Automatically Acquiring a Semantic Network of Related Concepts

Sean Szumlanski
seansz@cs.ucf.edu

Fernando Gomez
gomez@eecs.ucf.edu

Department of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL 32816, USA

ABSTRACT

We describe the automatic construction of a semantic network[1], in which over 3000 of the most frequently occurring monosemous nouns[2] in Wikipedia (each appearing between 1,500 and 100,000 times) are linked to their semantically related concepts in the WordNet noun ontology. Relatedness between nouns is discovered automatically from co-occurrence in Wikipedia texts using an information theoretic inspired measure. Our algorithm then capitalizes on salient sense clustering among related nouns to automatically disambiguate them to their appropriate senses (i.e., *concepts*). Through the act of disambiguation, we begin to accumulate relatedness data for concepts denoted by polysemous nouns, as well. The resultant concept-to-concept associations, covering 17,543 nouns, and 27,312 distinct senses among them, constitute a large-scale semantic network of related concepts that can be conceived of as augmenting the WordNet noun ontology with *related-to* links.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*semantic networks*; I.2.6 [**Artificial Intelligence**]: Learning—*concept learning, connectionism and neural nets, knowledge acquisition*

General Terms

Algorithms, Experimentation, Measurement

Keywords

semantic relatedness, semantic networks, lexical semantics, common sense knowledge, knowledge acquisition

[1]Available online: `http://www.cs.ucf.edu/~seansz/sem`
[2]This refers to the number of entries a noun has in WordNet, not Wikipedia.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
Copyright 2010 ACM 978-1-4503-0099-5/10/10 ...$10.00.

1. INTRODUCTION

While the contemplation of semantic networks and their usefulness for natural language understanding tasks dates back to the work of Quillian [26], we have yet to see the creation of a large-scale, viable model of semantic memory. The WordNet ontology [8] constitutes a partial realization of Quillian's dream through its instantiation of a variety of labeled edges indicating, *inter alia*, subsumptive *is-a* relationships between noun senses (concepts[3]). These relations constitute a rich taxonomy of semantic similarity.

Absent from the ontology, however, is a more general indication of semantic relatedness. From WordNet we can infer, for example, the similarity between *penguins* and *flamingos*: both share the superordinate concept *aquatic bird* (and are, by virtue of their similarity, also related). In contrast, the relatedness between *penguins* and *icebergs* (which certainly are not similar entities, although they are conceptually related), is not articulated in the ontology.

This distinction between similarity and relatedness is well established in the literature [28], and augmenting WordNet with *related-to* links would open new vistas for applications and research using the ontology (see Section 8). To understand the role of semantic relatedness in natural language understanding, consider, for example, the following sentences:

(1) The astronomer photographed the star.
(2) The paparazzi photographed the star.

Despite the syntactic equivalence of the two sentences, it is clear that the "star" in (1) denotes a celestial body, whereas the "star" in (2) refers to a celebrity. While it is conceivable that an astronomer would photograph a celebrity, or that a paparazzo would photograph a celestial object, the "stars" here are preferentially disambiguated by the strong semantic relatedness between *astronomer* and the *celestial body* sense of "star," and *paparazzi* and the *celebrity* sense of "star," respectively. Notice that if we relied on semantic similarity to disambiguate (1), the path in WordNet connecting *astronomer* and the *celebrity* sense of "star" (in that both are *people*) would lead us astray.

Implicit to this discussion so far is the assumption that the semantic network relates not just words, but concepts.

[3]Throughout this paper, we use the terms "concepts" and "noun senses" interchangeably. In distinguishing between words and the concepts they denote, we quote the former and italicize the latter.

Concept-level association is critical for natural language understanding, although we see in much of the existing literature a focus on surface-level relationships between words.

In this work, we automatically acquire a semantic network of related *concepts*. For our concepts, we use the noun senses defined in WordNet 3.0. This seems an obvious choice given the sophistication of WordNet's noun ontology and its ubiquitous use in computational linguistics and artificial intelligence. 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. However, this 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.

We first acquire relatedness between nouns by applying a novel adaptation of an information theoretic measure to co-occurrence data extracted from Wikipedia[4], chosen as our target corpus for its large size and coverage of the English language. However, our approach 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.

