Automatically Acquiring a Semantic Network of Related Concepts

by

Sean Szumlanski
B.S. University of Central Florida, 2004
M.S. University of Central Florida, 2005

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Electrical Engineering and Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2013

Major Professor:
Fernando Gomez

ABSTRACT

We describe the automatic acquisition of a semantic network in which over 7,500 of the most frequently occurring nouns in the English language are linked to their semantically related concepts in the WordNet noun ontology. Relatedness between nouns is discovered automatically from lexical co-occurrence in Wikipedia texts using a novel adaptation of an information theoretic inspired measure. Our algorithm then capitalizes on salient sense clustering among these semantic associates to automatically disambiguate them to their corresponding WordNet noun senses (i.e., *concepts*). The resultant concept-to-concept associations, stemming from 7,593 target nouns, with 17,104 distinct senses among them, constitute a large-scale semantic network with 208,832 undirected edges between related concepts. Our work can thus be conceived of as augmenting the WordNet noun ontology with *RelatedTo* links.

The network, which we refer to as the Szumlanski-Gomez Network (SGN), has been subjected to a variety of evaluative measures, including manual inspection by human judges and quantitative comparison to gold standard data for semantic relatedness measurements. We have also evaluated the network's performance in an applied setting on a word sense disambiguation (WSD) task in which the network served as a knowledge source for established graph-based spreading activation algorithms, and have shown: a) the network is competitive with WordNet when used as a stand-alone knowledge source for WSD, b) combining our network with WordNet achieves disambiguation results that exceed the performance of either resource individually, and c) our network outperforms a similar resource, WordNet++ (Ponzetto & Navigli, 2010), that has been automatically derived from annotations in the Wikipedia corpus.

Finally, we present a study on human perceptions of relatedness. In our study, we elicited quantitative evaluations of semantic relatedness from human subjects using a variation of the classical methodology that Rubenstein and Goodenough (1965) employed to investigate human perceptions of semantic similarity. Judgments from individual subjects in our study exhibit high average correlation to the elicited relatedness means using leave-one-out sampling (r = 0.77, σ = 0.09, N = 73), although not as high as average human correlation in previous studies of similarity judgments, for which Resnik (1995) established an upper bound of r = 0.90 (σ = 0.07, N = 10). These results suggest that human perceptions of relatedness are less strictly constrained than evaluations of similarity, and establish a clearer expectation for what constitutes human-like performance by a computational measure of semantic relatedness. We also contrast the performance of a variety of similarity and relatedness measures on our dataset to their performance on similarity norms and introduce our own dataset as a supplementary evaluative standard for relatedness measures.

That I am able to enjoy a life of safety, liberty, and broad acceptance as an openly gay man is a privilege that has been purchased for me by the suffering of countless other human beings. This dissertation is dedicated to those who found the courage to lead open and honest lives in the face of tremendous adversity, in memory of those who lost their lives for doing so, and to all who have stood with us in the protracted struggle for LGBT equality.

ACKNOWLEDGMENTS

I would first like to thank my committee for their countless contributions to my graduate career over the years. Charlie Hughes, Annie Wu, and Valerie Sims have been my teachers, collaborators, co-authors, and mentors. They have impressed me not only with their invaluable intellectual contributions to my work, but also with how tirelessly they work to set their students up for success. The encouraging and giving nature of these individuals has made me want to work harder and give more to my own students and colleagues.

I am particularly grateful to my advisor, Fernando Gomez. He has spent countless hours in conversation with me over the years, passing on his knowledge not only of artificial intelligence and computational linguistics, but also of art, philosophy, history, literature, and so many other things. In my early years under his advisement, one was as likely to find us discussing Chomsky as one was to find us discussing Lorca and Franco, Picasso's *Guernica*, Bach's cantatas, or Goya's witches. I found in him a veritable Abbé Faria (in the Dumasian sense) whose breadth and depth of knowledge helped make me a more educated and well-rounded person. My life is richer for our time together, and I am grateful.

Maxine Najle helped facilitate data collection for the perceptions of relatedness study that is included as part of this dissertation. I owe her huge thanks for that and for giving so generously of her time during one of the most demanding semesters of her undergraduate career.

I am also grateful to my colleagues from the UCF AI Lab whose discussions, counsel, distractions, and friendship over the years contributed to my education and personal wellbeing. In lexicographic order, they are: Adam Campbell, who inspired me with his diligence and unassuming competence; Adelein Rodriguez, who helped keep me grounded with her

perspectives on life, AI, and so many things in between; Andy Schwartz, who forged ahead of me on this incredible journey and, in doing so, cleared away some of the thorny brush along the trail and showed me it was possible to reach the goal; Chris Millward, who impressed me with his knack for coming up with creative solutions to problems large and small and by being one of the most level-headed and genuine people I know; Nadeem Mohsin, who made the AI Lab immeasurably more interesting, geeky, and fun by sharing just some of what is stored in his amazingly encyclopedic brain; and Ramya Pradhan, who reminded me that sometimes we have to make sacrifices for the things we want.

I would like to extend my thanks to Arup Guha and Ali Orooji, who have been long-time teaching mentors and who played critical roles in getting me into the classroom and eventually having me teach classes on my own. I would also like to thank Ali Orooji and the UCF Programming Team for welcoming me as a guest at their practices in the fall semesters of 2011 and 2012, where I acquired invaluable experience honing my craft as a programmer and benefited from their collective knowledge.

I am grateful to the Office of Student Conduct for providing me with so many opportunities to engage in challenging, meaningful work as a member of UCF's Student Conduct Review Board—work that broadened my mind, brought much-needed balance to my life, and offered me tremendous personal growth. I am particularly grateful to Dana Juntunen, Director of the Office of Student Rights and Responsibilities, for the special role she played as a mentor to me during my time on the Board.

I am profoundly grateful to my parents, who have always been there for me, who made incredible sacrifices to give me a good life, and who told me from a very young age that I could

achieve anything I set my mind to, and to my brother, who is a tremendous friend, whom I admire for the kindness and compassion in his heart, and who makes me feel understood and valued for the person I am.

Finally, I would like to thank NASA and the Division of Computer Science at UCF for their generous financial support over the years. This research was supported in part by the NASA Engineering and Safety Center under Grant/Cooperative Agreement NNX08AJ98A.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

JJ	Adjective (POS tag)
M&C	Miller and Charles (1991)
NELL	Never-Ending Language Learner
NLP	Natural Language Processing
NP	Noun Phrase
OMCS	Open Mind Common Sense
R&G	Rubenstein and Goodenough (1965)
SGN	Szumlanski-Gomez Network
VB	Verb (POS tag)
WN	WordNet
WN++	WordNet++
WSD	Word Sense Disambiguation

CHAPTER 1: INTRODUCTION

(1) *The astronomer photographed the star.*

When faced with a sentence like the one above, the human mind seamlessly navigates a complex mindscape of lexical and syntactic ambiguity in its quest to ascribe meaning—first to individual words, and then, ultimately, to the sentence as a whole. In our resultant understanding of (1), we know, for example, that "star" refers to a celestial body. We have excluded the possibility of "star" being an adjective or verb, or of it denoting a movie star, an asterisk, or any of the myriad other possible senses of the noun.

The human cognitive processes that give rise to this understanding are highly automatized. Few people even notice (consciously, at least) the lexical ambiguity of "star" in (1), despite cognitive evidence that the human mind accesses all possible meanings of ambiguous nouns during semantic interpretation, even when a sentence contains strong contextual clues as to the intended meaning of an ambiguous noun (Swinney, 1979), as is the case in (1), as well as the following, contrasting sentence:

(2) *The paparazzi photographed the star.*

Despite the syntactic equivalence of (1) and (2), it is clear that the "star" in (2) denotes not a celestial body, but a celebrity. While it is conceivable that a paparazzo would photograph a celestial object, or that an astronomer would photograph a celebrity, the "stars" here are preferentially disambiguated by the strong semantic relatedness between *paparazzi*[1] and the

1 In distinguishing between words and the concepts they denote, we adopt the convention of quoting the former and italicizing the latter. (See Section 1.8, "Style Conventions: Words and the Concepts They Denote," below.)

celebrity sense of "star," and *astronomer* and the *celestial body* sense of "star," respectively, à la mechanisms of spreading activation through semantic memory (Collins & Loftus, 1975; Quillian, 1968).

The ease with which the human mind resolves such natural language ambiguities belies the complexity of the cognitive processes and lexical semantic resources that drive semantic interpretation. Over half a century of artificial intelligence research has taught us as much. So intricate, so vast, and so deep is the semantic knowledge that resides in the human mind, that we have yet to see the creation of a comprehensive computational model of semantic memory, or an artificially intelligent agent that is capable of semantically interpreting arbitrary natural language utterances—a machine that we can confidently claim is able to process a sentence and subsequently *understand*.

1.1 Semantic Memory and WordNet

Quillian (1968) posited a theory of semantic memory that accounts for the disambiguation of our stars in the sentences above. In his model, concepts are represented as nodes in a semantic network and related to other concepts by way of labeled edges between nodes. These relations establish semantic relatedness between concepts and allow for the codification of attributes of individual concepts. During interpretation, concept nodes are activated by lexical stimuli, and that activation spreads in parallel to adjacent nodes in a breadth-first manner, with diminishing strength at each level of activation. For example, the word "astronomer" in *The astronomer photographed the star* causes activation of the *astronomer* node in memory, which then spreads to related concepts (e.g., *telescope, observatory*, the astronomer

Galileo, and several others,[2] including a concept node for the celestial body sense of "star"). When the word "star" is subsequently encountered in the sentence, the celestial body sense is already partially activated in memory. (In cognitive terms, the concept has been *primed*.) Its activation indicates an intersection of word meanings in the sentence, and the noun is disambiguated to its celestial body sense accordingly.[3]

An important feature of Quillian's model is that it allows for two concepts to be related by any other concept in the network. Typically, this relation takes the form of a verbal concept,[4] but Quillian also uses a canonical *IsA* relation to indicate superordinate and subordinate relationships between concepts (e.g., *IsA*(ASTRONOMER, SCIENTIST); *IsA*(PAPARAZZO, PHOTOGRAPHER)). Because relations themselves are concepts, they are also subject to the effects of spreading activation through the network. For example, if *astronomer* and *star* are conjoined in the network via a *study* relation (i.e., *study*(ASTRONOMERS, STARS)), spreading activation from the *astronomer* node activates not just the *star* node, but also the verbal concept, *study*.

The WordNet noun ontology (Miller, 1998) is one of the most sophisticated attempts to implement Quillian's ideas of semantic memory to date. It constitutes a partial realization of Quillian's dream through its instantiation of a variety of labeled edges indicating, *inter alia*, subsumptive *IsA* relationships between concepts. The lexical inventory of WordNet enumerates individual senses of English language nouns and relates them to synonymous senses of other

2 Quillian's theory assumes that the network expresses comprehensive semantic knowledge about the concepts it contains. An implementation with complete fidelity to Quillian's view of semantic memory would contain an inordinate amount of information about astronomers.

3 This is a simplified account of spreading activation. Extended versions call for sense assignment to occur only after candidate senses have been subjected to elaborate evaluative procedures that determine contextual and syntactic validity (Collins & Quillian, 1972; Quillian, 1969; for an overview of Quillian's various presentations of the model, and an extended account of spreading activation theory, see Collins & Loftus, 1975).

4 Because the model uses verbal concepts to coordinate related concepts, Quillian's semantic memory can be seen as an early attempt to create a common sense knowledge base (cf. Liu & Singh, 2004a; McCarthy, 1959; Minsky, Singh, & Sloman, 2004), although he does not explicitly frame his work in those terms.

nouns in the ontology. The resulting sets of synonyms, or *synsets*, form the basic concept nodes of WordNet. Each of these concepts is manually assigned a superordinate concept (*hypernym*) and, where applicable, subordinate concepts (*hyponyms*), resulting in a hand-crafted taxonomy of semantic classes. WordNet tells us, for example, that an *astronomer* is a *physicist*,[5] a *physicist* is a *scientist*, a *scientist* is a *person*, and so on, all the way up to *physical object*, which is an *entity*.[6] (*Entity* is the root node of the ontology, and the only node without a superordinate concept.) The ontology also codifies a small, closed set of additional relations, such as antonymy, holonymy and meronymy (part-whole relations), instance-of relationships, and domain terms.

The subsumptive architecture of the ontology serves as an indication of semantic similarity—which is a particular type of relatedness (Resnik, 1999)—between concepts. Hyponymic relationships reflect semantic similarity directly; that a *penguin* is an *aquatic bird* implies strong similarity between the two concepts. Through transitive subsumption, we can also infer the similarity of *penguin* and *animal*, although the distance between these nodes (they are interceded by *aquatic bird, bird, vertebrate*, and *chordate*) suggests weaker similarity than that of *penguin* to *aquatic bird*. From WordNet we can also infer the similarity of, e.g., *penguins* and *flamingos*, by virtue of their shared subsumption by the superordinate concept *aquatic bird*. Notably absent from the ontology, however, are indications of general relatedness, as with, e.g., *penguins* and *icebergs*, or *polar bears* and *global warming*.

In some cases, subsumption and similarity suffice to resolve lexical ambiguity, as in the following sentences:

5 This can be expressed equivalently by any of the following three binary relations: *IsA*(ASTRONOMER, PHYSICIST), *hyponym*(ASTRONOMER, PHYSICIST), and *hypernym*(PHYSICIST, ASTRONOMER).

6 The *IsA* relation is transitive; if an astronomer *IsA* physicist and a physicist *IsA* scientist, it follows that an astronomer *IsA* scientist, as well.

(3) *We tried it once in A-flat minor, but the key proved too difficult.*

(4) *There were no queens, rooks, or knights remaining.*

In (3), the subsumption of *A-flat minor* by the musical sense of "key" helps us disambiguate the latter term. (Miller (1998) points out that this results from "a linguistic convention that accepts anaphoric nouns that are hypernyms of the antecedent.") In (4), similarity (i.e., shared subsumption in WordNet) establishes that the "queens, rooks, or knights" being discussed are *chess pieces*. (Cf. *There were no kings or queens remaining*, which leaves us wondering whether the kings and queens are monarchs, chess pieces, or something else.)

In other cases, however, semantic interpretation requires more general indications of relatedness than those that are provided by WordNet; notice that if we relied on semantic similarity to disambiguate *The astronomer photographed the star*, the path in WordNet connecting *astronomer* and the *celebrity* sense of "star" (in that both are *people*) would lead us astray.

1.2 Our Contribution

The focus of this dissertation is the automatic, unsupervised acquisition of a semantic network that indicates general semantic relatedness between concepts denoted by nouns. This is the specific type of lexical semantic knowledge that enables interpretation of sentences like (1) and (2) above, and is a critical component of semantic memory and mechanisms of natural language understanding, such as word sense disambiguation.

Constructing such a network comprises two phases: association and disambiguation. The goal of association is to take as input a target noun (a *stimulus*), and return a list of strongly related nouns—nouns that one might reasonably expect would come to mind if a person were presented with the same stimulus in a word association game.

In the association phase of network acquisition, we establish semantic relatedness between nouns by applying a novel adaptation of an information theoretic measure to co-occurrence data extracted from Wikipedia. This is a context-sparse affair that takes place *in absentia* of the semantic annotations of Wikipedia, such as inter-article links, entries in disambiguation pages, the title of the article from which a sentence is extracted, and so on.

In the disambiguation phase, we capitalize on salient sense clustering among related nouns to automatically resolve them to their appropriate noun senses (i.e., concepts). For our concepts, we use the noun senses defined in WordNet 3.0; thus, our work can be conceived of as augmenting the WordNet noun ontology with *RelatedTo* links. This seems an obvious choice for our noun sense inventory, given the WordNet ontology's sophistication and ubiquitous use in computational linguistics and artificial intelligence.

The edges between concepts in our network indicate general semantic relatedness. Rather than tie edges to weights that we derive from co-occurrence data, which are susceptible to corpus biases, we create a network in which relatedness is represented categorically, without weight. This mirrors the unweighted structure of WordNet. However, our network could presumably be used as a kernel to infer quantitative relatedness scores, in the same way that WordNet has been used to derive semantic similarity scores between concepts (cf. Section 2.2 below).

Figure 1.1 offers a preview of the contribution of our research. In the semantic network we have created, *astronomer* is related to 45 distinct concepts. A sampling of those concepts is shown in Figure 1.1, denoted by medium blue nodes adjacent to the dark blue *astronomer#1* node.[7] In turn, those concepts are related to concepts in light blue, and those terms are related to concepts in white. This gives an idea of spreading activation through the network.

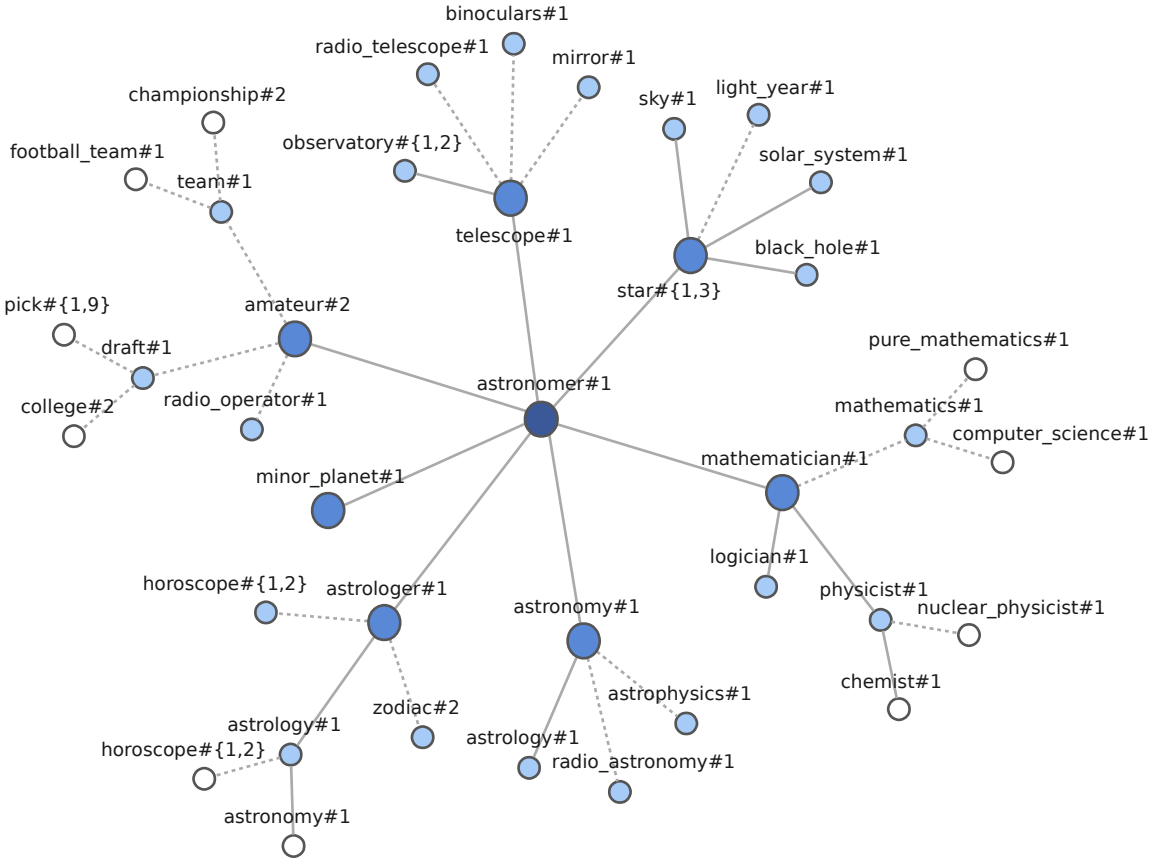


Figure 1.1:
Partial spreading activation view of the concepts related to *astronomer*.

7 We denote sense *n* of a noun in WordNet by *noun#n*, or multiple senses with, e.g., *noun#{m,n}*. (See Section 1.8, "Style Conventions: Words and the Concepts They Denote," below.)

A solid edge in our graphical depiction indicates that *astronomer#1* is related to the farther node incident to that edge. For example, the solid edge from *star#{1,3}* to *sky#1* indicates that *astronomer#1* is related to *sky#1*, too. The dotted edge from *astrology#1* to *horoscope#{1,2}* indicates that *astronomer#1* is not related to *horoscope#{1,2}* in our network. Many of the concepts in Figure 1.1 are interrelated in our network, but here we have omitted edges between them in order to avoid messy edge crossings.

1.3 Corpus Considerations and Co-occurrence

The correlation of lexical co-occurrence frequency to semantic association strength is well established in the literature (Church & Hanks, 1990; Spence & Owens, 1990; Wettler & Rapp, 1993); strongly related terms tend to co-occur more frequently in texts than unrelated or weakly related terms, and those that co-occur frequently tend to be related. However, investigations into this correlation typically compare adjusted co-occurrence counts to limited sets of *association norms*—quantitative measurements of how strongly humans judge two words to be related (cf. Palermo & Jenkins, 1964). These studies remain silent on the question of how to establish categorical relatedness, or how to deal with spurious cases where co-occurrence frequency is incongruent with relatedness. One of the primary contributions of our work is to resolve these limitations and adapt the measurement of co-occurrence frequency for building a large-scale network of semantic relatedness.

The corpus we use in our research, Wikipedia,[8] is an online encyclopedia that has been collaboratively constructed by volunteers and contains over 4 million articles. Stripped of all

8 http://en.wikipedia.org

markup, metadata, and duplicate sentences, our version of the corpus from August 2009 is 10 Gigabytes on disk and contains nearly 1.5 billion words. We have chosen Wikipedia as our target corpus from which to extract co-occurrence data primarily for its large size, broad coverage of the English language, and free availability for download on the Web. However, the co-occurrence approach we develop is not specific to Wikipedia; it can be applied to any large corpus, either to augment our existing semantic network, or to create a new one.

The encyclopedic nature of Wikipedia's text does, however, contribute to its appeal as a candidate for relatedness mining. In order to achieve its goal of giving informative overviews of the broad range of topics it covers, an encyclopedia must explicitly articulate relationships between many strongly related entities (and so, related entities must be mentioned together in the corpus; they must co-occur). Such is the case in the following sentences from Wikipedia, which establish the relationships between, e.g., woodpeckers and tree trunks (5), pianos and keys (6), astronomers and stars (7), and rooks and the game of chess (8):[9]

(5) *Many woodpeckers have the habit of tapping noisily on tree trunks with their beaks.*

(6) *The white keys of the piano correspond to the C major scale.*

(7) *By convention, astronomers grouped stars into constellations and used them to track the motions of the planets and the inferred position of the sun.*

(8) *In chess, a rook may move any distance along a row or column.*

9 Notice that many of these nouns are ambiguous: "trunk" can refer to the trunk of a car; "key" can refer to a device for opening locks (among many other things), "rook" can refer to a bird, "chess" can refer to a type of grass, and we have already discussed the ambiguity of "star." Yet, the intended meanings of these words are clear from the noun pairs listed above, even before we examine the sentential contexts in which they co-occur.

Incidental co-occurrence of unrelated terms, like that of "convention" and "position" in (7), is ubiquitous in natural language texts, but the relative infrequency with which any two *particular* unrelated terms co-occur will, in most cases, protect us from false indications of semantic relatedness. Furthermore, the related nouns in (5) through (8) must continue to co-occur throughout the corpus in order for their strong semantic association to be discovered. Their co-occurrence will likely take many different forms; related terms frequently appear together in contexts that do not explicitly articulate commonsense knowledge about their relationships, as in the co-occurrence of "ascension," "throne," and "regency" in the following sentence from Wikipedia:

(9) *Following the murder of King Henry IV and the ascension to the French throne by Louis XIII, under Marie de' Medici's regency, Biencourt and his father were authorized to return to Acadia.*

Although (9) is not intended to inform the reader about thrones or the act of ascension, one might argue that the sentence establishes a relation implicitly through frame semantics: Louis XIII is categorized as a monarch in WordNet, and the implicit relationship is [[Agent *Monarchs* monarch.n#1] ***ascend to*** ascend.v#3 [Goal *thrones* throne.n#3]]. Of course, this is not the only (or even necessarily the best) way to define the relationship between monarchs and thrones, and the sentence gives no clear indication of how either concept relates to regency. Ultimately, it is the co-occurrence of the terms in (9)—not the context in which they appear—that contributes to the cumulative evidence found within the corpus for their relatedness.

1.4 Corpus Context and the Limitations of Pre-Specified Relations

Whereas the lexical co-occurrence approach to relatedness discovery remains agnostic to the particular context in which that co-occurrence is manifest, previous methods have capitalized on context (both lexical and syntactic) to establish particular kinds of relatedness between nouns. Hearst (1992) established the tradition of using lexico-syntactic patterns to mine corpora for examples of a pre-specified relation (namely, in the case of Hearst, hyponymy). For example, she used the pattern ***NP***{, ***NP***}*{,} *or other* ***NP*** to establish all the former noun phrases (NPs) as hyponyms of the latter, as in "...wastebasket, trashcan, or other garbage receptacle," where we see that *wastebasket* and *trashcan* are hyponyms of *garbage receptacle*. Berland and Charniak (1999) used a similar method to discover meronymic part-whole relationships for a set of six hand-chosen wholes. Subsequent methods have used automatic pattern induction to induce search patterns from manually provided seed sets of noun pairs that typify a given relation. The induced patterns have then been used to discover new instances of the seeded relations, such as meronymy (Girju, Badulescu & Moldovan, 2006), hyponymy, and relations specific to the domain of chemistry, such as chemical *reaction* and *production* relations, among others (Pantel & Pennacchiotti, 2006).

Pattern matching is also the driving force behind current large-scale knowledge network acquisition projects: ConceptNet (Havasi, Speer, & Alonso, 2007; Liu & Singh, 2004a, 2004b) uses manually derived patterns to extract relationships from the semi-structured text of the Open Mind Common Sense corpus, which contains statements of commonsense knowledge acquired from over 10,000 contributors via a Web interface, often in forms that are particularly amenable to relation instance extraction via pattern matching (Singh et al., 2002). Similarly, the Never-

Ending Language Learner (henceforth NELL) (Carlson et al., 2010) automatically induces search patterns from example seed sets to learn relationships from unstructured text from the Web. Both of these projects rely on large, pre-determined sets of relations (e.g., *EffectOf*, *CapableOf*, and *LocationOf*, in the case of ConceptNet), and focus heavily on *IsA* relationships. Neither resource, however, attempts to establish a sophisticated ontology like that of WordNet, or to methodically delineate relationships according to individual noun senses (i.e., both resources relate words, not concepts, the name of ConceptNet notwithstanding).

The major limitation of the pattern matching approach is that it requires the pre-specification of the relation(s) to be mined. This ultimately precludes discovery of relatedness in the general case; as Quillian (1968) aptly points out, "in natural language text almost *anything* can be considered as a relationship, so that there is no way to specify in advance what relationships are to be needed" (p. 230, emphasis in original). Consider, for example, the strong semantic relationship between *penguin* and *tuxedo*, which defies labeling by any conventional relations. It seems unlikely that the relation that binds these concepts generalizes to a pattern that can capture other instances of the relation in a large corpus; indeed, it seems unlikely that many other examples of this particular relation exist at all. Yet, the pattern matching approach requires such examples if it is to have any chance of automatically discovering the relatedness between these two concepts.

Furthermore, Hearst (1992) found that some relations simply are not amenable to the pattern matching approach, either because they do not generalize well to patterns with broad coverage of the relation, or because they do not induce patterns that are exclusive enough to the relation to yield high precision extraction results.

In contrast to these context-driven pattern matching approaches, our focus on lexical co-occurrence allows us to establish relatedness between noun concepts regardless of the particular relation that binds them. While eschewing labeled relations in our network gives us the freedom and flexibility to associate nouns regardless of whether we can neatly articulate the relation between them, it also changes the fundamental nature of our contribution. The knowledge embedded in our network reflects a different type of commonsense knowledge than many of the associations in ConceptNet, NELL, and other networks that employ labeled relations, which explicitly codify statements of commonsense knowledge through binary relations (e.g., the correspondence of *study*(ASTRONOMERS, STARS) to the commonsense assertion *Astronomers study stars*). The relationships we discover reflect a different aspect of commonsense and lexical semantic knowledge, and can be thought of as a collection of relational kernels that underly commonsense assertions (e.g., the relatedness of *penguin* and *tuxedo*, which stems from, but does not fully express, the commonsense assertion that *The penguin's black and white coat of feathers makes it look like it is wearing a tuxedo*).

On the discovery of labeled relations between nouns, we defer to existing information extraction methods (cf. Fader, Soderland, & Etzioni, 2011); insofar as these relations tend to be expressed by verbs, discerning the relation between two arbitrary, related entities falls slightly outside the purview of our current research into more general semantic relatedness.

1.5 Using Semantic Resources to Discover Relatedness

In recent years, the availability of robust semantic resources has enabled new approaches to relatedness mining. Wikipedia has seen widespread use on this front. It might at first seem

unusual to classify Wikipedia not as a corpus, but as a semantic resource. Yet, it has several structural properties and annotations that qualify it as such and make it useful for relatedness mining, including: inter-article links that can be conceived of as edges connecting article nodes in a Wikipedia graph; disambiguation pages that enumerate distinct senses of articles that share the same title; the loose organization of its articles into an informal taxonomy (a *folksonomy*); and structured factual assertions in articles' info boxes, indicating, e.g., publication dates of books, movies in which actors have appeared, population sizes of cities, and so on.

These structural semantic attributes have been used to quantitatively measure relatedness between nouns or concepts (sometimes using disambiguation pages to derive concept inventories) (Gabrilovich & Markovitch, 2007; Milne & Witten, 2008a; Strube & Ponzetto, 2006; Zaragoza et al., 2007), as well as to learn categorical relationships between WordNet noun senses (Ponzetto & Navigli, 2010). Some of the semantic relations underlying Wikipedia have also been extracted to large-scale knowledge networks like DBpedia (Bizer et al., 2009) and YAGO (Suchanek, Kasneci, & Weikum, 2007). As with lexico-syntactic pattern matching, these approaches are limited by the restricted set of semantic relations expressed in Wikipedia. Relying on links between articles as an indications of relatedness is also problematic, given the ubiquity of cross-references to tangentially related topics in Wikipedia (e.g., the link from the *glacier* Wikipage to the article on *Vulgar Latin*).

Other approaches have turned to WordNet to search for relatedness. Navigli (2005) has developed a semi-automated method for creating a semantic network by disambiguating terms in collocations extracted from various semantically annotated resources, including WordNet and the Longman Language Activator, while Hughes and Ramage (2007) and Patwardhan and Pedersen

(2006) have used *IsA* relations and sense glosses from WordNet to quantitatively measure semantic relatedness between concepts. These WordNet-based approaches are inherently limited by the fact that, while the ontology serves as a rich taxonomy of semantic similarity, it lacks general indications of semantic relatedness. Consider, for example, how WordNet-based approaches would discover the strong semantic relationship between ontologically disparate entities like penguins and tuxedos. For this purpose, the minimalistic glosses of WordNet are simply insufficient; if we want to discover relatedness beyond semantic similarity, beyond the most obvious examples of relatedness, we need the assistance of a sizable corpus.

In general, hand-crafting useful ontologies and semantic resources is laborious work and requires some degree of training or expertise on the part of those who construct them. Such resources must provide massive amounts of data to be useful to learning algorithms, and require maintenance in order to remain relevant as language use shifts and changes over time. These limitations explain the field's predominant focus on unsupervised or weakly supervised learning algorithms for constructing semantic resources, and inform our lexical co-occurrence approach, which does not rely on a corpus that has been semantically annotated.

1.6 Other Parts of Speech

We have thus far limited our discussion to relatedness between concepts denoted by nouns. This is not to denigrate the lexical semantic contributions of other parts of speech to semantic interpretation. Rather, verbs and adjectives are excluded from consideration in our network on the grounds that their semantic associates typically take the form of entire semantic classes rather than lexical entries (cf. Katz & Fodor, 1963). For example, the verb "eat" has a

strong preference for themes (to use the technical term from semantic role labeling) that are categorized as *food*, and the adjective "tasty" has a similar preference with respect to its arguments (cf. Tanner & Gomez, 2010). Selectional constraints on arguments (often called selectional restrictions or selectional preferences) can override strong semantic relatedness between nouns, as in (10) and (11) below (from Waltz & Pollack, 1985, p. 53, and Charniak, 1983, p. 175, respectively):

(10) *The sailor ate a submarine.*

(11) *The astronomer married the star.*

In (10), the strong preference of the verb "eat" for *food* disambiguates the "submarine" to the *hoagie* sense, despite the strong relationship between sailors and the *warship* sense of "submarine." A similar restriction on the arguments of "marry" overrides the semantic relatedness between the astronomer and the *celestial body* sense of "star" in (11), selecting instead the sense that is a *person* (a movie star, celebrity, etc.). We do, however, see some interference from the *astronomer–star* and *sailor–submarine* relationships in these examples; Waltz and Pollack (1985) report that most people perform a "cognitive doubletake" (p. 62) when encountering these sentences, which initially lead us down a "semantic garden path" (p. 64) before selectional restriction ultimately resolves the ambiguities.

Some verbs constrain their arguments more weakly than others (Resnik, 1997). Such is the case with "photograph," which reveals some of the motivation behind our illustrative use of *The astronomer photographed the star* throughout this chapter: the verb's weak preference for people and landscapes is easily overridden by the association of the astronomer and the celestial

body sense of "star." Even the verb's ultimate preference for physical objects can be set aside in natural language use, as in the following sentence from Wikipedia, where the theme takes the form of an abstraction:

(12) *They photographed the fall of Sevastopol in September 1855.*

It is true that there are cases in which verbs strongly associate with specific nouns rather than entire noun classes, but these are typically collocative associations or lexicalized verb phrases that warrant their own entries in the lexicon. For example, the idiomatic "eat crow" has its own lexical entry in Wictionary, but not WordNet; it would be unusual to associate the verb "eat" in WordNet with the noun "crow," rather than instantiating a new, metaphorical sense of the verb. This, however, falls outside the aim of our research.

Clearly, establishing relatedness between noun senses is not a panacea for all our semantic interpretation problems. However, there is already a considerable body of research on associating verbs with selectional constraints. Resnik (1997) has had some success automatically abstracting from verb-noun lexical associations to verbs' selection preferences for entire classes of nouns from WordNet, while Gomez (2001, 2004) has hand-crafted an ontology of verbal predicates with over 3000 verbs, associating them with selectional restrictions (in the form of WordNet noun classes, with some modifications to the upper ontology; see Gomez, 2007) that are bound to thematic roles and the syntactic relations that realize them. Verb senses have been arranged into classes (Levin, 1993) and organized taxonomically in WordNet (Fellbaum, 1998). Chklovski and Pantel (2004) have used a lexico-syntactic pattern approach to automatically acquire a semantic network called VerbOcean, which indicates labeled relations between verbs,

while Baker, Fillmore, and Lowe (1998) have produced a semantically annotated corpus of verbal frames (cf. Fillmore, 1976) called FrameNet, which has enabled natural language processing tasks such as semantic role labeling (Gildea & Jurafsky, 2002).

1.7 Outline

The remainder of this dissertation is structured as follows.

In Chapter 2, "Literature Review," we present related research. We discuss WordNet in greater detail and present computational approaches to measuring similarity and relatedness that rely on the ontology. Other computational approaches are presented in relation to some of the cognitive literature on semantic associates and the relationship between corpus co-occurrence and semantic relatedness. We review pattern-based extraction methods for relation instance mining and examine how the semantic annotations and structural semantics of Wikipedia have been used in computational approaches to relatedness. We also discuss previous efforts to establish large-scale knowledge networks that exhibit characteristics of Quillian's semantic memory, such as ConceptNet, YAGO, DBpedia, CYC, and Freebase, to contextualize the novel contributions of our work.

In Chapters 3 and 4, "Constructing the Network: Semantic Associates of Nouns" and "Constructing the Network: From Nouns to Concepts," we detail the automatic, unsupervised acquisition of our semantic network. We begin in Chapter 3 with an examination of corpus co-occurrence and develop an information theoretic measure that gives a better indication of quantitative relatedness than simply counting the co-occurrence of words. An algorithm for determining categorical semantic association from these quantitative measurements of

relatedness is developed. Then, in Chapter 4, we present an elaborate suite of disambiguation methods that resolves related nouns in our network to noun senses in WordNet. Rather than defer evaluation of our network entirely to a separate chapter, we pause after each of these three steps (quantitative measurement of relatedness, categorical association, and disambiguation) to perform *in loco* evaluation of our progress so far.

