Evaluating Semantic Metrics on Tasks of Concept Similarity

Hansen A. Schwartz and **Fernando Gomez**
School of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL 32816, USA
{hschwartz, gomez}@cs.ucf.edu

Abstract

This study presents an evaluation of WordNet-based semantic similarity and relatedness measures in tasks focused on concept similarity. Assuming similarity as distinct from relatedness, the goal is to fill a gap within the current body of work in the evaluation of similarity and relatedness measures. Past studies have either focused entirely on relatedness or only evaluated judgments over words rather than concepts. In this study, first, concept similarity measures are evaluated over human judgments by using existing sets of word similarity pairs that we annotated with word senses. Next, an application-oriented study is presented by integrating similarity and relatedness measures into an algorithm which relies on concept similarity. Interestingly, the results find metrics categorized as measuring relatedness to be strongest in correlation with human judgments of concept similarity, though the difference in correlation is small. On the other hand, an information content metric, categorized as measuring similarity, is notably strongest according to the application-oriented evaluation.

Introduction

Semantic similarity and relatedness has a substantial history in computational linguistics signifying its importance to the field. However, an extensive evaluation of similarity and relatedness measures for the task of concept similarity has yet to be carried out. Such an evaluation could benefit applications of measures such as word sense disambiguation or query expansion for information retrieval. This study seeks to address this gap in the current body of work by providing results on the performance of various WordNet-based measures for tasks utilizing similarity judgments among concepts (word senses).

Two distinctions are important within this study: that between *words* and *concepts*, and that between between *relatedness* and *similarity*. Although many measures are designed for comparison of *concepts* (word senses), past comparisons of similarity and relatedness measures with human judgments have looked into similarity between *words* themselves, leaving some ambiguity. For example, while one would likely agree that 'bat' as in "a club used for hitting a ball" is similar to 'stick', one would be hard-pressed to agree that 'bat' as in "nocturnal mouselike mammal with forelimbs modified to form membranous wings" is also similar to 'stick' (definitions from WordNet (Miller et al. 1993)). On the other hand, while application-oriented studies have applied measures to concepts we have yet to see an evaluation utilizing an application calling for *similarity* judgments. This paper views *similarity* as a specific type of relatedness characterized by the relationships: synonymy, antonymy, and hyponymy. As an example, we would say a 'wooden stick' is *similar* and *related* to a 'baseball bat', while a 'baseball player' is only *related* to a 'baseball bat'. Although this similarity distinction has been noted previously (Resnik 1999; Patwardhan, Banerjee, and Pedersen 2003; Agirre et al. 2009), we believe this paper presents the first evaluation of measures for tasks of *concept similarity*.

After a brief review of similarity and relatedness measures, we present a summary of past evaluations. Our approach to evaluate measures is broken into two types of experiments. One type of experiment is based on existing human judgments of similarity which we annotated with senses. As a secondary contribution of this paper, we will make the sense annotated datasets available upon publication. The other experiment is application-oriented, integrating measures within a word sense disambiguation (WSD) algorithm that requires *similarity* judgments among *concepts*. Finally, the results are presented to demonstrate the effectiveness of each measure for tasks of concept similarity.

Similarity and Relatedness Measures

Our evaluation includes measures which are available and take input as concepts defined in WordNet(Miller et al. 1993). The measures are categorized based on types of relationships considered within the approach.

similarity measures: those not utilizing relatinoships beyond synonymy, antonymy, and hyponymy.

relatedness measures: those utilizing any relationships, but which may still give a good judgment of similarity.

The measures we included under *similarity* either rely entirely on the hyponymic *paths* in WordNet (Wu and Palmer 1994; Leacock, Chodorow, and Miller 1998; Schwartz and Gomez 2008), or rely additionally on the notion of *information content*: the negative log likelihood of a concept occurring (Resnik 1999; Jiang and Conrath 1997; Lin 1998).

