Extended Gloss Overlaps as a Measure of Semantic Relatedness

Satanjeev Banerjee
Carnegie Mellon University
Pittsburgh, PA 15213
satanjeev.banerjee@cs.cmu.edu

Ted Pedersen
University of Minnesota
Duluth, MN, 55812
tpederse@umn.edu

Abstract

This paper presents a new measure of semantic relatedness between concepts that is based on the number of shared words (overlaps) in their definitions (glosses). This measure is unique in that it extends the glosses of the concepts under consideration to include the glosses of other concepts to which they are related according to a given concept hierarchy. We show that this new measure reasonably correlates to human judgments. We introduce a new method of word sense disambiguation based on extended gloss overlaps, and demonstrate that it fares well on the SENSEVAL-2 lexical sample data.

1 Introduction

Human beings have an innate ability to determine if two concepts are related. For example, most would agree that the automotive senses of *car* and *tire* are related while *car* and *tree* are not. However, assigning a value that quantifies the degree to which two concepts are related proves to be more difficult [Miller and Charles, 1991]. In part, this is because *relatedness* is a very broad notion. For example, two concepts can be related because one is a more general instance of the other (e.g., a *car* is a kind of *vehicle*) or because one is a part of another (e.g., a *tire* is a part of a *car*).

This paper introduces *extended gloss overlaps*, a measure of semantic relatedness that is based on information from a machine readable dictionary. In particular, this measure takes advantage of hierarchies or taxonomies of concepts as found in resources such as the lexical database WordNet [Fellbaum, 1998].

Concepts are commonly represented in dictionaries by word senses, each of which has a definition or gloss that briefly describes its meaning. Our measure determines how related two concepts are by counting the number of shared words (overlaps) in the word senses of the concepts, as well as in the glosses of words that are related to those concepts according to the dictionary. These related concepts are explicitly encoded in WordNet as relations, but can be found in any dictionary via synonyms, antonyms, or also-see references provided for a word sense. To our knowledge, this work represents the first attempt to define a quantitative measure of relatedness between two concepts based on their dictionary definitions.

This paper begins with a brief description of WordNet, which was used in developing our measure. Then we introduce the extended gloss overlap measure, and present two distinct evaluations. First, we conduct a comparison to previous human studies of relatedness and find that our measure has a correlation of at least 0.6 with human judgments. Second, we introduce a word sense disambiguation algorithm that assigns the most appropriate sense to a target word in a given context based on the degree of relatedness between the target and its neighbors. We find that this technique is more accurate than all but one system that participated in the SENSEVAL–2 comparative word sense disambiguation exercise. Finally we present an extended analysis of our results and close with a brief discussion of related work.

2 WordNet

WordNet is a lexical database where each unique meaning of a word is represented by a synonym set or *synset*. Each synset has a gloss that defines the concept that it represents. For example the words *car*, *auto*, *automobile*, and *motorcar* constitute a single synset that has the following gloss: *four wheel motor vehicle, usually propelled by an internal combustion engine*. Many glosses have *examples* of usages associated with them, such as *"he needs a car to get to work."*

Synsets are connected to each other through explicit semantic relations that are defined in WordNet. These relations only connect word senses that are used in the same part of speech. Noun synsets are connected to each other through *hypernym*, *hyponym*, *meronym*, and *holonym* relations.

If a noun synset A is connected to another noun synset B through the *is–a–kind–of* relation then B is said to be a *hypernym* of synset B and B a *hyponym* of A. For example the synset containing *car* is a hypernym of the synset containing *hatchback* and *hatchback* is a hyponym of *car*. If a noun synset A is connected to another noun synset B through the *is–a–part–of* relation then A is said to be a *meronym* of B and B a *holonym* of A. For example the synset containing *accelerator* is a meronym of *car* and *car* is a holonym of *accelerator*. Noun synset A is related to adjective synset B through the *attribute* relation when B is a *value* of A. For example the adjective synset *standard* is a value of the noun synset *measure*.