In Chapter 5, "Coarse-Grained Word Sense Disambiguation: An Application," we evaluate the performance of our network on a word sense disambiguation task. The network is used as a plug-in knowledge source for two graph-based WSD algorithms. We compare the performance of our network on this task to that of two similar resources: WordNet (Miller, 1998) and WordNet++ (Ponzetto & Navigli, 2010). The evaluation of our network on this task serves as a supplement to the *in loco* evaluation performed throughout the network acquisition processes of Chapters 3 and 4.

In Chapter 6, "Measuring Human Perceptions of Relatedness," we present a study in which we establish a new set of relatedness norms for 122 noun pairs. The relatedness scores in our study are elicited from human participants using an established methodology that has been used in multiple studies to compile gold standard similarity norms. In this chapter, we also discuss existing gold standards for quantitative, computational measures of semantic relatedness and motivate the need for a new gold standard.

In Chapter 7, "Conclusions," we summarize the main contributions of our work. We close with a discussion of the current state of our semantic network, including elaborations on some of the relationships it contains and ideas for future work.

1.8 Style Conventions: Words and the Concepts They Denote

Throughout this dissertation, we use the terms "concepts" and "word senses" interchangeably. In distinguishing between words and the concepts they denote, we quote the former and italicize the latter, appending a sense number from WordNet when appropriate. For example, we might speak of "astronomer" co-occurring with "star" frequently in a corpus, or discuss the semantic relatedness of *astronomer#1* to both *star#1* and *star#3*, the two *celestial body* senses of the noun "star" in WordNet. We sometimes find it convenient to refer to multiple senses of a word in a more condensed format, in which case we adopt the convention of appending a set of sense numbers (as with, e.g., *star#{1,3}* to refer to both *star#1* and *star#3*).

In cases where a word's part of speech might not be clear from the context in which it appears, we append a tag before the sense number(s): *.n* for nouns, *.v* for verbs, *.a* for adjectives, and *.r* for adverbs (as with, e.g., *run.n#2* (a trial or test run) or *beach.v#1* (to land on a beach)).

For typographical reasons, we present the arguments of binary relations in small caps (e.g., *PartOf*(SPINDLE, SPINNING WHEEL)). This helps to distinguish arguments from surrounding copy text and from the relation itself, which is always italicized. This convention also frees us from the awkward and ungainly presentation of quoted arguments in roman type when dealing with relations in resources that associate words instead of concepts.

CHAPTER 2: LITERATURE REVIEW

In this chapter, we review existing literature on computational approaches to relatedness. We present several WordNet-based measures of semantic similarity and relatedness (Section 2.2), studies that establish the relationship between lexical co-occurrence and semantic association (Section 2.3), and corpus-based methods of relationship extraction that rely on lexico-syntactic pattern matching (Section 2.4) and the semantic annotations of the Wikipedia corpus (Section 2.5) to discover relatedness in semi-supervised and unsupervised settings.

In many cases, these methods have been used to acquire large-scale semantic networks. Networks that we discuss throughout this chapter include: VerbOcean (Section 2.4.1), the Never-Ending Language Learner (also Section 2.4.1), ConceptNet (Section 2.4.2), WordNet++ (Section 2.5.4), YAGO (Section 2.5.5), and DBpedia (also Section 2.5.5). The chapter concludes with a discussion of the hand-crafted knowledge networks CYC and Freebase (Section 2.6). We begin, however, with an overview of the WordNet noun ontology.

2.1 WordNet

The WordNet noun ontology (Miller, 1998), is a hand-crafted lexical database in which noun senses are organized into an inheritance system of semantic classes. Through its instantiation of a variety of labeled edges indicating, *inter alia*, subsumptive *IsA* relationships between noun senses, WordNet constitutes a partial realization of Qullian's (1968) dream of a computational model of semantic memory. It is by far the most extensive implementation of Qullian's ideas to date, and is one of the most widely used resources in natural language

processing (NLP). It has been employed in a variety of NLP applications, such as word sense disambiguation, coreference resolution, and measurement of semantic similarity and relatedness between words, word senses, and documents; incorporated into large-scale commonsense knowledge networks (most notably by ConceptNet); and recreated in several other languages.

Nouns in WordNet are broken up into noun senses that are then grouped into *synsets*—sets of noun senses grouped by synonymy.[10] These synsets form the basic concept nodes of WordNet. For example, "rook" has two senses in WordNet (see Figure 2.1 below). Its first sense, *rook#1*, is synonymous with *castle#3* (the chess piece), and the two noun senses compose the synset {*castle#3, rook#1*}. The second sense, *rook#2* is synonymous with *Corvus_frugilegus#1*, a species of bird resembling a crow; together, they compose the synset {*rook#2, Corvus_frugilegus#1*}. As with traditional dictionaries, each synset is associated with a gloss that provides a definition of the concept it denotes.

Synsets in WordNet are coordinated through a lexical inheritance hierarchy that indicates subsumptive relationships between concepts. Superordinate terms in the taxonomy are referred to as *hypernyms*, while subordinate terms are called *hyponyms*. For example, that a *rook* is a *bird* can be expressed by an *IsA* relationship between the concepts: *IsA*(ROOK, BIRD). We say that *rook* is a hyponym of *bird*, and *bird* is a hypernym of *rook*; the hypernym is a super-class of the hyponym, and is said to *subsume* the hyponym.

Hypernymy and hyponymy are transitive relations, so that if *rook* is a hyponym of *bird* and *bird* is a hyponym of *animal*, we also have that *rook* is a hyponym of *animal*. The inheritance structure of WordNet thus implies that *rook* should inherit all general properties of

10 Because indications of synonymy are incorporated into the structure of WordNet in such a fundamental way, the ontology is sometimes referred to in the literature as a thesaurus—a label that belies the power and sophistication of WordNet.

animals (they eat food to acquire energy, they reproduce, and so on), as well as properties specific to birds (they have wings, lay eggs, and so on). Although WordNet does not express all of these properties of animals and birds, the ontology establishes a general inheritance framework that can be incorporated into any knowledge base that does.

```
================================================================================
rook#1:

(chess) the piece that can move any number of unoccupied squares in a direction
parallel to the sides of the chessboard

castle, rook
 => chessman, chess piece
     => man, piece
         => game equipment
             => equipment
                 => instrumentality, instrumentation
                     => artifact, artefact
                         => whole, unit
                             => object, physical object
                                 => physical entity
                                     => entity

================================================================================
rook#2:

common gregarious Old World bird about the size and color of the American crow

rook, Corvus frugilegus
 => corvine bird
     => oscine, oscine bird
         => passerine, passeriform bird
             => bird
                 => vertebrate, craniate
                     => chordate
                         => animal, animate being, beast, brute, creature, fauna
                             => organism, being
                                 => living thing, animate thing
                                     => whole, unit
                                         => object, physical object
                                             => physical entity
                                                 => entity
================================================================================
```

Figure 2.1:
Lexical entries for “rook” in WordNet 3.0.

All concepts in the ontology are ultimately hyponyms of *entity*, which serves as the root of the hierarchy, and which has as its immediate hyponyms the dichotomous upper-level ontological concepts *physical_entity#1* and *abstraction#6* (as well as a third, more nebulous hyponym, *thing#8*, which can refer to physical or abstract things). From this dichotomous distinction between entities that are either physical or abstract flow further distinctions that categorize word senses into upper-level ontological concepts called "unique beginners." The taxonomic categorization of these unique beginners provides the basic framework for the categorization of all nouns in WordNet.

The subsumptive structure of WordNet serves as an indication of semantic similarity between concepts. Hyponymic relationships reflect similarity directly; that a *penguin* is an *aquatic bird* implies strong similarity between the two concepts. Through transitive subsumption, we can also infer the similarity of *penguin* and *bird*, although the increased distance between these nodes suggests weaker similarity than that of *penguin* to *aquatic bird*. Sister synsets in the ontology—those that are hyponyms of the same hypernym—also bear similarity to one another. For example, the shared subsumption of *flamingo* and *penguin* by *aquatic bird* suggests that they are similar entities, and that they are more similar to each other than either of them is to, say, a *crow*, which shares their subsumption by *bird*, but not by the more specific category, *aquatic bird*.

The noun ontology expresses other relationships between concepts in addition to synonymy and hyponymy, including antonymy, meronymy (part-whole relationships), attributes, derived forms, and domain terms, although these do not provide comprehensive indications of semantic relatedness. Miller (1998) points out, for example, that information about the game of

tennis is spread across the lexical database, with nothing to link together ontologically disparate concepts like tennis players, tennis equipment, tennis courts, and so on.

2.2 WordNet-Based Measures of Similarity and Relatedness

We have already informally observed that the WordNet ontology indicates semantic similarity between concepts through shared subsumption, and that it does not always indicate more general relatedness between concepts. This distinction—that similarity is only one particular type of relatedness, and that similarity relationships expressed in WordNet give us only a restricted view of the broader landscape of semantic relatedness—is a key idea that we return to throughout this dissertation, and a fact that is well established in the literature (Agirre, Alfonseca, et al., 2009; Budanitsky & Hirst, 2006; Resnik, 1999). There are many types of relatedness beyond semantic similarity, including, but not limited to, the antonymic and meronymic part-whole relations expressed in WordNet. As Budanitsky and Hirst (2006) observe, "any kind of functional relationship or frequent association" (p. 13) can relate two entities. In this section, we review several approaches that use the WordNet ontology to quantitatively measure semantic similarity and relatedness of words and concepts.

2.2.1 Preliminaries

WordNet-based measures of similarity and relatedness are typically divided into three categories: *path-based* measures that treat WordNet as a graph and examine the semantic distance between concept nodes (synsets); *information content* measures that incorporate corpus-

based probability frequencies; and *gloss-based* measures that turn to WordNet glosses for textual clues about relationships between synsets.

In the sections that follow, we denote measures of similarity between two concepts, c_1 and c_2, as $sim(c_1, c_2)$, with subscripts on the function name (sim) to distinguish between measures. Relatedness between concepts is similarly denoted $rel(c_1, c_2)$. For any such measure, it is common (Budanitsky & Hirst, 2006; Resnik, 1995) to find the similarity or relatedness between two words by choosing the two word senses that maximize the function's value, i.e.:

$$rel(w_1, w_2) = \max_{c_1 \in s(w_1),\, c_2 \in s(w_2)} rel(c_1, c_2) \tag{1}$$

where $s(w_i)$ is the set of w_i's word senses (restricted to the appropriate part of speech).

2.2.2 Path Length and the Uniformity Problem

The simplest path-based approach to similarity is to take the length of the shortest path between two concepts in a network as a direct measure of the semantic distance between them (Lee, Kim, & Lee, 1993; Rada and Bicknell, 1989; Rada, Mili, Bicknell, & Blettner, 1989). The shorter the semantic distance, the greater the conceptual similarity. If we denote this shortest path length $len(c_1, c_2)$, then we have a simple path length (PL) similarity measure:

$$sim_{PL}(c_1, c_2) = L_{max} - len(c_1, c_2) \tag{2}$$

where L_{max} is the maximum possible path length. In the case of WordNet, this is sometimes estimated as twice the depth of the hierarchy. Jarmasz and Szpakowicz (2003) notably used (2) not with WordNet, but with the hierarchical organization of classes in Roget's Thesaurus.

```
==========================================================================
flamingo#1:

large pink to scarlet web-footed wading bird with down-bent bill; inhabits
brackish lakes

flamingo
 => wading bird, wader
     => aquatic bird
         => bird
             => ...
==========================================================================
penguin#1:

short-legged flightless birds of cold southern especially Antarctic regions
having webbed feet and wings modified as flippers

penguin
 => sphenisciform seabird
     => seabird, sea bird, seafowl
         => aquatic bird
             => bird
                 => ...
==========================================================================
seagull#1:

mostly white aquatic bird having long pointed wings and short legs

gull, seagull, sea gull
 => larid
     => coastal diving bird
         => seabird, sea bird, seafowl
             => aquatic bird
                 => bird
                     => ...
==========================================================================
```

Figure 2.2:
Lexical entries for *flamingo#1*, *penguin#1*, and *seagull#1* in WordNet 3.0. The concepts are increasingly distant from the superordinate concept *aquatic_bird#1*, highlighting uniformity disparity in the ontology, where not all edges convey equal semantic distance between concepts.

As it relates to WordNet, a widely recognized problem with the path length approach is that different sections of the ontology make more fine-grained vertical distinctions between superordinate and subordinate classes. Resnik (1999) frames this as the *uniformity problem*,

because the underlying assumption of the path length measure in (2) is that all edges in the ontology indicate uniform semantic distance between concepts. This simply is not the case in WordNet. For example, *flamingo*, *penguin*, and *seagull* are increasingly distant from the superordinate *aquatic bird* class in WordNet, despite the intuitive notion that they ought to be (at least approximately) semantically equidistant from *aquatic bird* (see Figure 2.2 above). (It is also of peripheral interest that the distance of these concepts from *bird* is misaligned with the prototypicality effect (Rosch, 1978). Intuitively speaking, we would expect *seagull* to be closer to *bird* than *penguin* and *flamingo* are, because the latter two are less prototypical examples of birds.)

To address the uniformity problem, Wu and Palmer (1994) introduced a path-based measure that scaled path distance between concepts by the depth of their lowest common subsumer (LCS) in the ontology. The LCS of two concepts, denoted $lcs(c_1, c_2)$, is the deepest concept in the ontology categorizing both c_1 and c_2, where depth is defined as the distance of a concept from the root of the ontology (*entity*). The scaled similarity measure of Wu and Palmer is commonly given in the form:

$$sim_{WP}(c_1, c_2) = \frac{2 \times depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)} \quad (3)$$

The role of the LCS as a scaling factor may at first seem unclear in the presentation of (3), until we realize that the denominator accounts twice for the depth of $lcs(c_1, c_2)$ and once for the distance of the shortest path between c_1 and c_2, which must necessarily go through $lcs(c_1, c_2)$, and is therefore simply $len(c_1, c_2)$ (the distance of the shortest path between the two concepts). Thus, (3) can be rewritten as:

$$sim_{WP}(c_1, c_2) = \frac{2 \times depth(lcs(c_1, c_2))}{2 \times depth(lcs(c_1, c_2)) + len(c_1, c_2)} \quad (4)$$

In a similar vein, Leacock and Chodorow (1998) developed a normalized path length measure that took into account the maximum depth of the ontology:

$$sim_{LC}(c_1, c_2) = -\log \frac{len(c_1, c_2)}{2 \times \max_{c \in WordNet} depth(c)} \quad (5)$$

Hirst and St-Onge (1998) developed a more elaborate path-based measure that relied not only on the *IsA* taxonomy of WordNet, but also meronymic and antonymic relationships. Because the measure includes relations beyond hyponymy, it is often considered to capture not just similarity, but relatedness. Loosely speaking, it finds paths between concepts that have no more than five edges, and those edges are assigned orientations, or directions, based on the relations they denote: hypernymic and meronymic relations are considered upward links, hyponymic and holonymic relations are considered downward links, and antonymic relations are considered horizontal. A *turns* function establishes a shortest path between two concepts (subject to certain technical restrictions) and indicates the number of directional changes from edge to edge along the path. The resulting measure of relatedness is given as:

$$rel_{HS}(c_1, c_2) = C - len(c_1, c_2) - k \times turns(c_1, c_2) \quad (6)$$

In (6), C and k are constants; Hirst and St-Onge used 8 and 1, respectively.

2.2.3 Information Content

In contrast to the path length normalization approaches of Wu and Palmer (1994) and Leacock and Chodorow (1998), Resnik (1995) observed that similarity between concepts can be measured by "the extent to which they share information in common" (p. 448) and established *information content* (i.e., negative log likelihood) as a direct measure of similarity:

$$sim_R(c_1, c_2) = -\log p(lcs(c_1, c_2)) \tag{7}$$

where $p(c)$ is the probability of c or one of its instances (i.e., hyponymic terms) occurring in a corpus. To estimate the probability of each WordNet noun class's occurrence, Resnik used noun frequencies from the 100 million word Brown Corpus and, in the case of polysemous nouns, distributed occurrence frequency evenly across all possible senses of those nouns:

$$p(c) = \frac{freq(c)}{N} \tag{8}$$

$$freq(c) = \sum_{n \in words(c)} count(n) \tag{9}$$

where $words(c)$ is the set of all words categorized by c, $count(c)$ is the number of times c occurs in the corpus, and N is the total number of nouns in the corpus, excluding those that are not represented in WordNet. For example, since the root of the ontology, *entity*, categorizes all nouns in the ontology, $p(entity) = 1$. Thus, *entity* has no information content; i.e., $log(p(entity)) = 0$, and concepts that have *entity* as their lowest common subsumer in the ontology are considered maximally dissimilar, while those with more specific and less frequently occurring lowest common subsumers are considered to exhibit stronger similarity.

One of the nice features of Resnik's information content approach is that it can be adapted to any corpus while still leveraging the full power of the WordNet ontology, simply by recalculating $p(c)$ from the frequency distribution of the new corpus. One of the shortcomings of the approach, however, is the even distribution of occurrence frequency across all senses of polysemous nouns. The model does not take into account the fact that some senses of a noun are more common than others, or that polysemous nouns tend to keep the same meaning when repeated throughout the same discourse (Gale, Church, & Yarowsky, 1992; Yarowsky, 1993). Another limitation of the Resnik measure is that it does not account for semantic distance between concepts at all, so that two pairs of concepts with the same LCS are considered equally similar. For example, *lcs*(*currency*, *credit card*) = *lcs*(*currency*, *dissertation*) = *abstraction*, and therefore sim_R(*currency*, *credit card*) = sim_R(*currency*, *dissertation*).

In light of the latter shortcoming, Jiang and Conrath (1997) adjusted the information content framework of Resnik to re-weight the semantic distance between a child and its parent node in the WordNet graph in terms of the conditional probability $p(c|parent(c))$:

$$dist_{JC}(c, parent(c)) = -\log p(c|parent(c)) \tag{10}$$

Taken as a measure of semantic distance between two *arbitrary* concepts, Jiang and Conrath's function accounts once again for the distance of the shortest path between those concepts. In its most common form, the measure reduces to:

$$dist_{JC}(c_1, c_2) = 2 \times \log p(lcs(c_1, c_2)) - (\log p(c_1) + \log p(c_2)) \tag{11}$$

Since (11) is a measure of semantic *distance*, smaller values of $dist_{JC}(c_1, c_2)$ indicate greater semantic similarity.

Our final information content measure comes from Lin (1998), whose theoretical work, when applied to an *IsA* taxonomy like WordNet, yields the following:

$$sim_{Lin}(c_1, c_2) = \frac{2 \times \log p(lcs(c_1, c_2))}{\log p(c_1) + \log p(c_2)} \quad (12)$$

One might notice that Lin's measure bears striking similarity to Wu and Palmer's measure, particularly when viewed in the form of (3). Lin showed that Wu and Palmer's measure was actually a special case of (12) in which all edges between nodes in the *IsA* taxonomy are equally weighted. Because sim_{Lin} can be considered a generalization of sim_{WP}, some comparative studies of relatedness and similarity measures (e.g., Budanitsky & Hirst, 2006) evaluate only the former and not the latter directly.

2.2.4 Gloss-Based Measures

Lesk (1986) presented a dictionary-based approach for measuring relatedness that has since been applied to WordNet sense glosses to glean more general indications of relatedness than the similarity expressed through the ontology's *IsA* taxonomy. The sense glosses of WordNet serve as traditional dictionary definitions, and as such, they contain content words (nouns, adjectives, verbs, and adverbs) to which those noun senses otherwise have no direct links in the ontology. For example, the gloss of *penguin#1* mentions "Antarctic regions," providing a loose semantic link between the penguin and its natural habitat that is otherwise absent from the ontology. The relatedness measure of Lesk simply counts the number of content words in common between two dictionary definitions (in our case, WordNet sense glosses), with the

natural assumption that more strongly related concepts will have more words in common between their glosses:

$$rel_{Lesk}(c_1, c_2) = overlap(gloss(c_1), gloss(c_2)) \quad (13)$$

where $gloss(c_i)$ is the WordNet gloss of c_i, and $overlap(str_1, str_2)$ counts the number of content words in common between strings str_1 and str_2. The measure can of course be used to disambiguate nouns by maximizing $rel_{Lesk}(w_1, w_2)$ for all senses of w_1 and w_2.

Banerjee and Pedersen (2003) developed an extended version of Lesk's overlap measure, sometimes referred to as "Extended Lesk" or "ExtLesk," that works as follows: an extended gloss string, $gloss_{R_j}(c_i)$, is defined for relation R_j and concept c_i as the concatenated glosses of all concepts (synsets) related to c_i through the relation R_j (e.g., hypernymic, hyponymic, meronymic, holonymic, troponymic, attribute, and gloss relations in WordNet). (When R_j is the gloss relation, $gloss_{R_j}(c_i)$ returns the gloss of c_i.) Given two concepts, c_1 and c_2, the extended gloss strings $gloss_{R_1}(c_1)$ and $gloss_{R_2}(c_2)$ are compared for every pair of relations R_1 and R_2 from the set of relations given above. Instead of the traditional overlap measure of Lesk, which simply counts the number of words that the two strings have in common, Banerjee and Pedersen introduced an overlap function that awarded more points for multi-word substrings (such as open-form compound nouns) common to both strings (namely by squaring the number of words in each substring overlap and returning the sum of those as the relatedness score). Thus, we have:

$$rel_{BP}(c_1, c_2) = \sum_{R_i \in R} \sum_{R_j \in R} overlap_{sq}(gloss_{R_i}(c_1), gloss_{R_j}(c_2)) \quad (14)$$

where $overlap_{sq}(str_1, str_2)$ implements the multi-word square-of-word-count scoring mechanism described above, and R is the set of relations listed above. We discuss the ExtLesk algorithm in further detail in Section 5.3, where we use it in conjunction with our semantic network on a word sense disambiguation task.

Patwardhan and Pedersen (2006) also used WordNet glosses to measure relatedness between concepts, but they took a more geometric approach than the overlap methods described above. Patwardhan and Pedersen established a second-order vector space from the WordNet glosses, and measured the relatedness of two concepts as the cosine of their respective gloss vectors in that space. In a more graph-oriented approach, Hughes and Ramage (2007) used random walks on the WordNet graph (a Markov chain with transition probabilities between nodes) to measure relatedness between concepts. WordNet relations between synsets were used to establish edges in their graph, and glosses were also used to induce edges between nodes.

2.2.5 Evaluation

Resnik (1999) observed that "the worth of a similarity measure is in its fidelity to human behavior, as measured by predictions of human performance on experimental tasks" (p. 95). Toward this end, the most common gold standard evaluation of a similarity measure is its comparison to human judgments of similarity. For this purpose, most studies turn to the data from Rubenstein and Goodenough (1965) and Miller and Charles (1991). In these studies, participants rated the "similarity of meaning" of noun pairs on a scale of 0.0 ("semantically unrelated") to 4.0 ("highly synonymous"). Rubenstein and Goodenough had participants evaluate 65 word pairs in this manner. Miller and Charles then replicated the experiment using 30 of the

original 65 word pairs. Comparison to mean similarity scores from these studies has also emerged as a standard evaluation of relatedness measures in the literature, despite the fact that the authors specifically elicited similarity ratings of the noun pairs in their studies, and not ratings of semantic relatedness. We defer our critique of this particular method of evaluating relatedness measures to Chapter 6.

The similarity and relatedness measures we have discussed in this section are listed below in Table 2.1. Table 2.2 (below on page 36) presents the results of several comparative studies evaluating the correlation of these measures to the Miller and Charles (M&C) and Rubenstein and Goodenough (R&G) data.

Table 2.1:
Summary of similarity and relatedness measures presented in this section.

Type of Measure	**Measure**	**Authors**
Path-Based Similarity	sim_{PL}	Rada et al. (1989)
	sim_{WP}	Wu and Palmer (1994)
	sim_{LC}	Leacock and Chodorow (1998)
	sim_{JS}	Jarmasz and Szpakowicz (2003)
Information Content Similarity	sim_{R}	Resnik (1995)
	$dist_{\mathrm{JC}}$	Jiang and Conrath (1997)
	sim_{Lin}	Lin (1998)
Gloss-Based Relatedness	rel_{Lesk}	Lesk (1986)
	rel_{BP}	Banerjee and Pedersen (2003)
	rel_{PP}	Patwardhan and Pedersen (2006)
Graph/Path-Based Relatedness	rel_{HS}	Hirst and St-Onge (1998)
	rel_{HR}	Hughes and Ramage (2007)

Table 2.2:
Correlations of various similarity and relatedness measures to M&C and R&G similarity scores. Correlation data come from four comparative studies that replicated several measures. Self-reported results are included where available.

From Jarmasz and Szpakowicz (2003) (using Pearson's product-moment correlation, r)

Data	sim_{PL}	sim_{LC}	sim_{JS}	sim_{R}	$dist_{JC}$	sim_{Lin}	rel_{BP}	rel_{PP}	rel_{HS}	rel_{HR}
M&C	0.732	0.821	**0.878**	0.775	0.695	0.823	--	--	0.689	--
R&G	0.787	**0.852**	0.818	0.800	0.731	0.834	--	--	0.732	--

From Budanitsky and Hirst (2006) (using Pearson's product-moment correlation, r)

Data	sim_{PL}	sim_{LC}	sim_{JS}	sim_{R}	$dist_{JC}$	sim_{Lin}	rel_{BP}	rel_{PP}	rel_{HS}	rel_{HR}
M&C	--	0.816	--	0.774	**0.850**	0.829	--	--	0.744	--
R&G	--	**0.838**	--	0.779	0.781	0.819	--	--	0.786	--

From Patwardhan and Pedersen (2006) (using Spearman's rank correlation, ρ)

Data	sim_{PL}	sim_{LC}	sim_{JS}	sim_{R}	$dist_{JC}$	sim_{Lin}	rel_{BP}	rel_{PP}	rel_{HS}	rel_{HR}
M&C	--	0.74	--	0.72	0.73	0.70	0.81	**0.91**	--	--
R&G	--	0.77	--	0.72	0.75	0.72	0.83	**0.90**	--	--

From Hughes and Ramage (2007) (using Spearman's rank correlation, ρ)

Data	sim_{PL}	sim_{LC}	sim_{JS}	sim_{R}	$dist_{JC}$	sim_{Lin}	rel_{BP}	rel_{PP}	rel_{HS}	rel_{HR}
M&C	--	--	--	--	0.653	0.625	0.869	0.888	--	**0.904**
R&G	--	--	--	--	0.584	0.599	**0.829**	0.789	--	0.817

Self-Reported			*r*	*r*	*r*	*r*	*n/a*	*ρ*		*ρ*
Data	sim_{PL}	sim_{LC}	sim_{JS}	sim_{R}	$dist_{JC}$	sim_{Lin}	rel_{BP}	rel_{PP}	rel_{HS}	rel_{HR}
M&C	--	0.740	0.878	0.791	0.828	0.834	0.67	0.91	--	0.904
R&G	--	--	0.818	--	--	--	0.60	0.90	--	0.817

Where available, the self-reported results of the measures' original authors are also reported. (Some authors, such as Hirst and St-Onge, did not evaluate correlation to the M&C and

R&G ratings in their original studies.) In Table 2.2, the presentation of sim_{PL} from Jarmasz and Szpakowicz is their own implementation of a simple shortest path length measure using WordNet. Recall that the Jarmasz and Szpakowicz measure, sim_{JS}, is also a simple shortest path length measure, but that it evaluates paths through Roget's Thesaurus rather than WordNet.

Some of the studies cited in Table 2.2 use Spearman's rank correlation (ρ), while others use Pearson's product-moment correlation (r). Pearson's correlation measures the strength of the linear relationship between two datasets, while Spearman's evaluates the relationship between ordered rankings of the data points, without respect to a linear relationship between values. Both coefficients range from 0.0 to 1.0 inclusively, with higher values indicating better correlation to the human-assigned scores; 1.0 would indicate perfect correlation.

The results reported for individual measures vary, sometimes widely, across the literature. Budanitsky and Hirst (2006) offer possible explanations for these discrepancies between studies: a) variations in how authors count frequency with respect to compound nouns, b) the use of different version of WordNet, and c) the use of different corpora when harvesting word frequency data and probability distributions for WordNet classes.

In addition to examining correlation to human judgments, surveys of similarity and relatedness measures typically select an applied task to provide further evaluation of those measures. These tasks vary widely in the literature. Jarmasz and Szpakowicz (2003) used a standardized synonym test (for each question, the system attempted to identify which one of four multiple choice answers was "nearest in meaning" to a given target word), which is of course well suited to evaluating similarity measures, but does not provide a natural testbed for evaluating relatedness measures. Budanitsky and Hirst (2006) evaluated measures—both

similarity and relatedness—on a malapropism detection task (if a word like *dairy* bore no semantic relation to nearby words, but a word with a very similar spelling (e.g., *diary*) did, the latter was to be suggested as a spelling correction). Patwardhan and Pedersen (2006) employed similarity and relatedness measures in a word sense disambiguation task (Senseval-2). Hughes and Ramage (2007) restricted their evaluation to human judgment correlation on three datasets: M&C, R&G, and a third dataset, WordSim353 (Finkelstein et al., 2002), which has certain limitations as a gold standard (see, e.g., the critique of Jarmasz & Szpakowicz, 2003), and which we discuss in more detail in Chapter 6.

2.3 Lexical Co-occurrence and Semantic Association

Several studies have examined the relationship between lexical co-occurrence and semantic association. In this section, we review corpus-based approaches to semantic similarity (which tend to be distributional in nature), and semantic relatedness (which tend to rely on lexical co-occurrence frequency).

2.3.1 Distributional Approaches to Semantic Similarity

Distributional approaches have been widely employed in the literature to measure similarity of meaning as a function of the similarity of contexts in which words occur throughout a corpus (Gorman & Curran, 2006; Grefenstette, 1994; Lin, Zhao, Qin, & Zhou, 2003; Rubenstein & Goodenough, 1965; Sahlgren, 2008; Weeds, 2003; Weeds & Weir, 2006). The observation that "words which are similar in meaning occur in similar contexts" (Rubenstein &

Goodenough, 1965, p. 627) harks back to the Distributional Hypothesis of Harris (1954/1985) and the oft-quoted Firthian view that "[y]ou shall know a word by the company it keeps" (Firth, 1957/1968, p. 179).

In an early empirical investigation into the distributional hypothesis, Rubenstein and Goodenough developed a measure of word similarity based on contextual overlaps, as follows:

$$sim_{RG}(a,b) = \frac{|ctx(a) \cap ctx(b)|}{\text{MIN}\{|ctx(a)|, |ctx(b)|\}} \quad (15)$$

where $ctx(w)$ is the context of w (i.e., the set of all words in all sentences containing w).

To evaluate their measure, Rubenstein and Goodenough first had 51 human participants rate the "similarity of meaning" of 65 noun pairs on a scale of 0.0 to 4.0, with higher values indicating stronger similarity. The mean similarity scores assigned by participants for the 65 noun pairs are given below in Table 2.3. As we saw in the previous section, this dataset has become a time-honored gold standard for evaluation of computational similarity measures.

Rubenstein and Goodenough then had a separate group of individuals create a corpus of sentences containing the 48 distinct nouns represented in their 65 pairs from Table 2.3. The nouns were divided into two sets of equal size, *A* and *B*, such that the R&G noun pairs always contained exactly one term from *A* and one term from *B*. One group of participants was given the nouns in set *A* and asked to produce two sentences for each of them. A second group performed the same task using the nouns in set *B*. Participants were instructed to write sentences at least ten words in length and to use the words they were given as nouns. The resulting corpus consisted of 4,800 sentences and approximately 64,800 words.

Table 2.3:
Subjective similarity score judgments from Rubenstein and Goodenough (1965).

#	Word Pair		Score	#	Word Pair		Score
1	cord	smile	0.02	34	car	journey	1.55
2	rooster	voyage	0.04	35	cemetery	mound	1.69
3	noon	string	0.04	36	glass	jewel	1.78
4	fruit	furnace	0.05	37	magician	oracle	1.82
5	autograph	shore	0.06	38	crane	implement	2.37
6	automobile	wizard	0.11	39	brother	lad	2.41
7	mound	stove	0.14	40	sage	wizard	2.46
8	grin	implement	0.18	41	oracle	sage	2.61
9	asylum	fruit	0.19	42	bird	crane	2.63
10	asylum	monk	0.39	43	bird	cock	2.63
11	graveyard	madhouse	0.42	44	food	fruit	2.69
12	glass	magician	0.44	45	brother	monk	2.74
13	boy	rooster	0.44	46	asylum	madhouse	3.04
14	cushion	jewel	0.45	47	furnace	stove	3.11
15	monk	slave	0.57	48	magician	wizard	3.21
16	asylum	cemetery	0.79	49	hill	mound	3.29
17	coast	forest	0.85	50	cord	string	3.41
18	grin	lad	0.88	51	glass	tumbler	3.45
19	shore	woodland	0.90	52	grin	smile	3.46
20	monk	oracle	0.91	53	serf	slave	3.46
21	boy	sage	0.96	54	journey	voyage	3.58
22	automobile	cushion	0.97	55	autograph	signature	3.59
23	mound	shore	0.97	56	coast	shore	3.60
24	lad	wizard	0.99	57	forest	woodland	3.65
25	forest	graveyard	1.00	58	implement	tool	3.66
26	food	rooster	1.09	59	cock	rooster	3.68
27	cemetery	woodland	1.18	60	boy	lad	3.82
28	shore	voyage	1.22	61	cushion	pillow	3.84
29	bird	woodland	1.24	62	cemetery	graveyard	3.88
30	coast	hill	1.26	63	automobile	car	3.92
31	furnace	implement	1.37	64	midday	noon	3.94
32	crane	rooster	1.41	65	gem	jewel	3.94
33	hill	woodland	1.48				

Rubenstein and Goodenough used their manually-generated corpus to compare the results of their overlap measure to the mean similarity scores from their human subjects, and found that their overlap measure reliably predicted strong synonymy (values greater than 3.0 on their scale). However, the authors observed that dissimilar nouns exhibited too much homogeneity in their contexts for the measure to make useful distinctions between "low" and "medium" similarity.

Hindle (1990) showed that predicate-argument structure could also play a useful role in measuring semantic similarity. He found that nouns exhibiting high mutual information with many of the same verbs—and, importantly, via the same grammatical relation to those verbs, such as subject or object positions—tended to be semantically similar. For example, one can establish the similarity of *apples* and *peaches* by the fact that both can be *bought*, *sold*, *baked*, *harvested*, *grown*, *picked*, *sliced*, *eaten*, and so on.

Subsequent to these early approaches, one of the most common methods of measuring semantic similarity has been cosine similarity, often categorized in the literature as a geometric similarity measure (Sahlgren, 2008; Weeds, 2003): the context of a word is represented as a normalized co-occurrence frequency vector, and the similarity of two words is simply the cosine of the angle between their representative vectors, which ranges from 0.0 (completely dissimilar) to 1.0 (contextually identical).

2.3.2 Co-occurrence Approaches to Semantic Relatedness

Other studies have established co-occurrence as an indication of semantic association (Church & Hanks, 1990; McKoon & Ratcliff, 1992), and have shown that co-occurrence

frequency correlates to association strength (Chaudhari, Damani, & Laxman, 2011; Spence & Owens, 1990; Wettler & Rapp, 1993).

Spence and Owens (1990) first established this correlation by developing a relatedness measure that relied on adjusted co-occurrence frequencies from a large corpus, as follows:

$$rel_{SO}(x, y) = f_c(x, y) - f_c(x, y') \tag{16}$$

where $f_c(x, y)$ is the frequency with which y follows x within a window of c characters, and y' is a matched control word with corpus frequency and character count approximately equal to that of y.

Spence and Owens compared the values produced by their measure to a subset of the Palermo and Jenkins (1964) association norms. The Palermo and Jenkins data were one of the earliest collections of association norms, and were elicited from human participants in a free word association task. Individuals in their study were presented with lists of *stimulus* words (200 in total) and instructed to write the first *response* word that came to mind for each stimulus. Participants were restricted to one response word per stimulus. In their study, Palermo and Jenkins elicited participation from 500 students (250 male, 250 female) in each of grades 4 through 8, 10, and 12, as well as 1,000 college students (500 male, 500 female). The number of people responding to a given stimulus with a particular word was taken as a direct indication of the association strength between stimulus and response.

Spence and Owens restricted their consideration to the responses of college students. Thus, the theoretical maximum value of association strength was 1,000 (all college students responding to some stimulus with the same response). They also limited their consideration to a subset of 47 stimuli, choosing words that were concrete nouns, were not frequently used as

adjectives, and occurred above a frequency threshold of 1/100,000 in the 100 million word Brown Corpus. From these stimuli, Spence and Owens derived their 47 noun pairs by choosing the noun from Palermo and Jenkins with the greatest response frequency for each stimulus.