Once we have established relatedness between nouns, we automatically disambiguate them to their corresponding noun senses in WordNet, capitalizing on sense similarity clustering and high degrees of inter-relatedness that we have found to occur among related nouns.

We should note that, when faced with two polysemous nouns that are related, disambiguation often requires that we already have a *related-to* link established between the appropriate word senses. Consider, for example, the relation between "bus" and "horn." Here we think not of a computer's front-side bus, or of a rhinoceros's horn, but of the automobile and its car horn. This automatic disambiguation results from the semantic relatedness between the two senses denoted by these words. Of course, if our goal is to instantiate such a *related-to* link, then the link cannot be used for this initial act of disambiguation. Another approach is called for.

We instead focus on disambiguating words related to monosemous nouns. In doing so, the monosemous noun to which a polysemous noun is related provides an unequivocal context in which that disambiguation can take place (cf. "horn" and the monosemous "rhinoceros"). This makes disambiguation significantly more achievable. That is not to say, however, that we are not accumulating relatedness data for polysemous nouns. By disambiguating the "horn" to which "rhinoceros" is related and the "horn" to which "oboe" is related, we begin to accumulate relatedness data for individual senses of the polysemous "horn." Additionally, once this initial partition is formed, we can recover and disambiguate polysemous nouns related to individual senses of "horn," although a detailed discussion of this process is beyond the scope of this paper.

Thus, the contributions of this work are twofold: we offer (1) a novel approach to discovering semantic relatedness based on lexical co-occurrence data from a large corpus and (2) a first iteration of a semantic network of related concepts, automatically acquired by considering relatedness to over 3000 of the most frequently occurring monosemous nouns in Wikipedia, and currently relating 27,312 distinct senses from among 17,543 nouns.

[4]`http://www.wikipedia.org`

The rest of this paper proceeds as follows. In Section 2, we discuss our approach in the context of related work. In Sections 3 through 6, we explicate our approach to automatically acquiring the semantic network. At the end of each of these sections, we pause to present results and an evaluation of our algorithm's performance up to that point. We discuss two excerpts from the semantic network in Section 7, and present our conclusions and directions for future work in Section 8.

2. RELATED WORK

Our work bears strong relation to ConceptNet [19]. Notwithstanding the name, ConceptNet is a semantic network in which the nodes stand for words that are not disambiguated (not concepts). The network is constructed using a set of 20 predefined semantic relations (e.g., *EffectOf*, *CapableOf*, *LocationOf*) coupled with regular expression pattern matching to extract categorical relatedness between words from the Open Mind Common Sense project. The strengths of ConceptNet compared to our work are that it relates not only nouns, but also verbs, adjectives, and prepositional phrases, and it indicates the semantic relation that associates each pair of nodes.

However, its dependence on a finite set of predefined semantic relations precludes ConceptNet from discovering relatedness between words in the general case; as Quillian 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" (emphasis in original) [26]. Furthermore, the construction of ConceptNet is not fully automated, as it relies on common sense facts that are manually entered into its training corpus, and cannot discover relatedness from a corpus that is not hand-tailored for the purpose.

Several other methods have used pattern matching to discover specific semantic relations. Turney's Latent Relational Analysis (LRA) [34, 33] induces patterns automatically from a pair of relationally similar words, and solves SAT analogies, while Davidov and Rappoport's [7] unsupervised pattern clustering algorithm has been used to create categories of semantically similar words. Pantel and Pennacchiotti's Espresso algorithm induces search patterns automatically from a corpus, given small seed sets of related nouns, and has successfully discovered hyponymic (*is-a*) and meronymic (*part-of*) relations between nouns, several relations specific to the domain of chemistry (such as chemical *reaction* and *production* relations), among others [23].

Manually defined lexico-syntactic patterns have also been used to harvest relations from large corpora. Hearst [13] first used such patterns to automatically discover hyponymic relations not present in WordNet. 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." In a similar vein, Girju et al. [10] and Berland and Charniak [2] used manually defined lexico-syntactic patterns to mine large corpora for meronymic relations.

The major difference between these approaches and ours is that (with the exception of [7]) the pattern-based methods require a predetermination of the specific types of relations to be mined, whether through the articulation of exemplar seed sets, target noun pairs, or lexico-syntactic patterns, and

are not designed for the more general discovery of semantic relatedness that we are interested in.