Methods we included under *relatedness* rely on *paths* within WordNet including relationships outside the scope of similarity (Hirst and St Onge 1998; Yang and Powers 2006), or with a strong emphasis on *glosses* (Banerjee and Pedersen 2002; Patwardhan and Pedersen 2006). Table 1 lists all of the metrics used in our evaluation. All implementations were either downloaded from the respective authors or provided by the WordNet::Similarity package (Pedersen, Patwardhan, and Michelizzi 2004). Note that the notion of similarity in this work is applied over two *single* concepts; Other works have applied *similarity* over different terms, such as comparing two *pairs* of words when measuring analogy (Turney 2006).

Similarity - Path Based	
$S_{WuPalmer}$	(Wu and Palmer 1994)
$S_{LeacockChodorow}$	(Leacock, Chodorow, and Miller 1998)
$S_{SchwartzGomez}$	(Schwartz and Gomez 2008)
Similarity - Information Content	
S_{Resnik}	(Resnik 1999)
$S_{JiangConrath}$	(Jiang and Conrath 1997)
S_{Lin}	(Lin 1998)
Relatedness - Path Based	
$R_{HirstStOnge}$	(Hirst and St Onge 1998)
$R_{YangPowers}$	(Yang and Powers 2006)
Relatedness - Gloss Based	
$R_{BanerjeePedersen}$	(Banerjee and Pedersen 2002)
$R_{PartwardhanPedersen}$	(Patwardhan and Pedersen 2006)

Table 1: Categorized identifiers used for each metric.

Related Work

Several works have formulated experiments to compare the performance of similarity and relatedness measures in a variety of situations. Some evaluations, such as (Resnik 1999; Agirre and Soroa 2009), were based on manually crafted similarity data for *words* (Miller and Charles 1991; Rubenstein and Goodenough 1965) rather than *concepts*. Although the studies based on hand crafted data often found *information-content* measures outperform *path-based* measures (Resnik 1999; Agirre and Soroa 2009), our work focuses on *concepts* rather than words. In fact, we annotate the datasets of words used by others with senses.

One of the first comprehensive evaluations of WordNet semantic similarity and relatedness measures involved an application to a spell correction algorithm (Budanitsky and Hirst 2001; 2006). For a potential misspelling or malapropism (an incorrect spelling of a word that results in the correct spelling of another word), the algorithm determined if any of the senses are related to other words in context. When a word does not have any senses related to nearby words, the system determines if any senses of similarly spelled words are related to the other words in context. Budanitsky and Hirst (2006) write, "For example, if no nearby word in a text is related to diary but one or more are related to dairy, we suggest to the user that it is the latter that was intended." Their evaluation was run over the following metrics: $S_{JiangConrath}$, $S_{LeacockChodorow}$, S_{Resnik}, S_{Lin}, $R_{HirstStOnge}$. Overall they found that $S_{JiangConrath}$ showed significant improvement over the other measures (Budanitsky and Hirst 2006).

Patwardhan, Banerjee, and Pedersen (2003) developed a Lesk (1986) style *WSD* algorithm for nouns in which senses of the target word are compared to senses of the first three nouns on the left and right of the target word. It was previously shown that $R_{BanerjeePedersen}$ performed twice as well as the Lesk algorithm itself on Senseval-2 noun data (Banerjee and Pedersen 2002). In (Patwardhan, Banerjee, and Pedersen 2003), the authors focused on the following measures: $S_{LeacockChodorow}$, S_{Resnik}, S_{Lin}, $S_{JiangConrath}$, $R_{HirstStOnge}$, $R_{BanerjeePedersen}$. The $R_{BanerjeePedersen}$ measure performed best on the Senseval-2 set of 29 nouns, followed closely by $S_{JiangConrath}$. Additionally, with the exception of S_{Resnik}, *information content* measures outperformed the two *path-based* measures. The authors also found that alternative computations of *information content* did not lead to significant changes in performance. As part of the introduction to *gloss vectors*, Patwardhan and Pedersen (2006) presented an evaluation in conjunction with five other relatedness measures used in (Patwardhan, Banerjee, and Pedersen 2003) (omitting $R_{HirstStOnge}$). The experiment followed according to (Patwardhan, Banerjee, and Pedersen 2003), and found that the $R_{PartwardhanPedersen}$ measure performed just below that of $R_{BanerjeePedersen}$, and both were outscored by $S_{JiangConrath}$. Similarly, Rus et al (2009) used relatedness measures among words and found $R_{PartwardhanPedersen}$ on a task of text relatedness.