The associate pairs used by Spence and Owens are presented below in Table 2.4 along with their association strength from the Palermo and Jenkins norms. (For example, that the association strength of the pair *baby—boy* is 107 reflects the fact that 107 out of the 1,000 college students in Palermo and Jenkins' study gave "boy" as their first response to the stimulus word, "baby.") Frequency of co-occurrence of those nouns is derived from the Brown Corpus. For each stimulus, an unrelated control word is also given, along with its frequency of co-occurrence with the stimulus noun. Spence and Owens' measure of association strength (rel_{SO}) is given in the right-most column. Co-occurrence figures in Table 2.4 are derived using a 250-character window.

From their experiments, Spence and Owens established four key results: a) semantically related words tend to co-occur more frequently in a corpus than unrelated words; b) association strength correlates to adjusted co-occurrence frequency (rel_{SO}), albeit weakly ($r = 0.42$, $p < 0.01$); c) strongly associated nouns tend to co-occur more closely than weakly associated nouns (i.e., as association strength diminishes, lexical distance between stimulus and response in a corpus increases); and d) the effects of (a) and (c) are observable even when considering co-occurrence windows up to 1,000 characters in length, and the effect of (b) is observable up to window widths of 2,000 characters.

Table 2.4:
Corpus co-occurrence and adjusted co-occurrence (rel_{SO}) frequencies from Spence and Owens (1990) on select noun pairs from the Palermo and Jenkins (1964) association norms. Window size is 250 characters.

Stimulus Word	Response Word	Association Strength	Freq.	Unrelated Word	Freq.	rel_{SO}
Baby	Boy	107	1	Board	0	1
Bath	Water	264	2	Hand	0	2
Bible	God	316	10	Door	1	9
Boy	Girl	705	20	Land	0	20
Bread	Butter	466	0	Pistol	0	0
Butter	Bread	575	0	Seed	1	-1
Carpet	Rug	311	0	Map	0	0
Cars	Trucks	107	3	Bombers	0	3
Chair	Table	428	2	Road	0	2
Child	Baby	173	4	Wind	1	3
Children	Kids	188	2	Rice	0	2
City	Town	232	8	Table	1	7
Cottage	House	264	2	School	0	2
Doctor	Nurse	173	2	Basket	0	2
Dogs	Cats	679	2	Drops	0	2
Doors	Windows	358	0	Troops	0	0
Earth	Dirt	143	0	Meat	1	-1
Fingers	Hand	341	5	Night	1	4
Foot	Shoe	255	1	Purse	0	1
Fruit	Apple	450	0	Heel	0	0
Girl	Boy	598	12	Land	0	12
Hand	Foot	228	1	Song	2	-1
Head	Hair	194	13	Food	2	11
House	Home	230	21	Year	14	7
King	Queen	651	1	Seed	0	1
Lamp	Light	706	6	Church	0	6
Lion	Tiger	216	0	Canoe	0	0
Man	Woman	624	28	Court	5	23

Stimulus Word	Response Word	Association Strength	Freq.	Unrelated Word	Freq.	rel_{SO}
Moon	Star	236	0	Vein	0	0
Mountain	Hill	213	0	Boat	0	0
Music	Song	164	7	Dust	0	7
Needle	Thread	457	3	Stove	0	3
Ocean	Water	362	7	Hand	0	7
Priest	Church	225	0	Room	0	0
River	Water	286	8	Hand	0	8
Salt	Pepper	408	7	Posture	0	7
Sheep	Lamb	182	0	Lace	0	0
Shoes	Feet	358	3	Word	0	3
Soldier	Man	177	1	Time	1	0
Stem	Flower	398	0	Giant	0	0
Stomach	Food	242	1	Club	0	1
Street	Road	118	1	Table	1	0
Table	Chair	691	3	Dream	0	3
Tobacco	Smoke	482	2	Muscle	0	2
Whiskey	Drink	328	0	Team	0	0
Window	Glass	216	8	Bridge	1	7
Woman	Man	528	25	Years	5	20

The word counting approach of Spence and Owens was applied to a limited set of noun pairs, and one of its weaknesses was that it required that both nouns be given *a priori* in order to measure relatedness. Although they provided some evidence that unrelated pairs of nouns co-occurred infrequently, their approach gave no indication of how to deal with spurious cases of high co-occurrence frequency. Consider, e.g., the co-occurrence frequency in Table 2.4 of related nouns “house” and “woman” with the unrelated, matched control word, “year(s)” (14 and 5, respectively). In comparison, the relatively low co-occurrence frequency of, e.g., related nouns

"house" and "home," suggests that "year(s)" might have a tendency to co-occur frequently with unrelated nouns in the corpus.

Church and Hanks (1990), in contrast, were interested in mining corpora for semantic association with only the stimulus, or target, pre-specified. They treated recovery of associates as an open-ended task, examining values of all words co-occurring with a target of interest (with the restriction that word pairs must co-occur at least five times, as their measure was prone to error with lower co-occurrence frequencies). To measure association strength, and to aid in quashing noise from spurious co-occurrence of unrelated terms, Church and Hanks introduced an *association ratio*, which was essentially an estimate of mutual information:

$$rel_{CH}(x, y) = I(x, y) = \log_2 \frac{P_w(x, y)}{P(x)P(y)} \tag{17}$$

where $P_w(x, y)$ is the normalized co-occurrence frequency of x and y within a window of w words:

$$P_w(x, y) = \frac{f_w(x, y)}{N} \tag{18}$$

where $f_w(x, y)$ is the co-occurrence frequency of x and y (the frequency with which y follows x within a window of w words), and N is the size of the corpus (in words). Similarly, $P(x)$ and $P(y)$ are the normalized unigram (occurrence) frequencies of x and y:

$$P(x) = \frac{f(x)}{N} \tag{19}$$

where $f(x)$ is the raw occurrence frequency of x in the corpus.

Church and Hanks observed that their association ratio, when applied to the 15 million word 1987 AP corpus and the 36 million word 1988 AP corpus, produced results that seemed intuitively appealing. For example, some of the strongest and weakest associates of "doctor" inferred by their measure are presented in Table 2.5.

Table 2.5:
Strong and weak associates of "doctor" using the association ratio (rel_{CH}).
Data is taken from Church and Hanks (1990).

rel_{CH}(x, y)	*freq*(x, y)	*freq*(x)	x	*freq*(y)	y
11.3	12	111	honorary	621	doctor
11.3	8	1105	doctors	44	dentists
10.7	30	1105	doctors	241	nurses
9.4	8	1105	doctors	154	treating
9.0	6	275	examined	621	doctor
8.9	11	1105	doctors	317	treat
8.7	25	621	doctor	1407	bills
8.7	6	621	doctor	350	visits
8.6	19	1105	doctors	676	hospitals
8.4	6	241	nurses	1105	doctors
		. . .			
0.96	6	621	doctor	73785	with
0.95	41	284690	a	1105	doctors
0.93	12	84716	is	1105	doctors

Notice that some of the strong associates in Table 2.5 are verbs and adjectives; Church and Hanks observed that their association ratio was not just useful for discovering noun-noun associations, but that it was also useful for discovering associates of other parts of speech, and for function words in addition to content words (e.g., the preposition "to" was found to be

strongly associated with the verbs “alluding,” “amounted,” “relating,” “reverted,” “resorting,” and so on; the infinitival “to” was found to be strongly associated with verbs like “obligated,” “trying,” “compelled,” “supposed,” “vowing,” “tends,” “tries,” and so on). Their mutual information approach was successful at discovering meaningful verb-object relationships in both directions, as well. For example, Table 2.6 below shows the re-ordering effect of mutual information as compared with co-occurrence frequency for direct objects of the verb “drink,” as well as the same for verbs that appeared in their corpus with “telephone” as a direct object.

Table 2.6:
Direct objects of the verb “drink” and verbs with “telephone” as a direct object. The re-ordering effect of the association ratio is evident in comparison to corpus co-occurrence frequencies. Data are excerpted from Church and Hanks (1990).

“*What Can You Drink?*”

verb – x	**object – y**	rel_{CH}**(x, y)**	***freq*(x, y)**
drink	martinis	12.6	3
drink	cup water	11.6	3
drink	champagne	10.9	3
drink	beverage	10.8	8
drink	cup coffee	10.6	2
drink	cognac	10.6	2

“*What Can You Do to a Telephone?*”

verb – x	**object – y**	rel_{CH}**(x, y)**	***freq*(x, y)**
sit by	telephone	11.78	7
disconnect	telephone	9.48	7
answer	telephone	8.80	98
hang up	telephone	7.87	3
tap	telephone	7.69	15
pick up	telephone	5.63	11

Although Church and Hanks did not provide a quantitative analysis of their findings or compare their results to association norms, and although they provided limited anecdotal evidence for the success of their association ratio, their work established the usefulness of information theoretic measures in navigating the complexities and noise of a large corpus to discover indications of semantic relatedness.

2.4 Lexico-Syntactic Pattern Matching

Lexico-syntactic pattern matching has seen wide use in the literature for extracting semantic relationships from text (Berland & Charniak, 1999; Etzioni et al., 2004; Girju et al., 2006; Hearst, 1992; Moldovan, Badulescu, Tatu, Antohe, & Girju, 2004; Pantel & Pennacchiotti, 2006). The technique has also been employed in question answering systems (Fleischman, Hovy, & Echihabi, 2003; Ravichandran & Hovy, 2002) and used to generate semantic lexicons (Riloff & Jones, 1999), semantic relations between verbs (Chklovski & Pantel, 2004), and large-scale semantic networks (Carlson et al., 2004; Liu & Singh, 2004a).

An early approach to discovering relatedness between nouns saw the use of lexico-syntactic patterns to harvest specific types of relations from large corpora. Hearst (1992) was the first to embark on this approach, using pattern matching to automatically discover hyponymic relationships, many of which were not articulated in the WordNet ontology. For example, the pattern ***NP***{*, **NP***}*{,} *or other* ***NP*** was used to establish all the former NPs as hyponyms of the latter, as in, "... temples, treasuries, and other important civic buildings, …" where we see that *temple* and *treasury* are hyponyms of *civic building*. Hearst used a set of six lexico-syntactic patterns that she manually specified in a five-step process, as follows: a) decide on a relation of

interest (in Hearst's case, hyponymy); b) list examples that typify the relation, such as *hyponym*(ENGLAND, COUNTRY); c) search the corpus for sentences where these terms co-occur; d) inspect those sentences and infer general patterns that associate terms under this relation; and e) use those patterns to extract further examples that typify the relation. If additional patterns are desired, we can repeat the process from step (b) using the new examples garnered from step (e).

Table 2.7:
Heart's lexico-syntactic patterns for hyponymic relationships, with examples extracted from Wikipedia. $\boldsymbol{NP}_{\uparrow}$ indicates a hypernym; $\boldsymbol{NP}_{\downarrow}$ indicates a hyponym.

#	Pattern	Examples
(P1)	$\boldsymbol{NP}_{\uparrow}$ *such as* {$\boldsymbol{NP}_{\downarrow}$, }*{*or*|*and*} $\boldsymbol{NP}_{\downarrow}$	*creatures such as minotaurs, werewolves, and hags* → *hyponym*(MINOTAUR, CREATURE) → *hyponym*(WEREWOLF, CREATURE) → *hyponym*(HAG, CREATURE)
(P2)	*such* $\boldsymbol{NP}_{\uparrow}$ *as* {$\boldsymbol{NP}_{\downarrow}$, }*{*or*|*and*} $\boldsymbol{NP}_{\downarrow}$	*such cities as Berlin, Hamburg, Merseburg, Münster, Kassel, Hannover, and Cologne* → *hyponym*(BERLIN, CITY) → *hyponym*(HAMBURG, CITY) → *etc*…
(P3)	$\boldsymbol{NP}_{\downarrow}${, $\boldsymbol{NP}_{\downarrow}$}*{,} *or other* $\boldsymbol{NP}_{\uparrow}$	*wastebasket, trashcan, or other garbage receptacle* → *hyponym*(WASTEBASKET, GARBAGE RECEPTACLE) → *hyponym*(TRASHCAN, GARBAGE RECEPTACLE)
(P4)	$\boldsymbol{NP}_{\downarrow}${, $\boldsymbol{NP}_{\downarrow}$}*{,} *and other* $\boldsymbol{NP}_{\uparrow}$	*engineering, healthcare, and other professions* → *hyponym*(ENGINEERING, PROFESSION) → *hyponym*(HEALTHCARE, PROFESSION)
(P5)	$\boldsymbol{NP}_{\uparrow}$, *including* {$\boldsymbol{NP}_{\downarrow}$, }*{*or*|*and*} $\boldsymbol{NP}_{\downarrow}$	*block ciphers, including MARS, RC6, and Twofish* → *hyponym*(MARS, BLOCK CIPHER) → *hyponym*(RC6, BLOCK CIPHER) → *hyponym*(TWOFISH, BLOCK CIPHER)
(P6)	$\boldsymbol{NP}_{\uparrow}$, *especially* {$\boldsymbol{NP}_{\downarrow}$, }*{*or*|*and*} $\boldsymbol{NP}_{\downarrow}$	*fruits, especially peaches, apricots, and pears* → *hyponym*(PEACH, FRUIT) → *hyponym*(APRICOT, FRUIT) → *hyponym*(PEAR, FRUIT)

Table 2.7 above shows the six lexico-syntactic patterns created and used by Hearst. The table also shows sample sentence fragments we skimmed from the Wikipedia corpus using each pattern, along with the hyponymic relationships established from each of those fragments.

In a similar vein, Berland and Charniak (1999) manually derived two[11] lexico-syntactic patterns for extracting meronymic (part-whole) relationships from a corpus (see Table 2.8 below) and reported some success applying them to six hand-selected target wholes: "book," "building," "car," "hospital," "plant," and "school."

Table 2.8:
Berland and Charniak's lexico-syntactic patterns for meronymy, with examples extracted from Wikipedia. $\boldsymbol{NP}_{\uparrow}$ indicates a whole; $\boldsymbol{NP}_{\downarrow}$ indicates a part.

#	Pattern	Examples
(P7)	$\boldsymbol{NP}_{\uparrow}$ *'s* $\boldsymbol{NP}_{\downarrow}$	*car's radiator* → *meronym*(RADIATOR, CAR) *car's a-pillars* → *meronym*(A-PILLAR, CAR) *car's window* → *meronym*(WINDOW, CAR)
(P8)	$\boldsymbol{NP}_{\downarrow}$ *of* {*the*|*a*}{ ***JJ***|***NP***}* $\boldsymbol{NP}_{\uparrow}$	*Chassis of the car* → *meronym*(CHASSIS, CAR)

A novel contribution of Berland and Charniak's work was the use of log-likelihood (Dunning, 1993) and probability distribution metrics to quantify the likelihood that each extracted relation was valid (whereas Hearst took a single occurrence of a matched pattern as evidence of a hyponymic relationship). Using these metrics to rank the results they extracted

11 Berland and Charniak originally defined five lexico-syntactic patterns, but eliminated three of them when they discovered they were performing poorly in preliminary extraction trials.

from their corpus, for each target whole, Berland and Charniak took the 50 strongest meronym candidates and presented them to human judges alongside 50 unrelated terms for evaluation. They found that, among the top 50 meronyms they discovered for each of their six target wholes, 55% of them were valid meronyms. When restricting their consideration to only the top 20 results for each seed, they found their precision to be approximately 70%.

Table 2.9:
Hyponymic (P3) and meronymic (P7, P8) extraction patterns sometimes identify context-specific relationships or typify other relations, such as *PropertyOf*.

#	Pattern	Examples
(P3)	***NP***$_{\downarrow}${, ***NP***$_{\downarrow}$}*{,} *or other* ***NP***$_{\uparrow}$	The weft threads are usually wool or cotton, but may include *silk, gold, silver, or other alternatives.* → **hyponym*(SILK, ALTERNATIVE) → **hyponym*(GOLD, ALTERNATIVE) → **hyponym*(SILVER, ALTERNATIVE)
(P7)	***NP***$_{\uparrow}$*'s* ***NP***$_{\downarrow}$	*car's performance* → **meronym*(PERFORMANCE, CAR)
		car's history → **meronym*(HISTORY, CAR)
		car's ability → **meronym*(ABILITY, CAR)
(P8)	***NP***$_{\downarrow}$ *of* {*the*|*a*}{ ***JJ***|***NP***}* ***NP***$_{\uparrow}$	*velocity of the car* → **meronym*(VELOCITY, CAR)
		width of the car → **meronym*(WIDTH, CAR)

Hearst, in contrast, reported difficulty in mining meronymic relations, attributing the problem to the fact that "[t]he patterns for this [part-whole] relation do not tend to uniquely identify it, but can be used to express other relations as well" (p. 542). She also observed that extracted hyponymic relationships were sometimes invalid out of context, but typically reflected,

at the very least, some form of semantic relatedness. Table 2.9 above shows examples of both of these problems using one of Hearst's patterns (P3) and both of Berland and Charniak's patterns (P7 and P8) to extract relationships from Wikipedia. The categorization of silk, gold, and silver as *alternatives* is certainly a context-specific discovery, and the meronymic patterns exhibit a propensity for extracting *PropertyOf* relationships.

In light of Hearst's difficulty mining meronymic relationships, Berland and Charniak explicitly attributed their relative success to the size of their corpus (the 100 million word LDC North American News Corpus). However, the authors also pointed out that the generality of their approach could not be guaranteed because the six target wholes they used in their experiments were particularly amenable to part-whole relation mining. Specifically, the target wholes were selected by the authors on the grounds that they each exhibited high rates of occurrence in the corpus; each whole was, in fact, participant to part-whole relationships; and the authors perceived that there was a strong chance of those part-whole relationships being mentioned in the corpus.

Girju et al. (2006) observed the tendency, apparent from the results in Table 2.9, for extraction patterns to "express different semantic relations in different contexts" (p. 87). For example, the pattern ***X*** *with* ***Y*** can express relationships of meronymy (cf. *It was the girl with blue eyes*), possession (cf. *The baby with the red ribbon is cute*), or kinship (cf. *The woman with triplets received a lot of attention*) (examples from Girju et al., 2006, p. 94 & 96). To improve precision and recall of extraction patterns, they introduced an approach that incorporated selectional restrictions on WordNet classes. Their system was trained on positive and negative examples in which parts and wholes were manually annotated with their corresponding WordNet

noun senses. The authors used a classifier, the C4.5 decision tree induction algorithm (Quinlan, 1993), to learn rules on the form, "*if X is/is not of a WordNet semantic class A and Y is/is not of WordNet semantic class B, then the instance is/is not a part-whole relation*" (Girju et al., 2006, p. 96). Their approach automatically induced extraction patterns with precision and recall performance of approximately 80% on two large corpora (the *Wall Street Journal* and *LA Times* collections), but at the cost of large amounts of manually annotated training data.

Recognizing both the appeal and limitations of Girju et al.'s semi-automated approach, Pantel and Pennacchiotti (2006) developed Espresso, a "minimally supervised bootstrapping algorithm" (p. 114) for labeled relation learning and instance extraction. Their algorithm begins with manually specified seed sets of instances that exemplify a relation—typically between 10 and 15 instances. An automatic pattern induction phase then iteratively produces sets of generic patterns that, like the patterns of Berland and Charniak, are sometimes overly inclusive. However, relation instances extracted by those patterns are also subjected to what the authors call "reliable patterns" (p. 115 & 116), which are too exclusive to be used to harvest instances from a corpus (i.e., they have low recall), but which exhibit high precision and are therefore useful for evaluating the quality of extracted relation instances. Thus, Pantel and Pennacchiotti alleviated the acquisition bottleneck of lexico-syntactic pattern extraction methods in two ways: they eliminated the need to manually specify extraction patterns, and they reduced the amount of supervision or semantic annotation required for automatic pattern induction by creating a method that relied on very small sets of seed instances for each relation.

The precision of relation instance extraction by Espresso from two corpora is presented below in Table 2.10. One corpus is a collection of newswire articles (TREC), while the other is

an entire college-level chemistry textbook (CHEM). The authors extracted *IsA*, *PartOf*, and *succession* relationships from the TREC corpus. From the CHEM corpus, they extracted instances of the domain-specific *reaction* and *production* relations, as well as the *IsA* and *PartOf* relations. For extraction from a large corpus, no effective measure of recall is feasible, and so only precision results are reported in Table 2.10.

Table 2.10:
Precision of relation instance extraction by Espresso (Pantel & Pennacchiotti, 2006).

Corpus	***IsA***	***PartOf***	***succession***	***reaction***	***production***
TREC	36.2%	69.9%	49.0%	--	--
CHEM	76.0%	50.7%	--	91.4%	55.8%

2.4.1 VerbOcean and the Never-Ending Language Learner

In the context of constructing semantic networks, the pattern matching approach to discovering relation instances has certain limitations. The first is that one must begin by specifying a relation, or a set of relations, to be mined from the corpus. This seems to violate the intuition expressed by Quillian (1968) that "a memory model must provide a way to take any *two* tokens and relate them by any third token, which by virtue of this use becomes a relationship" (pp. 230-231, emphasis in original). The *a priori* articulation of a set of primitive relations upon which to focus one's mining efforts restricts one's ability to discover general semantic relatedness of the type that is either difficult to express through a binary relation, or has

too few examples to induce reliable extraction patterns. Consider, for example, the clear relationship between *penguin* and *tuxedo*, and the relative obscurity of the actual relation that binds them.

Nevertheless, the lexico-syntactic pattern matching approach has proven useful in constructing semantic networks that aim to express specific, restricted sets of semantic relations between entities. Chklovski and Pantel (2004), for example, used pattern matching and Web queries to create a semantic network of verb relations called VerbOcean, which expresses some relations that are not included in WordNet's verb ontology. In their work, 29,165 pairs of potentially associated verbs were identified by applying the DIRT algorithm of Lin and Pantel (2001) to a newspaper corpus. The authors then mined the Web for relations between those verb pairs using 35 manually constructed patterns typifying five specific verb relations (some of which were asymmetric): similarity (e.g., discover :: find), strength (e.g, muffle :: silence), antonymy (e.g, pass :: fail), enablement (e.g, try :: succeed), and happens-before (e.g., birth :: death). Their relation labeling algorithm also allowed for the possibility that no relationship existed between the verbs.

Their system achieved an estimated 65.5% accuracy in assigning "acceptable" relation labels to verb pairs based on the evaluations of two judges on a sample of their results. An estimated 53% of relation labels assigned by their system were the "preferred" labels indicated by the two judges. The accuracy on individual relation labels, as well as an estimated frequency of labeling based on the 100 randomly selected verb pairs evaluated by their judges, is given below in Table 2.11.

Table 2.11:
Accuracy of relation labeling on the five relations covered in VerbOcean.

Relation	Example	Frequency	Acceptable Tags	Preferred Tag
similarity	discover :: find	41%	63.4%	40.2%
strength	muffle :: silence	14%	75.0%	75.0%
antonymy	pass :: fail	8%	50.0%	43.8%
enablement	try :: succeed	2%	100%	100%
happens-before	birth :: death	17%	67.6%	55.9%
no relation	*n/a*	35%	72.9%	72.9%

Another notable application of lexico-syntactic pattern matching to network construction is the Never-Ending Language Learner (henceforth NELL) (Carlson et al., 2010). NELL covers 55 relations that have each been manually specified with 10 to 15 positive example instances and five negative instances. The network also includes 123 categories (semantic classes), each of which has been specified with 10 to 15 seeds (i.e., class members) and five manually defined lexico-syntactic patterns indicative of membership in that class. The learning algorithms of NELL use these training data to automatically induce patterns from Web texts, and use those patterns to extract new instances of relations and class membership. NELL then uses that newly acquired information to improve its extraction performance in subsequent iterations.

Many of the binary relations covered in NELL are quite specific. Examples include *athletePlaysForTeam*, *ceoOfCompany*, *teamWonTrophy*, and *cityInCountry*. Examples of classes specified in NELL include *scientist*, *restaurant*, *magazine*, *cardGame*, *mountain*, *lake*, *museum*, and *city*. The assertions expressed in NELL constitute a large-scale factual knowledge base, and, taken together, the class memberships expressed in NELL form a shallow *IsA* taxonomy. However, there is no methodical attempt to distinguish between concepts. There are, for

example, three distinct nodes for the aquatic bird sense of "flamingo," each expressing the discovery of a different class membership—*mammal:flamingo*, *animal:flamingo*, and *bird:flamingo*—but nothing to unify the nodes or to distinguish them from other senses of "flamingo" (*hotel:flamingo*, *river:flamingo*, and so on).

At the time of this writing, NELL has acquired over two million facts about which it has "high confidence," and is continuing to learn new facts from the Web—most of which take the form of *IsA* relationships between nouns, with strong emphasis on proper nouns.

2.4.2 ConceptNet and the Open Mind Common Sense Project

The Open Mind Common Sense (OMCS) project (Singh, 2002; Singh et al., 2002) was a commonsense acquisition project that leveraged the Web to crowdsource statements of commonsense knowledge from a community of online contributors. The type of knowledge expressed in the corpus might seem blatantly obvious or perhaps even trivial to humans, but is the sort of knowledge that might be necessary for computers to comprehend and reason with natural language. The OMCS corpus includes statements like *taking a shower will cause you to get wet*, *people often take pictures at special events*, and *helium balloons are used to decorate for parties*. In the first two years following its conception (September 2000 to August 2002), the OMCS project acquired over 450,000 such sentences from nearly 10,000 users. By 2004, ConceptNet had over 14,000 contributors and more than 700,000 commonsense knowledge assertions (Liu & Singh, 2004a).

OMCS contributors entered information through a Web interface using free-form natural language sentences (sometimes to tell short stories about concepts already represented in the

corpus, or to provide descriptions of photos or short movie clips). In all, 25 to 30 activities were used to elicit statements of commonsense knowledge from contributors. Some of those activities, instead of allowing free-form English responses, prompted users to complete fill-in-the-blank frames such as "[Something you find in *a pantry* is _____]." These frames expressed binary relations between entities and were developed by the OMCS authors in anticipation of using pattern matching to automatically extract relationships from their corpus (see Table 2.12). For example, the preceding frame was used to establish a spatial relationship, *AtLocation*, between *pantries* and things you might find there, such as *flour*, *cereal*, and *spices*.

Table 2.12:
Examples of OMCS frames and the relations they express (Singh, 2002).

Frame Type	Example Frame	Binary Relation
Functional	[A *hammer* is for _____]	*UsedFor*(HAMMER, ____)
Goals	[*People* want _____]	*DesireOf*(PERSON, ____)
Scripts	[The effect of *eating a sandwich* is _____]	*EffectOf*(EAT SANDWICH, ____)
Location	[Somewhere you find a *bed* is _____]	*LocationOf*(BED, ____)
Ontology	[A typical *activity* is _____]	*IsA*(____, ACTIVITY)

Whether responding to frames or providing free-form sentences, OMCS contributors were encouraged to use language that would be comprehensible even to children. However, because frames are particularly effective at enabling relation mining via regular expression pattern matching, and free-form sentences in the corpus were found to be less amenable to pattern matching extractions, the OMCS Web elicitation system was eventually modified to

encourage conformity to frames in all tasks; those that allowed free-form input were redesigned to inform users when their input matched a frame in the system (Singh et al., 2002).

ConceptNet (Havasi et al., 2007; Liu & Singh, 2004a, 2004b) is a large-scale semantic network of commonsense knowledge built primarily around knowledge extracted from the OMCS corpus. The current version of ConceptNet also incorporates external resources such as the WordNet ontology[12] (in its entirety), Wiktionary (a collaboratively built dictionary that, as the name implies, is a sister project to Wikipedia), and DBpedia (a semantic network derived from Wikipedia, which we discuss below in Section 2.5.5). Through the remainder of this work, when referring to ConceptNet, we restrict our consideration to the relationships derived automatically from OMCS, and not these independent projects that have been assimilated into the network.

Statements of commonsense knowledge, called "assertions" in ConceptNet, are expressed in the network using a limited set of relations, and are manifest as labeled edges (representing these relations) between adjacent nodes (representing "semi-structured natural language fragments" (Liu & Singh, 2004b, p. 293)). The textual fragments denoted by ConceptNet's nodes conform to certain syntactic constraints (hence they are "semi-structured"), and include not just first-order lexical entities (e.g., nouns and verbs—potentially compound—such as "penguin" or "piggy bank"), but also second-order phrases (e.g., "shop for food" and "capable of flight"). It is the ability of these second-order phrases to denote complex actions, entities, and properties that earn them the label of "concepts" in the ConceptNet literature, although they do not conform to the stricter definition of concepts used elsewhere in the literature and throughout this dissertation; neither the phrases nor phrasal constituents in ConceptNet are disambiguated to individual word senses. (E.g., the assertions *ConceptuallyRelatedTo*(MONEY, MINT) and

12 WordNet is assimilated, but not fully integrated; concepts in ConceptNet are not mapped to WordNet senses.

ConceptuallyRelatedTo(CANDY, MINT) are both present in the current version of ConceptNet, but there is no indication that the two "mints" here refer to different concepts.)

The assertions in ConceptNet are extracted from sentences in the OMCS corpus using a set of approximately 50 regular expression patterns that correspond to the fill-in-the-blank frames used to elicit sentences from users during the corpus's construction (see Table 2.13 below for examples). Extracted phrases undergo a normalization phase in which constituent words are stemmed, spelling is corrected, and modals and determiners are removed. Thus, a phrase like "eating a sandwich" is mapped to an "eat sandwich" node in ConceptNet.

Table 2.13:
Three extraction patterns for mining relation instances from OMCS (Singh et al., 2002).

Pattern	Relation Instance Example
{*a*|*an*|*the*}? ***NN*** {*is*|*are*} {*a*|*an*|*the*}? ***JJ***? ***NN***	Hurricanes are powerful storms → *IsA*(HURRICANE, POWERFUL STORM)
A person {*does not*}? *want*{*s*}? *to* ***VB JJ***	A person wants to be warm → *DesireOf*(PERSON, BE WARM)
NN *requires* {*a*|*an*} ***JJ***? ***NN***	Bathing requires water → *HasPrerequisite*(BATHING, WATER)

The original set of 20 relations expressed in ConceptNet 2.0 (Liu & Singh, 2004a) is given below in Table 2.14. This set of relations has subsequently been relaxed and extended. ConceptNet 5 at one point included relations on entities discovered automatically by ReVerb (Fader et al., 2011), which were subsequently filtered out of ConceptNet 5.1 for introducing too many unreliable statements into the network (R. Speer, personal communication, June 4, 2012). At the time of this writing, a new filter is being implemented to selectively reintegrate assertions

from ReVerb. Table 2.15 (below on page 63) lists the occurrence frequency of relations currently expressed in the core assertions of ConceptNet 5.1. The table excludes 23 relations that are expressed fewer than ten times in the network and tend to be resultant of extraction errors in the construction of the network (e.g., anomalous one-off relations with names like *e_size*, *nd_like*, and *d_of*).

Table 2.14:
Relations expressed in ConceptNet 2.0, with examples (Liu & Singh, 2004a).

Relation	**Type**	**Example**
ConceptuallyRelatedTo	K-Line	*bad breath—mint*
ThematicKLine	K-Line	*wedding dress—veil*
SuperThematicKLine	K-Line	*western civilisation—civilisation*
IsA	Thing	*horse—mammal*
PropertyOf	Thing	*fire—dangerous*
PartOf	Thing	*butterfly—wing*
MadeOf	Thing	*bacon—pig*
DefinedAs	Thing	*meat—flesh of animal*
CapableOf	Agent	*dentist—pull tooth*
PrerequisiteEventOf	Event	*read letter—open envelope*
FirstSubeventOf	Event	*start fire—light match*
SubeventOf	Event	*play sport—score goal*
LastSubeventOf	Event	*attend classical concert—applaud*
LocationOf	Spatial	*army—in war*
EffectOf	Causal	*view video—entertainment*
DesirousEffectOf	Causal	*sweat—take shower*
UsedFor	Functional	*fireplace—burn wood*
CapableOfReceivingAction	Functional	*drink—serve*
MotivationOf	Affective	*play game—compete*
DesireOf	Affective	*person—not be depressed*

Table 2.15:
Relations in ConceptNet 5.1 by frequency (core assertions only).

Frequency	Relation	Frequency	Relation
126450	IsA	5301	Desires
101642	HasProperty	5086	PartOf
60669	UsedFor	4287	NotDesires
54105	AtLocation	4242	HasFirstSubevent
52189	CapableOf	3967	HasLastSubevent
51198	RelatedTo	3834	NotIsA
46551	have_or_involve	2937	NotCapableOf
28576	HasSubevent	2701	SimilarSize
25660	HasA	1806	MadeOf
25178	HasPrerequisite	1406	DesireOf
23465	ConceptuallyRelatedTo	1313	NotHasProperty
18955	Causes	651	CreatedBy
17130	MotivatedByGoal	406	NotHasA
12256	be_in	330	InheritsFrom
11404	be_near	167	SymbolOf
11104	ReceivesAction	74	HasPainIntensity
11055	be_not	71	InstanceOf
6665	DefinedAs	45	LocationOfAction
6357	CausesDesire	34	HasPainCharacter
5487	LocatedNear	24	NotMadeOf

It has been observed (Tandon, Melo, & Weikum, 2011) that ConceptNet contains many unreliable assertions. These often result from the ambiguity of frames presented during OMCS elicitation tasks. For example, the frame "[*looking through a telescope* is for _____]" was used to elicit information about what looking through a telescope might be used for (e.g., observing stars or looking at faraway objects), but one person responded with "astronomers," giving rise to the assertion *UsedFor*(LOOK THROUGH TELESCOPE, ASTRONOMER) in ConceptNet. The frame "[You are

likely to find *the Moon* in _____]" similarly resulted in an unintended relationship when one contributor responded with "orbit around earth." While technically true when taken as a whole, the sentence does not reflect the stationary spatial relationship intended, and we find the assertion *AtLocation*(MOON, ORBIT AROUND EARTH) in ConceptNet. In some cases, the frames also resulted in ontologically infeasible assertions. For example, one contributor responded to the frame "[Something you find in *a quandry* is _____]" with "people."[13] The resulting assertion, *AtLocation*(PERSON, QUANDRY), still present in ConceptNet 5.1, does not indicate a place where people can be found, since a quandary is an abstraction, not a physical location.

Similarly, several instances of the frame "[_____ are sometimes _____]" demonstrate its unreliability for establishing hyponymic relationships (e.g., "[*tuxedos* are sometimes _____]," to which a contributor responded "called penguin suits;" *IsA*(TUXEDO, CALL PENGUIN SUIT) was introduced into the network accordingly).

2.5 Wikipedia-Based Approaches to Relatedness

Wikipedia has been the focus of a large body of NLP research in recent years. In addition to its free availability for download from the Web and the vast amount of natural language text it contains, its inclusion of a rich set of semantic annotations has contributed to the corpus's appeal among NLP researchers. These semantic annotations are largely derived from the structure of Wikipedia. For example, disambiguation pages enumerate distinct senses of articles that share the same title, giving rise to a new concept inventory for use in NLP applications; inter-article links induce relationships between articles that can be conceived of as establishing edges

13 ConceptNet has (separate) nodes for both QUANDARY and the commonly misspelled form, QUANDRY.

between concept nodes in a semantic network; and articles' infoboxes indicate factual assertions about named entities in a highly structured manner, giving the corpus the makings of a nascent knowledge network.

Wikipedia's semantic offerings also include the organization of its articles into a folksonomic taxonomy, which Strube and Ponzetto (2006) describe thusly: "[R]ather than being a well-structured taxonomy, the Wikipedia category tree is an example of a *folksonomy*, namely a collaborative tagging system that enables the users to categorize the content of the encyclopedic entries. Folksonomies as such do not strive for correct conceptualization in contrast to systematically engineered ontologies. They rather achieve it by collaborative approximation" (p. 1419, emphasis in original).

The use of Wikipedia in NLP tasks represents, in some cases, a departure from the field's reliance upon carefully hand-crafted ontologies for sense inventories, as well as semantic resources—like sense-tagged corpora—that make use of those ontologies. As researchers turn to Wikipedia, they cite the limitations of resources like WordNet; Gabrilovich and Markovitch (2007) point out that "such resources contain few proper names, neologisms, slang, and domain-specific technical terms. Furthermore, these resources have strong lexical orientation and mainly contain information about individual words but little world knowledge in general" (p. 1609).

