Adding Dense, Weighted Connections to WORDNET

Jordan Boyd-Graber and Christiane Fellbaum and Daniel Osherson and Robert Schapire
Princeton University

October 9, 2005

Abstract

WORDNET, a ubiquitous tool for natural language processing, suffers from sparsity of connections between its component concepts (synsets). Through the use of human annotators, a subset of the connections between 1000 hand-chosen synsets was assigned a value of "evocation" representing how much the first concept brings to mind the second. These data, along with existing similarity measures, constitute the basis of a method for predicting evocation between previously unrated pairs.

Submission Type: Long Article

Topic Areas: Extending WORDNET

Author of Record: Jordan Boyd-Graber, jbg@princeton.edu

Under consideration for other conferences (specify)? None

Adding Dense, Weighted Connections to WORDNET

Abstract

WORDNET, a ubiquitous tool for natural language processing, suffers from sparsity of connections between its component concepts (synsets). Through the use of human annotators, a subset of the connections between 1000 hand-chosen synsets was assigned a value of "evocation" representing how much the first concept brings to mind the second. These data, along with existing similarity measures, constitute the basis of a method for predicting evocation between previously unrated pairs.

1 Introduction

WORDNET is a large electronic lexical database of English. Originally conceived as a full-scale model of human semantic organization, it was quickly embraced by the Natural Language Processing (NLP) community, a development that guided its subsequent growth and design. WORDNET has become the lexical database of choice for NLP; Kilgariff (Kilgarriff, 2000) notes that "not using it requires explanation and justification." WORDNET's popularity is largely due to its free public availability and its broad coverage.

WORDNET already has a rich structure connecting its component synonym sets (synsets) to each other. Noun synsets are interlinked by means of hyponymy, the *super-subordinate* or *is-a relation*, as exemplified by the pair `[poodle]-[dog]`.[1] Meronymy, the *part-whole* or *has-a* relation, links noun synsets like `[tire]` and `[car]` (Miller, 1998). Verb synsets are connected by a variety of lexical entailment pointers that express manner elaborations `[walk]-[limp]`, temporal relations `[compete]-[win]`, and causation `[show]-[see]` (Fellbaum, 1998). The links among the synsets structure the noun and verb lexicons into hierarchies, with noun hierarchies being considerably deeper than those for verbs.

WORDNET appeals to the NLP community because these semantic relations can be exploited for word sense disambiguation (WSD), the primary barrier preventing the development of practical information retrieval, machine translation, summarization, and language generation systems. Although most word forms in English are monosemous, the most frequently occurring words are highly polysemous. Resolving the ambiguity of a polysemous word in a context can be achieved by distinguishing the multiple senses in terms of their links to other words. For example, the noun `[club]` can be disambiguated by an automatic system that considers the superordinates of the different synsets in which this word form occurs: `[association]`, `[playing card]`, and `[stick]`. It has also been noted that directly antonymous adjectives share the same contexts (Deese, 1964); exploiting this fact can help to disambiguate highly polysemous adjectives.

1.1 Shortcomings of WORDNET

Statistical disambiguation methods, relying on cooccurrence patterns, can discriminate among word senses rather well, but not well enough for the level of text understanding that is desired (Schütze, 1998); exploiting sense-aware resources, such as WORDNET, is also insufficient (McCarthy et al., 2004). To improve such methods, large manually tagged training corpora are needed to serve as "gold standards." Manual tagging, however, is time-consuming and expensive, so large semantically annotated corpora do not exist. Moreover, people have trouble selecting the best-matching sense from a dictionary for a given word in context. (Fellbaum and Grabowski, 1997) found that people agreed with a manually created gold standard on average 74% of the time, with higher disagreement rates for

[1] Throughout this article we will follow the convention of using a single word enclosed in square brackets to denote a synset. Thus, `[dog]` refers not just to the word dog but to the set – when rendered in its entirety – consisting of {`dog, domestic dog, canis familaris`}.

more polysemous words and for verbs as compared to nouns. Confidence scores mirrored the agreement rates.

In the absence of large corpora that are manually and reliably disambiguated, the internal structure of WORDNET can be exploited to help discriminate senses, which are represented in terms of relationships to other senses. But because WORDNET's network is relatively sparse, such WSD methods achieve only limited results.