The evaluations mentioned thus far used metrics for comparing a target word (or senses of a target word) to other words in context. The assumption is that concepts in context are *related*, but as we have previously mentioned *relatedness* does not imply *similarity*. Thus, the measures which are more appropriately categorized as measuring *similarity* (those which do not consider relationships beyond hyponymy, antonymy, and synonymy) may be at a disadvantage. The $S_{SchwartzGomez}$ measure was used in a noun *WSD* algorithm, where noun senses were compared with senses of words that are found to replace that noun in its context (a task calling for *similarity* comparisons) (Schwartz and Gomez 2008). They experimented over a few similarity and relatedness measures and found *path-based* measures to perform in line with *information content based* and *gloss-based* measures. However, unlike the previously mentioned *WSD* evaluations, this algorithm was focused on achieving top results for a *WSD* task rather than evaluating metrics, and the results were influenced by more than similarity comparisons. Our evaluation uses Schwartz and Gomez's algorithm with restrictions to limit influences beyond similarity comparisons. We also experiment on a wider variety of measures.

Experimental Setup

We implement two types of experiments over semantic similarity measures. The first is based on adding sense annotations to existing gold-standard judgments of similarity. The second evaluation is based on an application of the measures

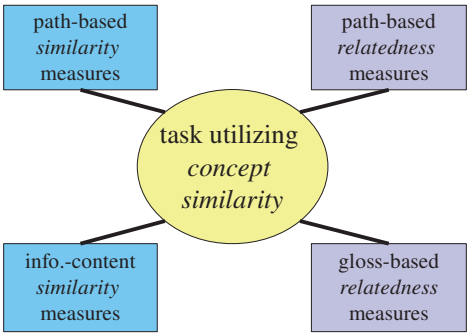

Figure 1: Depiction of the experimental setup, showing the similarity and relatedness measures as distinct from the task, which is solely focused on similarity.

to WSD. Note that although the task is focused on *similarity*, we include measures that are more correctly categorized as measuring relatedness. Because relatedness subsumes similarity, we do not want to exclude these measures from our study. Figure 1 shows this distinction between the task focus and the type of measure.

Datasets of Human Judgments

We use three datasets of human judgments of similarity, namely *RG* (Rubenstein and Goodenough 1965), *MC* (Miller and Charles 1991), and *WS-Sim* (Agirre et al. 2009). *RG* and *MC* were created specifically for *similarity* (*MC*'s 28 pairs, listed in (Resnik 1999), are a subset of *RG* with independent judgments). *WS-Sim* is a subset of the WordSim dataset (Finkelstein et al. 2001), which had subjects rate pairs on *relatedness* in general. Agirre et al., (2009) created the similarity subset by including pairs of words with relationships: identical, synonymy, antonymy, hyponymy, and unrelated.

As part of our work, two annotators marked the *RG*, *MC*, and *WS-Sim* datasets with the most similar pair of senses among each pair of words. The original scores of similarity between words were kept for the sense/concept annotated pairs. This approach is motivated by past works which have found the greatest correlation with human judgments by using the maximum similarity over all pairs of senses (Resnik 1999; Yang and Powers 2006).WordNet 3.0 served as the sense inventory (Miller et al. 1993). Annotators were able to indicate if a most similar sense was not present in WordNet, in which case the instance was dropped. For example 'jaguar' and 'car' were dropped because the automobile sense of 'jaguar' is not present in WordNet. Our *WS-Sim* dataset does not include the pairs which Agirre et al. marked as unrelated, because there was no basis for annotating senses of words considered unrelated.