Other approaches in the literature typically measure relatedness between two concepts or nouns quantitatively, and are distinct from our work in that they do not build databases or discover categorical relatedness. Whereas we attempt to answer the question, "What concepts are related to X?" the quantitative approaches attempt to answer the question, in which *both* concepts (or nouns) are given as priors, "*To what degree* are X and Y related?" (which clearly cannot be precomputed for every possible X and Y in the English language).

Some of these quantitative approaches attempt to measure relatedness using only information available from WordNet, such as *is-a* relations and sense glosses [24, 15]. These methods are inherently limited by the fact that, while WordNet serves as a rich taxonomy of semantic similarity, it lacks general indications of semantic relatedness (with its articulation of holonymic relationships being the notable exception). Consider, for example, how WordNet-based approaches would discover the strong semantic relationship, as our system does, between *penguin* and *tuxedo*. 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 sizeable corpus.

For this reason, many quantitative measures have turned to large corpora to measure relatedness, often relying on distributional similarity to establish synonymy and hypernym relations between nouns [12, 11]. Some measures have used the underlying structure of Wikipedia (i.e., disambiguation pages and links between articles) to measure semantic relatedness between nouns or concepts, sometimes grounding their work in the folksonomy of concepts constituted by titles of Wikipedia articles rather than measuring relatedness between WordNet synsets [30, 9, 35]. Suchanek et al. [31] derived a semantic network called YAGO from the underlying structure of Wikipedia articles. Over 73% of the facts in YAGO are encompassed by its *isCalled*, *type*, and *means* relations, which are indicative of semantic similarity. Among its most frequent relations beyond those indicating similarity are specific ones such as *bornOnDate*, *diedOnDate*, *hasPopulation*, *bornInLocation*, *actedIn*, *directed*, and *writtenInYear*.

Augmenting the structure of Wikipedia itself has been the subject of research, as well, and involves the discovery of relations between articles. Mihalcea and Csomai [20] augmented the underlying structure of Wikipedia by adding links between pages after automatically identifying keywords in each article and disambiguating those words to their appropriate Wikipedia concepts (article titles). Ponzetto and Navigli [25] used graph theoretic approaches to augment the taxonomic organization of Wikipedia articles.

Other quantitative approaches have leveraged the large amounts of data available on the Web to discover relatedness. Agirre and de Lacalle [1] employed web queries to associate WordNet synsets with representative context words, known as topic signatures. On average, a topic signature from their collection contains 6877 words and their associated weights. Cuadros and Rigau [6] have used these data to construct four KnowNets, semantic knowledge bases derived by disambiguating the top 5, 10, 15, and 20 nouns, respectively, from the topic signatures of Agirre and de Lacalle. Similarly, Navigli [22] 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.

In our approach, we rely solely on lexical co-occurrence between two nouns in a large corpus to discover semantic relatedness, rather than drawing on predetermined relations, lexico-syntactic patterns, distributional similarity (context), the underlying structure of Wikipedia or WordNet, or other semantically annotated resources. Because our approach is fully automated and we avoid relying on structured or semantically annotated resources, it can be applied to any large corpus, in any language, to discover new semantic relations, build new semantic networks, and augment existing ones with *related-to* links. Because we relate concepts categorically rather than quantitatively, and because we relate concepts, not nouns, the large-scale resource we have developed has potential for use in a wide variety of semantically driven tasks.

3. ACQUIRING THE SEMANTIC NETWORK: PRELIMINARIES

Our algorithm for acquiring the semantic network unfolds in three stages. First we measure the relational strength between nouns co-occurring in Wikipedia using an information theoretic measure. We then use this quantitative measure to make categorical assertions about relatedness between nouns. Finally, we disambiguate related nouns automatically, giving rise to a semantic network of related concepts.

To facilitate the extraction of co-occurrence data from Wikipedia, we have part-of-speech tagged the entire Wikipedia corpus (stripped of markup and metadata) using Brill's tagger [3]. Throughout the remainder of this work, co-occurrence of nouns[5] is considered only as sentence-level co-occurrence, 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 adopting a smaller or variable sized window, is generally quashed by the sheer magnitude of co-occurrence data available from the corpus.