In order to move beyond these meager beginnings, one must be able to use the entire context of a word to disambiguate it; instead of looking at only neighboring nouns, one must be able to compare the relationship between any two words via a complete comparison. Moreover, the character of these comparisons must be quantitative in nature. This paper serves as a framework for the addition of a complete, directed, and weighted relationship to WORDNET. As a motivation for this additon, we now discuss three fundamental limitations of WORDNET 's network.

No cross-part-of-speech links WORDNET consists of four distinct semantic networks, one for each of the major parts of speech. There are no cross-part-of-speech links.[2] The lack of syntagmatic relations means that no connection can be made between entities (expressed by nouns) and their attributes (encoded by adjectives); similarly, events (referred to by verbs) are not linked to the entities with which they are characteristically associated. For example, the intuitive connections among such concepts as `[traffic]`, `[congested]`, and `[stop]` are not coded in WORDNET.

Too few relations WORDNET's potential is limited because of its small number of relations. Increasing the number of arcs connecting a given synset to other synsets not only refines the relationship between that synset and other meanings but also allows a wider range of contexts (that might contain a newly connected word) to help disambiguation.

(Mel'cuk and Zholkovsky, 1998) propose several dozen lexical and semantic relations not included in WORDNET, such as "actor" (`[book]-[writer]`) and "instrument" (`[knife]-[cut]`). But many associations among words and synsets cannot be represented by clearly labeled arcs. For example, no relation proposed so far accounts for the association between pairs like `[tulip]` and `[Holland]`, `[sweater]` and `[wool]`, and `[axe]` and `[tree]`. It is easy to detect a relation between the members of these pairs, but the relations cannot be formulated as easily as hyponymy or meronymy. Similarly, the association between `[chopstick]` and `[Chinese restaurant]` seems strong, but this relation requires more than the kind of simple label commonly used by ontologists.

Some users of WORDNET have tried to make up for the lack of relations by exploiting the definition and/or the illustrative sentence that accompany each synset. For example, in an effort to increase the internal connectivity of WORDNET, Mihalcea and Moldovan (Mihalcea and Moldovan, 2001) automatically link each content word in WORDNET's definitions to the appropriate synset; the PrincetonWORDNET team is currently performing the same task manually. But even this significant increase in arcs leaves many synsets unconnected; moreover, it duplicates some of the information already contained in WORDNET, as in the many cases where the definition contains a monosemous superordinate. Another way to link words and synsets across parts of speech is to assign them to topical domains, as in (Magnini and Cavaglia, 2000). WORDNET contains a number of such links, but the domain labels are not a well-structured set. In any case, domain labels cannot account for the association of pairs like `[Holland]` and `[tulip]`. In sum, attempts to make WORDNET more informative by increasing its connectivity have met with limited success.

[2] WORDNET does contain arcs among many words from different syntactic categories that are semantically and morphologically related, such as `[operate]`, `[operator]`, and `[operation]` (Fellbaum and Miller, 2003). However, semantically related words like operation, perform, and dangerous are not interconnected in this way, as they do not share the same stem.

No weighted arcs A third shortcoming of WORDNET is that the links are qualitative rather than quantitative. It is intuitively clear that the semantic distance between the members of hierarchically related pairs is not always the same. Thus, the synset [run] is a subordinate of [move], and [jog] is a subordinate of [run]. But [run] and [jog] are semantically much closer than [run] and [move]. WORDNET currently does not reflect this difference and ignores the fact that words – labels attached to concepts – are not evenly distributed throughout the semantic space covered by a language. This limitation of WORDNET is compounded in NLP applications that rely on semantic distance measures where edges are counted, e.g., (Jiang and Conrath, 1997) and (Leacock and Chodorow, 1998). Recall that adjectives in WORDNET are organized into pairs of direct antonyms (e.g., long-short) and that each member of such a pair is linked to a number of semantically similar adjectives such as [lengthy] and [elongated], and [clipped] and [telescoped], respectively. The label "semantically similar," however, hides a broad scale of semantic relatedness, as the examples indicate. Making these similarity differences explicit would greatly improve WORDNET's content and usefulness for a variety of NLP applications.