Statistics of the datasets can be seen in Table 2. Inter-annotator agreement (ITA) was calculated as the mean percentage of senses agreed upon within a pair (1, 0.5, or 0 for *completely agreed*, *agreed on one word*, or *completely disagreed* respectively). The complete agreement figure (CPA) is the percentage of pairs with which both words were annotated identically. To finalize each dataset we asked the two annotators to come to an agreement on all instances which were not in complete agreements. There are two types of tests we run over the final datasets[1]:

	ITA	**CPA**	**pairs**	**drops**
MC	0.89	0.79	28	0
RG	0.93	0.86	65	0
WS-Sim	0.86	0.73	97	3

Table 2: The inter-annotator agreement (**ITA**) and complete pair agreement (**CPA**) over **pairs** number of pairs; **drops** indicates number of instances not annotated due to lack of WordNet sense.

wrd correlation of similarity values based on the word pairs (measures choose the max similarity over all pairs of senses).

cpt correlation of similarity values based on sense annotated (concept) pairs.

Application-Oriented Study

The second experiment is focused on evaluating similarity measures when applied to the task of WSD. We chose the *Web selectors* algorithm of Schwartz and Gomez (2008), since the algorithm relies on *similarity* judgments of concepts. As introduced by Lin (1997), selectors are words which take the place of an instance of a target word within its local context. The *Web selectors* algorithm performs disambiguation by comparing selectors acquired from the Web to senses of a target word.

In order to focus on the impact that a similarity measure has on the accuracy, restrictions are placed on the algorithm. First, senses are chosen by only considering target selectors, words which replace the target word that is being disambiguated. Target selectors are intended to be *similar* to the target sense, while other types of selectors within the algorithm are only intended to be *related*. The system is also setup to only attempt annotations of instances in which it acquires five or more selectors from queries of seven words or more in length. This restriction insures that there is both enough selectors and that the selectors are reliable. Finally, the use of a first sense heuristic as a backoff strategy is turned off to eliminate unnecessary bias.

Our testing corpus consisted of the training set from the SemEval-2007 Task 17: Lexical Sample (Pradhan et al. 2007). The lexical sample contained many instances of nouns and verbs, leaving the sample size quite large after the restrictions we placed on the algorithm. Note that the all-words portion of Task 17 contained fewer instances of nouns. The corpus, annotated with WordNet 2.1 senses, was also restricted to eliminate instances of monosemous words according to WordNet. This restriction in addition to those placed on the algorithm are likely to decrease disambiguation accuracy of the algorithm, in order to get a stronger comparison focused on each similarity measure.

[1]Datasets to be made available online upon publication.

	MC		*RG*		*WS-Sim*	
	wrd	**cpt**	**wrd**	**cpt**	**wrd**	**cpt**
$S_{WuPalmer}$	0.76 [.54, .88]	0.76 [.54, .88]	0.76 [.66, .86]	0.79 [.67, .87]	0.62 [.48, .73]	0.57 [.42, .69]
$S_{LeacockChodorow}$	0.75 [.52, .88]	0.75 [.52, .88]	0.78 [.67, .86]	0.80 [.69, .87]	0.62 [.48, .73]	0.58 [.44, .70]
$S_{SchwartzGomez}$	0.77 [.60, .90]	0.81 [.62, .91]	**0.82** [.71, .88]	0.77 [.65, .85]	0.61 [.47, .72]	0.54 [.38, .66]
S_{Resnik}	0.76 [.55, .88]	0.76 [.53, .88]	0.74 [.61, .83]	0.76 [.63, .84]	0.62 [.47, .73]	0.59 [.45, .71]
$S_{JiangConrath}$	0.82 [.65, .92]	0.85 [.70, .93]	0.78 [.66, .86]	0.80 [.69, .87]	0.60 [.45, .71]	0.51 [.34, .64]
S_{Lin}	0.77 [.56, .89]	0.80 [.61, .91]	0.77 [.64, .85]	0.78 [.66, .86]	**0.64** [.50, .74]	0.58 [.43, .70]
$R_{HirstStOnge}$	0.77 [.56, .89]	0.72 [.47, .86]	0.78 [.66, .86]	0.76 [.63, .85]	0.49 [.32, .63]	0.53 [.37, .66]
$R_{YangPowers}$	0.88 [.75, .94]	0.76 [.55, .88]	**0.82** [.72, .89]	0.78 [.66, .86]	**0.64** [.51, .75]	**0.63** [.49, .74]
$R_{BanerjeePedersen}$	0.81 [.62, .91]	0.76 [.54, .88]	0.72 [.58, .82]	0.69 [.54, .80]	0.49 [.32, .63]	0.46 [.29, .60]
$R_{PartwardhanPedersen}$	**0.92** [.83, .96]	**0.88** [.75, .94]	0.81 [.71, .88]	**0.81** [.71, .88]	0.57 [.41, .69]	0.55 [.39, .67]

Table 3: Correlation between similarity measure judgments and human judgments for each dataset (**wrd**, **cpt**).