4. FROM CO-OCCURRENCE TO RELATIONAL STRENGTH

We now adopt the following terminology. A **target** is any noun for which we would like to extract relatedness data. Nouns co-occurring with a target are called its **co-targets**, all of which are potentially semantically related to the target.

We define **relational strength** as a quantitative measure of the semantic relatedness of a target, t, to one of its co-targets, c. For this purpose, we adapt Resnik's selectional association metric [28], given here in the form of $S_{rel}(t, c)$, the relational strength of t to c:

$$S_{rel}(t, c) = \frac{1}{D_{KL}} P(c|t) \log \frac{P(c|t)}{P(c)}$$

where $P(c)$ is the relative frequency of c's occurrence in the corpus (the number of times c occurs, divided by the number

[5]Because the coverage of proper nouns in WordNet is minimal, we only consider common nouns here.

of noun tokens counted in the corpus). Similarly, $P(c|t)$ is the probability of encountering c in a sentence containing t (the number of times c occurs in sentences containing t, divided by the total number of nouns tokens co-occurring with t).

D_{KL} is the relative entropy, or Kullback-Leibler divergence, between probability distributions $P(C|t)$ and $P(C)$, where C is the class of all co-targets of t:

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

Intuitively speaking, D_{KL} indicates how likely we are to encounter c as a consequence of encountering t. Its highest values 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.

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. Accordingly, the metric is used only to sort the list of t's co-targets in order of decreasing relational strength, after which the usefulness of the metric 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. This leaves us with:

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

We also make this pragmatic change to our metric: to account for the relatedness of c to t, which certainly plays *some* role in the relational strength of t to c, we multiply $S_{rel}(t, c)$ by $P(t|c)$. This is particularly useful in suppressing words like "article," which tends 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" at all[6]. With this final modification, $S_{rel}(t, c)$ becomes:

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

Given a target of interest, we assemble its co-occurrence data (if it has not already been cached) and sort all 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 outright. This is done largely as a computational consideration. The presence of both $P(t|c)$ and $P(c|t)$ in $(*)$ requires us to have the co-occurrence data for both t and c to compute $S_{rel}(t, c)$, and, as there are often thousands of nouns co-occurring with a target below this frequency threshold, we save a considerable amount of processing time by eliminating them. This also protects us from false indications of relatedness that would arise if an incredibly rare word from the corpus were to co-occur with a semantically unrelated target just once or twice, as a matter of happenstance. We have found that lowering this threshold below 0.07% dramatically increases runtimes while producing negligible changes in our overall results. This makes intuitive sense, as we are effectively only eliminating from consideration those co-targets that account for fewer than 7 out of every 10,000 nouns co-occurring with a target.

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

Table 1: Coefficients of correlation with human similarity judgments. Starred rows are presented in [4].

Measure	M&C	R&G
Patwardhan and Pedersen [24]	.91	.90
Ranking by Relational Strength	.852	.824
Hughes and Ramage [15]	.838	.904
*Leacock and Chodorow [17]	.838	.816
Strube and Ponzetto [30]	.82	.86
*Lin [18]	.819	.829
*Hirst and St-Onge [14]	.786	.744
*Jiang and Conrath [16]	.781	.850
*Resnik [27]	.779	.774
Human Correlation [27]	.885	n/a

4.1 Evaluation

Although our aim is not to develop a quantitative measure of semantic relatedness, an objective evaluation is in order. In the relatedness literature, a standard approach is to measure correlation with mean similarity scores elicited from human subjects by Rubenstein and Goodenough [29] and Miller and Charles [21] (henceforth R&G and M&C, respectively). 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"). In R&G, participants evaluated 65 word pairs. M&C then replicated the experiment using 30 pairs from the 65 used in R&G.

Given that our S_{rel} function is used only to rank co-targets by their relative relatedness to a particular target, for this task we score relatedness between two words, a and b, as follows:

$$score(a, b) = 4.0 * \text{avg}\left(\frac{rank_a(b)}{|C_a|}, \frac{rank_b(a)}{|C_b|}\right)$$

where $rank_t(c)$ is the numerical rank of c among t's co-targets, as sorted by *increasing* value of relational strength to t, and $|C_t|$ is the number of t's co-targets. That is, the least related co-target of t has $rank_t(c) = 1$, and the most strongly related has $rank_t(c) = |C_t|$.