2 An Enrichment of WORDNET

To address these shortcomings, we are working to enhance WORDNET by adding a radically different kind of information. The idea is to add quantified, oriented arcs between pairs of synsets, e.g., from {car, auto} to {road, route}, from {buy, purchase} to {shop, store}, from {red, crimson, scarlet} to {fire, flame}, and also in the opposite direction. Each of these arcs will bear a number corresponding to the strength of the relationship. We chose to use the concept of evocation – how much one concept evokes or brings to mind the other – to model the relationships between synsets.

[cat] brings [dog] to mind, just as [swimming] evokes [water], and the word [cunning] evokes [cruel]. Such association of ideas has been a prominent feature of psychological theories for a long time (Lindzey, 1936). It appears to be involved in low-level cognitive phenomena such as semantic priming in lexical decision tasks (McNamara, 1992) and high-level phenomena like diagnosing mental illness (Chapman and Chapman, 1967). Its role in the on-line disambiguation of speech and reading has been explored by (Swinney, 1979), (Tabossi, 1988), and (Rayner et al., 1983), among others.

Evocation is a meaningful variable for all pairs of synsets and seems easy for human annotators to judge. In this sense our extension of WORDNET will have no overlap with knowledge repositories like CYC (Lenat, 1995) but can be viewed as complementary.

2.1 Collecting Ratings

We hired 20 Princeton undergraduates during the 2004-2005 academic year to rate evocation in 120,000 pairs of synsets. The synsets were drawn randomly from all pairs defined from a set of 1000 "core" synsets compiled by the investigators. The core synsets were compiled as follows. The most frequent strings (nouns, verbs, and adjectives) from the BNC were selected. For each string, the WORDNET synsets containing this string were extracted. Two of the authors then went over the list of synsets and selected those senses of a given string that seemed the most salient and basic. The initial string is the "head word" member of the synset; the synonyms function merely to identify the concept expressed by the central string. To reflect the distribution of parts of speech in the lexicon, we chose 642 nouns, 207 verbs, and 151 adjectives.

Our raters were first instructed about the evocation relation and were offered the following explications:

1. Evocation is a relation between meanings as expressed by synsets and not a relation between words; examples were provided to reinforce this point.

2. One synset evokes another to the extent that thinking about the first brings the second to mind. (Examples were given, such as [government] evoking [register] for the appropriate synsets including these terms.)

3. Evocation is not always a symmetrical relation (for example, [dollar] may evoke [green] more than the reverse).

4. The task is to estimate the extent to which one synset brings to mind another in the general undergraduate population of the United States; idiosyncratic evocations caused by the annotator's personal history are irrelevant.

5. It is expected that many pairs of synsets will produce no evocation at all (connections between synsets must not be forced).

6. There are multiple paths to evocation, e.g.:

[rose] - [flower]	(example)
[brave] - [noble]	(kind)
[yell] - [talk]	(manner)
[eggs] - [bacon]	(co-occurrence)
[snore] - [sleep]	(setting)
[wet] - [desert]	(antonymy)
[work] - [lazy]	(exclusivity)
[banana] - [kiwi]	(likeness)

7. In no case should evocation be influenced by the sounds of words or their orthographies (thus, [rake] and [fake] do not evoke each other on the basis of sound or spelling).

8. The integers from 0 to 100 are available to express evocation; round numbers need not be used.

Raters were familiarized with a computer interface that presented pairs of synsets (each as a list with the highest frequency word first and emphasized; we will refer to this word as the "head word"). The parts of speech corresponding to each synset in a pair were also shown. Presenting entire synsets instead of single words eliminates the risk of confusion between rival senses of polysemous words. Between the two synsets appeared a scale from 0 to 100; 0 represented "no mental connection," 25 represented "remote association," 50 represented "moderate association," 75 represented "strong association," and 100 represented "brings immediately to mind."