Results

Results are broken down according to the two types of experiments. First we present the correlation of the metrics with human judgments of *similarity*. Then, the results are presented for applying the metrics to the task of word sense disambiguation by *similarity*. Refer back to table 1 for the identifiers and categorization of each metric.

Human Judgments

Table 3 presents the results based on human judgments over all three datasets. Correlations are reported as Spearman rank correlations, avoiding issues arising from non-linear measure outputs as Agirre et al (2009) noted. Normal approximations of confidence intervals at 95% are also presented.

There was no single measure that performed best across all the datasets. When examining the results of the *MC* and *RG* datasets, we see that $R_{PartwardhanPedersen}$ had consistently high correlations. Keep in mind that the *MC* dataset contains a subset of the pairs in the *RG* dataset, with a different set of human judgments. For the *WS-Sim* dataset, which was a distinct set of words and concepts, it was $R_{YangPowers}$ with the highest correlations. In each case, a best performing metric was categorized under *relatedness*, but there is never a significant difference over the top performing metric categorized under *similarity*.

When examining the differences between the 'wrd' and 'cpt' tests, on average, *similarity* measures had higher correlations on the 'cpt' tests within the *MC* and *RG* datasets, while the *relatedness* measures had higher correlations on the 'wrd' tests. This suggests the similarity measures benefit from dealing specifically with concepts rather than ambiguous words, though the differences are small enough that a concrete conclusion can not be drawn. On the other hand, for the *WS-Sim* dataset, the *similarity* measures performed better at the 'wrd' test relative to the 'cpt' test. This difference between the *WS-Sim* dataset and the *MC*/*RG* dataset may have been due to *WS-Sim* containing more pairs of dissimilar words.

Application

Table 4 presents the results of the word sense disambiguation experiment. After the restrictions were placed on the corpus, we ended up with 795 instances (431 nouns and 364 verbs). The F1 values shown are calculated based on precision(P) and recall(R) as $F1 = 2 * \frac{P*R}{P+R}$.

Unlike the human judged experiment, we found one measure performs significantly better than any other measure in this experiment. The information-content similarity measure of Jiang and Conrath ($S_{JiangConrath}$) gives us the top results for both the noun and verb portions of the corpus. All of the relatedness measures ($R_{BanerjeePedersen}$, $R_{PartwardhanPedersen}$, $R_{YangPowers}$) along with the S_{Lin} measure performed approximately equally with over 10.4% more error than the $S_{JiangConrath}$ measure. The path-based similarity measures were all among the least effective for the task.

We saw improvement from all measures between noun and verb instances. Among the relatedness measures, the differences in values indicate that the $R_{YangPowers}$ measure may be better suited for nouns, while the $R_{PartwardhanPedersen}$ method may be stronger with verbs. We suspect the verb results were higher overall because the verb selectors were more often acquired with surrounding context, and were thus more reliable than noun selectors which were more often acquired at the beginning or end of a sentence. Had the algorithm not been restricted to focus on similarity, the noun results would have been higher as was

	noun	verb	both
$S_{WuPalmer}$	41.5	56.3	48.3
$S_{LeacockChodorow}$	44.1	59.3	51.1
$S_{SchwartzGomez}$	48.0	-	-
S_{Resnik}	46.3	51.1	48.5
$S_{JiangConrath}$	**59.6**	**65.1**	**62.1**
S_{Lin}	52.6	57.8	54.9
$R_{HirstStOnge}$	50.9	55.1	52.8
$R_{YangPowers}$	53.2	54.6	53.9
$R_{BanerjeePedersen}$	49.9	57.7	53.5
$R_{PartwardhanPedersen}$	50.6	61.5	55.6

Table 4: Results of the application-oriented experiment: F1 values (precision=recall) on the SemEval-2007 Task17, broken down by part of speech and combined.

reported originally by Schwartz and Gomez (2008).