Taxonomic or *is–a* relations also exist for verb synsets. Verb synset A is a hypernym of verb synset B if *to* B *is one way to* A. Synset B is called a *troponym* of A. For example the verb synset containing the word *operate* is a hypernym of *drive* since to drive is one way to operate. Conversely *drive* is a troponym of *operate*. The troponym relation for verbs is analogous to the hyponym relation for nouns, and henceforth we shall use the term hyponym instead of the term troponym. Adjective synsets are related to each other through the *similar to* relation. For example the synset containing the adjective *last* is said to be similar to the synset containing the adjective *dying*. Verb and adjective synsets are also related to each other through cross–reference *also–see* links. For example, the adjectives *accessible* and *convenient* are related through also–see links.

While there are other relations in WordNet, those described above make up more than 93% of the total number of links in WordNet. These are the measures we have employed in the extended gloss overlap measure.

3 The Extended Gloss Overlap Measure

Gloss overlaps were introduced by [Lesk, 1986] to perform word sense disambiguation. The Lesk Algorithm assigns a sense to a target word in a given context by comparing the glosses of its various senses with those of the other words in the context. That sense of the target word whose gloss has the most words in common with the glosses of the neighboring words is chosen as its most appropriate sense.

For example, consider the glosses of *car* and *tire*: *four wheel motor vehicle usually propelled by an internal combustion engine* and *hoop that covers a wheel, usually made of rubber and filled with compressed air*. The relationship between these concepts is shown in that their glosses share the content word *wheel*. However, they share no content words with the gloss of *tree*: *a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown*.

The original Lesk Algorithm only considers overlaps among the glosses of the target word and those that surround it in the given context. This is a significant limitation in that dictionary glosses tend to be fairly short and do not provide sufficient vocabulary to make fine grained distinctions in relatedness. As an example, the average length of a gloss in WordNet is just seven words. The extended gloss overlap measure expands the glosses of the words being compared to include glosses of concepts that are known to be related to the concepts being compared.

Our measure takes as input two concepts (represented by two WordNet synsets) and outputs a numeric value that quantifies their degree of semantic relatedness. In the sections that follow, we describe the foundations of the measure and how it is computed.

3.1 Using Glosses of Related Senses

There are two fundamental premises to the original Lesk Algorithm. First, words that appear together in a sentence will be used in related senses. Second, and most relevant to our measure, the degree to which senses are related can be identified by the number of overlaps in their glosses. In other words, the more related two senses are, the more words their glosses will share.

WordNet provides *explicit* semantic relations between synsets, such as through the *is–a* or *has–part* links. However such links do not cover all possible relations between synsets. For example, WordNet encodes no direct link between the synsets *car* and *tire*, although they are clearly related. We observe however that the glosses of these two synsets have words in common. Similar to Lesk's premise, we assert that such overlaps provide evidence that there is an *implicit* relation between those synsets. Given such a relation, we further conclude that synsets explicitly related to *car* are thereby also related to synsets explicitly related to *tire*. For example, we conclude that the synset *vehicle* (which is the hypernym synset of *car*) is related to the synset *hoop* (which is the hypernym synset of *tire*). Thus, our measure combines the advantages of gloss overlaps with the structure of a concept hierarchy to create an extended view of relatedness between synsets.

We base our measure on the idea of an extended set of comparisons. When measuring the relatedness between two input synsets, we not only look for overlaps between the glosses of those synsets, but also between the glosses of the hypernym, hyponym, meronym, holonym and troponym synsets of the input synsets, as well as between synsets related to the input synsets through the relations of attribute, similar–to and also–see. Not all of these relations are equally helpful, and the optimum choice of relations to use for comparisons is possibly dependent on the application in which the overlaps–measure is being employed. Section 6 compares the relative efficacy of these relations when our measure of relatedness is applied to the task of word sense disambiguation.