As final preparation, each rater was asked to annotate two sets of 500 randomly chosen pairs of synsets (distinct from the pairs to be annotated later). Both sets had been annotated in concert by two of the investigators. The first served as a training set: the response of the annotator-trainee to each pair was followed by the "gold standard" rating obtained by averaging the ratings of the investigators. The second served as a test set: no feedback was offered, and we calculated the Pearson correlation between the annotators rating versus our own. The median correlation obtained on the test set by the 24 annotators recruited for the project was .72; none scored lower than .64.

Unbeknownst to the annotators, some pairs were presented twice, on a random basis, always on different sessions. The average correlation between first and second presentations was .70 for those annotators who generated at least 100 test-retest pairs.

2.2 Analysis of Ratings

Every pair of synsets were evaluated by at least three people (additional annotations were sometimes collected to test consistency of annotator judgments), and as one might expect from randomly selecting pairs of synsets, most (67%) of the pairs were rated by every annotator as having no mental connection (see Figure 1). The ratings were usually consistent across different annotators; the average standard deviation for pairs where at least one rater labeled it as non-zero was 9.25 (on a scale from 0 to 100).

Because there is an active vein of research comparing the similarity of synsets within WORDNET, we present the Spearman rank order coefficient ρ for a variety of similarity measures. We use WORDNET::Similarity (Patwardhan et al., 2004) to provide WORDNET-based measures (e.g. (Leacock and Chodorow, 1998) and (Lesk, 1986) applied to WORDNET glosses). In addition, Infomap (Peters, 2005) is used to provide the cosine between LSA vectors (Landauer and Dumais, 1997) created from the British National Corpus (BNC). For every word for which the program computes a context vector, the 2000 words closest to it are stored. We only consider pairs where both words had context vectors and one was within the 2000 closest vectors to the other. Other words can be safely assumed to have a small value for the cosine of the angle between them.

The Leacock-Chodorow (LC) and Path measures

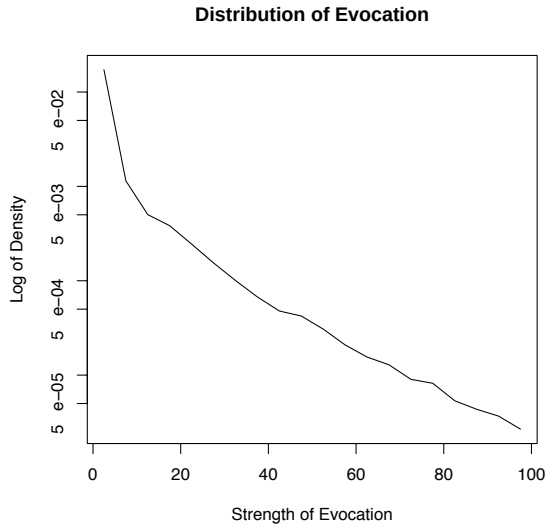

Figure 1: Logarithmic distribution of evocation ratings

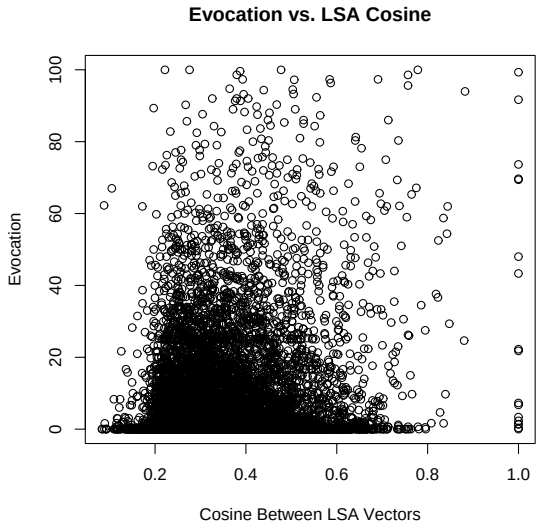

Figure 2: The relationship between LSA cosine vectors and evocation for pairs that were within the 1000 closest word vectors.

require connected networks to compute distance, so those values were only computed for noun-noun and verb-verb pairs. The correlations achieved by these various methods are displayed in the following table.