If neither rank is defined, then $score(a, b) = 0$. If exactly one of these ranks is defined, we take 75% of the defined term, rather than allowing it to be averaged with zero.

In Table 1, we compare our correlation results with those presented in a review by Budanitsky and Hirst [4] as well as three state of the art studies published since then. Higher values indicate better correlation with the human-assigned scores; 1.0 would indicate a perfect fit. The average correlation of ten individual human evaluations to the M&C scores comes from a replication of the study by Resnik [27].

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, and is comfortably within the realm of human performance. We caution, however, that high correlation on this task, and particularly scores that exceed average hu-

man correlation, might indicate that a measure is failing to capture semantic relatedness beyond that of similarity.

5. FROM RELATIONAL STRENGTH TO CATEGORICAL RELATEDNESS

We now present an algorithm for categorically determining 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 word 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 (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 their relational strength to some target, we have an exceptionally good idea of which nouns are strongly related to the target (those at the top of the list) and those which 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}), 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 Algorithm 1). 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 relations being discovered.

Algorithm 1 FINDRELATEDNOUNS(t)

Require: A target noun t.
Ensure: Set of pairs (t, c) such that t and c are semantically related.

```
S_0 ← ∅
noGain ← 0
for n = 20 to 100 do
    S ← {(t, c)|c ∈ m_x(t)}
    if |S| > |S_0| then
        noGain ← 0
    else
        noGain ← noGain + 1
    end if
    if noGain ≥ 5 then
        break
    end if
    S_0 ← S
end for
return S_0
```

Table 2: Summary of Statistics for the Semantic Network of Related Nouns

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)	120,588
Average Degree of Target Nodes (average number of nouns to which each target is related)	30.74
Average Threshold of Target Nouns (average of target thresholds determined in Algorithm 1)	28.19%

Upon termination of the algorithm, we admit all ordered pairs in S_0 to the network.

The algorithm exhibits several important properties worth mentioning. First, the algorithm accounts for the fact that some nouns are more promiscuous 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.

Secondly, 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). This stringent requirement causes us to miss some related noun pairs, but provides very strong evidence for the relatedness of pairs that do gain admission to the network.

5.1 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 above (Table 2).

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, 116,996 are redundant, in that they are the symmetric images of pairs already included in the network. Thus, the network has 120,588 distinct undirected edges. Each target noun is related, on average, to 30.74 other nouns.

To evaluate the precision of these relations, we asked three

Table 3: Judges' Evaluations of Accuracy on Related and Unrelated Noun Pairs

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

judges with backgrounds in computational linguistics 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 the following exemplars of semantic relatedness, which we hand picked from the network: (astronomer, observatory), (crime, prevention), (automobile, gasoline), (phone, signal), (penguin, tuxedo), (prison, lawyer), (tendon, cartilage), (string, output), and (desert, habitat).

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, none of whom had direct ties to this research. The results of their evaluations are summarized above in Table 3.

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 30.74 nouns while maintaining precision in excess of 95% is indicative of broad and accurate coverage of semantic relatedness.

To illustrate the quality of the relations discovered by our algorithm, we have included a discussion of the semantic network surrounding the monosemous nouns (concepts) *astronomer* and *tennis* in Section 7.

6. FROM NOUNS TO CONCEPTS

Once we have established relatedness between nouns, we automatically disambiguate them to their corresponding noun senses in WordNet 3.0. For this purpose, 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 three methods described below fail to disambiguate a noun, we default to its most frequent sense in WordNet.

6.1 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* in WordNet (planet#{1, 3}, star#{1, 3}, minor_planet#1, quasar#1), another under the purview of *scientist* (mathematician#1, physicist#1, chemist#1), and so on.[7]

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.

6.2 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 related nouns. 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 gloss for star#1: "(astronomy) a celestial body of hot gases that radiates energy derived from thermonuclear reactions in the interior."