Wikipedia has been employed in a wide variety NLP tasks over the past decade, such as question answering (Ahn et al., 2004), named entity disambiguation (Bunescu & Paşca, 2006), and topic identification (Coursey, Mihalcea, & Moen, 2009). Augmenting the structure of Wikipedia itself has been the subject of research as well. Mihalcea and Csomai (2007) investigated the possibility of enhancing Wikipedia articles by adding links between pages after

automatically identifying keywords in each article and disambiguating those words to their appropriate Wikipedia concepts (i.e., articles). Mihalcea (2007) developed a method for producing sense-tagged corpora using articles from Wikipedia as a sense inventory. Similarly, Milne and Witten (2008b) proposed a machine learning method for augmenting any document with relevant links to Wikipedia articles. Ponzetto and Navigli (2009) explored graph theoretic approaches for augmenting the taxonomic organization of Wikipedia articles.

Wikipedia has also been used in quantitative measures of semantic relatedness. In the sections that follow, we present three studies that have drawn on Wikipedia for that purpose: Strube and Ponzetto (2006) replaced WordNet with Wikipedia in several traditional quantitative relatedness measures (Section 2.5.1). Gabrilovich and Markovitch (2007) used Wikipedia article texts to establish a new vector space of Wikipedia concepts, and employed a common distributional approach to measure relatedness: the cosine of the angle between any two vectors represented in that space was used as a direct measure of the relatedness between natural language text fragments of arbitrary and unlimited length (Section 2.5.2). Milne and Witten (2008a) turned to inter-article links to measure relatedness between terms (Section 2.5.3).

We also present the work of Ponzetto and Navigli (2010), which used inter-article Wikipedia links to relate WordNet noun senses automatically, and then mapped those relationships to noun senses from WordNet (Section 2.5.4). We conclude our discussion of Wikipedia-based approaches with a presentation of two large-scale semantic networks that have been created by extracting semantic annotations from Wikipedia articles: YAGO (Suchanek et al., 2007) and DBpedia (Bizer et al., 2009) (Section 2.5.5).

2.5.1 Path-Based Relatedness Measures Using Wikipedia

Strube and Ponzetto (2006) used Wikipedia as the knowledge source for several relatedness measures designed for use with WordNet, including path-based (Leacock & Chodorow, 1998; Rada et al., 1989; Wu & Palmer, 1994), information content (Resnik, 1995), and gloss overlap measures (Banerjee & Pedersen, 2003). For their paths, they traversed folksonomic categorizations of Wikipedia articles, and for glosses, they experimented with using either the first paragraph of an article or its entire text.

Table 2.16 below shows how measures from Strube and Ponzetto's study correlated to similarity data from the M&C, R&G, and WordSim353 datasets. The results exclude noun pairs that are not covered in the WordNet ontology. The sim_{BP} results are based on ExtLesk using only the first paragraph as an article's gloss. (Performance was negligibly lower using the article's entire text as a gloss.)

Table 2.16:
Comparison of WordNet- and Wikipedia-based similarity measures in Strube and Ponzetto (2006) showing correlation (*r*-values) to human similarity judgments.

	WordNet-Based					**Wikipedia-Based**				
Data	sim_{PL}	sim_{WP}	sim_{LC}	sim_{R}	sim_{BP}	sim_{PL}	sim_{WP}	sim_{LC}	sim_{R}	sim_{BP}
M&C	0.71	0.77	**0.82**	0.78	0.37	0.49	0.45	0.46	0.29	0.47
R&G	0.78	0.82	**0.86**	0.81	0.34	0.56	0.52	0.54	0.34	0.47
WordSim353	0.27	0.32	0.36	0.36	0.21	0.46	**0.48**	**0.48**	0.38	0.20

Strube and Ponzetto found that the Wikipedia-based measures correlated weakly to the M&C and R&G similarity ratings; of all the approaches they tried, they achieved their best

Wikipedia-based results using a simple path length measure (sim_{PL}) that took the shortest folksonomic path between two articles as a measure of relatedness (r = 0.49 and 0.56 for M&C and R&G, respectively), but these fell drastically short of the top performing WordNet-based measure, sim_{LC} (r = 0.82). They also found fairly weak correlation to the WordSim353 data, in which their best results (r = 0.48) came from the Wikipedia-based adaptation of Leacock and Chodorow's path-based similarity function. However, the Wikipedia-based sim_{LC} outperformed all of the WordNet-based measures the authors evaluated on WordSim353, the best of which was the WordNet-based version of sim_{LC} (r = 0.36).

2.5.2 Relatedness via Explicit Semantic Analysis with Wikipedia

Gabrilovich and Markovitch (2007) introduced Explicit Semantic Analysis (ESA) to represent arbitrary words and text fragments of any length as vectors in Wikipedia concept-space, thus circumventing the need for carefully crafted semantic resources like WordNet to enable relatedness measurements. The authors first created a vector for each Wikipage with a distributional representation of its contents based on TF-IDF (Salton & McGill, 1983). TF-IDF essentially measures the relevance of an individual word to a document by taking the term's frequency within the document (TF) and multiplying by the inverse of the proportion of documents containing the term, or inverse document frequency (IDF). The measure is commonly given as:

$$tfidf(t, d, D) = \frac{f(t, d)}{\max_{w \in d} f(w, d)} \times \log \frac{|D|}{|\{d' \in D : f(t, d') > 0\}|} \tag{20}$$

where t is the term in question, d is the current document (a Wikipage), D is the collection of all documents (all pages in Wikipedia), and $f(t, d)$ is the frequency of t in d.

To represent arbitrary words and text fragments as vectors, Gabrilovich and Markovitch multiplied weightings for each word of a fragment's TF-IDF vector by pre-computed vectors indicating the word's "strength of association" (p. 1607) with each Wikipage, based on the TF-IDF representations constructed in the previous step. The resulting text fragment representation was a vector of N weights, where N is the number of Wikipedia articles represented in the system. The vector essentially oriented the text fragment in Wikipedia concept-space. Gabrilovich and Markovitch then calculated the relatedness between two text fragments by taking the cosine of their representative vectors in this concept space.

The authors showed that their quantitative measurements of semantic relatedness between nouns correlated strongly to human similarity judgments from the M&C and R&G data (r = 0.723 and 0.816, respectively, using Pearson's correlation) and the WordSim353 collection (ρ = 0.75 using Spearman's rank correlation). The authors also found strong correlation (r = 0.72) to human rankings of the relatedness of entire documents from the Australian Broadcasting Corporation's news mail service (Lee, Pincombe, and Welsh, 2005).

2.5.3 Measuring Relatedness from Inter-Article Links in Wikipedia

Milne and Witten (2008a) proposed two measures—one of similarity and one of relatedness—that, instead of relying on the folksonomic categorization of Wikipedia articles, capitalized exclusively on inter-article links. The first measure, like that of Gabrilovich and Markovitch (2007), represented Wikipedia articles as weighted term vectors. In the Milne and

Witten conception, an article's vector is a sequence of weighted link probabilities. If a source article *s* links to a target article *t*, the weight of that link in the vector representation of *s* is given as:

$$w(s \rightarrow t) = \log \frac{|D|}{|D_{\rightarrow t}|} \text{ if } s \in D_{\rightarrow t}, 0 \text{ otherwise} \quad (21)$$

where, as before, D is the set of all documents in Wikipedia (all Wikipedia articles), and $D_{\rightarrow t}$ is the collection of all documents linking to article *t*. The intuition behind (21) is that a link from *s* to *t* is more meaningful to the representation of *s* when there are few articles in Wikipedia that link to *t*. If links to *t* are common throughout the corpus, the weight of the link's significance to the representational vector of *s* is diminished. Of course, if there is some article *t* that *s* does not link to, its corresponding weight in the vector representation of *s* is zero. The cosine of the angle between any two such vectors is then taken as a measure of the similarity between the articles they represent.

For their second measure, Milne and Witten adapted the Normalized Google Distance measure of Cilibrasi and Vitanyi (2007) to measure the semantic distance between two articles as follows:

$$dist_{MW}(a, b) = \frac{\log(\text{MAX}(|D_{\rightarrow a}|, |D_{\rightarrow b}|)) - \log(|D_{\rightarrow a} \cap D_{\rightarrow b}|)}{\log(|D|) - \log(\text{MIN}(|D_{\rightarrow a}|, |D_{\rightarrow b}|))} \quad (22)$$

where *a* and *b* are two Wikipedia articles. As before, D is the set of all articles in Wikipedia, $D_{\rightarrow a}$ is the set of articles linking to *a*, and $D_{\rightarrow b}$ is the set of articles linking to *b*. The intuition behind (22) is that an article that links to both *a* and *b* provides evidence of the relatedness of *a* and *b*, whereas the frequent occurrence of articles linking to either *a* or *b*, but not the other, suggests

that articles *a* and *b* bear a weaker relationship to one another, or no relationship at all. Recall that semantic distance is inversely related to semantic relatedness.

To measure the relatedness between two terms, Milne and Witten took the average values of the two functions above (adjusted, of course, to invert values of $dist_{MW}$). Compared to Gabrilovich and Markovitch's (2007) ESA approach, the link-based method of Milne and Witten is faster because the text of articles is ignored. The authors point out that the use of inter-article links is also more reliable than measuring distributional similarity of article text because links are manual annotations that have been explicitly inserted and disambiguated by human contributors. However, Milne and Witten concede that an advantage of ESA is that it can measure the relatedness of natural language fragments of any length, and that the Strube and Ponzetto (2006) and Milne and Witten approaches "are not so easily extended" (Milne & Witten, 2008a, p. 29).

Table 2.17:
Comparison of three Wikipedia-based relatedness measures on the basis of their correlation to human similarity judgments.

Data	Strube and Ponzetto (2006)	Gabrilovich and Markovitch (2007)	Milne and Witten (2008a)
M&C	0.49	**0.72**	0.70
R&G	0.56	**0.82**	0.64
WordSim353	0.48	**0.75**[14]	0.69

A summary of results for the three Wikipedia-based relatedness studies presented in this section is given above in Table 2.17. Values reported are coefficients of correlation between the

14 Gabrilovich and Markovitch (2007) used Spearman's rank correlation to evaluate their results on WordSim353. All other values in the table are coefficients from Pearson's product-moment correlation (*r*-values).

authors' measures and the gold standard datasets of human similarity score judgments. Gabrilovich and Markovitch's correlation to the WordSim353 data was calculated using Spearman's rank correlation (ρ). All other coefficients of correlation are Pearson's *r*-values. For Strube and Ponzetto's results, we present the top performing measure for each dataset. We see that the high precision of the inter-article links used by Milne and Witten does not offer a competitive advantage over the vast amount of textual data harnessed by Gabrilovich and Markovitch's ESA approach, which achieves the best performance on each dataset.

2.5.4 WordNet++

Ponzetto and Navigli (2010) developed a semantic network called WordNet++ (henceforth WN++) that, like ours, categorically relates WordNet noun senses. They used semantic annotations underlying the Wikipedia corpus to build the network, first mapping the titles of Wikipages (i.e., Wikipedia articles) to WordNet noun senses, and then establishing semantic links between concepts in the following way: if some Wikipage, w_1, links to a second Wikipage, w_2, and the pages have been disambiguated to WordNet noun senses $\mu(w_1)$ and $\mu(w_2)$ respectively, then the edge $(\mu(w_1), \mu(w_2))$ is added to WN++.

Ponzetto and Navigli mapped Wikipages to WordNet concepts by first establishing *disambiguation contexts* for them. The disambiguation context of some Wikipage, *w*, is $ctx(w)$ and consists of sense labels, links, and categories from Wikipedia. *Sense labels* are the parenthetical categories that follow article titles to distinguish them from articles with the same name (e.g., the "operating system" in "Android_(operating system)," which distinguishes the

article from "Android_(robot)"). If w has a sense label, all words from that label are included in $ctx(w)$. *Links* are the lemmas (titles *without* sense labels) of all Wikipages to which w links. *Categories* are the syntactic heads of folksonomic Wikipedia classes to which an article belongs. For example, the article for "The Catcher in the Rye" is categorized by "Novels by J. D. Salinger," the syntactic head of which is simply "novel."

The disambiguation context of a WordNet noun sense s, denoted $ctx(s)$, consists of all nouns represented in s's synset, all nouns represented in its hypernymic, hyponymic, and sister synsets,[15] and all lemmatized content words (nouns, verbs, adjectives, and adverbs) from the gloss of s.

WN++ disambiguates w to $\mu(w)$, the sense of that noun in WordNet with the maximum number of content words in common between their respective contexts:

$$\mu(w) = \underset{s \in Senses_{WN}(w)}{argmax}\ p(s|w) = \underset{s}{argmax}\ \frac{p(s,w)}{p(w)} = \underset{s}{argmax}\ p(s,w) \tag{23}$$

In (23), $p(w)$ is a normalization factor that can be discarded without impacting which sense s is returned. The probability function in (24) is given as:

$$p(s,w) = \frac{score(s,w)}{\sum_{\substack{s' \in Senses_{WN}(w) \\ w' \in Senses_{Wiki}(w)}} score(s',w')} \tag{24}$$

where $score(s, w)$ returns the number of content words that strings $ctx(s)$ and $ctx(w)$ have in common, with an additive smoothing factor of 1:

$$score(s,w) = |ctx(s) \cap ctx(w)| + 1 \tag{25}$$

15 Sister concepts in WordNet are concepts that share the same immediate hypernym.

There are a few caveats to sense assignment by this method. First, (23) does not return a result in the event of a tie, and a link from article w_1 to w_2 in Wikipedia can only induce a relationship in WN++ if both $\mu(w_1)$ and $\mu(w_2)$ are non-empty. Second, if an article title w is unambiguous (i.e., monosemous) in both Wikipedia and WordNet, the Wikipage is mapped to the only possible WordNet noun sense. Furthermore, if d is such a page (monosemous in Wikipedia and WordNet), and d redirects to some Wikipage w, and w is one of the nouns represented in the synset $\mu(d)$, then $\mu(w) = \mu(d)$.

Ponzetto and Navigli evaluated their disambiguation algorithm by comparing their results to a set of 505 manually disambiguated Wikipedia page titles.[16] By this measure, the precision and recall of their disambiguation algorithm are P = 81.9% and R = 77.5%. Their algorithm mapped 81,533 Wikipages to WordNet noun senses and induced 1,902,859 links between WordNet concepts. These links, combined with the entire WordNet ontology, constitute the semantic network called WN++, but in this section we distinguish between the links derived from Wikipedia and the union of those links with WordNet, referring to the former as "WN++ (stand-alone)" and the latter as "WN++ (with WordNet)."

The authors evaluated the concept-to-concept relationships in their network by employing it in two word sense disambiguation tasks. The first task was the SemEval-2007 coarse-grained English all-words task (Navigli, Litkowski, and Hargraves, 2007), in which WN++ was used in two graph-based disambiguation algorithms: ExtLesk (Banerjee & Pedersen, 2003) and Degree Centrality (Navigli & Lapata, 2010). (These algorithms are described in detail in Chapter 5,

16 The authors originally selected 1,000 articles for manual disambiguation. 495 of those had no correct corresponding noun sense in WordNet and were excluded from the gold standard dataset on those grounds. There is no indication of how frequently the authors' approach assigns a WordNet noun sense to w when in fact no accurate mapping is possible.

where we subject our network to the same evaluation in order to compare its performance to that of WN++.)

Table 2.18 shows disambiguation results on this task for WordNet, WN++ (stand-alone), and WN++ (with WordNet). The F_1 measure given is the harmonic mean of precision and recall: $F_1 = {}^{(2PR)}/_{(P+R)}$. The results reported for Degree Centrality use a refined version of WN++ that contains only 79,422 of WN++'s strongest relationships. The refined subset was created in an unsupervised setting by Ponzetto and Navigli specifically for use with Degree Centrality when they discovered that WN++ had too many weak associations to perform well with the algorithm.

Table 2.18:
Results (F_1 scores) for WordNet and WN++ on the SemEval-2007 coarse-grained WSD task, as reported by Ponzetto and Navigli (2010).

Algorithm	Baselines	WordNet (stand-alone)	WN++ (stand-alone)	WN++ (with WordNet)
ExtLesk	--	68.3	72.0	75.4
Degree Centrality	--	74.5	57.4[17]	79.4[17]
MFS	77.4	--	--	--
Random	63.5	--	--	--

Ponzetto and Navigli's second experimental task used the same disambiguation algorithms, but involved WSD in domain-specific corpora (sports and finance) from Koeling, McCarthy, and Carroll (2005). WN++ (with WordNet) was evaluated, but not WN++ (stand-alone). The results on the domain-specific data, while markedly lower than those achieved on the SemEval-2007 task, were in line with the performance of other knowledge-based WSD

17 These results come from the use of the refined version of WN++.

algorithms on the same data, such as those of Agirre, de Lacalle, and Soroa (2009, as cited in Ponzetto & Navigli, 2010) (see Table 2.19).

Table 2.19:
Results (F_1 scores) for WN++ (with WordNet) on domain-specific WSD in the domains of sports and finance, as reported by Ponzetto and Navigli (2010).

Algorithm	Sports Domain Corpus	Finance Domain Corpus
ExtLesk	40.1	45.6
Degree Centrality	42.0	47.8
MFS Baseline	19.6	37.1
Random Baseline	19.5	19.6

2.5.5 Directly Extracting Semantic Relationships from Wikipedia

Many of the semantic annotations from Wikipedia have been extracted directly to large-scale knowledge networks. Suchanek et al. (2007), for example, derived a semantic network called YAGO from the underlying structure of Wikipedia articles. In particular, they derived facts from the *IsA* article folksonomy and assertions within articles' infoboxes.

Infoboxes in Wikipedia provide structured information about the subject of an article. For example, the infobox for Wikipedia's *The Catcher in the Rye* page explicitly lists the book's author (J. D. Salinger), cover artist, country of publication, the novel's original publication language (English), its genre (i.e., "Novel"), publisher, publication date, media type (i.e., "Print (hardback & paperback)"), number of pages, and identifying ISBN and OCLC numbers. Furthermore, *The Catcher in the Rye* is folksonomically categorized in Wikipedia under *1951*

novels; *American bildungsromans*; *Debut novels*; *Little, Brown and Company books*; *Novels by J.D. Salinger*; *Novels set in New York City*; *Novels set in Pennsylvania*; and *1949 in fiction*.

Suchanek et al. manually established heuristics for 170 frequently occurring infobox attributes that allowed for the extraction of those data to their network. WordNet classes are also incorporated into YAGO, but the network excludes WordNet's proper nouns, preferring instead to rely on Wikipedia as its source of information about named entities. YAGO then attempts to perform automatic hyponymic mappings of Wikipedia concepts to upper-level WordNet concepts.

Over 73% of the facts in YAGO are encompassed by its *isCalled*, *type*, and *means* relations, which are indicative of semantic similarity between entities. Among its most frequent relations beyond those indicating similarity are specific relationships such as *bornOnDate*, *diedOnDate*, *hasPopulation*, *bornInLocation*, *actedIn*, *directed*, and *writtenInYear*.

Bizer et al. (2009) similarly extracted structured information from Wikipedia into a semantic network called DBpedia. They established an ontology of infobox templates linking 350 frequently occurring infobox attributes into an ontology of 170 infobox classes, as some attributes are expressed in infoboxes in multiple ways. The resultant network establishes relationships between 2.6 million entities. Unlike YAGO, DBpedia does not incorporate WordNet classes or attempt to perform mappings between WordNet and Wikipedia concepts.

2.6 Hand-Crafted Knowledge Networks

We have so far seen that large-scale knowledge networks can be created by a variety of unsupervised and semi-supervised methods, including the application of lexico-syntactic pattern

matching to collaboratively constructed corpora (e.g, the acquisition of ConceptNet from the OMCS corpus) and the Web (as with the acquisition of VerbOcean and the on-going development of NELL). Other approaches have leveraged the semantic annotations of Wikipedia, such as inter-article links (as with WN++) and infobox attributes (as with YAGO and DBpedia), to establish semantic networks, sometimes performing mappings of concepts to WordNet classes (as with WN++ and, to some degree, YAGO). Despite the fact that many of these approaches rely on human contributions and manually annotated semantic resources, the machine learning methods used to construct them are error prone, as are, in some cases, the data being mined.

Many researchers have turned to hand-crafting knowledge bases, trading time-intensive knowledge crafting for assurances that their resources represent information with higher degrees of accuracy than automatically acquired resources. The WordNet ontology is perhaps the most obvious example of a hand-crafted semantic resource; it is the result of decades of careful knowledge crafting efforts, and has enjoyed ubiquitous use in the field of computational linguistics. In this section, we provide brief overviews of two other hand-crafted knowledge networks: CYC and Freebase.

CYC (Lenat, 1995) is a large-scale network with millions of assertions of commonsense knowledge. Much like ConceptNet, CYC expresses a variety of labeled relations between entities. However, CYC uses a deeper representation based on first-order predicate calculus and inference mechanisms, and therefore requires contributors to have some degree of expertise in knowledge engineering. The knowledge in CYC has been hand-coded into the network over the course of nearly three decades and is less prone to the kinds of acquisition and parsing errors that

give rise to malformed natural language expressions in ConceptNet and result in mislabeled relationships between entities. The expressive power of CYC is also more powerful than that of ConceptNet. It encodes commonsense assertions beyond ConceptNet's restricted set of binary relations, including, for example, "You have to be awake to eat" and "You can usually see people's noses, but not their hearts" (Lenat, 1995, p. 33). Another notable difference between CYC and ConceptNet is that CYC establishes an upper ontology that provides for the sound categorization of entities through *IsA* relationships. CYC also includes broad coverage of named entities, much like YAGO and DBpedia. Naturally, CYC does not restrict itself to WordNet's noun sense inventory, although preliminary attempts have been made at integrating WordNet into CYC (Reed & Lenat, 2002), as well as mapping CYC concepts to Wikipedia articles (Medelyan & Legg, 2008).

Freebase (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008; Bollacker, Tufts, Pierce, & Cook, 2007) is a large-scale knowledge base constructed collaboratively by online contributors through a Web interface. By crowdsourcing information from a vast array of contributors through the Web, Freebase has alleviated the acquisition bottleneck associated with the hand-crafting of knowledge networks and has seen tremendous growth in the short time since its conception. At the time of this writing, the knowledge base provides information about over 23 million entities, expressed through structured node properties and relationships. Entities in Freebase are organized into an upper ontology of classes (called "types" in Freebase) that rather resembles the folksonomic structure of Wikipedia; contributors are free to create new types as they see fit, and entities can be categorized by any number of types. In Freebase, "[r]ather than ontological correctness or logical consistency," the focus is on "collaborative creation of

structure" (Bollacker et al., 2007, p. 24); this concession to folksonomy and structural flexibility is the payment Freebase has made for the prodigious rate at which it has expanded.

Contributors create new entities in Freebase first by assigning them type categorizations in the upper ontology. Each type is associated with a schema that indicates attributes of that type, and, in some cases, type restrictions that operate like selectional restrictions on values for those attributes. Users are prompted to fill in attribute values for new entities, and auto-complete fields help them to assign permissible values.

Node structure in the Freebase graph is highly flexible, as well; node properties and relationships in Freebase are open classes that can be modified by contributors. Since contributors are explicitly prompted to assign attribute values to entities, assertions in Freebase tend to have high accuracy. However, all modifications to the Freebase graph are attributed to the individuals who make them, so that material provided by abusive contributors can be filtered out by end users. (Compare this to the indirect approach used to elicit information in OMCS, from which ConceptNet derives its assertions. Frames presented to users in the acquisition of OMCS were often ambiguous. For example, one response to the OMCS frame "[You are likely to find *the Moon* in _____]" was "orbit around the earth," which is true, but does not fulfill the frame's purpose of eliciting an instance of a spatial *AtLocation* relationship.)

CHAPTER 3: CONSTRUCTING THE NETWORK: SEMANTIC ASSOCIATES OF NOUNS

In this chapter, we present our methodology for acquiring a semantic network of related concepts. The acquisition process is fully automated and unsupervised, and comprises two stages: association, which is the subject of this chapter, and disambiguation, which we discuss in Chapter 4.

In the association phase, we discover semantic associates of common nouns using co-occurrence data extracted from Wikipedia. This discovery process is a context-sparse affair that takes place *in absentia* of the semantic annotations of Wikipedia, such as inter-article links, disambiguation page entries, the title of the article in which a sentence appears, and so on. The underlying assumption of our approach is that words that co-occur frequently in Wikipedia will bear semantic relation to one another, and, insofar as we consider the network to give a fairly comprehensive indication of semantic association, that the converse is true: that semantically related nouns will tend to appear in sentences together throughout the corpus. This correlation of lexical co-occurrence and semantic association is well established in the literature, most notably by Spence and Owens (1990) and Church and Hanks (1990). Of course, it is left to us to define what it means for two nouns to co-occur together "frequently" in the corpus, and to cull from consideration nouns that co-occur together frequently, yet do not bear semantic relation. Toward this end, we use a modified information theoretic approach to quantify the semantic relatedness of two nouns based on their frequency of co-occurrence in the corpus. These measurements are then used by an algorithm of our own creation that establishes relatedness between nouns categorically rather than quantitatively.

The second phase of network acquisition is a disambiguation phase, in which we capitalize on salient sense clustering among related nouns to automatically resolve them to individual senses from the WordNet 3.0 noun ontology (see Chapter 4).

Rather than defer evaluation of our methodology to a separate chapter, we pause after each step of the acquisition process to perform *in loco* evaluation of our progress so far. Following are the three sub-phases of acquisition that garner their own evaluation in this and the following chapter: quantitative measurement of semantic relatedness (Section 3.2), establishing categorical relatedness (Section 3.3), and noun sense disambiguation (Section 4.6). We conclude our discussion of network construction with a detailed explication of select excerpts from the network (Section 4.7).

3.1 Preliminaries: Corpus and Co-occurrence

To facilitate the extraction of co-occurrence data from Wikipedia, we have part-of-speech tagged the entire corpus (stripped of markup, metadata, and semantic annotation) using Brill's tagger (Brill, 1995). Throughout the remainder of this work, only intra-sentential co-occurrence of nouns is considered, and only between noun stems, rather than extracting separate data for distinct inflected forms. Any noise that results from considering co-occurrence at the sentence level, rather than employing a smaller or variable sized window, is generally quashed by the sheer magnitude of co-occurrence data available from the corpus.

Named entities are excluded from consideration in our research in part because WordNet lacks comprehensive coverage of proper nouns, which would leave many of our nouns without conceptual anchors in the ontology, or, worse yet, would anchor them to incorrect senses of

proper nouns that have only partial coverage in WordNet. Furthermore, the relation of a named entity to another concept typically represents a factual assertion that falls slightly outside the realms of commonsense knowledge and basic semantic relatedness and into the realm of encyclopedic knowledge. Accordingly, we restrict our consideration to co-occurrence between common nouns.

From the tagged corpus, we establish co-occurrence frequency distributions for each noun, n_i, indicating how many times every other noun occurs in sentences that contain n_i. Our measurement of co-occurrence is independent of word order and intermediary word distance, and our resulting data are asymmetric.

Consider, for example, the dual occurrence of the noun (stem) "astronomer" in the first of the following three sentences (in which the stems of all common nouns are highlighted):

(1) *Kamalakara (1616-1700), an Indian astronomer and mathematician, came from a family of astronomers.*

(2) *This quartic curve was studied by the Greek astronomer and mathematician Eudoxus of Cnidus.*

(3) *A school speed limit would be posted when entering the school zone.*

From (1) and (2), we have *frequency*(*astronomer*|*mathematician*) = 3, since there are three occurrences of "astronomer" in sentences containing "mathematician." However, *frequency*(*mathematician*|*astronomer*) = 2, and so the resulting frequency distributions are asymmetric (see Table 3.1 below).

Table 3.1:
Co-occurrence frequency distributions derived from sentences (1) and (2) above.

Source Noun	Co-occurring Noun	Freq.
astronomer	mathematician	2
	family	1
	curve	1
mathematician	astronomer	3
	family	1
	curve	1
family	astronomer	2
	mathematician	1
curve	astronomer	1
	mathematician	1

Table 3.2:
Co-occurrence frequency distributions derived from sentence (3) above.

Source Noun	Co-occurring Noun	Freq.
school	speed limit	1
	zone	1
speed limit	school	2
	zone	1
speed	school	2
	speed limit	1
	zone	1
limit	school	2
	speed limit	1
	zone	1
zone	school	2
	speed limit	1

Of interest is the fact that, in (2), our stemming algorithm reduces “quartic curve” to the head noun “curve” because the compound noun is not represented in the WN ontology. Consider,

in contrast, the occurrence of the compound "speed limit" in (3) (noting also that WN does not lexicalize "school zone," and so its constituents are marked as disjoint nouns).

In instances of open-form (multi-word) compound nouns, each unique noun (e.g., those highlighted above in (3): "school," "speed limit," and "zone") is counted in the co-occurrence frequency distribution for every other noun, as well as in the distributions for the constituents of any compound nouns. (Thus, the frequency distributions for nouns co-occurring with "school," "speed limit," "speed," "limit," and "zone" are updated with counts of "school," "speed limit," and "zone;" see Table 3.2 above.) We afford compound nouns this special treatment to ensure their constituents also benefit from semantic association with the nouns co-occurring in these sentences.

3.2 From Co-occurrence to Relational Strength

We now adopt the following terminology: a *target* is any noun for which we would like to discover a set of semantic associates. Nouns co-occurring intra-sententially with a target are called its *co-targets*, all of which come under consideration for semantic association to the target.

We define *relational strength*, $S_{rel}(t, c)$, as a quantitative measure of the semantic relatedness of a target, t, to one of its co-targets, c. To gauge relational strength, we measure the distance between two probability distributions: a prior distribution (giving the relative frequency of occurrence of every noun in the corpus), and a posterior distribution for our target (giving the relative frequency of its various co-targets with respect to all sentences containing the target).

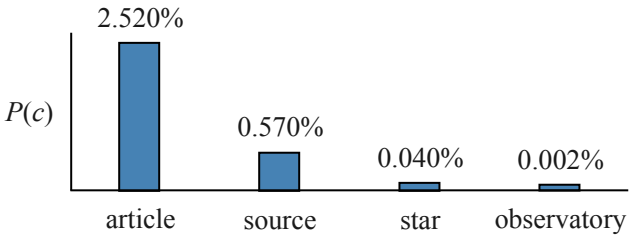


Figure 3.1:
Prior distribution sample from Wikipedia co-occurrence (not to scale).

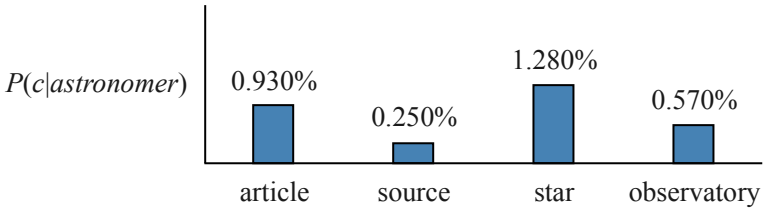


Figure 3.2:
Posterior distribution sample for co-targets of "astronomer" (not to scale).

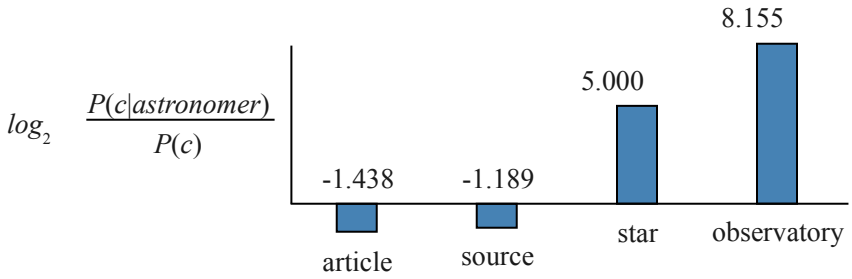


Figure 3.3:
Log ratio of the posterior and prior distributions (to scale).

Figure 3.1 shows a sample from our prior distribution.[18,19] We see that "article" (the most profuse noun in Wikipedia, accounting for 2.52% of all occurrences of common nouns) occurs

18 Figures 3.1, 3.2, and 3.3 are modeled after Resnik's (1997) presentation of prior and posterior distributions.

19 The graphical representations for Figures 3.1 and 3.2 are not to scale; they are smoothed with an additive factor of 0.1% for readability. The numerical values above each bar do not include this smoothing factor.

much more frequently than, for example, “star” and “observatory.” In contrast, “star” and “observatory” occur with significantly elevated relative frequency in sentences containing the noun “astronomer” (Figure 3.2), respectively accounting for 1.28% and 0.57% of all its co-occurring noun tokens. The posterior distribution for “astronomer” reveals that we cannot rely on co-occurrence as a direct measure of semantic relatedness. This is clear from the fact that “article” co-occurs more frequently with “astronomer” than does “observatory,” although the latter clearly bears stronger semantic relation to the astronomer.

Intuitively speaking, when $P(c|t)$ is greater than $P(c)$, c is co-occurring with t more frequently than dictated by chance, indicating heightened relational strength between the nouns. Conversely, if $P(c)$ is much greater than $P(c|t)$, we see a negative semantic relationship between t and c. As shown in Figure 3.3, dividing the posterior distribution by the prior distribution and taking the log (to ensure that $P(c) > P(c|t)$ yields negative values) gives us a reasonable initial view of relational strength.

We now formally define relational strength, the quantitative measure of the semantic relatedness of a target, t, to one of its co-targets, c, as follows:

$$S_{rel}(t,c) = P(t|c)\,P(c|t)\log\frac{P(c|t)}{P(c)} \tag{26}$$

$P(c)$ is the relative frequency of c’s occurrence in the corpus, and for $P(c|t)$ we use the relative frequency of c’s occurrence among all co-targets of t:

$$P(c) = \frac{frequency(c)}{\sum_{n\in W} frequency(n)} \tag{27}$$

$$P(c|t) = \frac{frequency(c|t)}{\sum_{n \in C_t} frequency(n|t)} \quad (28)$$

Here, W is the set of all nouns in Wikipedia, and C_t is the set of all co-targets of t.

Our formulation of $S_{rel}(t, c)$ is an adaptation of Resnik's (1999) selectional association measure:

$$A(w,c) = \frac{1}{D_{KL}} P(c|w) \log \frac{P(c|w)}{P(c)} \quad (29)$$

Resnik used (29) to measure the degree to which a word, w, selects a WordNet class, c, as an argument (e.g., the selectional preference of the adjective "wool" for nouns categorized as *clothing*, or the verb "eat" for *food*). In Resnik's formulation, D_{KL} is the relative entropy, or Kullback-Leibler divergence (Kullback & Leibler, 1951), between the probability distributions $P(C|t)$ and $P(C)$, where C is a set of WN noun classes:

$$D_{KL} = \sum_{c \in C} P(c|t) \log \frac{P(c|t)}{P(c)} \quad (30)$$

D_{KL} acts as a normalization factor in (29) and also gives an indication of how strongly the word w selects for its argument classes in general.

Bearing in mind that the selective power of a word reflects the degree to which it predicts the (co-)occurrence of a member of the class(es) for which it selects, Resnik's formula provides a good launching point for a measure of relational strength. We have, however, made two pragmatic changes to Resnik's formulation of $A(w, c)$ to derive our definition of $S_{rel}(t, c)$:

The first modification is the omission of the D_{KL} term. We are primarily interested in using $S_{rel}(t, c)$ to measure the relatedness of t to c relative to all other co-targets of t, rather than measuring relational strength in a global fashion. That is, given a target, t, and two of its co-targets, c_1 and c_2, we are interested in the comparative values of $S_{rel}(t, c_1)$ and $S_{rel}(t, c_2)$; they reveal which co-target bears the stronger relation from t. We are not, however, interested in the comparative values of $S_{rel}(t_1, c_1)$ and $S_{rel}(t_2, c_2)$, for two different targets, t_1 and t_2. To say that one is greater than the other reveals nothing about the association of t_1 to c_1 or of t_2 to c_2. Indeed, the extreme variability of D_{KL} from target to target, as well as the exponential decay of values of $S_{rel}(t, c)$ in practice, make it difficult to ascribe any meaning to the absolute values of the function. Accordingly, the function is used only to sort the list of t's co-targets in decreasing order of relational strength, after which the usefulness of the measure is exhausted, and its values are discarded. Thus, D_{KL}, which is constant with respect to c, can be dropped from the definition of $S_{rel}(t, c)$; the ordering of t's co-targets remains the same.

Our second modification is the inclusion of the $P(t|c)$ term in (26) in order to account for the relatedness of c to t, which certainly plays *some* role in the relational strength of t to c. This is particularly useful in suppressing words like "article" and "year," which tend to appear frequently with nouns that serve as titles of Wikipedia articles, despite the fact that those nouns are not generally semantically related to "article" or "year" at all.[20]

Intuitively speaking, $A(w, c)$ indicates how likely we are to encounter a noun categorized by c as a result of encountering w. $S_{rel}(t, c)$ follows suit, indicating how likely we are to

20 Although these problematic words are particular to our choice of corpus, our method for quashing them retains its generality for use with any corpus.

encounter c as a consequence of encountering t. The highest values of $S_{rel}(t, c)$ are assigned when c's relative frequency of co-occurrence with t is significantly higher than c's relative frequency of occurrence in the corpus.