Metric	**Subset (# Pairs)**	ρ
Lesk	All (119668)	0.008
Path	Verbs (4801)	0.046
LC	Nouns (49461)	0.130
Path	Nouns (49461)	0.130
LSA	Closest 2000 (15285)	0.131

Although there is evidence of a slight monotonic increase of evocation with these similarity measures, the lack of correlation shows that there is not a strong relationship. Our results therefore demonstrate that evocation is an empirical measure of some aspect of semantic interaction not captured by these similarity methods.

For each of the similarity measures, a wide range of possible evocation values is observed for the entire gamut of the similarity range. One typical relationship is shown in Figure 2, which shows evocation vs. the cosine between LSA vectors. The only exception is that the similarity measures tend to do very well in determining synsets with little evocation; for low values of similarity, the evocation is reliably low.

There are several reasons why these measures fail to predict evocation. Many of the WORDNET measures are limited to only a small subset of the synset pairs that are of interest to us; the path and Leacock-Chodorow metrics, for instance, are useable only *within* the components of WORDNET that have well defined *is-a* hierarchies.

Although the LSA metric is free of this restriction, it is really a comparison of the relatedness of *strings* rather than synsets. The vector corresponding to the string `fly`, for example, encompasses all of its meanings for multiple parts of speech; because many of the words under consideration are polysemous, LSA could therefore suggest relationships between synsets that correspond to meanings other than the intended one.

Finally, all these measures are symmetric, but the evocation ratings are not (see Figure 3). Of the 3302 pairs where both directions between pairs were rated by annotators and where one of the ratings was non-zero, the correlation coefficient between directions was 0.457. While there is a strong symmetric component, there are many examples where asymmetry is observed.

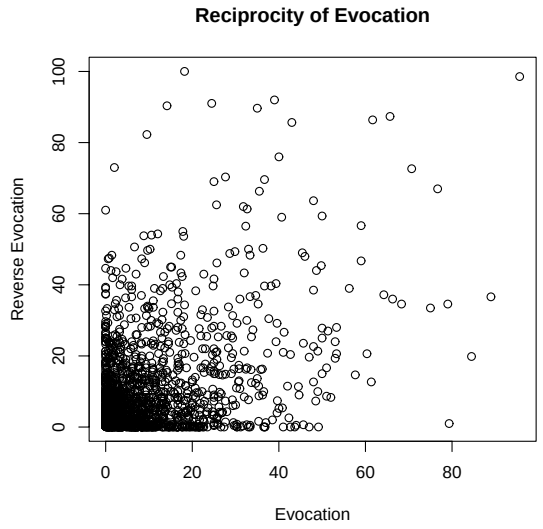

Figure 3: The evocation observed between synset pairs in opposite directions.

2.3 Extending Ratings

Before these data can be used for NLP applications, it must be possible to query the evocation between arbitrary synsets. Our goal is to create a means of automatically judging the evocation between synsets while avoiding the impractical task of hand annotating the links between all 10^{10} pairs of synsets. Our method attempts to leverage the disparate strengths of the measures of semantic distance discussed above in addition to measures of similarity between context vectors (culled from the BNC) for individual words.

These context vectors were created by searching the BNC for the head word of each synset and lemmatizing the results. Stop words were removed from the results, and frequencies were tabulated for words at most n words away (both right and left) from the head word found in the sentence for $n = 2, 4, 6, 8$. Because the BNC is tagged, it allows us to specify the part of speech of the target words. Although the tagging does not completely alleviate the problem of multiple senses being represented by the context vectors, it does eliminate the problem when the senses have different functions.

We created a range of features from each pair of context vectors including the relative entropy, cosine, L_1 distance, L_2 distance, and the number of words in the context vectors of both words (a full listing appears in the table below). Descriptive statistics for the individual context vectors were also computed. It is hoped that the latter information, in addition to relative entropy, would provide some asymmetric foundation for the prediction of evocation links.

WORDNET- based	BNC-derived
Jiang-Contrath	Relative Entropy
Path	Mean
Lesk	Variance
Hirst-St. Onge	L_1 Distance
Leacock-Chodorow	L_2 Distance
Part of Speech	Correlation
	Contextual Overlap
	LSA-vectors Cosine
	Frequency

These were exploited as features for the BoostTexter algorithm (Schapire and Singer, 2000), which learns how to automatically apply labels to each example in a dataset. In this case, we broke the range of evocations into five labels: $\{x \geq 0, x \geq 1, x \geq 25, x \geq 50, x \geq 75\}$. Because there are so many ratings with a value of zero, we created a special category for those values; the other categories were chosen to correspond to the visual prompts presented to the raters during the annotation process. Another option would have been to divide up the range to have roughly equal frequencies of evocation; given the large numbers of zero annotations, however, this would lead to very low resolution for higher – and more interesting – levels of evocation.

