1.09	6	0.48	0.8728	0.5794	1.2074	0.6476
cemetery	woodland	1.18	5	1.07	1.1249	1.0659	1.3772	0.7374
shore	voyage	1.22	8	0	0.5255	0	0.9355	0.6279
bird	woodland	1.24	7	0	0.6772	0.1056	1.0623	0.6309
coast	hill	1.26	3	2.24	1.8685	2.0911	1.8396	2.2327
furnace	implement	1.37	5	1.07	1.1249	1.0659	1.3772	0.7374
crane	rooster	1.41	7	0	0.6772	0.1056	1.0623	0.6309
hill	woodland	1.48	4	1.65	1.4498	1.5683	1.582	1.1513
car	journey	1.55	8	0	0.5255	0	0.9355	0.6279
cemetery	mound	1.69	4	1.65	1.4498	1.5683	1.582	1.1513
glass	jewel	1.78	5	1.07	1.1249	1.0659	1.3772	0.7374
magician	oracle	1.82	6	0.48	0.8728	0.5794	1.2074	0.6476
crane	implement	2.37	4	1.65	1.4498	1.5683	1.582	1.1513
brother	lad	2.41	4	1.65	1.4498	1.5683	1.582	1.1513
sage	wizard	2.46	5	1.07	1.1249	1.0659	1.3772	0.7374
oracle	sage	2.61	5	1.07	1.1249	1.0659	1.3772	0.7374
bird	cock	2.63	1	3.41	3.1036	3.2427	2.7261	3.4726

Continued on Next Page...

Table A.2 – Continued

Word #1	Word #2	Humans	distLen	relLinear	relDecay	relExp	relLog	relSig
bird	crane	2.63	3	2.24	1.8685	2.0911	1.8396	2.2327
food	fruit	2.69	3	2.24	1.8685	2.0911	1.8396	2.2327
brother	monk	2.74	1	3.41	3.1036	3.2427	2.7261	3.4726
asylum	madhouse	3.04	1	3.41	3.1036	3.2427	2.7261	3.4726
furnace	stove	3.11	2	2.83	2.4081	2.6429	2.1877	3.1739
magician	wizard	3.21	0	4	4	4	3.9754	3.5337
hill	mound	3.29	0	4	4	4	3.9754	3.5337
cord	string	3.41	1	3.41	3.1036	3.2427	2.7261	3.4726
glass	tumbler	3.45	1	3.41	3.1036	3.2427	2.7261	3.4726
grin	smile	3.46	0	4	4	4	3.9754	3.5337
serf	slave	3.46	3	2.24	1.8685	2.0911	1.8396	2.2327
journey	voyage	3.58	1	3.41	3.1036	3.2427	2.7261	3.4726
autograph	signature	3.59	1	3.41	3.1036	3.2427	2.7261	3.4726
coast	shore	3.6	1	3.41	3.1036	3.2427	2.7261	3.4726
forest	woodland	3.65	0	4	4	4	3.9754	3.5337
implement	tool	3.66	1	3.41	3.1036	3.2427	2.7261	3.4726
cock	rooster	3.68	0	4	4	4	3.9754	3.5337
boy	lad	3.82	1	3.41	3.1036	3.2427	2.7261	3.4726
cushion	pillow	3.84	1	3.41	3.1036	3.2427	2.7261	3.4726
cemetery	graveyard	3.88	0	4	4	4	3.9754	3.5337
automobile	car	3.92	0	4	4	4	3.9754	3.5337
gem	jewel	3.94	0	4	4	4	3.9754	3.5337
midday	noon	3.94	0	4	4	4	3.9754	3.5337
correlation:		1	-0.8877	0.8967	0.8896	0.8956	0.8568	0.8880
rank correlation:		1	-0.8566	0.8694	0.8694	0.8694	0.8694	0.8694