Given a target of interest, we sort all of its co-targets by descending order of $S_{rel}(t, c)$. The notable exception is that if $P(c|t) < 0.07\%$, we exclude c from consideration as one of t's semantic associates outright. We previously reported that this was done primarily out of computational considerations (Szumlanski & Gomez, 2010); in our preliminary investigation into co-occurrence methods for discovering semantic associates, we assembled co-occurrence frequency distributions on demand, and only for a limited number of nouns. Since then, we have extracted co-occurrence frequency distributions for all nouns in the corpus, but we maintain the 0.07% threshold because the performance of our function degrades as $P(c)$ approaches zero, assigning disproportionately high values of relational strength. (This is a known issue with related information theoretic measures. See, e.g., the remarks of Grefenstette (1994) regarding how mutual information "strongly favors rarely appearing words" (p. 31) when used to measure semantic similarity.)

Tables 3.3 and 3.4 (below on pages 92 and 93, respectively) demonstrate the re-ordering effect of our relational strength function. The first table shows co-targets of "astronomer" sorted by decreasing frequency of co-occurrence (the top 60 out of 224 nouns occurring above the 0.07% threshold); the second shows the top co-targets of "astronomer" sorted by decreasing value of relational strength. We observe that the reordering effect of our function is sometimes insignificant (e.g., the movement of "astronomy" from rank 6 to rank 5). In other cases, the re-ordering is more dramatic, acting to suppress frequently co-occurring nouns (e.g, the shift of

"article" from rank 7 to rank 174) or promote infrequently co-occurring nouns (as with the shift of "minor planet" from rank 64 to rank 4, or the movement of "astrophysicist" from rank 82 to rank 8). Moderate shifts occur, as well (e.g., the movement of "star" from rank 4 to rank 17).

The function does not provide a perfect measure of semantic relatedness. Certainly, few people would argue that "astronomer" bears stronger semantic relation to "geographer" than to "star," despite the results presented in Table 3.4. What is important, however, is that the overall ordering provided by the function is generally sound. Toward the top of the list, we see strongly related nouns, and in general this relatedness diminishes as we proceed through the list. Most importantly, the function washes out frequently co-occurring nouns that bear no semantic relation to the target. The most suspect nouns that co-occur frequently with "astronomer" are all removed to ranks greater than 60 when sorted by $S_{rel}(t, c)$; over half of the nouns from Table 3.3 do not appear in Table 3.4 because their resulting ranks in the re-ordering (indicated here in parentheses) place them so low in the sorted list of 224 co-targets: historian (62), light (73), work (78), model (79), definition (84), system (92), data (93), research (94), time (96), book (98), period (99), year (100), position (101), study (102), world (105), name (109), term (121), number (122), team (125), fact (126), use (127), way (128), group (130), reference (132), example (144), member (152), history (157), point (165), part (166), article (174), people (189), source (193).

Table 3.3:
60 nouns most frequently co-occurring with "astronomer" in Wikipedia.

#	Co-Target	Frequency	$P(c\|t)$	#	Co-Target	Frequency	$P(c\|t)$
1	mathematician	583	1.85%	31	world	113	0.36%
2	amateur	569	1.81%	32	fact	109	0.35%
3	planet	480	1.52%	33	asteroid	108	0.34%
4	star	403	1.28%	34	people	106	0.34%
5	century	329	1.04%	35	universe	103	0.33%
6	astronomy	318	1.01%	36	use	100	0.32%
7	article	292	0.93%	37	model	100	0.32%
8	physicist	274	0.87%	38	number	99	0.31%
9	time	244	0.77%	39	sky	96	0.30%
10	object	227	0.72%	40	light	94	0.30%
11	telescope	227	0.72%	41	research	93	0.30%
12	observation	222	0.71%	42	engineer	93	0.30%
13	years[21]	212	0.67%	43	definition	91	0.29%
14	work	204	0.65%	44	group	90	0.29%
15	observatory	180	0.57%	45	position	86	0.27%
16	theory	179	0.57%	46	period	86	0.27%
17	galaxy	174	0.55%	47	reference	86	0.27%
18	scientist	165	0.52%	48	historian	82	0.26%
19	discovery	153	0.49%	49	example	81	0.26%
20	name	149	0.47%	50	instrument	80	0.25%
21	philosopher	144	0.46%	51	part	80	0.25%
22	book	140	0.44%	52	source	79	0.25%
23	comet	139	0.44%	53	study	79	0.25%
24	science	137	0.44%	54	point	79	0.25%
25	astrologer	135	0.43%	55	sun	78	0.25%
26	year	123	0.39%	56	team	78	0.25%
27	way	122	0.39%	57	term	77	0.24%
28	system	117	0.37%	58	history	77	0.24%
29	moon	116	0.37%	59	data	75	0.24%
30	orbit	114	0.36%	60	member	75	0.24%

21 "Years" is lexicalized in WN, and is therefore morphologically ambiguous; we do not stem it further.

Table 3.4:
60 co-targets most strongly related to "astronomer" by $S_{rel}(t, c)$.

#	Co-Target	$S_{rel}(t,c)$	#	Co-Target	$S_{rel}(t,c)$
1	mathematician	6.622	31	geologist	0.113
2	amateur	2.634	32	moon	0.100
3	observatory	2.467	33	sky	0.085
4	minor planet	2.296	34	eclipse	0.083
5	astronomy	2.150	35	object	0.063
6	astrologer	1.906	36	chemist	0.060
7	telescope	1.716	37	dwarf	0.053
8	astrophysicist	1.596	38	scientist	0.052
9	physicist	1.273	39	cosmology	0.049
10	planet	0.768	40	century	0.044
11	comet	0.734	41	black hole	0.042
12	asteroid	0.565	42	theologian	0.040
13	geographer	0.525	43	biologist	0.038
14	supernova	0.367	44	engineer	0.038
15	cartographer	0.338	45	crater	0.036
16	galaxy	0.313	46	physician	0.034
17	star	0.276	47	sun	0.033
18	quasar	0.242	48	calendar	0.030
19	redshift	0.237	49	universe	0.028
20	constellation	0.232	50	inventor	0.023
21	cosmologist	0.225	51	treatise	0.019
22	solar system	0.205	52	calculation	0.019
23	observation	0.200	53	sphere	0.017
24	nebula	0.195	54	instrument	0.016
25	philosopher	0.167	55	educator	0.015
26	astrology	0.130	56	science	0.014
27	orbit	0.127	57	cluster	0.014
28	discovery	0.119	58	theory	0.014
29	discoverer	0.115	59	poet	0.012
30	meteorologist	0.114	60	motion	0.012

3.2.1 Evaluation

Although we ultimately discard values of $S_{rel}(t, c)$ in favor of constructing an unweighted semantic network, an objective evaluation of our function's performance is still in order. In the relatedness literature, a standard approach is to measure correlation with mean similarity scores elicited from human participants by Rubenstein and Goodenough (1965) and Miller and Charles (1991). In these studies, participants rated the "similarity of meaning" of noun pairs on a scale of 0.0 ("semantically unrelated") to 4.0 ("highly synonymous"). Rubenstein and Goodenough had participants evaluate 65 word pairs in this manner. Miller and Charles then replicated the experiment using only 30 of the original 65 word pairs.

Given that our measurement of relational strength, $S_{rel}(t, c)$, is used only to rank co-targets by their relative relatedness to a particular target, we now exploit those ranks to evaluate our function. We score relatedness between two words, a and b, as a scaled mean of their ranks in each other's list of co-targets, as follows:

$$score(a, b) = 4.0 \times \text{AVG}\left(\frac{|C_a| + 1 - rank_a(b)}{|C_a|}, \frac{|C_b| + 1 - rank_b(a)}{|C_b|}\right) \tag{31}$$

where $rank_t(c)$ is the numerical rank of some co-target c among all of t's co-targets, as sorted by decreasing[22] value of $S_{rel}(t, c)$, and $|C_t|$ is the number of t's co-targets. That is, the most strongly related co-target of t has $rank_t(c) = 1$, and the least related co-target has $rank_t(c) = |C_t|$.

22 This is a deviation from our definition of $rank_t(c)$ in previous work (Szumlanski & Gomez, 2010). We adopt the present form to maintain an internally consistent definition of ranking, which is used elsewhere in this dissertation. The *score*(a, b) function has been modified accordingly, so the values it produces are consistent with previous work.

In the event that neither rank is defined, we let $score(a, b) = 0$. If exactly one of these ranks is defined, we take 75% of the defined term, rather than allowing it to be averaged with zero. Recall that $rank_t(c)$ is undefined not only if t and c do not co-occur in the corpus, but also when $P(c|t) < 0.07\%$.

We evaluate the correlation of the scores produced by this function to the mean similarity scores of Rubenstein and Goodenough (henceforth R&G) and Miller and Charles (henceforth M&C). In Table 3.5, we compare our correlation results to those presented in a review by Budanitsky and Hirst (2006), as well as five semantic relatedness studies published since then (Gabrilovich & Markovitch, 2007; Hughes & Ramage, 2007; Milne & Witten, 2008a; Patwardhan & Pedersen, 2006; Strube & Ponzetto, 2006).

Table 3.5:
Coefficients of correlation with human similarity judgments. Figures in starred rows are taken from Budanitsky and Hirst (2006).

Measure	**M&C**	**R&G**
Patwardhan and Pedersen (2006)	0.910	0.900
Hughes and Ramage (2007)	0.904	0.817
Relational Strength: $S_{rel}(t, c)$	0.852	0.824
* Leacock and Chodorow (1998)	0.838	0.816
* Lin (1998)	0.819	0.829
* Hirst and St-Onge (1998)	0.786	0.744
* Jiang and Conrath (1997)	0.781	0.850
* Resnik (1995)	0.779	0.774
Gabrilovich and Markovitch (2007)	0.720	0.820
Milne and Witten (2008a)	0.700	0.640
Strube and Ponzetto (2006)[23]	0.490	0.560
Human Correlation (Resnik 1995)	0.885	n/a

23 These results are from the path length measure (sim_{PL}) and were misreported in Szumlanski and Gomez (2010).

Our results correlate strongly to both M&C ($r = 0.852$, $p < 0.01$) and R&G ($r = 0.824$, $p < 0.01$). The coefficients of correlation (r-values) are from Pearson's product-moment correlation, and measure the strength of the linear relationship between two sets of data. Higher values indicate better correlation to the human-assigned scores; 1.0 would indicate a perfect fit.

We find that, on this task, our lexical co-occurrence method produces results that are competitive with methods that draw on rich semantic resources like WordNet and the underlying structure of Wikipedia. Our results are also comfortably within the realm of human performance; the last row in Table 3.5 comes from a replication of the M&C study in which Resnik (1995) again had 10 human participants rate the similarity of the 30 word pairs used in the earlier study. He then measured the correlation of each individual participant's ratings to the M&C ratings. The figure presented in Table 3.5 ($r = 0.885$) is the arithmetic mean of the 10 resulting coefficients of correlation, which Resnik (1995) frames as "an upper bound on what one should expect from a computational attempt to perform the same task" (p. 450). Thus, we caution that high correlation on this task, and particularly scores that exceed average human correlation, might indicate that a measure is failing to capture semantic relatedness beyond that of similarity.

Below, we present the ratings from our *score*(a, b) function alongside the human ratings from the R&G (Table 3.6) and M&C (Table 3.7) studies.

Table 3.6:
Comparison of *score* function to subjective similarity score judgments from Rubenstein and Goodenough (1965) (R&G). Correlation: $r = 0.824$.

#	**Word Pair**		**R&G**	***score***	#	**Word Pair**		**R&G**	***score***
1	cord	smile	0.02	0.00	34	car	journey	1.55	2.28
2	rooster	voyage	0.04	0.00	35	cemetery	mound	1.69	2.27

#	Word Pair		R&G	*score*	#	Word Pair		R&G	*score*
3	noon	string	0.04	0.00	36	glass	jewel	1.78	0.00
4	fruit	furnace	0.05	0.00	37	magician	oracle	1.82	0.00
5	autograph	shore	0.06	0.00	38	crane	implement	2.37	0.00
6	automobile	wizard	0.11	0.00	39	brother	lad	2.41	1.87
7	mound	stove	0.14	0.00	40	sage	wizard	2.46	0.00
8	grin	implement	0.18	0.00	41	oracle	sage	2.61	0.00
9	asylum	fruit	0.19	0.00	42	bird	crane	2.63	2.65
10	asylum	monk	0.39	0.00	43	bird	cock	2.63	2.68
11	graveyard	madhouse	0.42	0.00	44	food	fruit	2.69	3.23
12	glass	magician	0.44	0.00	45	brother	monk	2.74	2.38
13	boy	rooster	0.44	1.68	46	asylum	madhouse	3.04	3.73
14	cushion	jewel	0.45	0.00	47	furnace	stove	3.11	3.65
15	monk	slave	0.57	0.00	48	magician	wizard	3.21	3.85
16	asylum	cemetery	0.79	0.00	49	hill	mound	3.29	3.49
17	coast	forest	0.85	2.48	50	cord	string	3.41	2.26
18	grin	lad	0.88	0.00	51	glass	tumbler	3.45	2.82
19	shore	woodland	0.90	0.00	52	grin	smile	3.46	2.96
20	monk	oracle	0.91	0.00	53	serf	slave	3.46	2.89
21	boy	sage	0.96	1.47	54	journey	voyage	3.58	3.55
22	automobile	cushion	0.97	0.00	55	autograph	signature	3.59	2.92
23	mound	shore	0.97	1.50	56	coast	shore	3.60	3.59
24	lad	wizard	0.99	0.00	57	forest	woodland	3.65	3.85
25	forest	graveyard	1.00	2.17	58	implement	tool	3.66	2.88
26	food	rooster	1.09	1.18	59	cock	rooster	3.68	3.97
27	cemetery	woodland	1.18	0.00	60	boy	lad	3.82	2.97
28	shore	voyage	1.22	1.96	61	cushion	pillow	3.84	3.89
29	bird	woodland	1.24	2.24	62	cemetery	graveyard	3.88	3.79
30	coast	hill	1.26	2.65	63	automobile	car	3.92	3.77
31	furnace	implement	1.37	0.00	64	midday	noon	3.94	3.75
32	crane	rooster	1.41	0.00	65	gem	jewel	3.94	3.85
33	hill	woodland	1.48	2.17					

Table 3.7:
Comparison of *score* function to subjective similarity score judgments from Miller and Charles (1991) (M&C). Correlation: $r = 0.852$.

#	Word Pair		M&C	*score*	#	Word Pair		M&C	*score*
1	noon	string	0.08	0.00	16	lad	brother	1.66	1.87
2	rooster	voyage	0.08	0.00	17	brother	monk	2.82	2.38
3	glass	magician	0.11	0.00	18	tool	implement	2.95	2.88
4	chord	smile	0.13	0.00	19	bird	crane	2.97	2.65
5	lad	wizard	0.42	0.00	20	bird	cock	3.05	2.68
6	coast	forest	0.42	2.48	21	food	fruit	3.08	3.23
7	monk	slave	0.55	0.00	22	furnace	stove	3.11	3.65
8	shore	woodland	0.63	0.00	23	midday	noon	3.42	3.75
9	forest	graveyard	0.84	2.17	24	magician	wizard	3.50	3.85
10	coast	hill	0.87	2.65	25	asylum	madhouse	3.61	2.73
11	food	rooster	0.89	1.18	26	coast	shore	3.70	3.59
12	cemetery	woodland	0.95	0.00	27	boy	lad	3.76	2.98
13	monk	oracle	1.10	0.00	28	journey	voyage	3.84	3.55
14	journey	car	1.16	2.28	29	gem	jewel	3.84	3.85
15	crane	implement	1.68	0.00	30	car	automobile	3.92	3.77

3.3 From Relational Strength to Categorical Relatedness

We now present an algorithm for discovering categorical semantic relatedness between nouns. We will write pairs of related nouns as, e.g., (astronomer, star), which indicates the relatedness of "astronomer" to "star;" the former is our target, and the latter is a co-target that we have found to be semantically related. The collection of all such pairs constitutes a semantic network of related nouns.

Intuitively speaking, the idea behind our algorithm is this: if *t* is strongly related to *c* and, conversely, *c* is strongly related to *t*, we include the ordered pair (*t*, *c*) in our semantic network. For this purpose we rely on our measure of relational strength: once we have sorted a list of co-targets by decreasing value of relational strength with respect to some target, we have a good idea of which nouns are strongly related to the target (those at the top of the list) and which ones are strongly unrelated to the target (those at the bottom).

More formally, we introduce the notion of *mutual relatedness* between nouns, defined as follows: if *c* is in the top *x*% of *t*'s most strongly related co-targets (sorted by $S_{rel}(t, c)$), and *t* is in the top *x*% of *c*'s most strongly related co-targets, we say that *t* and *c* are mutually related within *x*%. The set of all nouns mutually related to *t* within *x*% is denoted $m_x(t)$.

To find the nouns categorically related to a target, *t*, we let $x = 20$ and find the initial set $m_x(t)$. We then expand this set by incrementing *x* until 5 iterations pass without *t* being related to any additional co-targets (see Figure 3.4 below). Our experiments have shown that varying these parameters has negligible effects on the results of our algorithm, even if we allow the algorithm to proceed until as many as 10 iterations have passed without any new relationships being discovered.

Upon termination of the algorithm, we admit to the network all ordered pairs (*t*, *c*) such that *c* is in $m_x(t)$ (for the final value of *x*, which we call the *admittance threshold* of *t*). In our algorithm, this set of ordered pairs is denoted S_0.

Input: A target noun, t.

Returns: Set of noun pairs (t, c) such that t and c are semantically related.

```
 1:  FindRelatedNouns( t ) {
 2:      S_0 = {}
 3:      NoGain = 0
 4:
 5:      for x = 20 to 100 do
 6:          S = {(t, c) | c ∈ m_x(t)}
 7:          if |S| > |S_0| then
 8:              NoGain = 0
 9:          else
10:              NoGain++
11:
12:          if NoGain ≥ 5 then
13:              break
14:          end if
15:
16:          S_0 = S
17:      end for
18:
19:      return S_0
20:  }
```

Figure 3.4:
Algorithm for establishing categorical relatedness from mutual relatedness.

The mutual relatedness algorithm exhibits several important properties worth mentioning. First, it accounts for the fact that some nouns are more permissive with their semantic relatedness than others, and relates each target to as many or as few nouns as it deems fit, rather than using a single, arbitrary threshold to restrict relatedness to all targets.

Second, the algorithm is resilient to the gradated nature of the relational strength of a target to its co-targets. This gradation makes it impossible even for human judges to find a clear cutoff above which we can consider all nouns to be related to the target, and below which we can

comfortably exclude their relatedness. However, our algorithm makes incisive decisions about relatedness without being lured down the slippery slope of over-inclusiveness.

A third notable feature of our algorithm is that it admits (*t*, *c*) only when the strength of *t*'s relatedness to *c* is reciprocated from *c* to *t* (as with "penguin" and "iceberg," which are strongly related in both directions; compare this with "ice" and "penguin," which are far more strongly related in one direction (penguin to ice) than the other (ice to penguin) and are therefore excluded from relation in the network).

3.3.1 Evaluation

We have constructed a semantic network of related nouns with this algorithm, using as our target nouns all those occurring between 1,500 and 100,000 times in Wikipedia. An overview of the resultant network is given in Table 3.8.

Table 3.8:
Summary of statistics for the semantic network of related nouns.

Property	**Description**	**Count**
Target Nouns	Number of nouns occurring between 1,500 and 100,000 times in Wikipedia.	7,593
Nodes	Number of nouns represented in network; includes both targets and co-targets.	25,142
Edges	Number of related word pairs; (*a*, *b*) and (*b*, *a*) are not counted as distinct word pairs.	155,180
Average Threshold of Target Nouns	Mutual relatedness algorithm's average admittance threshold for target nouns in network.	28.19%
Average Degree of Target Nodes	Average number of nouns to which each target is related.	31.29

We restrict our consideration to nouns occurring between 1,500 and 100,000 times in the corpus primarily because of the limitations of our information theoretic approach mentioned above: our approach often assigns disproportionately high values of relational strength when considering nouns that occur infrequently in the corpus. In the case of nouns occurring fewer than 1,500 time, we thus avoid false positive associations that arise under conditions of data sparsity. In the case of nouns occurring more than 100,000 times in the corpus (of which there are 430), we avoid false positives resultant of their high rates of co-occurrence with nouns that occur *comparatively* rarely in the corpus.

For the 7,593 target nouns in our restricted range, our algorithm produces a semantic network relating 25,142 distinct nouns (most of which appear as co-targets, but not targets themselves, because of their low frequency of occurrence in the corpus), derived from 237,584 noun pairs. Of these noun pairs, 82,404 are redundant, in that they are the symmetric images of pairs already included in the network. Thus, the network has 155,180 distinct undirected edges. Each target noun is related, on average, to 31.29 other nouns.

To evaluate the precision of the related noun pairs discovered by this procedure, we asked three judges with backgrounds in computational linguistics, none of whom had direct ties to this research, to evaluate 150 noun pairs and determine whether they would consider the nouns in those pairs to be semantically related or not. To prepare them for this task, we presented the judges with several exemplars of relatedness, which we hand picked from the network (see Table 3.9 below), and which exemplify a variety of relations (*AtLocation*, *TypeOf*, *UsedFor*, *ConceptuallyRelatedTo*, other functional relationships, collocations, and so on).

Of the 150 noun pairs presented to the judges for evaluation, 100 were chosen at random from the related pairs in our network. Additionally, 50 pairs of unrelated nouns were generated at random from among the nouns currently represented in the network. The 150 pairs were presented in random order to the judges. The results of their evaluations are summarized below in Table 3.10.

Table 3.9:
Exemplars of semantic relatedness, hand-picked from our network.

#	Pair
1	(astronomer, observatory)
2	(crime, prevention)
3	(automobile, gasoline)
4	(phone, signal)
5	(penguin, tuxedo)
6	(prison, lawyer)
7	(tendon, cartilage)
8	(string, output)
9	(desert, habitat)

Table 3.10:
Judges' evaluations of precision on related and unrelated noun pairs.

Judge	Related Pairs Judged as Related	Unrelated Pairs Judged as Unrelated
#1	99%	72%
#2	93%	80%
#3	95%	90%
Averages	95.66%	80.66%

On average, the judges evaluated 95.66% of the pairs from our network to be semantically related. They also judged 80.66% of the unrelated pairs to be unrelated. (That is, they identified an average of 19.34% of the unrelated (randomly paired) nouns as being related.)

This domain is too open-ended for there to be any feasible measure of recall. However, the fact that our target nouns are related to an average of 31.29 nouns while maintaining precision in excess of 95% is indicative of broad and accurate coverage of semantic relatedness. To further illustrate the quality of the relationships discovered by our approach, we have included a discussion of the semantic network surrounding the monosemous nouns (concepts) *astronomer* and *tennis* in the following chapter (see Section 4.7) and employed our network in a word sense disambiguation task to verify its utility as an applied resource (see Chapter 5).

CHAPTER 4: CONSTRUCTING THE NETWORK: FROM NOUNS TO CONCEPTS

Once we have established relatedness between nouns, we turn our attention to automatically disambiguating them to their corresponding noun senses in WordNet 3.0. In this chapter, we present our methodology for disambiguating nouns in our network (Sections 4.2 through 4.5) and provide an evaluation of our results (Section 4.6). In Section 4.7, we provide an explication of the semantic network surrounding the monosemous nouns (concepts) *astronomer* and *tennis*. In Section 4.8, we provide a discussion of the special considerations involved with disambiguating polysemous-to-polysemous pairs of associate nouns.

4.1 Preliminaries

To disambiguate the nouns in our network, we use a complex suite of disambiguation methods that work in tandem to support or refute one another's results. Because each of these methods has certain weaknesses, a noun sense has to be verified by at least two of them in order to be admitted to the network when the methods produce conflicting results. Preference is given to results produced by these methods in order of their presentation below. If all the methods described below fail to disambiguate a noun, we default to its most frequent sense in WordNet.

It is possible for multiple senses of a noun to be verified by these methods and admitted to the network. This is often desirable; rather than restricting ourselves to one sense, we allow for the possibility of ambiguity within the relationship (e.g., the relationship of *tax#1* (monosemous) to "administration," which could be either a presidential administration or, in the case of the

nominalized form, the act of administering a tax), and to some degree ameliorate the problem of fine-grained polysemic distinctions in WordNet (e.g., the relationship of *astronomer#1* (monosemous) to both *star#1* ("a celestial body of hot gases that radiates energy derived from thermonuclear reactions in the interior") and *star#3* ("any celestial body visible (as a point of light) from the Earth at night"), both of which are celestial bodies).

4.2 Subsumption Method

Our first disambiguation method capitalizes on the sense similarity clustering that we have found to occur among related nouns. For example, concepts related to *astronomer* form one cluster beneath the umbrella of *celestial_body#1* in WordNet (*planet#{1,3}*, *star#{1,3}*, *minor_planet#1*, *quasar#1*), another under the purview of *scientist#1* (*mathematician#1*, *physicist#1*, *chemist#1*), and so on.[24]

Accordingly, we determine the most frequently occurring immediate hypernyms for all the senses of the nouns related to a given target, and allow them to disambiguate the concepts they subsume. Although accidental inclusion of fringe senses categorized by common hypernyms occurs in rare cases, this is the strongest of our methods for disambiguation.

4.3 Gloss Method

Our gloss method gathers all monosemous nouns related to a target, as well as the target itself, and searches for these terms in the WordNet glosses of the target's polysemous associates.

24 Recall that we denote sense *n* of a noun by *noun#n*, or multiple senses with, e.g., *noun#{m, n}*.

Search terms may be pluralized, and suffixes from the set {-y, -er, -ist, -ing} may be replaced with any suffix from the set {-s, -es, -ies, -y, -er, -ist, -ing}, so that, e.g., "biologist" can also be matched by the occurrence of "biology," or "engineering" by "engineers."

This method returns a list of all noun senses with at least one of the search terms occurring in their glosses. Even with target nouns that have a large number of related terms, this list is surprisingly concise, although the results are less reliable than those of the previous method. However, these results do not require verification by another method if a search term matches a topic word in a sense gloss, as with "astronomy" in the glosses of *planet#1*, *galaxy#3*, and *star#1* (see Figure 4.1). Thus, any noun related to both "astronomy" (monosemous) and "star" will take *star#1* as an intended meaning of "star." However, this does not preclude us from including additional senses of "star" if there is strong evidence from the other disambiguation methods to support their inclusion.

planet#1: (astronomy) any of the nine large celestial bodies in the solar system that revolve around the sun and shine by reflected light
—from "astronomer"→"astronomy," "astronomy," and "solar system"

planet#3: any celestial body (other than comets or satellites) that revolves around a star
—from "comet"→"comets"

galaxy#3: (astronomy) a collection of star systems; any of the billions of systems each having many stars and nebulae and dust
—from "astronomer"→"astronomy" and "astronomy"

cosmology#2: the branch of astrophysics that studies the origin and evolution and structure of the universe
—from "astrophysicist"→"astrophysics"

star#1: (astronomy) a celestial body of hot gases that radiates energy derived from thermonuclear reactions in the interior
—from "astronomer"→"astronomy" and "astronomy"

Figure 4.1:
Inflected variants of monosemous associates of "astronomer" occurring in glosses of polysemous associates of "astronomer."

4.4 Selectional Preference Method

Next we use Resnik's (1999) selectional association measure to build selectional preferences for the nouns related to a given target. Formally, we define the selectional association, $A(t, c)$, of a target noun t with a WordNet class c as:

$$A(t,c) = \frac{1}{D_{KL}} P(c|t) \log \frac{P(c|t)}{P(c)} \tag{32}$$

As before, D_{KL} is the Kullback-Leibler divergence between probability distributions $P(C|t)$ and $P(C)$:

$$D_{KL} = \sum_{c \in C} P(c|t) \log \frac{P(c|t)}{P(c)} \tag{33}$$

Here, C is the set of WordNet classes denoted by monosemous associates of t, along with all the concepts in their hypernymic traces (all hypernyms of those concepts up to and including the root of the hierarchy, *entity#1*).

The posterior distribution, $P(C|t)$, derives from the frequency of co-occurrence of t's monosemous related nouns. To compute the prior distribution, $P(C)$, we use the frequency data for all monosemous nouns occurring between 1,500 and 100,000 times in Wikipedia. This is a departure from the approach of Resnik, who includes polysemous nouns (and their hypernymic traces) in both probability distributions and apportions credit for a noun evenly across all its senses. By focusing only on monosemous nouns in this approach, we eliminate the noise introduced by the ambiguity of polysemous nouns.

Each concept in C, a category in WordNet, is thereby associated with a numerical value indicating the strength of its selectional association with the target, t. Higher values indicate

stronger association. Once we have the selectional preferences derived from *t*'s monosemous associates, we use them to preferentially disambiguate *t*'s polysemous associates.

Table 4.1:
All selectional preferences derived from monosemous associates of "unicorn."

Rank	WordNet Class (c)	A(t, c)	Rank	WordNet Class (c)	A(t, c)
1	monster#1	12.350	20	chordate#1	4.355
2	mythical_being#1	12.350	21	vertebrate#1	4.355
3	mythical_monster#1	12.350	22	animal#1	4.040
4	mermaid#1	10.734	23	container#1	3.470
5	goblin#1	10.519	24	cognition#1	3.265
6	utensil#1	9.113	25	content#5	2.499
7	imaginary_being#1	8.703	26	psychological_feature#1	2.160
8	imagination#1	8.703	27	activity#1	2.129
9	creativity#1	8.237	28	act#2	0.920
10	vessel#3	7.495	29	event#1	0.639
11	evil_spirit#1	7.327	30	abstraction#6	0.523
12	spirit#4	7.327	31	instrumentality#3	0.412
13	spiritual_being#1	6.763	32	entity#1	0.000
14	ability#2	6.547	33	organism#1	-0.971
15	creation#1	6.490	34	living_thing#1	-0.980
16	implement#1	5.763	35	whole#2	-1.038
17	placental#1	5.419	36	artifact#1	-1.070
18	mammal#1	5.090	37	object#1	-1.281
19	belief#1	4.688	38	physical_entity#1	-1.491

Consider, for example, the categories in WordNet with the highest selectional association with the monosemous noun "unicorn" (Table 4.1). Among these selectional preferences we find *mythical_monster#1*, *imaginary_being#1*, and *spiritual_being#1*, which do not appear as

semantic associates of "unicorn," but do categorize many of the monosemous associates of "unicorn," such as "griffin," "goblin," "mermaid," "leprechaun," and "minotaur," among others. (The complete list of semantic associates of "unicorn" is presented in Table 4.2.)

Table 4.2:
All semantic associates of "unicorn" in our network.

Polysemous Associates		***Monosemous Associates***	
#	**Noun**	**#**	**Noun**
1	lion	1	griffin
2	dragon	2	goblin
3	nerd	3	mermaid
4	pony	4	origami
5	beast	5	teapot
6	satyr	6	leprechaun
7	phoenix	7	minotaur
8	tapestry	8	mythical creature
9	centaur	9	manticore
10	li	10	legendary creature
11	horn	11	narwhal

The selectional preferences from Table 4.1 are applied, in decreasing order of selectional strength, to each sense of the target's polysemous associates, which are disambiguated to the sense or senses categorized by the first such selection preference that subsumes them. Thus, "phoenix" (as it relates to "unicorn") is disambiguated to *phoenix#3* in WordNet ("a legendary Arabian bird said to periodically burn itself to death and emerge from the ashes as a new phoenix") by virtue of its subsumption by *mythical_being#1*. (No senses of "phoenix" are subsumed by the stronger selectional preference, *monster#1*.) The three senses of "phoenix" that

are excluded here are *phoenix#1* (the capital city of Arizona), *phoenix#2* (the taxonomic group *genus Phoenix*, which classifies many palm trees, including the date palm), and *phoenix#4* (a constellation). These selectional preferences similarly succeed in disambiguating the polysemous "lion" to *lion#1* (a feline, as opposed to the celebrity, astrological categorization of a person, or sign of the zodiac, denoted by senses 2, 3, and 4 of "lion," respectively), "beast" to *beast#1* (the animal, as opposed to a cruel person, which is sense 2 of "beast"), and "satyr" to *satyr#2* (the mythical woodland deity, as opposed to sense 1 of "satyr," which refers to a lecherous man).

If an upper-level ontological concept like *physical_entity#1* or *abstract_entity#1* performs the disambiguation in this method, we automatically dismiss the result as being too general to be reliable. More specifically, if c_0 is the strongest selectional preference from our list that disambiguates some polysemous noun related to t, and $A(t, c_0) < \overline{A(t, c)}$ (the mean value of $A(t, c)$ for all $c \in C$), then we discard the result and this method fails to disambiguate the polysemous noun in question. (For t = "unicorn," for example, $\overline{A(t, c)} = 4.682$. Thus, all WN classes in the right-hand column of Table 4.1 are prohibited from performing disambiguation by selectional preference.)

This method sometimes assigns disproportionately strong selective power to hypernyms that are particularly rare in the prior distribution. As such, this method defers to the subsumption and gloss methods when its results conflict with theirs.

4.5 Extended Gloss Method

In the event that none of the methods above produce verifiable results, we extend our gloss method from Section 4.3 by using as our search terms all semantic associates of the target (including polysemous associates), and all of *their* monosemous associates, in turn. In this case, we do not allow noun senses to be disambiguated by topic word matches, as the list of search terms has become too bloated. We do, however, allow this method to validate the results of the subsumption method, or, failing that, to support the results of the selectional preference method, or, as a last resort, to support the results of the original gloss method if it is supporting only one or two of the noun senses given by that method, and only if the list of noun senses given by this extended method is only larger than that of the gloss method by one or two terms.

That is to say, we treat the results of this extended gloss method with skepticism, and they are admitted to the semantic network only in rare cases. Barring the ability of this method, if it is called upon, to support a disambiguation result of one of the other methods given above, we default to the most frequent noun sense for the polysemous noun in question.

4.6 Evaluation

In our initial investigation into automatic semantic network construction (Szumlanski & Gomez, 2010), we only used these methods to disambiguate the polysemous associates of monosemous target nouns in our network. That is, we restricted our concept-network to pairs from the noun-network that included at least one monosemous noun. The intuition behind our approach was that monosemous nouns provide an unambiguous context in which disambiguation

of a polysemous associate can take place (cf. the monosemous "rhinoceros" and the polysemous "horn," which brings to mind an animal appendage, but not a car horn or the kind of horn that is a musical instrument). We deferred the resolution of ambiguity in polysemous—polysemous relationships to later work (Szumlanski & Gomez, 2011).

There are 3,024 monosemous target nouns in our network, heading up 76,264 of our related noun pairs from the previous section. 36,385 of these pairs associate two monosemous nouns and are admitted to our network of related concepts without need for disambiguation. The remaining 39,879 noun pairs connect our monosemous targets to polysemous nouns that we disambiguated using the subsumption, gloss, and selectional preference methods described above. Statistics for the resulting semantic network of related concepts are given below in Table 4.3.

Table 4.3:
Summary of statistics for the semantic network of related concepts (monosemous targets only).

Property	**Description**	**Count**
Target Nouns	Number of monosemous nouns occurring between 1,500 and 100,000 times in Wikipedia.	3,024
Nouns	Number of nouns represented in network; includes both targets and co-targets.	17,543
Nodes	Number of noun senses represented in network; includes both target and co-target noun senses.	24,547
Edges	Number of related noun sense pairs; (*a*, *b*) and (*b*, *a*) are not counted as distinct pairs.	74,166
Average Degree of Target Nodes	Average number of noun senses to which each monosemous target is related.	27.80

To evaluate the precision of our disambiguation results, we randomly selected 50 monosemous-to-polysemous noun pairs from our network and presented them to our three judges, along with the gloss and taxonomic categorization for every sense of the polysemous noun in the pair. The judges were asked to grade the relation of each noun sense of the polysemous associate to the monosemous target using the scale presented below (Table 4.4). Figure 4.2 (below on page 115), shows how the data was presented to the judges, and gives one judge's ratings for all senses of "dissociation" as it relates to the monosemous noun "nucleotide."

Table 4.4:
Scale used by judges to rate acceptability of disambiguation results.

Rating	Description
4	Primary intended sense or one of its synonyms.
3	Strongly related sense, but not the primary intended meaning.
2	Weakly related sense; could reasonably be included or excluded from relation to the target.
1	Unrelated sense.