6.3 Selectional Preference Method

Next we use Resnik's selectional association measure [28] 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:

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

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

$$\begin{aligned} D_{KL} &= D(P(C|t)||P(C)) \\ &= \sum_{c \in C} P(c|t) \log \frac{P(c|t)}{P(c)} \end{aligned}$$

Here, however, C is no longer the class of nouns co-occurring with t. Rather, C is the set of concepts in WordNet denoted by the *monosemous* nouns that are related to 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 in both probability distributions and apportions credit for a noun evenly across all its senses. By focusing only on

[7]We denote sense n of a noun by noun#n, or multiple senses with, e.g., noun#$\{m, n\}$.

Table 4: Selectional Preferences Derived from Monosemous Co-Targets of "Unicorn"

WordNet Class (c)	A(unicorn, c)
monster#1	12.3501575599
mythical_being#1	12.3501575599
mythical_monster#1	12.3501575599
mermaid#1	10.7337066300
goblin#1	10.5188861760
utensil#1[8]	9.1129996410
imaginary_being#1	8.7033691549
imagination#1	8.7033691549
creativity#1	8.2372712617
vessel#3[8]	7.4946279622
evil_spirit#1	7.3265447433
spirit#4	7.3265447433
spiritual_being#1	6.7626973121
...	...
whole#2	-1.0379049622
artifact#1	-1.0701173602
object#1	-1.2809861573
physical_entity#1	-1.4906802554

monosemous nouns in this approach, we eliminate the noise introduced by the ambiguity of polysemous nouns. Once we have the selectional preferences derived from our target's monosemous nouns, we use them to preferentially disambiguate our polysemous nouns.

Consider, for example, the categories in WordNet with the highest selectional association with the monosemous noun "unicorn" (Table 4). Among these selectional preferences we find mythical_monster#1, imaginary_being#1, and spiritual_being#1, which do not appear as co-targets of "unicorn," but do categorize many of the monosemous co-targets of "unicorn," such as "griffin," "goblin," "mermaid," "leprechaun," and "minotaur," among others.

These selectional preferences are applied, in decreasing order of selectional strength, to each sense of the target's polysemous related nouns, which are disambiguated to the sense or senses categorized by the first such selectional 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. The three senses of "phoenix" that are excluded here are phoenix#1 (the capital city of Arizona), phoenix#2 (the taxonomic group *genus Phoenix*), 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_1 is the strongest selectional preference from our list that disambiguates some polysemous noun related to t, and $A(t, c_1)$ is less than the average 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.

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.

[8]From "teapot," vis-à-vis Russell's teapot and pink unicorns.

Table 5: Summary of Statistics for the Semantic Network of Related Concepts

Target Nouns (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 senses represented in network; includes both target and co-target senses)	27,312
Edges (number of related sense pairs; (a, b) and (b, a) are not counted as distinct sense pairs)	84,086
Average Degree of Target Nodes (average number of noun senses to which each monosemous target is related)	27.81%

6.4 Evaluation

We have used these methods to disambiguate the polysemous nouns related to monosemous targets occurring at least 1,500 times in the corpus. There are 3,024 such target nouns, heading up 76,264 of our related noun pairs from the previous section. 36,385 of these pairs associate two monosemous nouns. The remaining 39,879 connect our monosemous targets to polysemous nouns that must be disambiguated. Statistics for the resulting semantic network of related concepts are given above in Table 5.

To test our precision at disambiguation, we randomly selected 50 pairs from among those used to build this network and presented them to our three judges with the gloss and taxonomic categorization of each sense of the polysemous nouns. The judges were asked to grade the relation of each sense to its monosemous target, using the following scale: (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 6, below).

The first column ($grade \geq 4$) 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.