Word #1	Word #2	Humans	relHS	distJC	simLC	simL	simR	distS
noon	string	0.08	0	22.6451	1.5025	0	0	3
rooster	voyage	0.08	0	26.908	0.9175	0	0	3
glass	magician	0.11	0	22.829	1.9175	0.0788	0.9764	2.143
chord	smile	0.13	0	20.2418	1.628	0.1808	2.2341	3
coast	forest	0.42	0	20.2206	2.2801	0.1299	1.5095	1.6047
lad	wizard	0.42	94	16.5177	2.7655	0.2349	2.535	1.0557
monk	slave	0.55	94	18.9192	2.7655	0.2113	2.535	1.0864
shore	woodland	0.63	93	19.3361	2.5025	0.1351	1.5095	1.5559
forest	graveyard	0.84	0	24.573	1.7655	0	0	1.3458
coast	hill	0.87	94	10.8777	2.7655	0.5326	6.1974	0.5442
food	rooster	0.89	0	17.4637	1.387	0.1006	0.9764	1.0435
cemetery	woodland	0.95	0	25.0016	1.7655	0	0	1.3458
monk	oracle	1.1	0	22.7657	2.0875	0.1821	2.535	1.7251
journey	car	1.16	0	16.3425	1.2801	0	0	3
lad	brother	1.66	94	16.3583	2.7655	0.2366	2.535	1.0718
crane	implement	1.68	94	15.6813	2.7655	0.2704	2.9062	0.9074
brother	monk	2.82	93	19.2087	2.5025	0.2088	2.535	0.1809
tool	implement	2.95	150	1.1777	4.0875	0.9133	6.2034	0.2197
bird	crane	2.97	97	7.403	3.0875	0.706	8.8872	0.484
bird	cock	3.05	150	5.403	4.0875	0.7669	8.8872	0.179
food	fruit	3.08	0	10.2695	2.2801	0.2272	1.5095	0.698
furnace	stove	3.11	0	20.5459	2.0875	0.1342	1.8563	0.2814
midday	noon	3.42	200	0	5.0875	1	15.9683	0
magician	wizard	3.5	200	0	5.0875	1	13.5898	0
asylum	madhouse	3.61	150	0.263	4.0875	0.9917	15.7052	0.1026
coast	shore	3.7	150	0.8845	4.0875	0.9618	11.1203	0.1975
boy	lad	3.76	150	5.3942	4.0875	0.7285	8.2987	0.1851
gem	jewel	3.84	200	0	5.0875	1	14.3833	0
journey	voyage	3.84	150	5.2133	4.0875	0.7476	7.7194	0.2092
car	automobile	3.92	200	0	5.0875	1	8.6231	0
correlation:		1	0.7444	-0.85	0.8157	0.8292	0.7736	-0.8356
rank correlation:		1	0.7614	-0.8128	0.7622	0.7904	0.7357	-0.8629

Table A.3: Results of previous similarity and relatedness measures for the MC data set

Word #1	Word #2	Humans	distLen	relLinear	relDecay	relExp	relLog	relSig
noon	string	0.08	8	0	0.5255	0	0.9355	0.6279
rooster	voyage	0.08	8	0	0.5255	0	0.9355	0.6279
glass	magician	0.11	7	0	0.6772	0.1056	1.0623	0.6309
chord	smile	0.13	8	0	0.5255	0	0.9355	0.6279
coast	forest	0.42	6	0.48	0.8728	0.5794	1.2074	0.6476
lad	wizard	0.42	4	1.65	1.4498	1.5683	1.582	1.1513
monk	slave	0.55	4	1.65	1.4498	1.5683	1.582	1.1513
shore	woodland	0.63	5	1.07	1.1249	1.0659	1.3772	0.7374
forest	graveyard	0.84	5	1.07	1.1249	1.0659	1.3772	0.7374
coast	hill	0.87	3	2.24	1.8685	2.0911	1.8396	2.2327
food	rooster	0.89	6	0.48	0.8728	0.5794	1.2074	0.6476
cemetery	woodland	0.95	5	1.07	1.1249	1.0659	1.3772	0.7374
monk	oracle	1.1	7	0	0.6772	0.1056	1.0623	0.6309
journey	car	1.16	8	0	0.5255	0	0.9355	0.6279
lad	brother	1.66	4	1.65	1.4498	1.5683	1.582	1.1513
crane	implement	1.68	4	1.65	1.4498	1.5683	1.582	1.1513
brother	monk	2.82	1	3.41	3.1036	3.2427	2.7261	3.4726
tool	implement	2.95	1	3.41	3.1036	3.2427	2.7261	3.4726
bird	crane	2.97	3	2.24	1.8685	2.0911	1.8396	2.2327
bird	cock	3.05	1	3.41	3.1036	3.2427	2.7261	3.4726
food	fruit	3.08	3	2.24	1.8685	2.0911	1.8396	2.2327
furnace	stove	3.11	2	2.83	2.4081	2.6429	2.1877	3.1739
midday	noon	3.42	0	4	4	4	3.9754	3.5337
magician	wizard	3.5	0	4	4	4	3.9754	3.5337
asylum	madhouse	3.61	1	3.41	3.1036	3.2427	2.7261	3.4726
coast	shore	3.7	1	3.41	3.1036	3.2427	2.7261	3.4726
boy	lad	3.76	1	3.41	3.1036	3.2427	2.7261	3.4726
gem	jewel	3.84	0	4	4	4	3.9754	3.5337
journey	voyage	3.84	1	3.41	3.1036	3.2427	2.7261	3.4726
car	automobile	3.92	0	4	4	4	3.9754	3.5337
correlation:		1	-0.8978	0.9129	0.9098	0.9109	0.8676	0.9329
rank correlation:		1	-0.8594	0.8658	0.8652	0.8652	0.8652	0.8652