3.2 Scoring Mechanism

We introduce a novel way of finding and scoring the overlaps between two glosses. The original Lesk Algorithm compares the glosses of a pair of concepts and computes a score by counting the number of words that are shared between them. This scoring mechanism does not differentiate between single word and phrasal overlaps and effectively treats each gloss as a "bag of words". For example, it assigns a score of 3 to the concepts *drawing paper* and *decal*, which have the glosses *paper that is specially prepared for use in drafting* and *the art of transferring designs from specially prepared paper to a wood or glass or metal surface*. There are three words that overlap, *paper* and the two–word phrase *specially prepared*.

There is a Zipfian relationship [Zipf, 1935] between the lengths of phrases and their frequencies in a large corpus of text. The longer the phrase, the less likely it is to occur multiple times in a given corpus. A phrasal n–word overlap is a much rarer occurrence than an single word overlap. Therefore, we assign an n word overlap the score of n^2. This gives an n–word overlap a score that is greater than the sum of the scores assigned to those n words if they had occurred in two or more phrases, each less than n words long.

For the above gloss pair, we assign the overlap *paper* a score of 1 and *specially prepared* a score of 4, leading to a total score of 5. Note that if the overlap was the 3–word phrase *specially prepared paper*, then the score would have been 9.

Thus, our overlap detection and scoring mechanism can be formally defined as follows: When comparing two glosses, we define an *overlap* between them to be the longest sequence of one or more consecutive words that occurs in both glosses such that neither the first nor the last word is a function word, that is a pronoun, preposition, article or conjunction. If two or more such overlaps have the same longest length, then the overlap that occurs earliest in the first string being compared is reported. Given two strings, the longest overlap between them is detected, removed and in its place a unique marker is placed in each of the two input strings. The two strings thus obtained are then again checked for overlaps, and this process continues until there are no longer any overlaps between them. The sizes of the overlaps thus found are squared and added together to arrive at the score for the given pair of glosses.

3.3 Computing Relatedness

The extended gloss overlap measure computes the relatedness between two input synsets A and B by comparing the glosses of synsets that are related to A and B through explicit relations provided in WordNet.

We define RELS as a (non-empty) set of relations that consists of one or more of the relations described in Section 2. That is, RELS $\subset \{r \mid r$ is a relation defined in WordNet$\}$. Suppose each relation r ($r \in$ RELS) has a function of the same name that accepts a synset as input and returns the gloss of the synset (or synsets) related to the input synset by the designated relation.

For example, assume r represents the hypernym relation. Then $r(A)$ returns the gloss of the hypernym synset of A. r can also represent the gloss "relation" such that $r(A)$ returns the gloss of synset A, and the example "relation" such that $r(A)$ returns the example string associated with synset A. If more than one synset is related to the input synset through the same relation, their glosses are concatenated and returned. We perform this concatenation because we do not wish to differentiate between the different synsets that are all related to the input synset through a particular relation, but instead are only interested in all their definitional glosses. If no synset is related to the input synset by the given relation then the null string is returned.

Next, form a non–empty set of *pairs* of relations from the set of relations above. The only constraint in forming such pairs is that if the pair (r_1, r_2) is chosen, ($r_1, r_2 \in$ RELS), then the pair (r_2, r_1) must also be chosen so that the relatedness measure is reflexive. That is, $relatedness(A, B) = relatedness(B, A)$. Thus, we define the set RELPAIRS as follows:

$$\text{RELPAIRS} = \{(R_1, R_2) \mid R_1, R_2 \in \text{RELS}; \\ \text{if } (R_1, R_2) \in \text{RELPAIRS, then } (R_2, R_1) \in \text{RELPAIRS}\}$$

Finally, assume that $score()$ is a function that accepts as input two glosses, finds the phrases that overlap between them and returns a score as described in the previous section. Given all of the above, the relatedness score between the input synsets A and B is computed as follows:

$$relatedness(A, B) = \sum score(R_1(A), R_2(B)) \\ \forall (R_1, R_2) \in \text{RELPAIRS}$$