Given that 47.7% of the edges in our network connect two

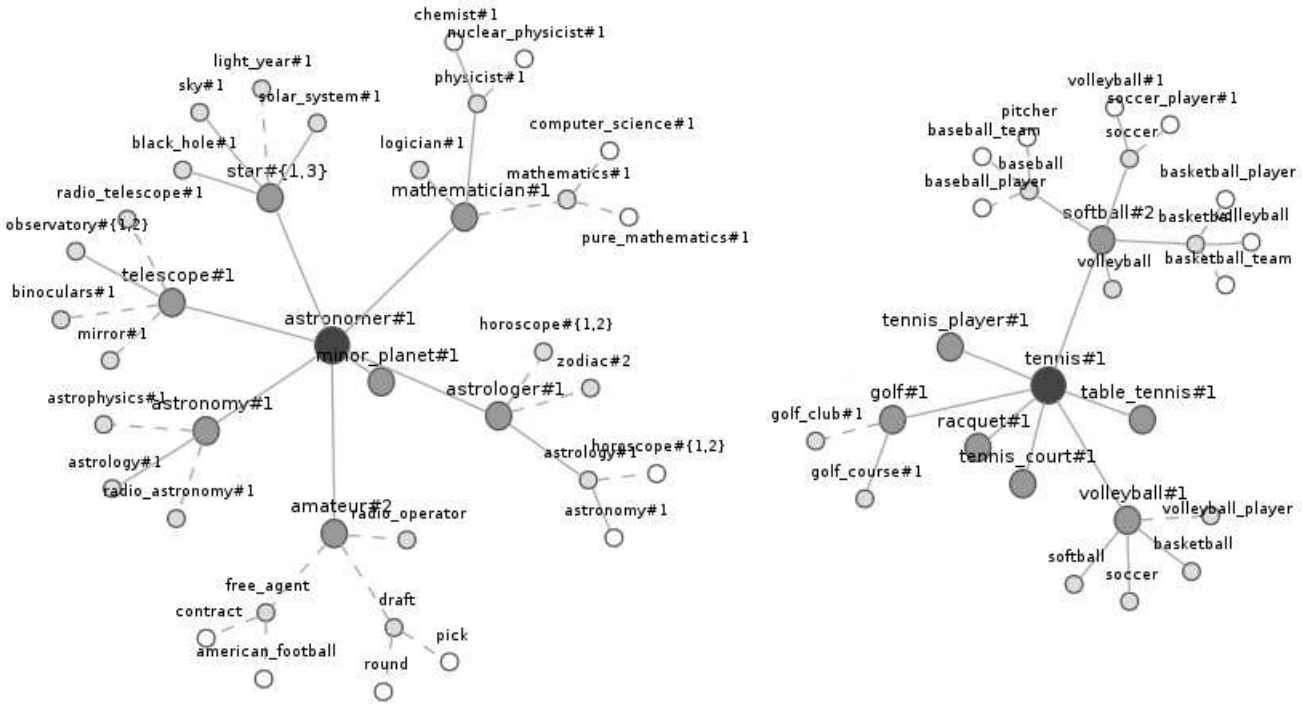

Figure 1: A partial spreading activation view of the concepts related to the monosemous *astronomer* and *tennis*, explicated in Section 7.

Table 6: Precision of Our System's Disambiguation Results, as Compared with Judges' Manually Disambiguated Senses

Judge	$grade \geq 4$	≥ 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%

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 accuracy of the concept-to-concept associations in our semantic network to be 92.15%.

7. EXCERPTS FROM THE SEMANTIC NETWORK: A DISCUSSION

The graphs in Figure 1 are abbreviated excerpts from the semantic network of related concepts for the monosemous nouns "astronomer" and "tennis." *Astronomer* is related to 44 distinct concepts in our semantic network (listed in full in Table 7, below), and *tennis* is related to 80. For the sake of clarity, we present only a small sampling of those related nouns graphically. Furthermore, to avoid messy edge crossings in the graphs, we do not show the inter-relatedness between the concepts related to each 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.)

The target concepts' nodes in the graph are dark gray (astronomer#1 and tennis#1). We provide a sampling of their related terms in medium gray. In turn, those concepts are related to concepts in light gray, 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 smaller 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}.

Some nouns are not yet disambiguated because they are related to concepts denoted by polysemous nouns. We have included a sampling of these nouns to give a fair indication of the current state of the network. We also see how these might be easily 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 relation between softball#2 and "volleyball" as (softball#2, volleyball#1).

There are also cases in which polysemous nouns are related to disambiguated concepts in the graph, 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 relations of semantic similarity (e.g., the relation of astronomer#1 to mathematician#1 and astrophysicist#1), it also discovers many relations beyond similarity, including concepts re-

Table 7: All concepts related to *astronomer*

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

lated 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 relations 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 the WordNet ontology would erroneously associate 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.