Given the probabilities for membership in each of the range of values, this allows us to compute an estimate of the expected predicted evocation from our coarse probability distribution. We randomly held out 20% of the labeled data and trained the learning algorithm on the remaining data. Because it is reasonable to assume that different parts of speech will have different models of evocation and because WORDNET and WORDNET-derived similarity measures provide different data for different parts of speech, we trained the algorithm on each of the six pairs of parts of speech as well as the complete, undivided dataset. The mean squared errors (the square of the predicted minus the correct level

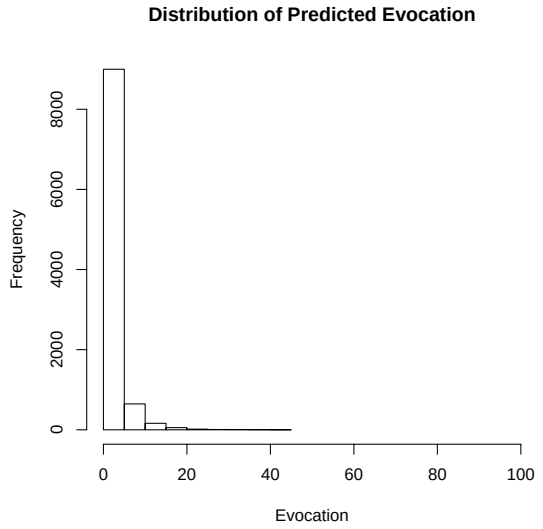

Figure 4: After training, the learning algorithm predicted a distribution of evocation on the test data consistent with the log-transformed distribution of evocations displayed in Figure 1.

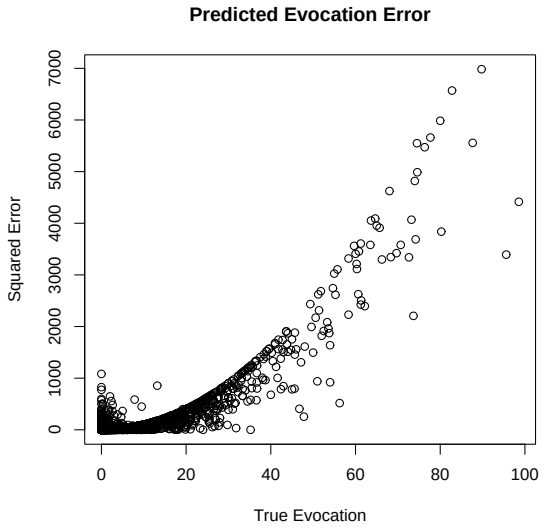

Figure 5: Although most of the data are clustered around $(0, 0)$, there are many data points for which high levels of evocation were assigned to zero evocation data (the spike to the left) and some data of high evocation that was assigned to zero levels of evocation (the line following x^2). For high levels of evocation, the predictions become less accurate.

of evocation) and the sizes of the training corpora for each are provided below.

Dataset	Mean Squared Error	Training Size
AA	89.9	2202
VV	83.2	3861
NV	80.7	25483
AN	67.2	18471
All	63.0	95603
AV	53.8	6022
NN	49.8	39564

A naïve algorithm would simply label all pairs as having no evocation; this would yield a 73.0 mean squared error for the complete data set, higher than our mean squared error of 63.0 on the complete dataset and 53.8 on noun-noun pairs (as above). Not only does the algorithm perform better using this metric, but these predictions also, as one would hope, have a distribution much like that observed for the original evocation data (see Figure 4). Taken together, it confirms that we are indeed exploiting the features to create a reasonably consistent model of evocation, particularly on noun-noun pairs, which have the richest set of features.