Our relatedness measure is based on the set of all possible pairs of relations from the list of relations described in section 3.1. For purposes of illustration, assume that our set of relations RELS = $\{$gloss, hype, hypo$\}$ (where hype and hypo are contractions of hypernym and hyponym respectively). Further assume that our set of relation pairs RELPAIRS = $\{$(gloss, gloss), (hype, hype), (hypo, hypo), (hype, gloss), (gloss, hype)$\}$. Then the relatedness between synsets A and B is computed as follows:

$$relatedness(A, B) = score(gloss(A), gloss(B)) + \\ score(hype(A), hype(B)) + score(hypo(A), hypo(B)) + \\ score(hype(A), gloss(B)) + score(gloss(A), hype(B))$$

Observe that due to our pair selection constraint as described above, $relatedness(A, B)$ is indeed the same as $relatedness(B, A)$.

4 Comparison to Human Judgements

Our comparison to human judgments is based on three previous studies. [Rubenstein and Goodenough, 1965] presented human subjects with 65 noun pairs and asked them how similar they were on a scale from 0.0 to 4.0. [Miller and Charles, 1991] took a 30 pair subset of this data and repeated this experiment, and found results that were highly correlated (.97) to the previous study. The results from the 30 pair set common to both studies were used again by [Budanitsky and Hirst, 2001] in an evaluation of five automatic measures of semantic relatedness that will be mentioned in Section 7. They report that all of the measures fared relatively well, with the lowest correlation being .74 and the highest .85. When comparing our measure to these 30 words, we find that it has a correlation of .67 to the Miller and Charles human study, and one of .60 to the Rubenstein and Goodenough experiment.

We do not find it discouraging that the correlation of extended gloss overlaps is lower than those reported by Budanitsky and Hirst for other measures. In fact, given the complexity of the task, it is noteworthy that it demonstrates some correlation with human judgement. The fact that the test set contains only 30 word pairs is a drawback of human evaluation, where rigourous studies are by necessity limited to a small number of words. Automatic measures can be evaluated relative to very large numbers of words, and we believe such an evaluation is an important next step in order to establish where differences lie among such measures. As a final point of concern, concepts can be related in many ways, and it is possible that a human and an automatic measure could rely on different yet equally well motivated criteria to arrive at diverging judgements.

5 Application to WSD

We have developed an approach to word sense disambiguation based on the use of the extended gloss overlap measure.

In our approach, a window of context around the target word is selected, and a set of candidate senses is identified for each content word in the window. Assume that the window of context consists of $2n + 1$ words denoted by w_i, $-n \leq i \leq +n$, where the target word is w_0. Further let $|w_i|$ denote the number of candidate senses of word w_i, and let these senses be denoted by $s_{i,j}$, $1 \leq j \leq |w_i|$.

Next we assign to each possible sense k of the target word a $SenseScore_k$ computed by adding together the relatedness scores obtained by comparing the sense of the target word in question with every sense of every non–target word in the window of context. The SenseScore for sense $s_{0,k}$ is computed as follows:

$$SenseScore_k = \sum_{i=-n}^{n} \sum_{j=1}^{|w_i|} relatedness(s_{0,k}, s_{i,j}), i \neq 0$$

That sense with the highest SenseScore is judged to be the most appropriate sense for the target word. If there are on average a senses per word and the window of context is N words long, there are $a^2 \times (N-1)$ pairs of sets of synsets to be compared, which increases linearly with N.

5.1 Experimental Data

Our evaluation data is taken from the English *lexical sample* task of SENSEVAL–2 [Edmonds and Cotton, 2001]. This was a comparative evaluation of word sense disambiguation systems that resulted in a large set of results and data that are now freely available to the research community.

This data consists of 4,328 *instances* each of which contains a sentence with a single target word to be disambiguated, and one or two surrounding sentences that provide additional context. A human judge has labeled each target word with the most appropriate WordNet sense for that context. A word sense disambiguation system is given these same instances (minus the human assigned senses) and must output what it believes to be the most appropriate senses for each of the target words. There are 73 distinct target words: 29 nouns, 29 verbs, and 15 adjectives, and the part of speech of the target words is known to the systems.