We then measured how often the senses chosen by our disambiguation algorithm fell into each of these categories, and compared our results to the standard baseline of randomly selecting noun senses (see Table 4.5 below on page 116). The first column of the table (*grade* $\geq$ 4) indicates how frequently our system disambiguated to senses the judges considered to be the primary intended meanings of the related nouns. The last column (*grade* = 1) indicates how often our system selected senses that were unacceptable to the judges. The next-to-last column (*grade* $\geq$ 2) indicates how frequently our system chose senses that were acceptable to our judges.

```
==========================================================================
   TARGET (monosemous):  nucleotide
 COTARGET (polysemous):  dissociation
==========================================================================
 [1]  dissociation#1:
      the act of removing from association

      dissociation
       => separation
           => change of integrity
               => change
                   => action
                       => act, deed, human action, human activity
                           => event
                               => psychological feature
                                   => abstraction, abstract entity
                                       => entity

 [1]  dissociation#2:
      a state in which some integrated part of a person's life becomes
      separated from the rest of the personality and functions
      independently

      dissociation, disassociation
       => psychological state, psychological condition, mental state
           => condition, status
               => state
                   => attribute
                       => abstraction, abstract entity
                           => entity

 [4]  dissociation#3:
      (chemistry) the temporary or reversible process in which a
      molecule or ion is broken down into smaller molecules or ions

      dissociation
       => chemical process, chemical change, chemical action
           => natural process, natural action, action, activity
               => process, physical process
                   => physical entity
                       => entity
==========================================================================
```

Figure 4.2:
Sample judge's evaluation indicating the degree to which each sense of "dissociation" relates to "nucleotide."

Table 4.5:
Disambiguation precision, as compared to judges' manual sense annotations.

Judge	***grade*** **≥ 4**	**≥ 3**	**≥ 2**	**= 1**
#1	77%	79%	83%	17%
#2	65%	77%	90%	10%
#3	71%	79%	83%	17%
Average	71%	78%	85%	15%
Baseline	44%	53%	62%	38%

Given that 47.7% of the edges in our network connect two monosemous nouns (where there is no room for disambiguation error) and the remaining 52.3% have an average rate of acceptability of 85% as evaluated by our judges, we estimate the precision of the concept-to-concept associations in our semantic network to be 92.15%.

4.7 Excerpts and Explication: Selected Views of the Semantic Network

We now present abbreviated excerpts from the semantic network of related concepts for the monosemous nouns "tennis" (Figure 4.3) and "astronomer" (Figure 4.4). These excerpts come from the version of the network in which only associate pairs involving at least one monosemous noun have been disambiguated. In this network, *astronomer#1* is related to 45 distinct concepts (Table 4.6), and *tennis#1* is related to 80. For the sake of clarity, we present only a small sampling of those related concepts graphically. Furthermore, to avoid messy edge crossings in the graphs, we do not show interrelatedness between semantic associates of our

targets. (For example, *astronomy#1* and *astrologer#1* are both related to *astrology#1*, but we instantiate the latter node twice in the graph to preserve clarity.)

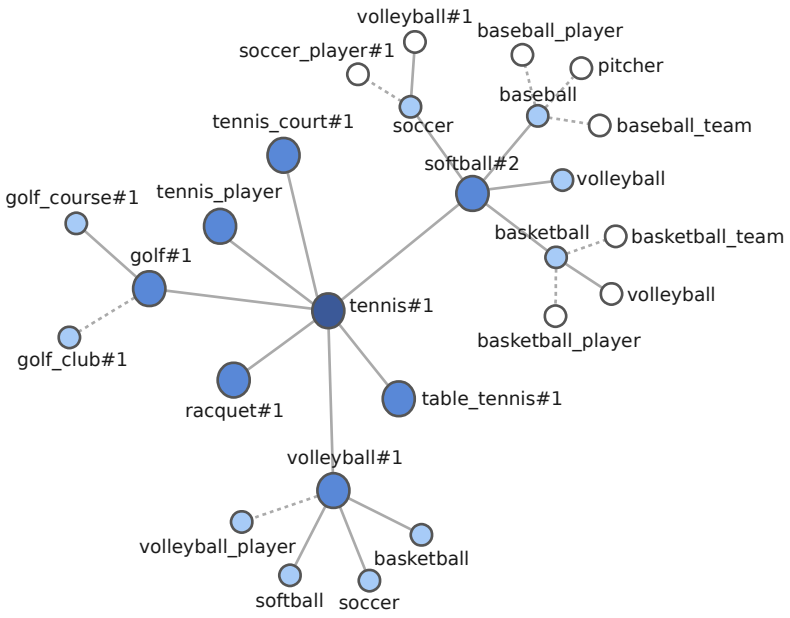


Figure 4.3:
Partial spreading activation view of concepts related to *tennis* in our network.

The target concepts' nodes in the graphs are dark blue (*astronomer#1* and *tennis#1*). We provide a sampling of their related terms in medium blue. In turn, those concepts are related to concepts in light blue, and those terms are related to concepts in white. This gives an idea of spreading activation through the semantic network.

In all cases, solid edges indicate that the target is related to the farther node incident to that edge. For example, the solid edge from *star#{1,3}* to *sky#1* in Figure 4.4 indicates that

astronomer#1 is related to *sky#1*, too. The dotted edge from *astrology#1* to *horoscope#{1,2}* indicates that *astronomer#1* is not related to *horoscope#{1,2}* in our network.

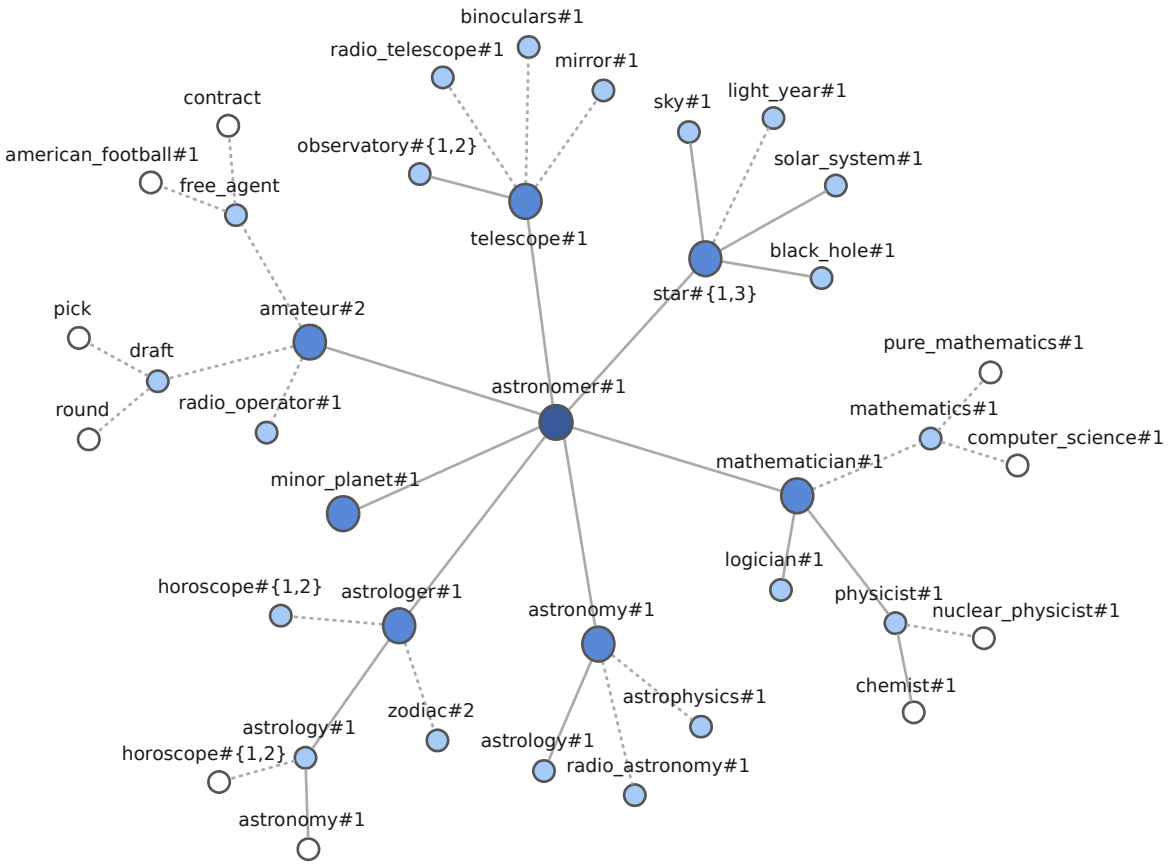


Figure 4.4:
Partial spreading activation view of concepts related to *astronomer* in our network.

Some nouns are not yet disambiguated in these graphs because they are related to concepts denoted by polysemous nouns, but we see how these might easily be disambiguated. Notice, for example, that *tennis#1* is related to *softball#2* (the *game* of softball, as opposed to the ball itself), which is in turn related to some (as yet undetermined) sense of "volleyball." Because

tennis#1 is related to *volleyball#1* (again, the game as opposed to the ball), this can be propagated through the network to disambiguate the relationship between *softball#2* and "volleyball" as (*softball#2*, *volleyball#1*). Although this is not the approach we will take as we resolve remaining ambiguities in the following section, it provides insight into how we might subsequently resolve disambiguation errors in the network.

Table 4.6:
All semantic associates of *astronomer* in our network.

minor_planet#1	constellation#2	orbit{#1,4}	discovery#1
geographer#1	astrologer#1	cosmologist#1	moon#6
theologian#1	asteroid#1	supernova#1	geologist#1
astronomy#1	astrophysicist#1	nebula#3	galaxy#3
quasar#1	biologist#1	observation#1	redshift#1
telescope#1	black_hole#1	solar_system#1	amateur#2
cartographer#1	astrology#1	meteorologist#1	cosmology#2
comet#1	mathematician#1	physicist#1	eclipse#1
planet#{1,3}	observatory#1	chemist#1	treatise#1
sky#1	philosopher#1	star#{1,3}	dwarf#2
discoverer#1	crater#3		

There are also cases in which polysemous nouns are related to disambiguated concepts in the graphs, such as with the relation of *star#{1,3}* to *solar_system#1*. "Solar system" is monosemous in WordNet, and our disambiguation algorithm found it to be semantically related to *star#{1,3}*.

We note that while our algorithm discovers some semantic similarity relationships (e.g., the relation of *astronomer#1* to *mathematician#1* and *astrophysicist#1*), it also discovers many

relationships beyond similarity, including concepts related through collocation (as with *amateur#2*, which, incidentally, is incorrectly disambiguated) and more general semantic relatedness (*telescope#1*, *star#{1,3}*, *planet#{1,3}*, *galaxy#3*, *observatory#1*, *redshift#1*, etc.).

Equally important is the absence of relationships to semantically similar concepts to which the targets are not strongly semantically related. Consider, for example, the fact that *astronomer#1* is related to some hyponyms of *scientist#1* (*physicist#1*, *mathematician#1*, *chemist#1*), but not others (*linguist#1*, *psychologist#1*, *medical_scientist#1*, etc.), despite the fact that quantitative relatedness measures based on their taxonomic categorizations in WordNet would erroneously relate *astronomer#1* to all these terms with nearly equal strength.

The network also associates *astronomer#1* with *astrologer#1*, which is clearly related, but is surprisingly far removed from *astronomer#1* in WordNet. (Their first shared hypernym in the ontology is *person#1*.)

Finally, notice the relation of *astronomer#1* to *astrophysicist#1* and *mathematician#1*, but neither *astrophysics#1* nor *mathematics#1*, although it is transitively related to the latter concepts by way of the former, as well as by way of *astronomy#1*. Similarly, mechanisms of spreading activation transitively relate *astronomer#1* to additional concepts like *light_year#1* by way of *star#{1,3}*, *radio_astronomy#1* by way of *astronomy#1*, and so on. This is arguably quite ontologically sound. The *astronomer* himself is more strongly related to the *astrophysicist* and the *celestial body* senses of "star" than to the *light year* or the study of *astrophysics*, although he is indirectly related to the latter concepts.

4.8 Completing the Network: Resolving Ambiguity with Polysemous Noun Targets

With a monosemous target, the disambiguation methods described above (Sections 4.2 to 4.5) benefit from the fact that all semantic associates under consideration are related through the same sense (the *only* sense) of the target noun in question. High degrees of interrelatedness and shared subsumption among those co-targets thus ameliorate the disambiguation task. In the case of a polysemous target, semantic associates are no longer bound together by that single common monosemous associate. Thus, the associate nouns, no longer necessarily being interrelated, exhibit greater entropy in terms of their ontological categorizations. In this section, we discuss the special considerations that therefore arise during the disambiguation of polysemous targets and their semantically related nouns in the network.

We first note that when dealing with a polysemous target, *t*, and a monosemous associate, *c*, sometimes it so happens that *c* is treated as a target in its own right elsewhere in the network (i.e., *c* occurs between 1,500 and 100,000 times in the corpus and has been associated with other nouns in addition to *t*). Since *c* is monosemous, we have already disambiguated all of its co-targets in the previous sections. Thus, there is nothing to be done for the noun pair (*t*, *c*); the ambiguity of *t* was resolved when considering the pair's symmetric image, (*c*, *t*) (and *c*, being monosemous, requires no disambiguation).

This forms an initial partitioning of nouns by their relation to individual senses of our polysemous target, *t*. Consider, for example, the polysemous "virus," which can refer to a computer virus (*virus#3*) or a microorganism (*virus#1*). In Figure 4.5 below, we show all monosemous associates of "virus" that also occur as targets in our network (light blue nodes). Among them is the monosemous *spyware#1*, shown in relation to its own semantic associates

(white nodes in the figure). Many of the nouns related to "spyware" have senses categorized as *software* in WordNet. These include "freeware," "computer virus," "trojan," "trojan horse," and "virus." Our subsumption method (Section 4.2) disambiguates "virus" to *virus#3*, the computer virus, accordingly. Our disambiguation methods similarly relate the biologically oriented associates of "virus" to *virus#1*, the infectious agent, given their relatedness to other nouns that fall under various biological categorizations in WordNet.

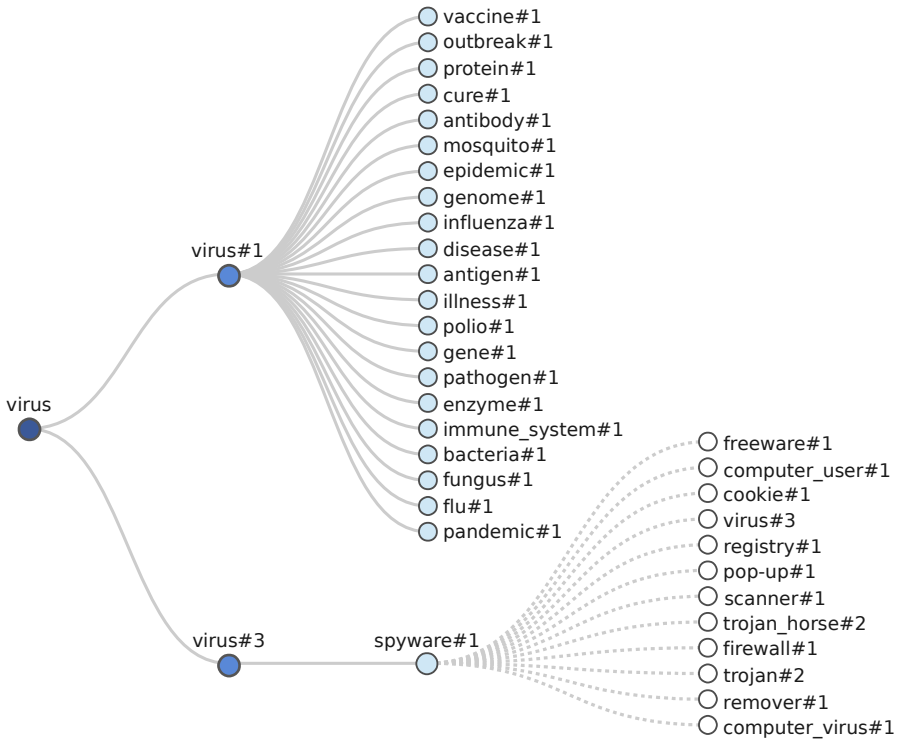


Figure 4.5:
Monosemous associates of "virus" that also appear as targets in our network.

Table 4.7:
All semantic associates of "batter" in our network.

at-bat	dough	hitter	perfect game	strike zone
baking	dugout	home plate	pitch	strikeout
ball	fastball	home run	pitcher	swing
base	fielder	homer	pitching	tempura
baseball	fielding	infield	plate	third base
bat	first base	infielder	pudding	third baseman
batsman	first baseman	inning	reliever	throw
bowler	flour	major league	runner	thrower
bunt	fly	mound	second base	triple
cake	fly ball	no-hitter	second baseman	umpire
catcher	foul ball	outfield	shortstop	waffle
center field	glove	outfielder	shutout	walk
center fielder	ground ball	pan	strike	wild pitch
double play	hit	pancake		

If, on the other hand, *c* is a polysemous associate of *t*, our task is slightly more complex. Without a monosemous semantic anchor for the pair, we no longer have an unequivocal context in which disambiguation can take place. We have found, however, that semantic clusters still form among the semantic associates of polysemous nouns.

Consider, for example, the semantic associates of the polysemous "batter," which can refer to a baseball player (*batter#1*) or the kind of batter used to make cakes and other baked goods (*batter#2*). A list of all nouns related to "batter" in our network is given above in Table 4.7. A subset of these associates is shown below in Figure 4.6, where we see the clusters that form from shared hypernymic relationships between individual senses of these nouns. In the graph, blue nodes denote semantic associates of "batter;" white nodes are their hypernyms and

are not semantic associates of "batter" in our network. The gray node in the center, *entity#1*, is the root of the hierarchy, and categorizes all adjacent concepts. Subsumption radiates outward from that node, so that, for example, *food#1* subsumes *dessert#1*, which in turn subsumes *pudding#{2,3}*, and so on. Solid edges in the graph represent immediate subsumption by the more central node (e.g., the solid edge from *cake#3* to *waffle#1* establishes *cake#3* as the immediate hypernym of *waffle#1*), whereas dotted edges represent *eventual* hypernymy (as with the edge from *sports_equipment#1* to *glove#3*; *sports_equipment#1* is a hypernym of *glove#3*, although there are other concepts between them in the ontology).

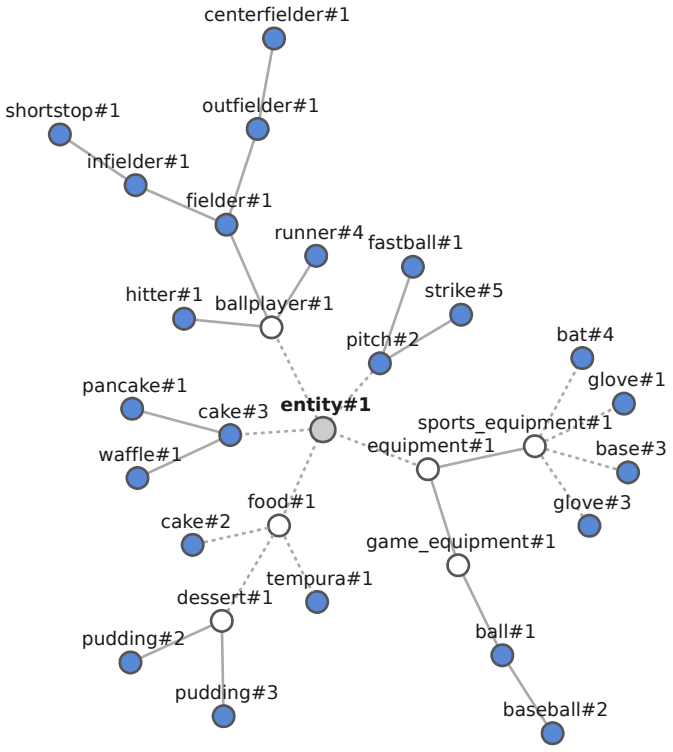


Figure 4.6:
Partial view of the WordNet graph, showing subsumption clusters formed by a subset of the semantic associates of "batter" in our network.

We see from Figure 4.6 that many of the nouns related to "batter" have senses categorized by *food#1*, *cake#3*, *pitch#2*, *ballplayer#1*, or *equipment#1*—the heads of five distinct clusters by semantic similarity.

It is worth noting that some nouns related to "batter" (such as "baking," "swing," and "umpire") do not fall into any of these semantic clusters. In these cases, the WordNet glosses serve as our primary tool for disambiguation. (For example, the glosses of both *swing#8* and *umpire#1* include mention of "baseball," which is also related to "batter.")

Conversely, some of the polysemous nouns in our example have senses that join semantic clusters unintendedly. For instance, *cake#2* (a "small flat mass of chopped food," according to WordNet) falls under the cluster headed by *food#1*. Although this is potentially problematic, *cake#2* is discarded in this particular case in favor of *cake#3* (the baked good), which has a greater mass because of its subsumption of *waffle#1* and *pancake#1*, and is indeed the intended meaning of "cake" as it relates to "batter."

Another example of unintended cluster membership comes from *bat#4* (the cricket bat), which is categorized by *sports_equipment#1*. In contrast, the baseball bat does not have its own entry in WordNet, and the most reasonable sense choice, *bat#5* ("a club used for hitting a ball in various games"), is categorized as a stick (*stick#1*), and not as equipment, sports equipment, or game equipment.

These unintended cluster memberships are bound to cause minor errors in our disambiguation efforts. However, we do not find such high entropy among the relatives of a polysemous noun that the semantic clustering effect (which is necessary for the success of the disambiguation algorithms described above in Sections 4.2 to 4.5) is diminished. Thus, when

confronted with a pair of semantically related polysemous nouns, we apply our disambiguation mechanism in both directions, and then fuse the results together. So, in one direction, the various baked goods related to "batter" help us to properly disambiguate "cake" to *cake#3*, yielding the pair (batter, *cake#3*). A similar scenario yields (cake, *batter#2*) when disambiguating in the other direction. We fuse the results together into the properly disambiguated pair (*batter#2*, *cake#3*).

This process assumes that we have already acquired the semantic associates of the co-target, *c*. Otherwise, our disambiguation methods have no way to resolve the meaning of the polysemous target. However, if *frequency*(*c*) < 1,500, then we have no associates for *c* other than those incidental targets (like our current *t*) that found association to *c*. In these cases, we use our mutual relatedness algorithm (Section 3.3) to derive a temporary set of associate nouns for *c*. These associates are not admitted to the network; they are simply used for disambiguation and then discarded, the idea being that if association is over-inclusive in the case of infrequently occurring nouns, we will still see some clustering effects among an inflated set of temporary associates.

Using this method, we have resolved all nouns in our network to noun senses, giving rise to a semantic network that has 208,832 pairs of related noun senses—the most extensive semantic network between WordNet noun senses to be derived from a lexical co-occurrence measure. A summary of relevant statistics is given below in Table 4.8. Of the 7,593 target nouns for which we have acquired semantic associates, 3,024 are monosemous and represented by a single node in the network. The remaining 4,569 are polysemous and are represented by 17,104 distinct concepts. In all, the network contains 25,142 unique nouns, with 38,249 distinct senses

among them. On average, target nodes in the network (those that represent individual senses of our 7,593 target nouns) are related to 19.06 other concepts.

Table 4.8:
Summary of statistics for the semantic network of related concepts (SGN).
Includes monosemous and polysemous targets.

Property	Description	Count
Target Nouns	Number of nouns occurring between 1,500 and 100,000 times in Wikipedia.	7,593
Monosemous Target Nouns	Number of monosemous target nouns for which our system has acquired relatedness data.	3,024
Polysemous Target Nouns	Number of polysemous target nouns for which our system has acquired relatedness data.	4,569
Target Nodes	Number of target noun senses represented in the network.	17,104
Target Nodes (From Monosemous Nouns Only)	Number of target noun senses represented in the network that are derived from monosemous target nouns.	3,024
Target Nodes (From Polysemous Nouns Only)	Number of target noun senses represented in the network that are derived from polysemous target nouns.	14,080
Nouns	Number of nouns represented in network; includes both targets and co-targets.	25,142
Nodes	Number of noun senses represented in network; includes both target and co-target noun senses.	38,249
Edges	Number of related noun sense pairs; (a, b) and (b, a) are not counted as distinct pairs.	208,832
Average Degree of Target Nouns	Average number of noun senses to which each (possibly ambiguous) target noun is related.	42.93
Average Degree of Target Nodes (From Monosemous Nouns Only)	For all nodes derived from monosemous target nouns, the average number of adjacent nodes.	28.33
Average Degree of Target Nodes (From Polysemous Nouns Only)	For all nodes derived from polysemous target nouns, the average number of adjacent nodes.	17.06
Average Degree of Target Nodes	Average number of noun senses to which each target noun sense is related.	19.06

For the remainder of this dissertation, we will refer to our network as the Szumlanski-Gomez Network, or SGN. In the following chapter, we evaluate our network by examining its performance on a word sense disambiguation task that relies on the concept-to-concept associations in SGN.

CHAPTER 5: COARSE-GRAINED WORD SENSE DISAMBIGUATION: AN APPLICATION

In the preceding chapters, we presented a method for automatically acquiring a semantic network of related concepts, or noun senses, from lexical co-occurrence in a large corpus. We applied our approach to Wikipedia, giving rise to a network that has relatedness data for over 7,500 of the most frequently occurring nouns in the corpus. The target nouns represented in our network are related to an average of 19.06 distinct noun senses. It consists of 208,832 undirected edges indicating general semantic relatedness between concepts. We refer to the network as the Szumlanski-Gomez Network (henceforth SGN).

In this chapter, we evaluate the performance of our semantic network on a word sense disambiguation (WSD) task and show: a) the network is competitive with WordNet when used as a stand-alone plug-in knowledge source for two graph-based WSD algorithms, b) combining our network with WordNet achieves disambiguation results that exceed the performance of either resource individually, and c) our network outperforms a similar resource, WordNet++ (Ponzetto & Navigli, 2010), that has been automatically derived from semantic annotations in the Wikipedia corpus.

5.1 WordNet++

WordNet++ (henceforth WN++) (Ponzetto & Navigli, 2010) is constructed automatically from the semantic annotations and structural properties of Wikipedia. Links in WN++ are established from inter-article links in the encyclopedia. For example, the article on *astronomy* in

Wikipedia links to the article on *celestial navigation*, so we find an edge from *astronomy#1* to *celestial_navigation#1* in WN++. The nouns related in WN++ are disambiguated automatically using further semantic annotations and metadata from Wikipedia, including sense labels, the titles of other pages linked to by any two related nouns, and the folksonomic categories to which articles belong. These serve as context words that are compared with context words from various WordNet relations in order to map the nouns to their appropriate WordNet senses. The resulting resource contains 1,902,859 unique edges between noun senses. The construction of WN++ is discussed in detail above in Chapter 2 (see Section 2.5.4).

Ponzetto and Navigli use "WN++" to refer to the union of all edges in WordNet and the set of additional edges they derived from Wikipedia. That is, WN++ is an augmented version of WordNet and contains the entire WordNet noun ontology. We depart from this convention for the remainder of this dissertation, instead using "WN++" to refer strictly to the *RelatedTo* links contributed by Ponzetto and Navigli. This gives us a convenient way to identify their resource as we evaluate it in comparison to SGN and in isolation from WordNet.

5.2 Coarse-Grained WSD Experiments

To evaluate our semantic network, and to provide fair comparison to related work, we take our cue from Ponzetto and Navigli (2010), who evaluated the performance of WN++ on the SemEval-2007 (Navigli et al., 2007) coarse-grained all-words WSD task using the extended gloss overlaps measure (Banerjee & Pedersen, 2003) and the graph-based degree centrality algorithm (Navigli & Lapata, 2010).

In this particular SemEval task, we are presented with 237 sentences in which target words have been lemmatized (that is, reduced from morphologically inflected forms to their canonical WordNet forms and tagged with parts of speech) and flagged for disambiguation (see Figure 5.1). For example, the sentence *In quoting from our research, you emphasized the high prevalence of mental illness and alcoholism* has the following lemmatized target words to be disambiguated: quote.v, research.n, emphasize.v, high.a, prevalence.n, mental.a, illness.n, and alcoholism.n.

d001.s006: In quoting from our research, you emphasized the high prevalence of mental illness and alcoholism.
— Lemmas: quote.v, research.n, emphasize.v, high.a, prevalence.n, mental.a, illness.n, alcoholism.n

d001.s011: The interactions between health and homelessness are complex, defying sweeping generalizations as to "cause" and "effect."
— Lemmas: interaction.n, health.n, homelessness.n, be.v, complex.a, defy.v, sweeping.a, generalization.n, cause.n, effect.n

Figure 5.1:
Example sentences from SemEval-2007, showing target words to be disambiguated (highlighted in blue) and their lemmatized forms.

In our experiments, we disambiguate nouns only (as did Ponzetto and Navigli), since both SGN and WN++ relate only concepts denoted by nouns, and no other parts of speech. In our experimental setup, each sentence is considered in isolation from the rest, and all lemmatized content words in a sentence are provided to the disambiguation algorithm; the verbs, adjectives, and adverbs, although we do not resolve their senses, lend additional context to the disambiguation algorithms.

The coarse-grained nature of the SemEval-2007 task provides that there may be more than one acceptable sense assignment for many of the targets. In the coarse-grained setting, an algorithm's sense assignment is considered correct when it appears in the list of acceptable senses for the given target word. These lists of acceptable senses are provided with the dataset.

Both of the algorithms below allow for multiple disambiguation results to be returned in the event of a tie. In these cases (although they are rare), we adopt the approach of Banerjee and Pedersen (2003), who award partial credit and discredit proportionally for all the senses returned by the algorithm.

5.3 WSD with Extended Gloss Overlaps (ExtLesk)

The first disambiguation algorithm we employ is the extended gloss overlaps measure (henceforth ExtLesk) of Banerjee and Pedersen (2003), which is an extension of the Lesk (1986) gloss overlaps measure. The algorithm takes a target (our target noun to be disambiguated) and its surrounding context (in our case, all other lemmatized targets in the sentence under consideration), and proceeds as follows:

For each sense s_i of the target noun n, we find all word senses related to s_i in WordNet via some specific relation, R_x. We then concatenate the glosses of these noun senses into a single string. Let us denote the concatenation of the glosses of all noun senses related to s_i by the relation R_x as $gloss_{R_x}(s_i)$. Then, for each sense s_j of each word in our surrounding context, we take all the word senses related to s_j in WordNet via a particular relation R_y (which may or may not be the same relation used above), and concatenate the glosses of those word senses into a

string that is, following our notation above, denoted $gloss_{R_y}(s_j)$. (For example, see Figure 5.2 below, which shows all hyponyms of *celestial_body#1* and $gloss_{\text{HYPO}}$(*celestial_body#1*) (the concatenation of their glosses).

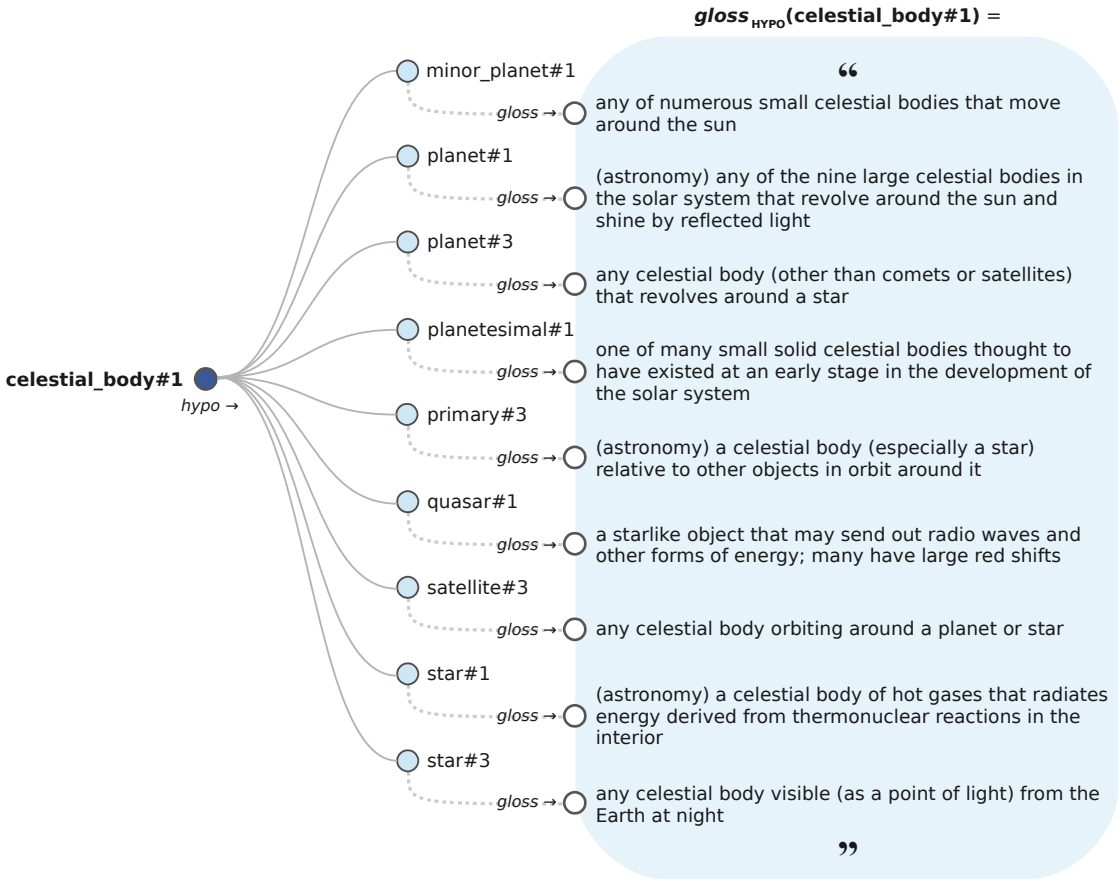


Figure 5.2:
All hyponyms of *celestial_body#1* in WordNet and their concatenated glosses, $gloss_{\text{HYPO}}$(*celestial_body#1*).

We then count how many content words (nouns, verbs, adjectives, and adverbs) are common to both $gloss_{R_x}(s_i)$ and $gloss_{R_y}(s_j)$. More formally, we define a function *overlap*(*a*, *b*)

that tells us how many content words two strings, *a* and *b*, have in common.[25] We perform stemming and part-of-speech tagging in this function, so that, for example, an occurrence of "wheel" in one gloss will match the occurrence of "wheels" in another, provided they both have the same part-of-speech tag (for an example, see Figure 5.3).[26] If a content word is repeated in both strings, multiple points are awarded accordingly.[27]

---

planet#1: (astronomy) any of the nine large celestial bodies in the solar system that revolve around the sun and shine by reflected light; Mercury, Venus, Earth, Mars, ...

star#1: (astronomy) a celestial body of hot gases that radiates energy derived from thermonuclear reactions in the interior

$$\textit{overlap}(\textit{gloss}_{\text{GLOS}}(\textit{planet\#1}), \textit{gloss}_{\text{GLOS}}(\textit{star\#1})) = 3$$

planet#1: (astronomy) any of the nine large celestial bodies in the solar system that revolve around the sun and shine by reflected light; Mercury, Venus, Earth, Mars, ...

star#2: someone who is dazzlingly skilled in any field

$$\textit{overlap}(\textit{gloss}_{\text{GLOS}}(\textit{planet\#1}), \textit{gloss}_{\text{GLOS}}(\textit{star\#2})) = 0$$

planet#1: (astronomy) any of the nine large celestial bodies in the solar system that revolve around the sun and shine by reflected light; Mercury, Venus, Earth, Mars, ...

star#3: any celestial body visible (as a point of light) from the Earth at night

$$\textit{overlap}(\textit{gloss}_{\text{GLOS}}(\textit{planet\#1}), \textit{gloss}_{\text{GLOS}}(\textit{star\#3})) = 4$$

Figure 5.3:
The overlap function counts content words common to two strings.

25 This is a slight departure from the traditional ExtLesk implementation, which awards more points for multi-word string overlaps. We have found that our approach offers substantial savings in running time while having only negligible effects on our overall results. In a subset of experimental runs of ExtLesk in which we used the traditional scoring mechanism, we found that F_1 values varied on average by a mere 0.32% (absolute change) (0.42% relative change).