Table A.4: Results of proposed relatedness measures for the *MC* data set

Word #1	Word #2	Humans	relHS	distJC	simLC	simL	simR	distS
fruit	furnace	0.05	0	18.5264	2.2801	0.1482	1.8563	1.5691
autograph	shore	0.06	0	22.724	1.387	0	0	3
automobile	wizard	0.11	0	17.8624	1.5025	0.0986	0.9764	3
mound	stove	0.14	0	17.2144	2.2801	0.2204	2.9062	0.9963
grin	implement	0.18	0	16.6232	1.2801	0	0	3
asylum	fruit	0.19	0	19.5264	2.2801	0.1425	1.8563	1.571
asylum	monk	0.39	0	25.6762	1.628	0.0707	0.9764	3
graveyard	madhouse	0.42	0	29.7349	1.1806	0	0	3
boy	rooster	0.44	0	17.8185	1.5025	0.2112	2.3852	3
cushion	jewel	0.45	0	22.9386	2.2801	0.1393	1.8563	1.5307
asylum	cemetery	0.79	0	28.1499	1.5025	0	0	3
grin	lad	0.88	0	20.8152	1.2801	0	0	3
boy	sage	0.96	93	19.934	2.5025	0.2028	2.535	1.2026
automobile	cushion	0.97	98	15.0786	2.0875	0.2782	2.9062	0.4315
mound	shore	0.97	91	12.492	2.7655	0.498	6.1974	0.8089
shore	voyage	1.22	0	23.738	1.387	0	0	3
bird	woodland	1.24	0	18.1692	2.0875	0.1382	1.5095	2.0093
furnace	implement	1.37	93	15.8742	2.5025	0.1895	1.8563	1.3538
crane	rooster	1.41	0	12.806	2.0875	0.5812	8.8872	1.0914
hill	woodland	1.48	93	18.2676	2.5025	0.1418	1.5095	1.0123
cemetery	mound	1.69	0	23.8184	1.9175	0	0	0.7791
glass	jewel	1.78	0	22.0185	2.0875	0.1443	1.8563	1.1184
magician	oracle	1.82	98	1	3.5025	0.9645	13.5898	1.4945
sage	wizard	2.46	93	22.8275	2.5025	0.1817	2.535	1.2552
oracle	sage	2.61	0	26.2251	2.0875	0.162	2.535	1.0381
hill	mound	3.29	200	0	5.0875	1	12.0807	0
cord	string	3.41	150	2.2707	4.0875	0.8907	9.2513	0.1911
glass	tumbler	3.45	150	5.9425	4.0875	0.7925	11.3477	0.1889
grin	smile	3.46	200	0	5.0875	1	10.4198	0
serf	slave	3.46	0	19.8021	2.2801	0.348	5.2844	0.5692
autograph	signature	3.59	150	2.415	4.0875	0.9221	14.2902	0.1514
forest	woodland	3.65	200	0	5.0875	1	11.2349	0
cock	rooster	3.68	200	0	5.0875	1	14.2902	0
cushion	pillow	3.84	150	0.7004	4.0875	0.9749	13.5898	0.213
cemetery	graveyard	3.88	200	0	5.0875	1	13.7666	0
correlation:		1	0.7887	-0.7216	0.8371	0.8164	0.8106	-0.7898
rank correlation:		1	0.759	-0.5398	0.7324	0.7452	0.7414	-0.7973

Table A.5: Results of previous similarity and relatedness measures for the $RG \setminus MC$ data set