5.2 Experimental Results

For every instance, function words are removed and then a window of words is defined such that the target word is at the center (if possible). Next, for every word in the window, candidate senses are picked by including the synsets in WordNet that the word belongs to, as well as those that an uninflected form of the word belong to (if any). Given these candidate senses, the algorithm described above finds the most appropriate sense of the target word.

It is possible that there be a tie among multiple senses for the highest score for a word. In this case, all those senses are reported as answers and partial credit is given if one of them prove to be correct. This would be appropriate if a word were truly ambiguous in a context, or if the meanings were very closely related and it was not possible to distinguish between them. It is also possible that no sense gets more than a score of 0 – in this case, no answer is reported since there is no evidence to choose one sense over another.

Given the answers generated by the algorithm, we compare them with the human decided answers and compute *precision* (the number of correct answers divided by the number of answers reported) and *recall* (the number of correct answers divided by the number of instances). These two values can be summarized by the *F–measure*, which is the harmonic mean of the precision and recall:

$$F\text{–}measure = \frac{2 \times precision \times recall}{precision + recall}$$

Table 1: WSD Evaluation Results

Baselines and Other Systems

Algorithm	Prec.	Recall	F-Meas.
Sval-First	0.402	0.401	0.401
Overall*	0.351	0.342	0.346
Sval-Second	0.293	0.293	0.293
Sval-Third	0.247	0.244	0.245
Original Lesk	0.183	0.183	0.183
Random	0.141	0.141	0.141

Extended Gloss Overlaps w 3 word window

POS	Prec.	Recall	F-Meas.
Noun	0.429	0.416	0.422
Adj.	0.367	0.346	0.356
Verb	0.270	0.266	0.268
Overall*	0.351	0.342	0.346

Table 1 lists the precision, recall and F–measure for all the SENSEVAL-2 words when disambiguated using a window size of 3. The overall results for our approach are shown as *Overall**, and these are also broken down based on the part of speech (POS) of the target word. This table also displays results from other baseline or representative systems. The *Original Lesk* results are based on utilizing the glosses of only the input synsets and nothing else. While this does not exactly replicate the original Lesk Algorithm it is quite similar. The *random* results reflect the accuracies obtained by simply selecting randomly from the candidate senses.

The *Sval–First*, *Sval–Second*, and *Sval–Third* results are from the top three most accurate fully automatic unsupervised systems in the SENSEVAL-2 exercise. This is the class of systems most directly comparable to our own, since they require no human intervention and do not use any manually created training examples. These results show that our approach was considerably more accurate than all but one of the participating systems.

These results are significant because they are based on a very simple algorithm that relies on assigning relatedness scores to the senses of a target word and the senses of its immediately adjacent neighbors. While the disambiguation results could be improved via the combination of various techniques, our focus is on developing the extended gloss overlap measure of relatedness as a general tool for Natural Language Processing and Artificial Intelligence.

6 Discussion

Table 1 shows that the disambiguation results obtained using the extended gloss overlap measure of semantic relatedness are significantly better than both the random and Original Lesk baselines. In the Original Lesk Algorithm, relatedness between two synsets is measured by considering overlaps between the glosses of the candidate senses of the target word and its neighbors. By adding the glosses of related synsets, the results improve by 89% relative (16.3% absolute). This shows that overlaps between glosses of synsets explicitly related to the input synsets provide almost as much evidence about the implicit relation between the input synsets as do overlaps between the glosses of the input synsets themselves.

Table 1 also breaks down the precision, recall and F–measure according to the part of speech of the target word. Observe that the noun target words are the easiest to disambiguate, followed by the adjective target words. The verb target words prove to be the hardest to disambiguate. We attribute this to the fact that the number of senses per target word is much smaller for the nouns and adjectives than it is for the verbs. Nouns and adjective target words have less than 5 candidate senses each on average, whereas verbs have close to 16. Thus, when disambiguating verbs there are more choices to be made and more chances of errors.