We hope to further refine the algorithm and the feature set to improve prediction further; even for the best prediction scheme, for noun-noun pairs, many pairs with zero evocation were assigned to high levels of evocation (see Figure 5). It is heartening, however, to note that the algorithms successfully predicted many pairs with moderate levels of evocation.

3 Future Work

Although our work with developing a learning algorithm to predict evocation is still in its preliminary stages, a foundation is in place for creating a complete, directed, and weighted network of interconnections between synsets within WORDNET that represent the actual relationships observed by real language users. Once our automatic system has shown itself to be reliable, we intend to extend its application beyond the 1000 synsets selected for this study: first by extending it to 5000 synsets judged central to a basic vocabulary and then to the rest of WORDNET.

The real test of the efficacy of any addition to WORDNET remains how well it performs in tasks that require sense disambiguation. It is hoped that an enriched network in WORDNET will be able to improve disambiguation methods that use WORDNET as a tool.

4 Acknowledgments

The authors would like to acknowledge the support of the National Science Foundation (grant Nos. 0414072 and 0530518) for the funding to gather the initial annotations and Princeton Computer Science for fellowship support. The authors wish to especially acknowledge the work of our undergraduate annotators for their patience and hard work in collecting these data and Ben Haskell, who helped compile the initial list of synsets.

References

L. Chapman and J. Chapman. 1967. Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, (72):193–204.

J. Deese. 1964. The associative structure of some english adjectives. *Journal of Verbal Learning and Verbal Behavior*, (3):347–357.

C. Fellbaum and J. Grabowski. 1997. Analysis of a hand-tagging task. In *Proceedings of the ACL/Siglex workshop*. Association for Computational Linguistics.

C. Fellbaum and G. A. Miller. 2003. Morphosemantic links in WordNet. *Traitement automatique de langue*.

C. Fellbaum, 1998. *WordNet : An Electronic Lexical Database*, chapter A semantic network of English verbs. MIT Press, Cambridge, MA.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, Taiwan.

A. Kilgarriff. 2000. Review of WordNet : An electronic lexical database. *Language*, (76):706–708.

T. Landauer and S. Dumais. 1997. Solutions to Plato's problem: The latent semantic analsyis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, (104).

C. Leacock and M. Chodorow, 1998. *WordNet : An Electronic Lexical Database*, chapter Combining local context and WordNet similarity for word sense identification. MIT Press, Cambridge, MA.

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

M. Lesk. 1986. Automatic sense disambiguation using machine-readable dictionaries. In *Proceedings of SIGDOC*.

G. Lindzey, editor. 1936. *History of Psychology in Autobiography*. Clark University Press, Worcester, MA.

B. Magnini and G. Cavaglia. 2000. Integrating subject field codes into WordNet. In *Proceedings of LREC*, pages 1413–1418, Athens, Greece.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Barcelona, Spain.

T. P. McNamara. 1992. Priming and constraints it places on theories of memory and retrieval. *Psychological Review*, pages 650–662.

I. Mel'cuk and A. Zholkovsky, 1998. *Relational Models of the Lexicon*, chapter The explanatory combinatorial dictionary. Cambridge University Press, Cambridge.

R. Mihalcea and D. Moldovan. 2001. Extended WordNet : Progress report. In *Proceedings of the NAACL Workshop on WordNet and other lexical resources*, pages 95–100, Pittsburgh, PA.

G. Miller, 1998. *WordNet : An Electronic Lexical Database*, chapter Nouns in WordNet. MIT Press, Cambridge, MA.

S. Patwardhan, T. Pedersen, and J. Michelizzi. 2004. WordNet::similarity–measuring the relatedness of concept. In *Proceedings of the 19th National Conference on Artificial Intelligence*, page 25.

S. Peters. 2005. Infomap NLP software: an open-source package for natural language processing. Webpage, October.

K. Rayner, M. Carlson, and L. Frazier. 1983. The interaction of syntax and semantics during sentence processing — Eye-movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior*, 22(3):358–374.

R. Schapire and Y. Singer. 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

D. Swinney. 1979. Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, (18):645–659.

P. Tabossi. 1988. Acessing lexical ambiguity in different types of sentential context. *Journal of Memory and Language*, (27):324–340.