26 For clarity, Figures 5.2 and 5.3 do not show the stemmed, part-of-speech tagged text of these glosses.

27 Notice that in Figure 5.3 we have overloaded the *gloss* function so that, e.g., $\textit{gloss}_{\text{GLOS}}(\textit{star\#1})$ simply returns the gloss of *star#1*. (Contrast this with the behavior of $\textit{gloss}_{\text{HYPO}}(\textit{celestial\_body\#1})$ in Figure 5.2.) That is to say, the *glos* relation in WordNet returns the gloss of a synset, which we use for direct comparison.

Finally, we say that the *score* of sense s_i of our target noun, denoted $score(s_i)$, is the sum of these values for every possible s_j from the surrounding context, and every possible relation R_x and R_y available to us:

$$score(s_i) = \sum_{R_x \in R} \sum_{R_y \in R} \sum_{s_j \in S} overlap(gloss_{R_x}(s_i), gloss_{R_x}(s_j)) \quad (34)$$

In (34), S is the context of s_i (all senses of all surrounding content words in the sentence), and R is our set of relations. In our implementation of ExtLesk, we use a standard, comprehensive set of relations from WordNet, R = {*hype*, *hypo*, *holo*, *mero*, *attr*, *also*, *sim*, *enta*, *caus*, *pert*, *glos*, *example*, *syns*},[28] corresponding to the following relations from WordNet, respectively: *hypernymy*, *hyponymy*, *holonymy*, *meronymy*, *attributes* (for nouns and adjectives), *also see* (denoted within synset glosses), *similar to* (also taken from synset glosses), *entailment* (for verbs), *cause to* (also for verbs), *pertainymy* (adjectives and adverbs), gloss (synset glosses, without *also see* and *similar to* or *example* annotations), *example* (examples taken directly from synset glosses), and other words represented in the concept's synset. With SGN and WN++, ExtLesk uses the single relation expressed by the networks: *RelatedTo*.

The sense of our target noun with the highest score from this function is used for sense assignment. In the event of a tie, multiple senses may be returned. ExtLesk does not attempt to perform sense assignment if the score for every sense of a target noun is zero, except when dealing with a monosemous noun, in which case we default to the only sense possible.

28 $gloss_{\text{GLOS}}(s_i)$ simply yields the gloss of s_i, since WordNet's *glos* relation returns a string, not a synset which we can gloss further. $gloss_{\text{EXAMPLE}}(s_i)$ behaves similarly, and $gloss_{\text{SYNS}}(s_i)$ returns a concatenated string of part-of-speech tagged nouns that constitute the synset of s_i, rather than repeatedly concatenating the gloss of s_i.

5.3.1 Results

In our experimental setup, we use ExtLesk to disambiguate the nouns in the SemEval-2007 dataset with five combinations of semantic resources: WordNet only, SGN only, SGN and WordNet combined (that is, the union of all links contained in both networks), WN++ only, and WN++ combined with WordNet. In our results (see Table 5.1), we include the traditional baselines of most frequent sense (MFS) assignment and random sense assignment for comparison, and measure precision (number of correct sense assignments divided by the number of attempted sense assignments), recall (number of correct sense assignments divided by the number of target nouns to be disambiguated), and the harmonic mean of the two, F_1, defined as:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{35}$$

Table 5.1:
ExtLesk disambiguation results on the SemEval-2007 all-words coarse-grained WSD task (nouns only).

Resource	**Precision** (%)	**Recall** (%)	$\boldsymbol{F_1}$ (%)
WordNet	78.80	74.82	76.76
SGN	78.64	72.82	75.62
SGN and WordNet	**82.35**	**78.11**	**80.18**
WN++	74.67	61.87	67.67
WN++ and WordNet	77.35	73.38	75.31
Baseline: Most Frequent Sense	77.40	77.40	77.40
Baseline: Random	63.50	63.50	63.50

On this task, our results with SGN as a stand-alone network (F_1 = 75.62%) rival the performance of WordNet (F_1 = 76.76%).[29] This result is particularly impressive given the fact that the relationships in SGN are derived automatically from a context-sparse lexical co-occurrence measure.

Equally impressive is the ability of SGN and WordNet, when used in combination, to achieve results (F_1 = 80.18%) that exceed what either network is able to accomplish as a stand-alone knowledge source. When combined, we see improvements of 3.42% and 4.56% (absolute F_1 values) over WordNet and SGN as stand-alone resources, respectively. It is also only with these resources combined that we are able to outperform the redoubtable MFS baseline of F_1 = 77.40%, and we do so by 2.78%.[30]

In contrast, WN++ (F_1 = 67.67%) fails to perform as a stand-alone resource, falling behind the MFS baseline by 9.73%. Of all the resources tested, WN++ yields the lowest results. When combined with WordNet, WN++ actually diminishes (rather than bolstering) the ability of

29 Ponzetto and Navigli (2010) report results of F_1 = 68.3% and 72.0% using WordNet and WN++, respectively, as stand-alone knowledge sources for ExtLesk. In contrast, our experimentally derived values for those resources are F_1 = 76.76% and 67.67%. In light of this disparity, we verified our results (as they pertain to WordNet as a stand-alone resource) using the WordNet::Similarity Perl module (Pedersen, Patwardhan, and Michelizzi, 2004). The Perl module, which implements ExtLesk, produced results with P = 78.27%, R = 72.90%, and F_1 = 75.49% on this task. Enabling and disabling stemming had a negligible impact on results, as did running the experiments with and without an extensive list of stop words. The WordNet::Similarity results vary slightly from those we obtained using our own implementation of ExtLesk with WordNet (P = 78.80%, R = 74.82%, F_1 = 76.76%), but this difference can be explained by differences in parsing and stemming algorithms, as well as our use of the traditional overlap counting approach of Lesk (1986). Furthermore, working backward from the results reported by Ponzetto and Navigli for ExtLesk with WordNet reveals that their implementation only produced disambiguation results for 764 out of the 1108 nouns to be disambiguated, and provided no disambiguation results for the remaining 31% of target nouns in the task. In contrast, our experiments with WordNet::Similarity produced results for 1032 of the 1108 (some correct, some incorrect, of course). Intuitively, the 31% figure seems excessively high, because ExtLesk only fails to produce a disambiguation result for some noun s_i if there are no content words in common between its extended glosses (i.e., the glosses of all concepts related to s_i through every possible edge type in WordNet) and the extended glosses of *any* of the content words co-occurring in the sentence where s_i appears.

30 Other systems have obtained better results on the same dataset, but we focus only on SGN and WN++ because our aim is to compare the resources themselves.

WordNet to perform on this WSD task by 1.45%. We defer our discussion of factors impacting the performance of WN++ to Section 5.5.

5.4 WSD with Degree Centrality

The second disambiguation algorithm we use in our experiments, Degree Centrality, is a graph-based measure of semantic relatedness (Navigli & Lapata, 2010). The algorithm searches through a semantic network (using all possible relations as edges) for paths of length $l \leq$ *maxLength* between all sense nodes of all lemmas in our context. The edges along all such paths are added to a new graph, G', and for each target noun to be disambiguated, the sense node with the greatest number of incident edges (highest vertex degree) in G' is taken as its intended sense. In these graphs, nodes represent synsets, as opposed to instantiating separate nodes for different members of the same synset and allowing edges to be constructed between them. We include all lemmas from a sentence in our context, but only return disambiguation results for the nouns.

With SGN and WN++, the implementation of this algorithm is straightforward. We initiate a breadth-first search (BFS)[31] at each target sense node in the network, and proceed through $\lfloor (maxLength + 1)/2 \rfloor$ iterations of spreading activation. Whenever the tendrils of this spreading activation from one target sense node in the graph connect to those of another,[32] we add the path between the nodes to our new graph, G', potentially incrementing the degree of the involved target sense nodes in G' as we do so.

31 This is in contrast to the DFS implementation of Navigli and Lapata (2010).

32 When maxLength is odd, this requires an additional check to ensure that the intersection is not taking place at a node that is exactly $\lfloor (maxLength + 1)/2 \rfloor$ degrees removed from each of the two target nodes it is connecting, as this would result in a path with overall length ($maxLength$ + 1) between the target nodes.

BFS, as an admissible algorithm, is guaranteed to find the shortest path from an initial state to a goal (e.g., from one target sense node in our graph to another). Therefore, because any node on a path of length $l \leq maxLength$ between two target nodes is at most $\lfloor l/2 \rfloor$ nodes removed from at least one of those target sense nodes, we only need to perform a BFS of depth $\lfloor (maxLength + 1)/2 \rfloor$ from every target sense node in order to guarantee that every such path between them will be discovered. Since the time complexity of BFS is exponential with respect to the depth of the search, cutting this depth in half (in comparison to performing a BFS of depth *maxLength*) greatly reduces the running time of our algorithm.

We take the same approach in traversing the WordNet noun graph, using all possible sense relations as edges. There is, however, one complication:

In keeping with the approach of Navigli and Lapata (2010), an edge is also induced between synsets if the gloss of one synset contains a monosemous content word. For example, the gloss for *leprechaun#1*, "a mischievous elf in Irish folklore," contains the monosemous noun "folklore;" thus, we have an edge between *leprechaun#1* and *folklore#1* in the WordNet graph.

Unlike the other edges in these semantic graphs, this gloss relation cannot be discovered bidirectionally, even though the edge, once we encounter it and add it to our graph representation of WordNet, is considered undirected. Gloss edges can therefore spontaneously introduce a short path between two nodes if they are encountered along much longer paths, deep within a BFS.

Thus, when traversing the WordNet graph, we perform a preliminary BFS of depth *maxLength* in an expedition to discover these gloss relations. This still does not guarantee that all possible paths of length *maxLength* between two target sense nodes will be discovered. Some node lying along a path of length (*maxLength* + 1) from a target synset node could easily have

directed links (via its gloss) to two other target synset nodes, providing a hidden path of length two that cannot be discovered without traversing all paths of length (*maxLength* + 1) from our target nodes. However, our approach reduces our chances of missing a short, gloss-induced bridge in the graph.

Once we have our new graph, G', constructed in this manner, the vertex degree is considered an indication of the semantic relatedness of a particular synset to all other lemmas in our context. For each target noun, we use the sense node(s) with the highest degree in G' for sense assignment.

5.4.1 Results

In our experimental setup, we examine the performance of the Degree Centrality algorithm with the following combinations of semantic resources: WordNet, SGN, WN++, Refined WN++, SGN and WordNet combined, and Refined WN++ and WordNet combined. Refined WN++ consists of 79,422 of WN++'s strongest relations, and was created in an unsupervised setting by Ponzetto and Navigli specifically for use with Degree Centrality when they discovered that WN++ had too many weak relationships to perform well with the Degree Centrality algorithm.

We have observed that the performance of Degree Centrality rapidly levels off as *maxLength* increases. Navigli and Lapata (2010) also reported this so-called "plateau" effect, and employ a *maxLength* of 6 in their experiments, despite finding that results leveled off around *maxLength* = 4. We, too, find that performance levels off around *maxLength* = 4 in almost all cases, and so only continue up to *maxLength* = 5.

Table 5.2:
Degree Centrality disambiguation results on the SemEval-2007 all-words coarse-grained WSD task (nouns only) with maximum path lengths $1 \leq L_{max} \leq 5$.

Resource	L_{max}	P (%)	R (%)	F_1 (%)	Resource	L_{max}	P (%)	R (%)	F_1 (%)
WordNet	1	96.9	16.8	28.6	WN++	1	87.2	23.5	37.1
(*stand-alone*)	2	77.6	45.1	57.0	(*stand-alone*)	2	71.6	60.2	65.4
	3	76.7	65.6	70.7		3	70.7	64.3	67.3
	4	769	71.0	73.9		4	70.4	64.5	67.3
	5	76.6	71.6	74.0		5	70.4	64.5	67.3
SGN	1	79.7	32.9	46.6	Refined WN++	1	**98.3**	15.3	26.5
(*stand-alone*)	2	72.0	64.6	68.4	(*stand-alone*)	2	91.4	23.4	37.3
	3	68.7	63.5	66.0		3	88.7	29.9	44.7
	4	68.0	63.9	65.9		4	83.7	32.3	46.7
	5	68.0	64.2	66.1		5	80.2	35.3	49.0
SGN	1	77.4	52.4	62.5	Refined WN++	1	83.3	31.2	45.4
(*with WordNet*)	2	74.7	70.7	72.7	(*with WordNet*)	2	77.5	66.6	71.6
	3	70.3	67.1	68.7		3	77.6	**73.6**	75.5
	4	70.5	67.4	68.9		4	74.7	71.4	73.0
	5	70.1	67.0	68.5		5	74.7	71.4	73.0
MFS Baseline	--	77.4	77.4	**77.4**	Rand. Baseline	--	63.5	63.5	63.5

We find that, in all cases tested, Degree Centrality is unable to outperform the MFS baseline (with respect to F_1) (see Table 5.2). SGN and WN++ exhibit comparable performance with this algorithm, with maximum F_1 values of 68.4% (at *maxLength* = 2) and 67.3% (at *maxLength* = 3 to 5), respectively. Neither achieves the performance of WordNet with Degree

Centrality (F_1 = 74.0%), which under-performs the MFS baseline (F_1 = 77.4%) by 3.4%.[33] Ponzetto and Navigli (2010) reported that only performing sense assignment when the maximum degree exceeded an empirically derived but non-disclosed threshold improved performance, but we have found that implementing such a threshold universally lowers results for all resources we tested with Degree Centrality.

The lowest performance using Degree Centrality comes from Refined WN++ as a stand-alone resource. We attribute this to the fact that Refined WN++ is so semantically sparse. On average, noun senses in Refined WN++ are related to only 3.42 other noun senses, while those in WN++ and SGN relate to an average of 44.59 and 10.92 noun senses, respectively. Accordingly, the success of Refined WN++ and WordNet, when combined, is attributable mostly to the success of WordNet as a stand-alone resource; as *maxLength* increases, the contributions made by the sparse Refined WN++ network rapidly become negligible in comparison to those provided by the WordNet ontology.

5.5 Discussion

The fact that the performance of Degree Centrality quickly plateaus hints at the root cause of its weak performance compared to ExtLesk and the MFS baseline. As the maximum path length is increased in a dense semantic network, all possible edges from our target sense nodes rapidly find themselves involved with paths to other target sense nodes. This is particularly true of WN++ (notice its rapid and stable convergence), where certain "sticky" nodes form

33 Although Ponzetto and Navigli (2010) reported similar results with WordNet (F_1 = 74.5%), we have been unable to reproduce their results using Refined WN++, either combined with WordNet (F_1 = 79.4% vs. 75.5%) or as a stand-alone resource (F_1 = 57.4% vs. 49.0%).

bridges between seemingly unrelated concepts. For example, the frequent appearance of "United States" in Wikipedia articles, and its tendency to be linked to the *United States* Wikipage when it occurs, causes the term to serve as a bridge between such diverse concepts as *automaton#2* and *burrito#1*, which one would typically expect to be far removed from one another in a model of semantic relatedness (and which also bear questionable relatedness to *United_States#1*).

If it is indeed true that Degree Centrality's plateau effect is a result of each target sense node's edges rapidly finding themselves participant to paths to other sense nodes, then one would expect the algorithm to perform comparably to performing sense assignment based on the most semantically well connected sense of each target noun in the network. That is, as path length increases, the results of Degree Centrality should converge to the results obtained by foregoing the algorithm altogether and simply disambiguating each noun to the sense with the most edges in the network (regardless of whether those edges ultimately connect two word senses from the disambiguation context). This is, in fact, the case: the expected values of convergence attained by defaulting to the semantically most well-connected sense of each target noun in each network are F_1 = 66.3%, 67.5%, and 74.6% for SGN, WN++, and WordNet, respectively, as compared to the experimentally derived Degree Centrality results of F_1 = 66.1%, 67.3%, and 74.0%.

5.6 Summary

We have evaluated our semantic network, SGN, in a coarse-grained WSD experiment setting (SemEval-2007) using two graph-based algorithms: ExtLesk and Degree Centrality. We found that our network performs comparably to WordNet using ExtLesk (F_1 = 75.62% and

76.76%, respectively), and that combining SGN and WordNet for use with ExtLesk yields results that exceed the performance that either resource is able to attain individually (F_1 = 80.18%).

With both ExtLesk and Degree Centrality, we observed that the performance of WN++ falls short of that of WordNet, and that combining the two resources negatively impacts the performance of WordNet. With Degree Centrality in particular, we discovered that the spurious relationships in WN++ hamper its performance, and that the smaller version of the network, Refined WN++, is too semantically sparse to perform well as a stand-alone knowledge source or to significantly impact the performance of WordNet when the two resources are combined.

With regard to Degree Centrality, we observed that the algorithm has a strong disambiguation bias toward the sense of a word with the most incident edges in a semantic network, and that this bias accounts for the rapid convergence of its performance (i.e., plateau effect) as the algorithm's maximum path length is increased.

CHAPTER 6: MEASURING HUMAN PERCEPTIONS OF RELATEDNESS

In this chapter, we present the results of our investigation into human perceptions of semantic relatedness. We have elicited quantitative human judgments of relatedness for 122 noun pairs. The mean relatedness scores have been compiled into a new dataset that can be used to supplement existing evaluative standards for computational measures of semantic relatedness.

In Section 6.1, we provide some background and motivation for this study: we discuss related work on gold standards for evaluating relatedness measures, address some shortcomings of those standards, and explain the need for datasets like the one we present here. In Section 6.2, we lay out our experimental procedure and explain how we chose the noun pairs in our dataset. In Section 6.3, we provide analysis and discussion of our experimental results: the mean relatedness scores elicited from human participants are presented, and we evaluate the performance of a variety of similarity and relatedness measures on the new dataset. We then summarize the contributions of this study in Section 6.4.

6.1 Mean Similarity Scores as Gold Standard Datasets

With 65 and 30 noun pairs respectively, the Rubenstein and Goodenough (1965) (R&G) and Miller and Charles (1991) (M&C) datasets (discussed above in Section 2.2.5), are widely considered to be too small to provide adequate evaluation of similarity measures (Banerjee & Pedersen, 2003; Budanitsky & Hirst, 2006; Milne & Witten, 2008a; Strube and Ponzetto, 2006). Nonetheless, we have observed their ubiquitous use in the literature for evaluating not only similarity measures, but also relatedness measures. Only minor credit for their continued use as a

gold standard can be attributed to the fact that they provide a common point of comparison to previous work in the field. Resnik (1999) observed that "the worth of a similarity measure is in its fidelity to human behavior, as measured by predictions of human performance on experimental tasks" (p. 95). Budanitsky and Hirst similarly remarked that "comparison with human judgments is the ideal way to evaluate a measure of similarity or relatedness" (p. 32). Thus, comparison to the kind of data provided by R&G and M&C enjoys a certain gold standard primacy in the literature, and we continue to employ these two particular datasets because no other datasets have yet emerged as reasonable candidates to replace them.

This is particularly problematic in the evaluation of relatedness measures, where perhaps the most obvious concern about the use of R&G and M&C as gold standards is that subjects were asked to evaluate "similarity of meaning" (Rubenstein & Goodenough, 1965, p. 628) in those studies—not semantic relatedness.

6.1.1 WordSim353 as a Gold Standard

WordSim353[34] (Finkelstein et al., 2002) has recently emerged as a potential surrogate dataset for evaluating relatedness measures. Several studies have reported correlation to the WordSim353 data as part of their standard evaluation procedures, with some studies explicitly referring to it as a collection of human-assigned relatedness scores (Gabrilovich & Markovitch, 2007; Hughes & Ramage, 2007; Milne & Witten, 2008a). Finkelstein et al. reported that, in creating the dataset, they "employed 16 subjects to estimate the 'relatedness' of the word pairs on a scale from 0 (totally unrelated words) to 10 (very much related or identical words)" (p. 13).

34 http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html

However, the status of WordSim353 as a relatedness gold standard remains unclear because the instructions given to participants in its creation emphasized similarity, not relatedness (see Figure 6.1 below). The instructions opened with an explanation that the study was "aimed at estimating the *similarity* [emphasis added] of various words in the English language," and that participants would "assign *similarity* [emphasis added] scores to pairs of words, so that machine learning algorithms [could] be subsequently trained and adjusted using human-assigned scores." The full instructions that Finkelstein et al. provided to participants in their study repeatedly framed the task as one in which participants were expected to assign word similarity scores, and only twice mentioned relatedness. Furthermore, the Web page for downloading the WordSim353 collection frequently refers to it as a set of "similarity scores" (Gabrilovich, 2006), and the name of the dataset itself stands for "Word Similarity."

Jarmasz and Szpakowicz (2003) have notably raised methodological concerns about the acquisition of WordSim353 data, observing that: a) relatedness is rated on a scale of 0 to 10, which is intrinsically more difficult for humans to manage than the scale of 0 to 4 used by R&G and M&C; b) a certain amount of cultural bias is introduced into the data, particularly with respect to the inclusion of proper nouns (e.g., the evaluation of the pair *Arafat–terror*); and c) there is no indication of how the 353 word pairs were chosen, other than the fact that the 30 M&C pairs were included as a subset. We add to these concerns the fact that the instructions obfuscate whether subjects were expected to evaluate relatedness in the general case, or simply to extend their definition of similarity to encompass antonymy (thus using the term "relatedness" to denote two particular types of relatedness: similarity and antonymy).

Estimation of word similarity

Hello,

We kindly ask you to assist us in a psycholinguistic experiment, aimed at estimating the similarity of various words in the English language. The purpose of this experiment is to assign similarity scores to pairs of words, so that machine learning algorithms can be subsequently trained and adjusted using human-assigned scores.

Below is a list of pairs of words. For each pair, please assign a numerical similarity score between 0 and 10 (0 = words are totally unrelated, 10 = words are VERY closely related). By definition, the similarity of the word to itself should be 10. You may assign fractional scores (for example, 7.5).

Specific instructions:

1) The questionnaire starts on the next page.

2) Please fill in your full name at the beginning of the questionnaire. We need the names to ensure individual estimations do not get mixed, and to be able to contact you should any clarifications become necessary.

3) Please fill in the similarity scores in the appropriate column of the table. To facilitate processing your questionnaire, please do not print the document but rather type in the values in the table provided.

4) If you do not know the meaning of a particular word - please use a dictionary, or ask a native English speaker.

5) Please DO NOT consult your friends on assigning the similarity scores - it is highly important that the scores you assign be independent of someone else's assessment.

6) When estimating similarity of antonyms, consider them "similar" (i.e., belonging to the same domain or representing features of the same concept), rather than "dissimilar".

If you have any questions or require further clarifications (or if you have a suggestion), please do not hesitate to contact us.

Thank you for your assistance!

Figure 6.1:
Instructions for assigning scores for the WordSim353 word pairs of Finkelstein et al. (2002). The task is framed as being intended to elicit similarity scores.

Agirre, Alfonseca, et al. (2009) recently attempted to disentangle the pairs of similar nouns in WordSim353 from those that were related but not similar, but did not assess the validity of the scoring distribution in the resulting relatedness subset to ensure that strongly related word pairs were not penalized by human subjects for being dissimilar. Perhaps not surprisingly, the highest scores in WordSim353 (all ten ratings between 9.0 and 10.0, inclusively) were assigned to pairs that Agirre, Alfonseca, et al. placed in their similarity subset. Agirre, Alfonseca, et al. showed that similarity and relatedness measures alike correlated better to the subset of similar entities than they did to the subset of related entities from WordSim353.

6.1.2 The R&G Methodology and the Reliability of Human Judgments

In contrast to the methodology of Finkelstein et al., the instructions of Rubenstein and Goodenough were straightforward in their intent (see Figure 6.2 below). These instructions were also used in replications of the study by Miller and Charles (1991) and Resnik (1995).

Several studies have shown that human judgments of similarity are consistent within subjects, between subjects, and across groups of subjects using these instructions. Rubenstein and Goodenough established intra-judge reliability by having one group of subjects perform similarity evaluations twice—once using a set of 48 noun pairs, and again two weeks later using the full set of 65 R&G noun pairs. The two sets had 36 noun pairs in common. Rubenstein and Goodenough measured how well each subject's judgments correlated on those 36 pairs between the two experimental trials. Among 15 judges, the average intra-judge correlation was $r = 0.85$.

There were 65 pairs of nouns (*theme pairs*) presented for judgment. Each subject was given a shuffled deck of 65 slips of paper, each slip containing a different theme pair. The subject was given the following instructions:

1. After looking through the whole deck, order the pairs according to amount of "similarity of meaning" so that the slip containing the pair exhibiting the greatest amount of "similarity of meaning" is at the top of the deck and the pair exhibiting the least amount is on bottom.

2. Assign a value from 4.0–0.0 to each pair—the greater the "similarity of meaning," the higher the number. You may assign the same value to more than one pair.

Figure 6.2:
Procedure used by Rubenstein and Goodenough (1965) to elicit similarity scores for their 65 word pairs.

Rubenstein and Goodenough also had two separate groups of judges evaluate the full set of 65 noun pairs, and found the two resulting sets of mean similarity scores had very high correlation ($r = 0.99$). The Miller and Charles replication of Rubenstein and Goodenough's study using 30 noun pairs from the R&G data also had mean similarity scores that correlated strongly to the results of R&G ($r = 0.97$).

Resnik (1995) replicated Miller and Charles' study and found strong correlation ($r = 0.96$) to the M&C means. Resnik also assessed inter-judge reliability, and found that the average of individual judges' correlations to the M&C data was $r = 0.885$ ($\sigma = 0.08$). Within the data from his own replication, using leave-one-out sampling, Resnik found that individual judges' scores correlated to mean similarity scores with $r = 0.903$ ($\sigma = 0.07$). Finkelstein et al.

(2002) included the 30 M&C noun pairs in their study, which had a total of 353 word pairs, and found their results correlated to M&C with $r = 0.95$.

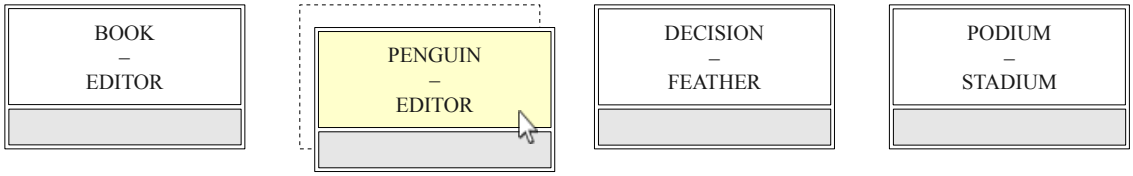
These strong correlations hold even with wide variation in the number of subjects participating in each study. Rubenstein and Goodenough used 15 and 36 subjects in the two groups described above. Miller and Charles employed the help of 38 subjects. Resnik performed his replication using only 10 subjects, and each pair of words in Finkelstein et al.'s study was evaluated by 13 to 16 subjects.

6.2 Methodology

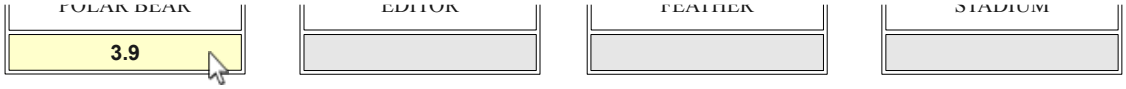
In our experiments, we elicited human ratings of semantic relatedness for 122 noun pairs. In doing so, we followed the methodology of R&G (Figure 6.2 above) as closely as possible: participants were instructed to read through a set of noun pairs, sort them by how strongly related they were, and then assign each pair a relatedness score on a scale of 0.0 (completely unrelated) to 4.0 (very strongly related). We made two notable modifications to the experimental procedure of R&G. First, instead of asking participants to judge "amount of 'similarity of meaning,'" we asked them to judge "how closely related in meaning" each pair of nouns was. Second, we used a Web interface to collect data in our study; instead of reordering a deck of cards, participants were presented with a grid of cards that they were able to rearrange interactively with the use of a mouse or any touch-enabled device, such as a mobile phone or tablet PC.

Figure 6.3 below shows the instructions as they were presented to participants in our study, including excerpted screen captures that show how elements of the interface (e.g., the "cards" containing each noun pair) were presented to and manipulated by users.

STEP 1. On the following page, you will be presented with a grid of cards, each containing a pair of words. After looking through the whole grid of cards, order the pairs according to how closely related in meaning each pair of words is, so that the card containing the most closely related pair is at the start of the grid (top-left) and the pair that is least closely related is at the end (bottom-right). To move a card, simply click and drag the part of the card containing the word pair, as shown below:



STEP 2. When you have finished rearranging the cards, assign a value from 0.0 (completely unrelated) to 4.0 (very strongly related) to each pair. The more closely related in meaning the words are, the higher the number. You may assign the same value to more than one pair. To assign a value, click the gray field at the bottom of a card and type a number. You can use the TAB key to quickly move to the next field, or SHIFT-TAB to return to the previous field.



Examples: For example, *cup* and *coffee* are strongly related, and would be given a high score. *Umbrella* and *rain* are also strongly related. In contrast, *printer* and *hippopotamus* are strongly *unrelated*, and would be given a very low score, as would *soil* and *telephone*.

We want to gauge your gut reaction to how closely related each pair of nouns is. Therefore, we ask that you complete this task in one sitting, within **20 minutes**, and **do not consult a dictionary** or ask others for help.

Figure 6.3:
Instructions presented to participants in our study.

In early usability testing of our Web interface, we observed that large datasets (e.g., 40 to 65 noun pairs) made the sorting task too difficult for users to manage. This was a limitation not of the interface itself, but of the time and attention required to reorder so many pairs of nouns. With large datasets, users were overwhelmed by the need to make so many fine-grained distinctions in their relatedness judgments. For this reason, we chose to present each user with only 32 noun pairs for evaluation. We have already seen that sets of noun pairs can be split into smaller subsets for evaluation by different groups of participants in order to keep the task to a manageable size without significantly impacting subjects' score distributions; when Miller and Charles replicated Rubenstein and Goodenough's study with a subset of only 30 of the 65 original noun pairs, the resulting means from the two experiments exhibited strong correlation ($r = 0.97$).

6.2.1 Experimental Conditions

Each participant in our study was randomly assigned to one of four conditions. Each condition contained 32 noun pairs for evaluation. Of those pairs, 10 were randomly selected from our network (SGN), 10 from WN++, and 10 were generated by randomly pairing words from a list of all nouns occurring in Wikipedia. All pairs were matched for frequency using the 100 nearest neighbors for each noun, as sorted by frequency of occurrence in Wikipedia. We manually selected two additional pairs that appeared across all four conditions: *leaves–rake* and *lion–cage*. These control pairs were included to ensure that each condition contained examples of strong semantic relatedness, and potentially to help identify and eliminate data from participants who assigned random relatedness scores. Within each condition, the 32 word pairs were

presented to all subjects in the same random order. Across conditions, the two control pairs were always presented in the same positions in the word pair grid.

Each word pair was subjected to additional scrutiny before being included in our dataset. We eliminated any pairs falling into one or more of the following categories: (1) pairs containing proper nouns (e.g., *grape–Europe* and *Italian–kilobyte*[35]); (2) pairs in which one or both nouns might easily be mistaken for adjectives or verbs (e.g., *cement–pact* and *second–knight*, where "cement" might be taken as a verb, or "second" as an adjective or verb); (3) pairs with advanced vocabulary or words that might require domain-specific knowledge in order to be properly evaluated (e.g., *soprano–falsetto*, which might be difficult for anyone without a musical background to evaluate, and *baronet–privy council*, which might require basic knowledge of British hereditary titles and monarchic government); and (4) pairs with shared stems or common head nouns (e.g., *first cousin–second cousin* and *sinner–sinning*). The latter were eliminated to prevent subjects from latching onto superficial lexical commonalities as indicators of strong semantic relatedness without reflecting upon meaning.

6.2.2 Participants

Participants in our study were recruited from introductory undergraduate courses in psychology and computer science at the University of Central Florida. Students from the psychology courses participated for course credit and accounted for 89% of respondents.

35 These pairs in particular were drawn randomly from WN++. Pairs with proper nouns account for at least 20% of relationships in the WN++ network—a lower bound that accounts only for proper nouns categorized using the *InstanceOf* relation in WordNet. The actual occurrence in WN++ of pairs with proper nouns is necessarily higher.

A total of 92 participants provided data for our study. Of these, we identified 19 as outliers, and their data were excluded from the results reported below to prevent interference from individuals who appeared to be assigning random scores to noun pairs. Here we consider an outlier to be any individual whose numeric relatedness ratings fell outside two standard deviations from the means for more than 10% of the word pairs they evaluated (i.e., for at least four word pairs, since each condition contained 32 word pairs). For outlier detection, means and standard deviations were computed using leave-one-out sampling. That is, data from individual *J* were not incorporated into means or standard deviations when considering whether to eliminate *J* as an outlier. We used this sampling method to prevent extreme outliers from masking their own aberration during outlier detection, which is potentially problematic when dealing with small populations. Without leave-one-out sampling (i.e., comparing to means established from the whole population), we would have identified fewer outliers (14 instead of 19), but the resulting means would still have correlated strongly to the dataset presented below ($r = 0.991$, $p < 0.01$).

Of the 73 participants remaining after outlier elimination, there was a near-even split between males (37) and females (35), with one individual declining to provide any demographic data. The average age of participants was 20.32 ($\sigma = 4.08$, N = 72). Most students were freshmen (49), followed in frequency by sophomores (16), seniors (4), and juniors (3). The most common majors represented were computer science, computer engineering, and information technology (12); psychology (11); and engineering (mechanical, electrical, aerospace, and industrial) (10). Participants earned an average score of 42% on a standardized test of advanced vocabulary ($\sigma = 16\%$, N = 72) (Test I – V-4 from Ekstrom, French, Harman, and Dermen, 1976).

6.3 Results

6.3.1 Mean Relatedness Scores

The mean relatedness scores (μ) and standard deviations (σ) for all 122 noun pairs in our study are reported below in Table 6.1. Initially, each word pair was evaluated by at least 20 individuals. After outlier removal (described above), each word pair retained evaluations from 14 to 22 individuals.

In Table 6.1, we also indicate the source of each randomly selected word pair, although occurrence of these pairs is not necessarily exclusive to any one network. For example, the pairs *apparel–jewellery* and *underwear–lingerie* were randomly selected from SGN and WN++ respectively, but appear in both networks.

Table 6.1:
Mean relatedness scores for the 122 noun pairs in our study.