The results in table 1 are based on a 3 word window of context. In other experiments we used window sizes of 5, 7, 9 and 11. Although this increase in window size provides more data to the disambiguation algorithm, our experiments show that this does not significantly improve disambiguation results. This suggests that words that are in the immediate vicinity of the target word are most useful for disambiguation, and that using larger context windows is either adding noise or redundant data. The fact that small windows are best corresponds with earlier studies on human subjects that showed that humans often only require a window of one or two surrounding words to disambiguate a target word [Choueka and Lusignan, 1985].

We also tried to normalize the overlap scores by the maximum score that two glosses can generate, but that did not help performance. We believe that the difference between the sizes of various glosses in terms of number of words is small enough to render normalization unnecessary.

6.1 Evaluating Individual Relation Pairs

Our measure of relatedness utilizes pairs of relations picked from the list of relations in section 3.1. In this section we attempt to quantify the relative effectiveness of these individual relation pairs. Specifically, given a set of relations RELS, we create all possible *minimal* relation pair sets, where a minimal relation pair set is defined as the set that contains either exactly one relation pair $\{(R_1, R_1)\}$ or exactly two relation pairs $\{(R_1, R_2), (R_2, R_1)\}$, where $R_1 \neq R_2$. For example $\{$(gloss, gloss)$\}$ and $\{$(hype, gloss), (gloss, hype)$\}$ are both minimal relation pair sets.

We evaluate each of these minimal relation pair sets by performing disambiguation using only the given minimal relation pair set and computing the resulting precision, recall and F–measure. The higher the F–measure, the "better" the quality of the evidence provided by gloss overlaps from that minimal relation pair set. In effect we are decomposing the extended gloss overlap measure into its individual pieces and assessing how each of those pieces perform individually.

Recall that each part of speech has a different set of relations associated with it. The difference in the numbers and types of relations available for the three parts of speech leads us to expect that the optimal minimal relation pair sets will differ with the part of speech of the input synsets. Table 2 lists the top 5 minimal relation pair sets for target words belonging to the three parts of speech, where relation pair sets are ranked on the F–measure achieved by using them in disambiguation. Note that in this table, hypo, mero, also, attr, and hype stand for the relations hyponym, meronym, also–see, attribute, and hypernym respectively. Also in the table the relation pair r_1–r_2 refers to the minimal relation pair set $\{(r_1, r_2), (r_2, r_1)\}$ if $r_1 \neq r_2$ and $\{(r_1, r_1)\}$ otherwise.

Table 2: Best Relation Pair Sets

Nouns

Relation pair	Prec.	Recall	F-Meas.
hypo-mero	0.263	0.091	0.136
hypo-hypo	0.168	0.111	0.134
glos-mero	0.272	0.087	0.132
gloss-gloss	0.161	0.108	0.129
example-mero	0.314	0.074	0.120

Adjectives

Relation pair	Prec.	Recall	F-Meas.
also-gloss	0.220	0.084	0.122
attr-gloss	0.323	0.072	0.117
gloss-gloss	0.146	0.094	0.114
example-gloss	0.138	0.094	0.112
gloss-hype	0.164	0.083	0.110

Verbs

Relation pair	Prec.	Recall	F-Meas.
example-example	0.061	0.048	0.053
example-hype	0.060	0.046	0.052
hypo-hypo	0.061	0.042	0.050
gloss-hypo	0.053	0.046	0.049
example-gloss	0.054	0.045	0.049

Perhaps one of the most interesting observations is that no single minimal relation pair set achieves F–measure even close to that achieved using all the relation pairs (0.42, 0.35, and 0.26 for nouns, verbs, and adjectives respectively), suggesting that there is no single relation pair that generates a lot of evidence for the relatedness of two synsets. This finding also implies that the richer the set of explicit relations between synsets in WordNet, the more accurate the overlap based measure of semantic relatedness will be. This fact is borne out by the comparatively high accuracy attained by nouns which is the best developed portion of WordNet.