Word #1	Word #2	Humans	distLen	relLinear	relDecay	relExp	relLog	relSig
fruit	furnace	0.05	6	0.48	0.8728	0.5794	1.2074	0.6476
autograph	shore	0.06	8	0	0.5255	0	0.9355	0.6279
automobile	wizard	0.11	8	0	0.5255	0	0.9355	0.6279
mound	stove	0.14	5	1.07	1.1249	1.0659	1.3772	0.7374
grin	implement	0.18	7	0	0.6772	0.1056	1.0623	0.6309
asylum	fruit	0.19	6	0.48	0.8728	0.5794	1.2074	0.6476
asylum	monk	0.39	8	0	0.5255	0	0.9355	0.6279
graveyard	madhouse	0.42	8	0	0.5255	0	0.9355	0.6279
boy	rooster	0.44	7	0	0.6772	0.1056	1.0623	0.6309
cushion	jewel	0.45	6	0.48	0.8728	0.5794	1.2074	0.6476
asylum	cemetery	0.79	8	0	0.5255	0	0.9355	0.6279
grin	lad	0.88	8	0	0.5255	0	0.9355	0.6279
boy	sage	0.96	5	1.07	1.1249	1.0659	1.3772	0.7374
automobile	cushion	0.97	3	2.24	1.8685	2.0911	1.8396	2.2327
mound	shore	0.97	4	1.65	1.4498	1.5683	1.582	1.1513
shore	voyage	1.22	8	0	0.5255	0	0.9355	0.6279
bird	woodland	1.24	7	0	0.6772	0.1056	1.0623	0.6309
furnace	implement	1.37	5	1.07	1.1249	1.0659	1.3772	0.7374
crane	rooster	1.41	7	0	0.6772	0.1056	1.0623	0.6309
hill	woodland	1.48	4	1.65	1.4498	1.5683	1.582	1.1513
cemetery	mound	1.69	4	1.65	1.4498	1.5683	1.582	1.1513
glass	jewel	1.78	5	1.07	1.1249	1.0659	1.3772	0.7374
magician	oracle	1.82	6	0.48	0.8728	0.5794	1.2074	0.6476
sage	wizard	2.46	5	1.07	1.1249	1.0659	1.3772	0.7374
oracle	sage	2.61	5	1.07	1.1249	1.0659	1.3772	0.7374
hill	mound	3.29	0	4	4	4	3.9754	3.5337
cord	string	3.41	1	3.41	3.1036	3.2427	2.7261	3.4726
glass	tumbler	3.45	1	3.41	3.1036	3.2427	2.7261	3.4726
grin	smile	3.46	0	4	4	4	3.9754	3.5337
serf	slave	3.46	3	2.24	1.8685	2.0911	1.8396	2.2327
autograph	signature	3.59	1	3.41	3.1036	3.2427	2.7261	3.4726
forest	woodland	3.65	0	4	4	4	3.9754	3.5337
cock	rooster	3.68	0	4	4	4	3.9754	3.5337
cushion	pillow	3.84	1	3.41	3.1036	3.2427	2.7261	3.4726
cemetery	graveyard	3.88	0	4	4	4	3.9754	3.5337
correlation:		1	-0.8679	0.878	0.8688	0.8757	0.8383	0.8682
rank correlation:		1	-0.7937	0.8118	0.8072	0.8072	0.8072	0.8072

Table A.6: Results of proposed relatedness measures for the $RG \setminus MC$ data set

A.2 Algorithms

The following algorithms are described in Chapter 3, with informal complexity analyses.

description: unidirectional breadth-first search for shortest path in a graph
input : a source node and a target node
output : the shortest path between the input nodes

```
1  queue ← {(start_node )}
2  while queue ≠ ∅ do
3      path ← pop(queue)
4      node ← last(path)
5      neighbours ← expand(node)
6      forall node ∈ neighbours do
7          if node ∈ path then
               /* avoid cycles                                    */
8          else if node = end_node then
               /* found a solution                                */
9              return push(path,node)
10         else
11             path ← push(path,node)
12             queue ← push(queue,path)
13 return null
```

Algorithm A.1: Unidirectional breadth-first search for shortest path in a graph

```
description: bidirectional asymmetric breadth-first search for shortest path
             in a graph
input      : a source node and a target node
output     : the shortest path between the input nodes
1  sourceQ ← {(startNode )}
2  targetQ ← {(endNode )}
3  while sourceQ ≠ ∅ and targetQ ≠ ∅ do
       /* Use bigQ and smallQ as aliases                                  */
4      if sourceQ > targetQ then
5          bigQ ← sourceQ
6          smallQ ← targetQ
7      else
8          bigQ ← targetQ
9          smallQ ← sourceQ
10     path ← pop(smallQ)
11     node ← last(path)
12     neighbours ← expand(node)
13     foreach node ∈ neighbours do
14         if node ∉ path then
               /* Check for a solution                                    */
15             foreach targetPath ∈ targetQ do
16                 targetNode ← last(targetPath)
17                 if targetNode = node then
18                     solution ← join(path,reverse(targetPath))
19                     return solution
20             path ← push(path,node)
21             smallQ ← push(smallQ,path)
22 return null
```

Algorithm A.2: Bidirectional asymmetric breadth-first search for shortest path in a graph

Bibliography

[1] Aristotle. On memory and reminiscence. In Richard McKeon, editor, *The Basic Works of Aristotle*, pages 607–618. Random House, New York, 1941.