#	Word Pair		μ	σ	Source
1	underwear	lingerie	3.94	0.14	WN++
2	digital camera	photographer	3.85	0.32	WN++
3	tuition	fee	3.85	0.24	SGN
4	leaves	rake	3.82	0.34	*control*
5	symptom	fever	3.79	0.33	SGN
6	fertility	ovary	3.78	0.23	WN++
7	beef	slaughterhouse	3.78	0.34	WN++
8	broadcast	commentator	3.75	0.32	SGN
9	apparel	jewellery	3.72	0.49	SGN
10	arrest	detention	3.69	0.28	SGN
11	hardware	pc	3.61	0.46	SGN
12	street	neighborhood	3.60	0.59	WN++
13	pixel	digital camera	3.57	0.60	WN++
14	vehicle	trailer	3.54	0.45	SGN
15	mathematics	method	3.47	0.48	SGN
16	draft	manuscript	3.46	0.92	SGN
17	flavor	pepper	3.45	0.68	SGN
18	defense	soldier	3.39	0.60	WN++
19	seller	profit	3.39	0.91	WN++
20	lion	cage	3.36	0.68	*control*
21	treasure	hunter	3.35	0.75	SGN
22	translation	meaning	3.33	0.84	WN++
23	bread	egg	3.24	0.53	WN++
24	garage	opener	3.21	0.66	SGN
25	prohibition	rum	3.17	1.07	WN++
26	fax	e-mail	3.17	0.46	WN++
27	captive	custody	3.15	0.63	SGN
28	solar system	sphere	3.13	0.54	WN++
29	vegetation	pastureland	3.13	1.00	SGN

#	Word Pair		μ	σ	Source
30	leather	pouch	3.12	0.47	SGN
31	terrace	pavilion	3.10	0.61	WN++
32	recycling	landfill	3.09	0.79	WN++
33	garden	art	3.07	0.58	WN++
34	recording studio	loudspeaker	3.01	0.45	WN++
35	strike	enemy	3.00	0.73	SGN
36	ethanol	benzene	3.00	0.83	SGN
37	pepper	corn	2.99	0.95	SGN
38	murder	gang	2.93	0.93	SGN
39	multiple	coefficient	2.92	0.90	WN++
40	infantry	reconnaissance	2.90	1.16	WN++
41	density	electron	2.90	0.75	SGN
42	representation	gender	2.86	0.74	WN++
43	palm	anatomy	2.79	0.80	*random*
44	poem	singer	2.78	0.79	*random*
45	maintenance	aviation	2.74	0.82	SGN
46	mushroom	herb	2.73	1.29	SGN
47	yeast	lager	2.71	1.08	SGN
48	headache	caffeine	2.71	0.97	SGN
49	truss	cantilever bridge	2.71	1.40	SGN
50	workplace	discrimination	2.67	1.09	SGN
51	hunter	squirrel	2.66	1.05	WN++
52	disorder	abuse	2.66	1.01	SGN
53	contract	legislation	2.65	0.58	WN++
54	motivation	morality	2.52	0.92	WN++
55	sewer	overflow	2.51	1.02	SGN
56	blindness	placebo	2.48	1.22	WN++
57	agility	fox hunting	2.46	0.92	*random*
58	proportion	stability	2.39	1.12	WN++
59	drug	public school	2.34	0.92	*random*
60	resentment	instigator	2.28	0.69	*random*

#	Word Pair		μ	σ	Source
61	pow	combatant	2.26	1.23	SGN
62	banana	salad	2.25	0.87	WN++
63	emphasis	newspaper	2.24	0.88	*random*
64	forestry	urban area	2.23	1.24	WN++
65	hypocrisy	condemnation	2.22	1.34	SGN
66	facility	activity	2.05	1.11	SGN
67	rendezvous	convoy	2.04	1.09	SGN
68	propagation	radio wave	2.01	1.13	SGN
69	puppetry	slapstick	2.00	1.35	WN++
70	public domain	brand	1.99	1.04	WN++
71	cartridge	lid	1.97	1.11	*random*
72	enclosure	mental health	1.92	1.15	*random*
73	credit	foundation	1.91	1.37	*random*
74	guardian	livestock	1.91	1.23	SGN
75	coronation	majority rule	1.89	1.10	*random*
76	enzyme	depression	1.88	1.23	*random*
77	precursor	prevention	1.86	1.17	*random*
78	arbitration	committee	1.84	1.38	SGN
79	lemon	coriander	1.75	1.13	SGN
80	stock	bull	1.66	1.34	*random*
81	vicar	archdiocese	1.66	1.29	SGN
82	scrap	chemical element	1.57	1.12	WN++
83	paranoia	newsroom	1.48	1.14	*random*
84	phase	consistency	1.47	1.17	*random*
85	hope	psychology	1.44	0.91	WN++
86	juvenile	rope	1.39	1.15	*random*
87	half-hour	weeknight	1.37	0.90	SGN
88	evolution	publicity	1.36	1.19	*random*
89	contestant	donor	1.31	1.15	*random*
90	sheet	window	1.21	1.13	WN++
91	robbery	mobile phone	1.21	1.00	WN++

#	Word Pair		μ	σ	Source
92	metre	semifinal	1.20	0.84	SGN
93	relief	total	1.18	0.97	*random*
94	public service	array	1.18	0.93	*random*
95	fresco	modern times	1.11	0.98	*random*
96	summit	canal	1.09	1.02	SGN
97	outlet	silk	1.09	1.04	*random*
98	greed	vest	1.08	0.98	*random*
99	eyeball	flatworm	1.05	1.10	WN++
100	spreadsheet	silk	0.99	0.95	WN++
101	distinction	sword	0.94	0.92	*random*
102	inclusion	career	0.83	0.82	*random*
103	feud	programmer	0.81	0.80	*random*
104	switch	glass	0.79	0.71	WN++
105	penalty	programming	0.78	0.77	*random*
106	duty	verb	0.77	0.71	*random*
107	catering	loan	0.77	1.01	*random*
108	musical group	confession	0.76	0.71	*random*
109	complication	harp	0.74	0.84	*random*
110	female	insect	0.74	0.69	WN++
111	seminar	fern	0.71	1.00	*random*
112	home run	surfer	0.71	0.94	*random*
113	mishap	cube root	0.70	0.83	*random*
114	thriller	sunlight	0.61	0.66	WN++
115	crusade	catwalk	0.59	0.74	*random*
116	jumper	furnishing	0.59	0.76	*random*
117	eclipse	cord	0.57	0.80	*random*
118	fork	combination	0.55	0.66	WN++
119	madness	nest	0.46	0.55	*random*
120	gas	algebra	0.41	0.51	WN++
121	hotel	bibliography	0.37	0.49	*random*
122	gladiator	plastic bag	0.13	0.32	*random*

6.3.2 Distribution of Standard Deviations

The standard deviations for relatedness scores ranged from 0.13 to 1.40 in our study, with strongly related and strongly unrelated pairs exhibiting the lowest variation (see Figure 6.4). These results are consistent with the findings of Rubenstein and Goodenough, who reported standard deviations ranging from 0.70 to 1.30 for word pairs with similarity means from 1.0 to 3.0. In our data, pairs with relatedness means from 1.0 to 3.0 had standard deviations ranging from 0.58 to 1.40. This indicates that human perceptions of relatedness vary widely for moderately and weakly related nouns, but does not reveal the source of variation—whether some individuals are simply more liberal or more conservative than others with their relatedness ratings, or if the relative ordering of pairs' relatedness also varies widely between individuals.

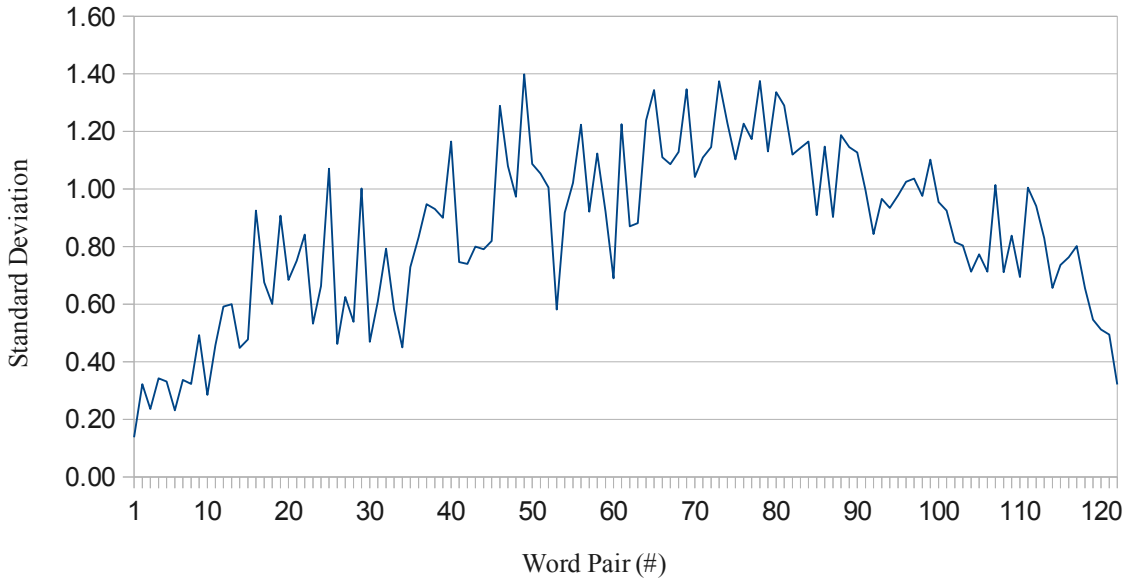


Figure 6.4:
Standard deviations of relatedness scores from our study range from 0.13 to 1.40 and are lowest for pairs that are strongly related or strongly unrelated.

6.3.3 Human Correlation to Relatedness Means

Within each of our four experimental conditions, we computed how strongly each participant's data correlated to the mean relatedness scores, again using leave-one-out sampling. The means of these correlations are presented below in Table 6.2. Individual correlations for all 73 participants were significant at $p < 0.01$.

We find that judgments from individual subjects in our study exhibit high average correlation to the elicited relatedness means ($r = 0.769$, $\sigma = 0.09$, N = 73). Resnik, in his replication of the M&C study, reported average individual correlation of $r = 0.90$ ($\sigma = 0.07$, N = 10) to similarity means elicited from a population of 10 graduate students and postdoctoral researchers. Presumably Resnik's subjects had advanced knowledge of what constitutes semantic similarity, as he established $r = 0.90$ as an upper bound for expected human correlation on that task. The fact that average human correlation in our study is weaker than in previous studies suggests that human perceptions of relatedness are less strictly constrained than perceptions of similarity, and that a reasonable computational measure of relatedness might only approach a correlation of $r = 0.769$ to relatedness norms.

Table 6.2:
Mean human correlation to relatedness norms from each condition.

Condition	r	σ	N
1	0.774	0.09	20
2	0.773	0.08	22
3	0.802	0.06	17
4	0.759	0.10	14
All	0.769	0.09	73

6.3.4 Correlation of Similarity and Relatedness Measures to Rel-122 Norms

In Table 6.3 we present the performance of a variety of relatedness and similarity measures on our new set of relatedness means, listed here as Rel-122. With the exception of our own measure, which is the *score* function presented above in Section 3.2.1, the measures listed in Table 6.3 are all discussed in detail above in Section 2.2, "WordNet-Based Measures of Similarity and Relatedness." Figures in starred rows are traditionally considered to be relatedness measures; the remaining rows are similarity measures. (For a summary of these measures, see Table 2.1.) Coefficients of correlation are given for Pearson's product-moment correlation (r), as well as Spearman's rank correlation (ρ). For comparison, we include results for the correlation of these measures to the M&C and R&G similarity means.

Table 6.3:
Coefficients of correlation to mean relatedness scores (Rel-122) and mean similarity scores (M&C, R&C) for various measures. Pearson's product-moment correlations (r-values) and Spearman's rank correlations (ρ-values) are reported.

	Rel-122		**M&C**		**R&G**	
Measure	**r**	**ρ**	**r**	**ρ**	**r**	**ρ**
* Szumlanski and Gomez (2010)	**0.654**	**0.534**	0.852	0.859	0.824	**0.841**
* Patwardhan and Pedersen (2006)	0.341	0.364	**0.865**	**0.906**	0.793	0.795
Path Length	0.225	0.183	0.755	0.715	0.784	0.783
* Banerjee and Pedersen (2003)	0.210	0.258	0.356	0.804	0.340	0.718
Resnik (1995)	0.203	0.182	0.806	0.741	0.822	0.757
Jiang and Conrath (1997)	0.188	0.133	0.473	0.663	0.575	0.592
Leacock and Chodorow (1998)	0.173	0.167	0.779	0.715	**0.839**	0.783
Wu and Palmer (1994)	0.187	0.180	0.764	0.732	0.797	0.768
Lin (1998)	0.145	0.148	0.739	0.687	0.726	0.636
* Hirst and St-Onge (1998)	0.141	0.160	0.667	0.782	0.726	0.797

Aside from the implementation of our own measure, the results reported above are derived from the standard implementations of these algorithms in version 2.05 of the WordNet::Similarity Perl module (Pedersen et al., 2004), using WordNet version 3.0.

We note that the strength of our own measure's correlation to the relatedness norms, $r = 0.654$, is encouraging, especially in light of the fact that our measure was only developed to produce a relative reordering of co-targets by relational strength to a target, and not to provide globally meaningful measurements of semantic relatedness.

The generally weak performance of the WordNet-based measures on this task is not surprising, given our observation that WordNet's minimalistic sense glosses and strong disposition toward codifying semantic similarity make it an impoverished resource for discovering general semantic relatedness. Even the three measures that have been touted in the literature as relatedness measures (Banerjee & Pedersen, 2003; Hirst & St-Onge, 1998; Patwardhan & Pedersen, 2006) have been hampered by their reliance upon WordNet.[36]

6.4 Summary

In this chapter, we presented a new set of relatedness norms for 122 noun pairs. The norms are offered as a new evaluative standard for quantitative computational measures of semantic relatedness, which have seen strong reliance on comparison to R&G and M&C

36 Recall that Hirst and St-Onge's path-based measure (Section 2.2.2) is largely dependent on similarity relationships denoted by WordNet's *IsA* relations, and that it earned more general classification as a relatedness measure for its incorporation of antonymic and meronymic (part-whole) relationships (see, e.g., Budanitsky & Hirst, 2006). We should note that there is some question about the accuracy of this classification, as some sources (cf. the comments of Resnik, 1999, p. 95) have pointed out that meronymic relations can be considered indications of similarity. Antonymy is similarly considered by some to capture notions of strong similarity, albeit with negative polarity.

similarity norms—despite widespread acknowledgement in the literature that similarity is only one specific type of relatedness. Our relatedness norms were elicited from human participants with minor modifications to an established methodology that has been used to acquire gold standard similarity norms, and which has been shown to yield consistent results across multiple similarity studies. The dissemination of this new dataset is the primary contribution of this study.

In analyzing the results of our study, we also arrived at several key findings: first, human participants exhibit lower degrees of correlation to relatedness norms than to similarity norms, suggesting that human perceptions of relatedness are less strongly constrained than those of similarity. Second, the average human correlation of $r = 0.769$ to our relatedness norms suggests that in order to achieve human-like performance at measuring semantic relatedness, a computational measure need not aspire to the same degree of alignment with relatedness norms as with similarity norms. Finally, we observed that WordNet-based measures of similarity and relatedness are indeed inhibited from discovering relatedness by the network's strong emphasis on codifying similarity relationships, vis-à-vis its sophisticated *IsA* ontology. In order to achieve Quillian's dream of a computational model of semantic memory, we must look beyond WordNet to find more general indications of semantic relatedness.

CHAPTER 7: CONCLUSIONS

In this chapter, we review the contributions and central findings of this dissertation. First we give an overview of how the network was automatically acquired (Section 7.1) and evaluated (Section 7.2). Then we discuss our investigation into human perceptions of relatedness and the implications our findings have for assessing computational relatedness measures (Section 7.3). We conclude with a discussion of the current state of the network, the kinds of relationships it expresses, and a few directions for future work (Section 7.4).

7.1 Acquisition

The primary limitation of WordNet as a model of semantic memory (cf. Quillian, 1968) is its strong focus on codifying semantic similarity, which, as we have seen, is only one particular type of semantic relatedness. While other semantic networks have attempted to represent a wider variety of semantic relations, they typically focus on surface relationships between words instead of concepts (cf. ConceptNet and NELL), or only attempt to measure relatedness quantitatively instead of constructing networks (vis-à-vis the Wikipedia-based measures discussed in Sections 2.5.1 through 2.5.3). A notable exception is WordNet++ (WN++), which derives semantic links between WordNet concepts from inter-article links in Wikipedia. However, links in Wikipedia are often capricious, which gives rise to many spurious relationships in WN++ (cf. the relationship of *prostitution#1* to *English_language#1* in WN++, or of *United_States#1* to *burrito#1*; spurious relationships in the network are particularly common with proper nouns,

which have a tendency to appear in articles about entities to which they bear no strong semantic relationship).

Cognizant of the limitations of these approaches to semantic relatedness, and of the importance of concept-level relationships to mechanisms of natural language understanding, we embarked upon the acquisition of a new semantic network. We first leveraged the tremendous amount of data available in the Wikipedia corpus to automatically discover the semantic associates of over 7,500 of the most common nouns in the English language. Toward this end, we adapted an information theoretic measure that took higher-than-expected co-occurrence frequencies as indications of relatedness between words. Relying on our asymmetric, quantitative measure, we then found pairs of nouns that exhibited strong, mutual relatedness and admitted them to a semantic network of related nouns. This first phase of network acquisition saw the creation of a semantic network with 155,180 edges indicating semantic relatedness between nouns.

In the second phase of network acquisition, we automatically disambiguated those nouns to noun senses (i.e., *concepts*) from WordNet 3.0. To do so, we employed a suite of disambiguation algorithms that capitalized on salient sense clustering (vis-à-vis categorization in WordNet) among related nouns. For example, the noun "pie" is strongly associated with several nouns that have senses categorized by *baked_goods#1* in WordNet, which serves as a strong cue to preferentially disambiguate "cookie" (as it relates to "pie") to *cookie#1* (the baked good), as opposed to *cookie#2* (a cook) or *cookie#3* (a web browser cookie). Similarly, the higher-than-expected relationship of "astronomer" to nouns categorized by *celestial_body#1* helps us disambiguate "star" to its *celestial body* senses, and the relationship of "unicorn" to several

nouns categorized by *mythical_being#1* informs our disambiguation of, e.g., "phoenix" to the legendary bird that rises from the ashes to be born anew (*phoenix#3*), as opposed to the constellation of the same name (*phoenix#4*), the capital city of Arizona (*phoenix#1*), or *genus Phoenix* (*phoenix#2*).

The resulting network, which we call the Szumlanski-Gomez Network (SGN), indicates semantic relatedness between concepts from the noun sense inventory of WordNet. It articulates 208,832 relationships between 38,249 distinct concepts. It is derived from the automatic discovery of semantic associates for over 7,500 target nouns, with 17,104 distinct senses among them. Mirroring the structure of WordNet, concepts are related categorically, rather than quantitatively. Furthermore, following the observation of Quillian that any concept can serve as a relationship between two entities, we have not restricted ourselves to mining instances of specific, labeled relations. Instead, our network represents relatedness in an unlabeled manner. The addition of labels to a network like ours is a potentially exciting avenue for future work, albeit a challenging one, as even humans sometimes have tremendous difficulty verbalizing the relation that binds strongly related entities.

7.2 Evaluation

Following standard procedure in the relatedness literature, we evaluated our network's performance on two tasks: comparison to similarity norms and an applied task. With respect to similarity norms, we found high correlation of our quantitative relatedness scoring mechanism to the similarity norms of Rubenstein and Goodenough (1965) and Miller and Charles (1991) (r = 0.852 and r = 0.824, respectively). With respect to an applied task, we evaluated the

performance of our network on the SemEval-2007 coarse-grained word sense disambiguation (WSD) task (Navigli et al., 2007) using two graph-based disambiguation algorithms: the extended gloss overlaps measure of Banerjee and Pedersen (2003) and the degree centrality algorithm of Navigli and Lapata (2010). We compared our results on this task to those achieved using WordNet and WN++ with the same algorithms, and presented three central findings with respect to results from the extended gloss overlaps measure: first, our network's performance was comparable to that of WordNet. Second, the combination of SGN and WordNet produced results that out-performed what either network achieved as a stand-alone resource. Third, our network outperformed WN++, which we attributed to spurious relationships found in WN++.

With respect to the degree centrality algorithm, we found that neither SGN nor WN++ outperformed WordNet, and that WordNet itself was unable to surpass the most frequent sense baseline using that algorithm. We attributed the shortcomings of the degree centrality algorithm to the fact that semantic networks like WordNet, WN++, and SGN tend to be dense, and so it is often possible to find short paths between any two concepts in the networks. Thus, the degree centrality algorithm, which searches for short paths between co-occurring nouns in a context and disambiguates nouns to the sense(s) that are participant to the greatest number of such paths, essentially serves to disambiguate a noun to whichever sense has the greatest number of relationships in a network. This explains the "plateau effect" of the algorithm that we observed, and which Navigli and Lapata (2010) also reported, and is one of the novel findings of our research.

In addition to these standard evaluative procedures, we subjected our network to manual inspection by independent judges who evaluated the precision of the noun-noun relationships

that were admitted to our network, as well as the precision of our disambiguation results. Although manual inspection is tedious and does not allow for comprehensive evaluation of a network, the small samples of data that our judges evaluated yielded promising results. On average, they judged the precision of noun-noun relationships in our network to stand at 95.66% (out of 100 pairs evaluated), and an average of 85% of our disambiguation results were deemed acceptable to our judges (out of 50 pairs evaluated).

7.3 Human Perceptions of Relatedness

Although comparison to human judgments of semantic similarity has long served as a gold standard for evaluating similarity and relatedness measures, the distinction between similarity and relatedness (in that similarity is one particular type of relatedness) is well established in the field. To date, no viable gold standard of relatedness norms has emerged to supplement or supplant comparison to the similarity norms of Rubenstein and Goodenough (1965) (R&G) and Miller and Charles (1991) (M&C). Thus, we embarked on the creation of a set of relatedness norms. In our study, we followed the established methodology of R&G, adapted in our case to elicit relatedness scores instead of similarity scores. Our resulting set of relatedness norms for 122 noun pairs is the primary contribution of that study.

In analyzing the results of our study, we also presented three key findings with respect to relatedness norms. First, individuals in our study exhibited strong correlation to mean relatedness scores using leave-one-out sampling (so individuals were never compared to their own data) ($r = 0.77$, $\sigma = 0.09$, $N = 73$). Second, we found individual correlation to our relatedness norms to be lower than the expected human correlation to similarity norms, for which Resnik (1995)

established an upper bound of $r = 0.90$. This suggests that human perceptions of relatedness are less strictly constrained than perceptions of similarity, and that quantitative, computational measures of semantic relatedness need not aspire to $r = 0.90$ correlation with relatedness norms in order to claim human-like performance. Third, we evaluated WordNet-based quantitative measures of semantic similarity and relatedness that exhibited high average correlation to the R&G and M&C similarity norms ($r = 0.711$ and $r = 0.689$ on the two datasets, respectively, for the nine measures evaluated) and found that they exhibited low average correlation to our relatedness norms ($r = 0.201$). In comparison, our adapted scoring measure correlated to the relatedness norms from our study with $r = 0.654$. These results support our claim that WordNet —despite its sophisticated *IsA* ontology and the additional relations and glosses it provides—is insufficient to indicate semantic relatedness between concepts in the general case.

7.4 Discussion

The relationships in SGN reflect broad coverage of general human knowledge and perceptions of relatedness. The network codifies commonsense relationships, such as (*lock#1*, *key#1*), (*pen#1*, *pocket#1*), (*camping#1*, *tent#1*), and (*camping#1*, *campfire#1*), as well as basic, essential relationships, such as (*elephant#1*, *tusk#{1,2}*). (In WordNet, *elephant#1* is the pachyderm sense of "elephant;" *elephant#2* is the symbol of the United States' Republican Party.) Of particular interest in the case of (*elephant#1*, *tusk#{1,2}*) is the fact that our network allows for relationships between multiple senses of the same words. In this case, our disambiguation methods have taken both *tusk#1* and *tusk#2* for relation to *elephant#1*. Indeed, both senses seem strongly related:

(*tusk#1*) a hard smooth ivory colored dentine that makes up most of the tusks of elephants and walruses

(*tusk#2*) a long pointed tooth specialized for fighting or digging; especially in an elephant or walrus or hog

The conflation of these senses of "tusk" in relation to *elephant#1* is similar to the conflation of *star#1* and *star#3* that we saw earlier in relation to *astronomer#1*, in that it demonstrates the ability of our disambiguation methods to cope with the kinds of fine-grained polysemic distinctions we often see in WordNet. The concepts *star#1* and *star#3* are sisters in WordNet, both sharing *celestial_body#1* as their immediate hypernym, and the distinction between the two is often difficult for humans to pinpoint:

(*star#1*) (astronomy) a celestial body of hot gases that radiates energy derived from thermonuclear reactions in the interior

(*star#3*) any celestial body visible (as a point of light) from the Earth at night

Compare this to WordNet's delineation of the following senses of "key," which can be viewed as an instance of homonymy rather than one of systemic polysemy, as the concepts bear no similarity:

(*key#1*) metal device shaped in such a way that when it is inserted into the appropriate lock the lock's mechanism can be rotated

(*key#4*) any of 24 major or minor diatonic scales that provide the tonal framework for a piece of music

In addition to basic, commonsense relationships, our network also gives broad indication of relationships that are explicitly historical or cultural in nature, such as the pairwise association of *communist#1*, *atheist#1*, and *homosexual#1*, or the relationship of *evolution#1* to both *creationism#1* and *public_school#1*. Upon close examination, the network can even be found to contain traces of prevailing attitudes toward certain entities, such as the one reflected in the relationship of *concentration_camp#1* to *atrocity#1* in SGN. These subtle indications of how the human mind classifies or associates certain entities might eventually be useful in automatically assessing the polarity of nouns in the ontology (i.e., the positive and negative connotations of certain words).

Of course, for a complete understanding of why two entities are related in the network, we often have to analyze sentences in which they co-occur. This is particularly true of relationships in SGN that represent some of the specific, technical knowledge articulated in Wikipedia. For example, the pair (*mansion#1*, *constellation#1*), which at first glance seems spurious, represents a relationship from the domain of astrology; the gloss for *mansion#1* in WordNet is "(astrology) one of 12 equal areas into which the zodiac is divided," and the concept is synonymous with *star_sign#1* and *sign_of_the_zodiac#1*. Similarly, the pair (*canal#3*, *summit#1*) might seem spurious and nonsensical to those of us without technical knowledge of the workings of canal locks; a *summit level canal* is a particular type of canal, and a *summit pound* is the highest pound (i.e., body of water between two canal locks) of a particular route along a canal. Unfortunately, neither *summit level canal* nor *summit pound* have lexical entries in WordNet, and so our network cannot relate them to *canal#3*. We are instead left with the association of "canal" and "summit" from their frequent co-occurrence in Wikipedia. However,

the augmentation of WordNet with new concepts, signified by significant collocation throughout a corpus like Wikipedia, is one exciting avenue for future research. For example, the frequent co-occurrence of "elephant" and "graveyard" in Wikipedia is represented in our network with the relationship (*elephant#1*, *graveyard#1*), since WordNet has no lexical entry for "elephant graveyard." However, the terms' collocative occurrence throughout the corpus as "elephant graveyard," combined with the strong evidence for their relatedness discovered in our research, suggests the phrase should garner its own entry in the ontology.

The method we have presented for network acquisition is general enough that it can be applied not only to Wikipedia, but to other large corpora, as well. This provides several avenues for future research. Of particular interest is the continued development and expansion of the network to include new relationships not yet discovered from our version of the Wikipedia corpus. This can perhaps be achieved by applying our methodology to new versions of Wikipedia, or even to other corpora. The former suggests another avenue for future research, which is an analysis of how semantic relatedness, as reflected by usage in a large corpus, changes over time. Because we have restricted our consideration to semantic associates of the most common nouns in the English language, our research also leaves room to explore the discovery of semantic associates of infrequently occurring nouns—possibly using the current network, which already contains relationships for some nouns outside of our target range, to bootstrap a new phase of acquisition.

LIST OF REFERENCES

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL),* 19–27.

Agirre, E., de Lacalle, O. L., & Soroa, A. (2009). Knowledge-based WSD on specific domains: Performing better than generic supervised WSD. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, 1501–1506.

Ahn, D., Jijkoun, V., Mishne, G., Müller, K., de Rijke, M., & Schlobach, S. (2004). Using Wikipedia at the TREC QA track. In *Proceedings of the 13th Text Retrieval Conference (TREC 2004)*.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, 86–90.

Banerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, 805–810.

Berland, M., & Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, 57–64.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyaganiak, R., & Hellmann, S. (2009). DBpedia – A crystallization point for the Web of Data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 7*, 154–165.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD 2008 International Conference on Management of Data*, 1247–1250.

Bollacker, K., Tufts, P., Pierce, T., & Cook, R. (2007). A platform for scalable, collaborative, structured information integration. In *Proceedings of the 6th International Workshop on Information Integration on the Web, Association for the Advancement of Artificial Intelligence (AAAI)*, 22–27.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics, 21*, 543–565.

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

Bunescu, R., & Paşca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 9–16.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., Jr., & Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the 24th Conference on Artificial Intelligence (AAAI)*, 1306–1313.

Charniak, E. (1983). Passing markers: A theory of contextual influence in language comprehension. *Cognitive Science, 7*(3), 171–190.

Chaudhari, D. L., Damani, O. P., & Laxman, S. (2011). Lexical co-occurrence, statistical significance, and word association. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1058–1068.

Chklovski, T., & Pantel, P. (2004). VerbOcean: Mining the Web for fine-grained semantic verb relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 33–40.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22–29.

Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering, 19*(3), 370–383.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407–428.

Collins, A. M., & Quillian, M. R. (1972). How to make a language user. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory* (pp. 309–351). New York: Academic Press.

Coursey, K., Mihalcea, R., & Moen, W. (2009). Using encyclopedic knowledge for automatic topic identification. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL)*, 210–218.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for Kit of Factor-Referenced Cognitive Tests*. Princeton, NJ: Educational Testing Service.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., . . . Yates, A. (2004). Web-scale information extraction in KnowItAll. In *Proceedings of the 13th International World Wide Web Conference*, 100–110.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1535–1545.

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, 280*, 20–32.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems (TOIS), 20*(1), 116–131.

Firth, J. R. (1957). A synopsis of linguist theory 1930-1955. In *Studies in Linguistic Analysis*, 1–32. Oxford: Philological Society. Reprinted in F. R. Palmer (Ed.), *Selected Papers of J. R. Firth 1952-1959* (pp. 168–205). Bloomington and London: Indiana University Press (1968).

Fleischman, M., Hovy, E., & Echihabi, A. (2003). Offline strategies for online question answering: Answering questions before they are asked. In *Proceedings of the 41st Annual Meetings of the Association for Computational Linguistics (ACL)*, 1–7.

Gabrilovich, E. (2006, October 4). *The WordSimilarity-353 Test Collection*. Retrieved December 6, 2012, from http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 1606–1611.

Gale, W. A., Church, K. W., & Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 233–237.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics, 28*(3), 245–288.

Girju, R., Badulescu, A., & Moldovan, D. (2006). Automatic discovery of part-whole relations. *Computational Linguistics, 32*(1), 83–135.

Gomez, F. (2001). An algorithm for aspects of semantic interpretation using an enhanced WordNet. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 87–94.

Gomez, F. (2004). Building verb predicates: A computational view. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL)*, 351–358.

Gomez, F. (2007). Semantic interpretation and the WordNet upper-level ontology. *Journal of Intelligent Systems, 16*(2), 93–116.

Gorman, J., & Curran, J. R. (2006). Scaling distributional similarity to large corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, 361–368.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Boston, MA: Kluwer Academic Publishers.

Harris, Z. S. (1954). Distributional structure. *Word, 10*(1), 146–162. Reprinted in J. J. Katz (Ed.), *The Philosophy of Linguistics* (pp. 26–47). Oxford University Press (1985).

Havasi, C., Speer, R., & Alonso, J. B. (2007). ConceptNet 3: A flexible, multilingual semantic network for commonsense knowledge. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, 539–545.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL)*, 268–275.

Hirst, G., & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 305–332). MIT Press.

Hughes, T., & Ramage, D. (2007). Lexical semantic relatedness with random graph walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CONLL)*, 581–589.

Jarmasz, M., & Szpakowicz, S. (2003). Roget's Thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, 212–219.

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

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language, 39*(2), 170–210.

Koeling, R., McCarthy, D., & Carroll, J. (2005). Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, 60–67.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*(1), 79–86.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 265–283). MIT Press.

Lee, J. H., Kim, M. H., & Lee, Y. J. (1993). Information retrieval based on conceptual distance in is-a hierarchies. *Journal of Documentation, 49*(2), 188–207.

Lee, M. D., Pincombe, B., & Welsh, M. (2005). An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Cognitive Science Conference (CogSci)*, 1254–1259.

Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM, 33*(11), 33–38.

Lesk, M. (1986). 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 (SIGDOC)*, 24–26.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL: University of Chicago Press.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning (ICML),* 296–304.

Lin, D., & Pantel., P. (2001). Discovery of inference rules for question answering. *Natural Language Engineering, 7*(4), 343–360.

Lin, D., Zhao, S., Qin, L., & Zhou, M. (2003). Identifying synonyms among distributionally similar words. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, 1492–1493.

Liu, H., & Singh, P. (2004a). ConceptNet – a practical commonsense reasoning toolkit. *BT Technology Journal, 22*(4), 211–226.

Liu, H., & Singh, P. (2004b). Commonsense reasoning in and over natural language. In *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES)*, 293–306.

McCarthy, J. (1959). Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, 756–791.

McKoon, G., & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(6), 1155–1172.

Medelyan, O., & Legg, C. (2008). Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence*, 13–18.

Mihalcea, R. (2007). Using Wikipedia for automatic word sense disambiguation. In *Proceedings of Human Language Technology: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 196–203.

Mihalcea, R., & Cosmai, A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM)*, 233–242.

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

Miller, G. A. (1998). Nouns in WordNet. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 23–46). MIT Press.

Milne, D., & Witten, I. (2008a). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI)*, 25–30.

Milne, D., & Witten, I. (2008b). Learning to link with Wikipedia. In *Proceedings of the 17th ACM International Conference on Information and Knowledge Management (CIKM)*, 509–518.

Minsky, M., Singh, P., & Sloman, A. (2004). The St. Thomas common sense symposium: Designing architectures for human-level intelligence. *AI Magazine, 25*(2), 113–124.

Moldovan, D., Badulescu, A., Tatu, M., Antohe, D., & Girju, R. (2004). Models for the semantic classification of noun phrases. In *Proceedings of the HLT-NAACL 2004 Workshop on Computational Lexical Semantics*, 60–67.

Navigli, R. (2005). Semi-automatic extension of large-scale linguistic knowledge bases. In *Proceedings of the 18th Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 548–553.

Navigli, R., & Lapata, M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(4), 678–692.

Navigli, R., Litkowski, K. C., & Hargraves, O. (2007). SemEval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval '07)*, 30–35.

Palermo, D., & Jenkins, J. (1964). *Word association norms: Grade school through college*. Minneapolis, MN: University of Minnesota Press.

Pantel, P., & Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, 113–120.

Patwardhan, S., & Pedersen, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics Workshop on Making Sense of Sense*, 1–8.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet::Similarity – Measuring the relatedness of concepts. In *Proceedings of the 5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 38–41.

Ponzetto, S. P., & Navigli, R. (2009). Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, 2083–2088.

Ponzetto, S. P., & Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1522–1531.

Quillian, M. R. (1968). Semantic Memory. In M. Minsky (Ed.), *Semantic Information Processing*. MIT Press.

Quillian, M. R. (1969). The Teachable Language Comprehender: A simulation program and theory of language. *Communications of the ACM, 12*(8), 459–476.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Rada, R., & Bicknell, E. (1989). Ranking documents with a thesaurus. *Journal of the American Society for Information Science (JASIS) 40*(5), 304–310.

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

Ravichandran, D., & Hovy, E. (2002). Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 41–47.

Reed, S. L., & Lenat, D. B. (2002). Mapping ontologies into Cyc. In *Proceedings of the AAAI Workshop on Ontologies and the Semantic Web*.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 448–453.

Resnik, P. (1997). Selectional preference and sense disambiguation. In *Proceedings of the ANLP Workshop on Tagging Text with Lexical Semantics: Why What, and How?* 52–57.

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

Riloff, E., & Jones, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence and 11th Innovative Applications of Artificial Intelligence Conference (AAAI/IAAI)*, 474–479.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (27–48). Hillsdale, NJ: Erlbaum.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8*(10), 627–633.

Sahlgren, M. (2008). The Distributional Hypothesis. *Rivista di Linguistica, 20*(1), 33–53.

Salton, G., & McGill, M. J. (1983). *An Introduction to Modern Information Retrieval*. McGraw-Hill.

Singh, P. (2002). The public acquisition of commonsense knowledge. In *Proceedings of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, 47–52.

Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., & Zhu, W. L. (2002). Open Mind Common Sense: Knowledge acquisition from the general public. In *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*, 1223–1237.

Spence, D. P., & Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research, 19*, 317–330.

Strube, M., & Ponzetto, S. P. (2006). Wikirelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 1419–1424.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). *YAGO: A large scale ontology from Wikipedia and WordNet.* (Research Report MPI-I-2007-5-003).

Szumlanski, S., & Gomez, F. (2010). Automatically acquiring a semantic network of related concepts. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM)*, 19–28.

Szumlanski, S., & Gomez, F. (2011). Evaluating a semantic network automatically constructed from lexical co-occurrence on a word sense disambiguation task. In *Proceedings of the 15th Conference on Computational Natural Language Learning (CoNLL)*, 190–199.

Tandon, N., de Melo, G., & Weikum, G. (2011). Deriving a Web-scale common sense fact database. In *Proceedings of the 25th Conference on Artificial Intelligence (AAAI)*, 152–157.

Tanner, J. J., & Gomez, F. (2010). Extracting ontological selectional preferences for non-pertainym adjectives from the Google Corpus. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI)*, 1033–1038.

Waltz, D. L., & Pollack, J. B. (1985). Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science, 9*(1), 51–74.

Weeds, J. (2003). *Measures and Applications of Lexical Distributional Similarity* (PhD thesis). University of Sussex, UK.

Weeds, J., & Weir, D. (2006). Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics, 31*(4), 439–475.

Wettler, M., & Rapp, R. (1993). Computation of word associations based on the co-occurrences of words in large corpora. In *Proceedings of the 1st Workshop on Very Large Corpora*, 84–93.

Yarowsky, D. (1993). One sense per collocation. In *Proceedings of the ARPA Workshop on Human Language Technology*, 266–271.

Zaragoza, H., Rode, H., Mika, P., Atserias, J., Ciaramita, M., & Attardi, G. (2007). Ranking very many typed entities on Wikipedia. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM)*, 1015–1018.