8. CONCLUSIONS

We have automatically acquired a semantic network of related concepts. The network is derived from relatedness between nouns co-occurring in Wikipedia texts, which are automatically disambiguated to their corresponding WordNet 3.0 noun senses (i.e., concepts). At present, monosemous noun targets form the basis of the network, each being related to an average of 27.81 concepts (denoted both by monosemous and polysemous nouns). The network currently relates 17,543 nouns, with 27,312 distinct noun senses among them, and is available for download on-line.

There are several potential applications for this resource, including semantic interpretation, exploration of spreading activation mechanisms [5], contextual frameworks for computer vision (cf. Torralba et al. [32]), noun sense disambiguation, question answering systems, query prediction, and user profiling for providing recommendations in multimedia content delivery systems.

In future work, we expect to continue expanding and refining the semantic network. Polysemous targets and targets that occur fewer than 1,500 times in Wikipedia need to be incorporated into both the network of nouns and the network of related concepts. We are investigating the feasibility of applying our algorithm to these targets and using the existing semantic network to guide (i.e., bootstrap) the process, which is more error prone with nouns that occur infrequently in the corpus and does not currently resolve ambiguity of polysemous-to-polysemous noun relations.

9. ACKNOWLEDGMENTS

This research was supported in part by the NASA Engineering and Safety Center under Grant/Cooperative Agreement NNX08AJ98A.

10. REFERENCES

[1] E. Agirre and O. L. de Lacalle. Publicly available topic signatures for all WordNet nominal senses. In *Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC)*, pages 1123–1126, Lisbon, Portugal, 2004.

[2] M. Berland and E. Charniak. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 57–64, College Park, MD, 1999.

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

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

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

[6] M. Cuadros and G. Rigau. KnowNet: building a large net of knowledge from the Web. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 161–168, 2008.

[7] D. Davidov and A. Rappoport. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 297–304, Sydney, Australia, 2006. Association for Computational Linguistics.

[8] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

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

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

[11] J. Gorman and J. R. Curran. 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)*, pages 361–368, Sydney, Australia, 2006. Association for Computational Linguistics.

[12] Z. S. Harris. Distributional structure. In J. J. Katz, editor, *The Philosophy of Linguistics*, pages 26–47. Oxford University Press, 1985.

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

[14] G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332. MIT Press, 1998.

[15] T. Hughes and D. Ramage. Lexical semantic relatedness with random graph walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 581–589, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

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

[17] C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 265–283. MIT Press, 1998.

[18] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pages 296–304, Madison, WI, 1998. Morgan Kaufmann.

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

[20] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM)*, pages 233–242, Lisbon, Portugal, 2007. ACM.

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

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

[23] P. Pantel and M. Pennacchiotti. 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)*, pages 113–120, Sydney, Australia, 2006. Association for Computational Linguistics.

[24] S. Patwardhan and T. Pedersen. 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*, pages 1–8, Trento, Italy, 2006.

[25] S. P. Ponzetto and R. Navigli. Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence (IJCAI)*, pages 2083–2088, Pasadena, CA, 2009. Morgan Kaufmann.

[26] M. R. Quillian. *Semantic Memory*. In M. Minsky (ed.), Semantic Information Processing, MIT Press, Cambridge, MA, 1968.

[27] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 448–453, Montreal, QC, 1995. Morgan Kaufmann.

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

[29] H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[30] M. Strube and S. P. Ponzetto. Wikirelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 1419–1424, Boston, MA, 2006. AAAI Press.

[31] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago : a large ontology from Wikipedia and WordNet. Research Report MPI-I-2007-5-003, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, December 2007.

[32] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Proceedings of the 9th International Conference on Computer Vision (ICCV)*, pages 273–280, Nice, France, 2003.

[33] P. D. Turney. Expressing implicit semantic relations without supervision. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 313–320, Sydney, Australia, 2006. Association for Computational Linguistics.

[34] P. D. Turney. Similarity of semantic relations. *Computational Linguistics*, 32:379–416, 2006.

[35] H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, and G. Attardi. Ranking very many typed entities on Wikipedia. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1015–1018, Lisbon, Portugal, 2007. ACM.