[2] Satanjeev Banerjee and Ted Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, August 2003.

[3] Alexander Budanitsky. Lexical semantic relatedness and its application in natural language processing. Technical Report CSRG390, University of Toronto, 1999.

[4] Alexander Budanitsky and Graeme Hirst. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1), March 2006.

[5] Patrick Cassidy. An investigation of the semantic relations in the Roget's Thesaurus: Preliminary results. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLING-00)*, pages 13–19, Mexico City, Mexico, February 2000.

[6] Allan M. Collins and Elizabeth F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82:407–428, 1975.

[7] Courtney Corley and Rada Mihalcea. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan, June 2005.

[8] D. Alan Cruse. *Lexical Semantics*. Cambridge University Press, 1986.

[9] James Deese. *The structure of associations in language and thought*. The Johns Hopkins Press, Baltimore, Maryland, 1965.

[10] Ann Devitt and Carl Vogel. The topology of WordNet: Some metrics. In *Proceedings of the Second International WordNet Conference*, pages 106–111, Brno, Czech Republic, January 2004.

[11] Edsger W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, December 1959.

[12] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts, 1998.

[13] Sandiway Fong. Semantic opposition and WordNet. *Journal of Logic, Language and Information*, 13(2):159–171, 2004.

[14] W. Nelson Francis and Henry Kucera. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston, 1982.

[15] 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*, chapter 13, pages 305–332. The MIT Press, Cambridge, Massachusetts, 1998.

[16] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, pages 19–33, Taiwan, August 1997.

[17] Upali Sathyajith Kohomban and Wee Sun Lee. Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 05)*, pages 34–41, Ann Arbor, Michigan, June 2005.

[18] Claudia Leacock and Martin Chodorow. Combining local context and WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 265–283. The MIT Press, Cambridge, Massachusetts, 1998.

[19] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26, Toronto, June 1986.

[20] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, pages 296–304, Madison, Wisconsin, July 1998.

[21] Donald Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11(2):431–441, June 1963.

[22] Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Barcelona, Spain, July 2004.

[23] Arthur Mettinger. *Aspects of Semantic Opposition in English*. Clarendon Press, Oxford, England, 1994.

[24] Stanley Milgram. The small-world problem. *Psychology Today*, 2:60–67, 1967.

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

[26] Saif Mohammad and Graeme Hirst. Distributional measures as proxies for semantic relatedness. Unpublished, 2005.

[27] Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–43, 1991.

[28] Jane Morris and Graeme Hirst. Non-classical lexical semantic relations. In Dan Moldovan and Roxana Girju, editors, *Workshop on Computational Lexical Semantics (HLT-NAACL 2004)*, pages 46–51, Boston, May 2004.

[29] Victoria Lynn Muehleisen. *Antonymy and Semantic Range in English*. PhD thesis, Northwestern University, Evanston, Illinois, June 1997.

[30] Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-03)*, pages 241–257, Mexico City, Mexico, February 2003.

[31] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet::Similarity – measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, pages 1024–1025, San Jose, July 2004.

[32] James Pustejovsky. Events and the semantics of opposition. In Carol L. Tenny and James Pustejovsky, editors, *Events as grammatical objects*, pages 445–482. CSLI Publications, California, 2000.

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

[34] Philip Resnik. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, August 1995.

[35] 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*, 11:95–130, 1999.

[36] Graeme Ritchie. *The Linguistic Analysis of Jokes*. Routledge, New York, 2004.

[37] Peter M. Roget. *Roget's International Thesaurus, Fourth Edition*. Harper and Row Publishers Inc., New York, 1977.

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

[39] Nuno Seco. Computational models of similarity in lexical ontologies. Master's thesis, University College, Dublin, Ireland, 2005.

[40] David St-Onge. Detecting and correcting malapropisms with lexical chains. Master's thesis, University of Toronto, 1995.

[41] Mark Stevenson and Mark Greenwood. A semantic approach to IE pattern

induction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 05)*, pages 379–386, Ann Arbor, Michigan, June 2005.

[42] Mark Steyvers and Joshua B. Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29:41–78, 2005.

[43] Steven Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[44] 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, Washington, D.C., November 1993.

[45] Amos Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.

[46] Duncan Watts and Steven Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[47] Dongqiang Yang and David M.W. Powers. Measuring semantic similarity in the taxonomy of WordNet. In V. Estivill-Castro, editor, *Proceedings of the 28th Australasian Computer Science Conference*, pages 315–322, Newcastle, Australia, February 2005.