For nouns, Table 2 shows that comparisons between the glosses of the hyponyms and meronyms of the input synsets and also between the glosses of the input synsets are most informative about the relatedness of the synsets. Interestingly, although both hyponyms and hypernyms make up the is–a hierarchy, the hypernym relation does not provide an equivalent amount of information. In WordNet, a noun synset usually has a single hypernym (parent) but many hyponyms (children), which implies that the hyponym relation provides more definitional glosses to the algorithm than the hypernym relation. This assymetry also exists in the holonym–meronym pair of relations. Most noun synsets have less holonym (is–a–part–of) relations than meronyms (has–part) resulting in more glosses from the meronym relation. These further confirm that the accuracy of the relatedness measure depends at least partly on the number of glosses that we can access for a given pair of synsets.

This finding also applies to adjectives. The two most frequent relations, the also–see relation and the attribute relation, rank highest among the useful relations for adjectives. Similarly for verbs, the hyponym relation again appears to be extremely useful. Interestingly, for all three parts of speech, the example "relation" (which simply returns the example string associated with the input synset) seems to provide useful information. This is in keeping with the SENSEVAL–2 results where the addition of example strings to a Lesk–like baseline system improves recall from 16% to 23%.

7 Related Work

A number of measures of semantic relatedness have been proposed in recent years. Most of them rely on the noun taxonomy of the lexical database WordNet. [Resnik, 1995] augments each synset in WordNet with an information content value derived from a large corpus of text. The measure of relatedness between two concepts is taken to be the information content value of the most specific concept that the two concepts have in common. [Jiang and Conrath, 1997] and [Lin, 1997] extend Resnik's measure by scaling the common information content values by those of the individual concepts. Our method of extended gloss overlaps is distinct in that it takes advantage of the information found in the glosses. The other measures rely on the structure of WordNet and corpus statistics. In addition, the measures above are all limited to relations between noun concepts, while extended gloss overlaps can find relations between adjectives and verbs as well.

8 Conclusions

We have presented a new measure of semantic relatedness based on gloss overlaps. A pair of concepts is assigned a value of relatedness based on the number of overlapping words in their respective glosses, as well as the overlaps found in the glosses of concepts they are related to in a given concept hierarchy. We have evaluated this measure relative to human judgements and found it to be reasonably correlated. We have carried out a word sense disambiguation experiment with the SENSEVAL-2 lexical sample data. We find that disambiguation accuracy based on extended gloss overlaps is more accurate than all but one of the participating SENSEVAL-2 systems.

Acknowledgements

Thanks to Jason Rennie for his WordNet::QueryData module, and to Siddharth Patwardhan for useful discussions, experimental help, and for integrating the extended gloss overlap measure into his WordNet::Similarity module. Both of these modules are freely available from the Comprehensive Perl Archive Network (search.cpan.org).

This work has been supported by a National Science Foundation Faculty Early CAREER Development award (#0092784) and NSF Grant no. REC–9979894. Any opinions, findings, conclusions, or recommendations expressed in this publications are those of the authors and do not necessarily reflect the views of the NSF or the official policies, either expressed or implied, of the sponsors or of the United States Government.

References

[Budanitsky and Hirst, 2001] A. Budanitsky and G. Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, June 2001.

[Choueka and Lusignan, 1985] Y. Choueka and S. Lusignan. Disambiguation by short contexts. *Computers and the Humanities*, 19:147–157, 1985.

[Edmonds and Cotton, 2001] P. Edmonds and S. Cotton, editors. *Proceedings of the Senseval–2 Workshop*. Association for Computational Linguistics, Toulouse, France, 2001.

[Fellbaum, 1998] C. Fellbaum, editor. *WordNet: An electronic lexical database*. MIT Press, 1998.

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

[Lesk, 1986] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In *Proceedings of SIGDOC '86*, 1986.

[Lin, 1997] D. Lin. Using syntactic dependency as a local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 64–71, Madrid, July 1997.

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

[Resnik, 1995] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, August 1995.

[Rubenstein and Goodenough, 1965] H. Rubenstein and J.B. Goodenough. Contextual correlates of synonymy. *Computational Linguistics*, 8:627–633, 1965.

[Zipf, 1935] G. Zipf. *The Psycho-Biology of Language*. Houghton Mifflin, Boston, MA, 1935.