Conclusion

We presented evaluations of WordNet-based semantic similarity and relatedness measures focused on *concept similarity*. One type of experiment was based on human judgments and the other was an application-oriented task. The measures of Patwardhan and Pederson (2006), and Yang and Powers (2006) had consistently high correlations with human judgments. Both of these measures were categorized as more broad *relatedness* measures, though the best performing *similarity* measures were not significantly lower for any of the datasets. For the application-oriented experiment, the *similarity* measure of Jiang and Conrath (1997) clearly gave us the best results with an error reduction of 10.4% over the next best measure.

There are several possible extensions to this work to provide additional insights about similarity measures. The existing gold-standard judgments of similarity that we annotated with senses only included nominal concepts. To address this drawback, a human annotated dataset of verb pairs could be created. Additionally, one could replicate experiments over different versions of WordNet as an evaluation of the WordNet improvements. Never the less, the results of this study alone are intended to impact work in computational linguistics when a task calls for similarity judgments over concepts.

Acknowledgement

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

References

Agirre, E., and Soroa, A. 2009. Personalizing PageRank for word sense disambiguation. In *Proc. of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 33–41. Athens, Greece: Association for Computational Linguistics.

Agirre, E.; Alfonseca, E.; Hall, K.; Kravalova, J.; Pasca, M.; and Soroa, A. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *The Annual Conference of the NAACL*, 19–27.

Banerjee, S., and Pedersen, T. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proc. of the 3rd CICLing*.

Budanitsky, A., and Hirst, G. 2001. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other lexical resources*. Morristown, NJ, USA: Association for Computational Linguistics.

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

Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G.; and Ruppin, E. 2001. Placing search in context: The concept revisited. In *ACM Trans. on Information Systems*.

Hirst, G., and St Onge, D. 1998. Lexical chains as representation of context for the detection and correction malapropisms. In *WordNet: An Electronic Lexical Database*. The MIT Press.

Jiang, J., and Conrath, D. 1997. Semantic similarity on corpus statistics and lexical taxonomy. In *Proc. of ROCLING X*.

Leacock, C.; Chodorow, M.; and Miller, G. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics* 24(1):147–165.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proc. of the 5th SIGDOC*, 24–26.

Lin, D. 1998. An information-theoretic definition of similarity. In *Proc. of the 15th ICML*, 296–304. Madison, WI, USA: Morgan Kaufmann.

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

Miller, G.; Beckwith, R.; Fellbaum, C.; Gross, D.; and Miller, K. 1993. Five papers on wordnet. Technical report, Princeton University.

Patwardhan, S., and Pedersen, T. 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *EACL 2006 Workshop Making Sense of Sense*, 1–8.

Patwardhan, S.; Banerjee, S.; and Pedersen, T. 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *Proc. of the 4th CICLing*, 241–257.

Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. In *HLT Conference of the NAACL Demonstrations*, 38–41.

Pradhan, S.; Loper, E.; Dligach, D.; and Palmer, M. 2007. Semeval-2007 task 17: English lexical sample, srl and all words. In *SemEval '07*, 87–92.

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.

Rubenstein, H., and Goodenough, J. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8:627–633.

Rus, V.; Lintean, M.; Graesser, A.; and McNamara, D. 2009. Assessing student paraphrases using lexical semantics and word weighting. In *Proceeding of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care*, 165–172.

Schwartz, H. A., and Gomez, F. 2008. Acquiring knowledge from the web to be used as selectors for noun sense disambiguation. In *Proc. of the Twelfth CoNLL*, 105–112.

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

Wu, Z., and Palmer, M. 1994. Verb semantics and lexical selection. In *Proc. of the 32nd. ACL*, 133–138.

Yang, D., and Powers, D. M. W. 2006. Verb similarity on the taxonomy of wordnet. In *Proc. of GWC-06*.